Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ
ГИДРОМЕТЕОРОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ

В.Н. Малинин

СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ГИДРОМЕТЕОРОЛОГИЧЕСКОЙ ИНФОРМАЦИИ

Том 1. Первичный анализ и построение эмпирических зависимостей

Рекомендовано учебно-методическим объединением по образованию
в области гидрометеорологии в качестве учебного пособия
для студентов высших учебных заведений, обучающихся
по направлению «Гидрометеорология», и специальности «Океанология»

Издание второе, исправленное и дополненное

Санкт-Петербург
РГГМУ
2020

УДК [551.46+551.5+556]:519.23(075.8)
ББК 26.22/26.23:22.172я73
М19

Рецензенты: Г.В. Алексеев, д.г.н., проф., зав отделом взаимодействия океана и атмосферы ААНИИ; Ю.А. Трапезников, д.ф.-м.н., вед. научн. сотр. Института озероведения РАН.

Малинин В.Н.

М19 Статистические методы анализа гидрометеорологической информации: учебник. В 2 томах: Том. 1. Первичный анализ и построение эмпирических зависимостей. – Издание 2, испр. и доп. – СПб.: РГГМУ, 2020. – 256 с.

Дается систематизированное изложение начальных основ математической статистики применительно к решению гидрометеорологических задач. Теоретические сведения иллюстрируются значительным числом конкретных примеров. В данном томе рассматриваются первичный анализ данных и методы построения эмпирических зависимостей.

Книга предназначена для студентов гидрометеорологических и географических специальностей, а также может быть полезна аспирантам и специалистам указанных профилей.

Given is the systematized enunciation of initial elements of mathematical statistics relating to solution of hydrometeorological problems. Theoretical data is illustrated by significant number of concrete examples. Considered are the primary analysis of data and methods of empirical constraint creation.

The textbook is intended for students of hydrometeorological and geographical specialties as well as it can be useful for post-graduate students and specialists of noted profiles.

УДК [551.46+551.5+556]:519.23(075.8)
ББК 26.22/26.23:22.172я73

Введение

В общем случае причинно-следственные связи между различными явлениями и процессами, происходящими в природной среде, можно рассматривать как детерминированные и вероятностные. Детерминированные связи являются функциональными и базируются на решении системы дифференциальных или интегральных уравнений, выражающих законы сохранения массы различных субстанций, энергии, импульса, газов. Для детерминированных связей свойственно, что каждому значению какой-либо переменной соответствует одно и только одно значение другой переменной.

Однако вследствие изменчивости природных процессов, обусловленных наличием прямых и обратных связей между ними, а также разнообразным, подчас противоположным, действием большого числа вынуждающих факторов (сил), оказывается невозможным построение строгих (полных) детерминированных моделей. По существу, это означает, что уже каждому значению какой-либо переменной будет соответствовать с определенной вероятностью (достоверностью) значение другой переменной.

Таким образом, приходим к вероятностному описанию природных процессов, основой которого служит представление о том, что характеристики этих процессов меняются произвольным образом, т. е. являются случайными величинами.

Естественно, что для описания свойств и закономерностей случайных величин необходимо использование разнообразного математического аппарата. Поэтому *раздел математики, направленный на изучение общих закономерностей случайных явлений вне зависимости от их конкретной природы, получил название теории вероятностей.*

Принципиально важным является то, что практически все выводы и результаты, получаемые в теории вероятностей, относятся к генеральной совокупности, т. е. ко всему мыслимо возможному диапазону, в пределах которого может меняться конкретная случайная величина. Именно здесь происходит главный «водораздел» между теорией вероятности и математической статистикой, одна из главных задач которой как раз состоит в том, чтобы по ограниченным данным (выборке) восстановить с определенной степенью достоверности характеристики, присущие всей генеральной совокупности. Другими словами, теорию вероятностей можно рассматривать

как теоретическую базу или своего рода фундамент для математической статистики. Экспериментальной базой для нее служат эмпирические данные, полученные в результате измерений, наблюдений или расчетов, которые естественно считать случайными величинами. Таким образом, приходим к следующему определению: *математическая статистика – это математические методы обработки и анализа эмпирической информации, представленной в виде совокупности случайных величин.*

Немного об истории. Термин «статистика» происходит от латинского слова status, что означает «*состояние*». В средние века этот термин означал политическое состояние государства. В науку данный термин введен немецким ученым Г. Ахенвалем в 1749 г., который читал в университетах Германии учебный курс с таким названием. Основным содержанием этого курса было описание политического состояния и достопримечательностей государства. Развитие с тех пор статистики как науки привело к изменению содержание самого понятия «статистика».

В настоящее время в общественных науках и экономике термин «статистика» используется в трех значениях:

1) под статистикой понимают отрасль практической деятельности, которая имеет своей целью сбор, обработку, анализ и публикацию цифровых данных о самых различных явлениях и процессах общественной жизни;

2) статистикой называют совокупность цифровых сведений, статистические данные, представляемые в отчетности предприятий, организаций, отраслей экономики, а также публикуемые в сборниках, справочниках, периодической печати и являющиеся результатом статистической работы;

3) статистикой называют отрасль знания – науку, занимающуюся разработкой теории и методов, используемых для обработки и анализа эмпирических данных.

В естественных науках (науках о Земле) в отличие от общественных наук нет такого «многогранного» толкования термина «статистика». Она понимается обычно в более «узком» смысле – как *научное направление, связанное с обработкой, анализом и интерпретацией эмпирических (гидрологических, метеорологических, геологических и др.) данных.* Естественно, что его основой является использование методов математической статистики.

Так как процесс измерений и наблюдений за гидрометеорологическими параметрами осуществляется уже в течение многих

десятилетий, а для некоторых характеристик (например, уровень моря или температура воздуха) – даже в течение нескольких столетий, то понятно, что к настоящему времени накоплены очень большие объемы экспериментальных данных, статистический анализ которых позволяет решать широкой круг самых разнообразных научных и практических задач.

Однако выполнение статистических расчетов, особенно для больших выборок, в настоящее время немыслимо без непосредственного использования пакетов прикладных статистических программ (ППСП). Действительно, поскольку процесс обработки цифровой информации связан обычно с трудоемкими вычислениями, то это предполагает применение компьютерной техники. Особенно удобно обработку информации осуществлять в рамках ППСП, реализующих комплекс стандартных статистических методов и предназначенных в основном для усредненного пользователя.

Можно перечислить десятки пакетов как иностранных, так и отечественных. Например, широкое распространение получили иностранные пакеты Statistica, SPSS, Statgraphics. Из отечественных можно отметить Stadia, Мезозавр, Сигамд. Практика показывает, что если пользователь хорошо справляется с выполнением расчетов в одном пакете, то он довольно легко может справиться с аналогичными расчетами в других пакетах. Кроме того, во всех пакетах существует файл «Помощь» (Help), позволяющий разобраться в сути поставленной задачи и способах ее решения. В некоторых ППСП содержание Help столь полно, что его можно рассматривать как оперативное руководство по выполнению расчетов, адаптированное к конкретному программному продукту. В то же время, для большинства ППСП, как правило, иностранных, зачастую отсутствует математическое описание используемых алгоритмов и методов. Вследствие этого данные пакеты представляют собой своеобразные «черные ящики».

Особое место среди ППСП занимает табличный процессор Microsoft Excel, так как он интегрирован в пакет Microsoft Office (начиная с Microsoft Excel 7.0 for Windows 95). Правда, следует иметь в виду, что статистические возможности Microsoft Excel значительно уступают другим ППСП. Тем не менее, его библиотека, содержащая около сотни статистических функций, оказывается вполне достаточной для выполнения большинства стандартных методов обработки информации. Отметим, что хотя математическая «начинка» многих статистических алгоритмов является весьма упрощенной,

это полностью компенсируется простотой и удобством в эксплуатации Excel. Если рассматривать статистические методы, приведенные в данной книге, то Excel не позволяет осуществлять расчеты только для некоторых из них в последних разделах книги.

Цель настоящего учебного пособия – систематизированное изложение начальных основ математической статистики применительно к решению гидрометеорологических задач. Теоретические сведения иллюстрируются значительным числом конкретных примеров, причем большинство из них носит оригинальный характер и специально подготовлено для данного пособия. Изложение текста ведется с учетом того, что у читателя отсутствуют знания по статистике, т. е. изучение предлагаемого материала в данном пособии может осуществляться с «чистого листа». Приводимый материал не претендует на исчерпывающую полноту изложения математической статистики, для этого есть обширная специальная литература. Для подготовленного в математическом отношении и знакомого с основами статистики читателя можно рекомендовать, например, двухтомник В.А. Рожкова «Теория и методы статистического оценивания вероятностных характеристик случайных величин и функций с гидрометеорологическими приложениями».

В данном пособии рассматриваются главным образом те аспекты математической статистики, которые, по мнению автора, нашли широкое применение в гидрометеорологии и соответственно входят в Программу курса по статистическим методам анализа гидрометеорологической информации. Исходя из этого, содержание пособия включает в себя 4 части:

– первичный анализ данных;
– построение эмпирических зависимостей;
– анализ временных рядов;
– анализ пространственных полей.

Поскольку учебным планом предусмотрено изучение дисциплины «Методы статистической обработки и анализа гидрометеорологической информации» в течение двух семестров, то в соответствии с этим учебное пособие разделено на 2 тома. В первый том входят части «Первичный анализ данных» и «Построение эмпирических зависимостей», а во второй том – «Анализ временных рядов и случайных полей».

Отметим, что, учитывая выборочный характер эмпирических данных, не следует получаемые результаты считать окончательными и истиной в последней инстанции. Более того, используя

статистические методы, необходимо обязательно соотносить получаемые результаты со здравым смыслом. Противоречие между ними может быть вызвано:

1) ненадежными исходными данными,

2) неверной постановкой статистического эксперимента (выбора статистических методов анализа данных),

3) отсутствием здравого смысла у исследователя.

То, что точность измерений многих гидрометеорологических величин не является высокой, известно хорошо. Погрешности расчетов многих характеристик, которые затем используются в статистических оценках, могут находиться в пределах точности самих расчетов. Но еще хуже, что во временных рядах могут присутствовать грубые ошибки или, другими словами, выбросы. В этом случае можно получать заведомо искаженные статистические результаты, особенно при коротких объемах исходных данных.

Что касается последнего третьего положения, то оно, вообще говоря, вполне уместно. Погоня за нужными результатами заставляет некоторых исследователей игнорировать основные постулаты и допущения, лежащие в основе тех или иных статистических методов. Действительно, если исследователь хочет получить, например, статистическую связь между числом солнечных пятен и числом самоубийств в каком-нибудь городе N, то он ее обязательно получит. Для этого достаточно взять узкое окно полосовой фильтрации и затем, не проверив отфильтрованные ряды на нормальность, рассчитать между ними корреляцию. Недаром про статистику существует столько анекдотов и изречений. Особенно часто грешат этим при использовании статистических методов в общественных науках (социологии, политологии, экономике), в частности, при обработке данных социологических опросов. Возможно, поэтому еще на рубеже XIX и XX вв. известный английский политический деятель и литератор Бенджамин Дизраэли сказал, что *«есть три вида лжи: ложь, наглая ложь и статистика»*. Безусловно, это не более чем остроумное изречение, но статистики своими работами не должны давать повода думать, что в нем есть хотя бы зерно правды.

Излагаемый материал полностью соответствует программе дисциплины «Методы статистической обработки и анализа гидрометеорологической информации», которая на протяжении уже нескольких десятилетий читается автором на океанологическом факультете, а затем в Институте гидрологии и океанологии РГГМУ. Поскольку в рамках данной дисциплины предусмотрено выполнение

цикла практических работ, посвященных лучшему усвоению теоретических знаний, то в 2010 г. был издан Практикум, автором которого является к.г.н., доцент Гордеева С.М., в течение многих лет ведущая занятия со студентами по данной дисциплине.

Теоретический материал иллюстрируется большим числом оригинальных примеров, многие из которых взяты из научно-исследовательских работ, а некоторая часть подготовлена специально для книги. Автор безмерно благодарен своим многолетним помощникам: доценту, к.г.н. С.М. Гордеевой, подготовившей большинство примеров, и к.г.н., доценту кафедры океанологии О.И. Шевчуку за помощь в вычислениях и компьютерной подготовке рисунков к печати, а также аспирантам и студентам, общение с которыми постоянно стимулирует автора к более доступному изложению даже довольно сложных статистических методов. В конце каждого тома пособия дается словарь статистических терминов, который является очень полезным для читателей, впервые начинающих знакомиться с основами статистики.

Автор признателен рецензентам: д.г.н., проф., руководителю Отдела взаимодействия океана и атмосферы Арктического и антарктического научно-исследовательского института Г.В. Алексееву; д.ф.-м.н., вед. научн. сотр. Института озероведения РАН Ю.А. Трапезникову за конструктивные советы и полезные замечания по улучшению рукописи учебного пособия.

Предлагаемое читателю пособие является вторым изданием, первое вышло в 2008 г. Изменения по сравнению с первым изданием не столь существенные. Дело в том, что первое издание, если использовать статистический язык, доказало свою состоятельность и эффективность. Во многих главах первого тома приводятся новые примеры и рисунки, в текст внесены технические правки и выполнено более тщательное редактирование.

Часть 1.
Первичный анализ данных

Глава 1. Основные понятия случайной величины

1.1. Классификация случайных величин

В статистике понятие случайной величины является одним из центральных. Вообще говоря, под *случайной величиной понимают такую переменную величину, которая в результате испытания (измерения) в одинаковых условиях может принимать то или иное заранее неизвестное значение.* Случайные величины обычно обозначаются прописными буквами латинского алфавита, т. е. X, Y, Z, а их конкретные значения, называемые иногда вариантами, обозначаются строчными буквами с индексом. Например, если случайная величина X имеет n возможных значений, то они будут обозначены как: $x_1, x_2, ..., x_n$.

Рассмотрим классификацию случайных величин (рис. 1.1). Если в результате измерения (испытания, наблюдения) регистрируется только одно число, то такую случайную величину принято называть *одномерной*. Она всегда является скалярной. Если же результатом измерения (испытания) является регистрация целого набора характеристик, то случайную величину называют *многомерной*, которая всегда является уже векторной. Например, многомерной величиной является вертикальное распределение температуры в океане или в атмосфере. Действительно, температура, измеренная на стандартных горизонтах в период выполнения многосуточной гидрологической станции, имеет две шкалы измерения: глубина и время и, следовательно, не может быть отнесена к одномерной случайной величине.

Различают два типа одномерных случайных величин: непрерывные и дискретные (прерывные). *Случайная величина называется непрерывной, если она может принять любое значение из некоторого определенного диапазона числовой оси,* который, в частности, может быть бесконечным. Ее примером могут служить многие

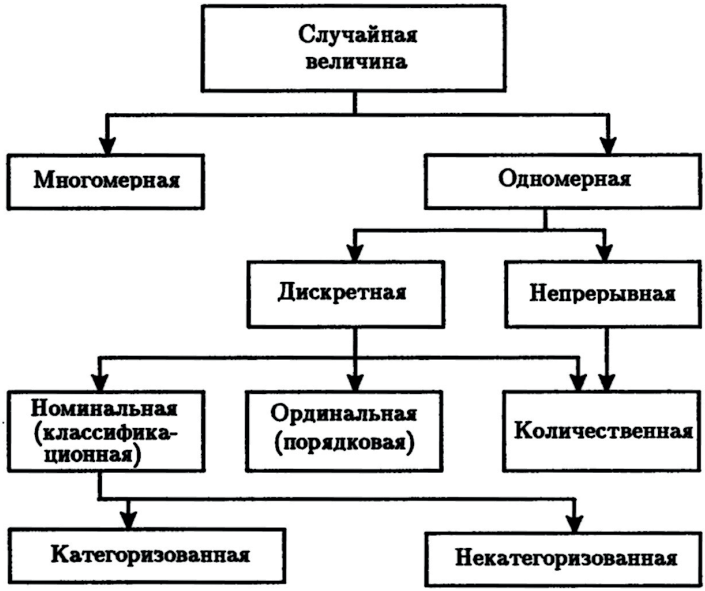

Рис. 1.1. Классификация случайной величины

гидрофизические (температура воды, плотность, скорость течения и т.д.), гидрохимические (соленость, содержание кислорода, углекислого газа и т.д.) и иные характеристики.

Случайная величина называется дискретной, если она принимает не любые значения, а только их конечное или счетное множество. В качестве примера дискретной случайной величины можно привести шкалы облачности, ледовитости, степень волнения в баллах и т.п. Например, сплоченность ледовитости измеряется в пределах от 0 до 10 баллов. Поэтому ряд наблюдений ледовитости может состоять только из целых цифр этого диапазона. Заметим, что на практике непрерывные случайные величины за счет точности измерений и округлений или за счет дискретности измерений непосредственно приборами заменяются дискретными случайными величинами. Например, температура воздуха измеряется с точностью до десятых градуса Цельсия, а уровень моря – до сантиметров. Поэтому в дальнейшем будем рассматривать только дискретные случайные величины.

Кроме того, в зависимости от своей природы и способов описания одномерные дискретные случайные величины подразделяются

на количественные, ординальные (порядковые) и номинальные (классификационные). В основе каждой из указанных случайных величин находится та или иная *шкала* наблюдения (измерения). Так, *количественная* случайная величина характеризуется метрической шкалой, в которой традиционно принятыми единицами измерения являются системы СИ или СГС. Информация в метрической шкале представляется в виде вещественных чисел.

Ординальная случайная величина соответствует порядковой (ординальной) шкале, которая выражает оценку интенсивности явления или процесса в квантованном (дискретном) виде. Например, в таком виде задается состояние на поверхности моря: от 0 баллов при полном штиле до 12 баллов при урагане. Информация при использовании порядковой шкалы выражается целыми числами баллов в принятых интервалах.

Номинальная случайная величина, соответствующая номинальной шкале, характеризует класс или тип явления, принадлежность к которому определяется по совокупности признаков. Так, вертикальные градиенты температуры, солености и плотности воды задают тип вертикальной стратификации в океане: устойчивый, неустойчивый и безразличный, а форма волнения и его зависимость от ветра позволяют подразделить волны на свободные и вынужденные. Следует отметить, что в номинальной шкале, как правило, фиксируется наличие или отсутствие явления, но не его интенсивность. Поэтому при проведении массовых расчетов на ЭВМ значения элементов в номинальной шкале кодируются как признак качества *да* (1), *нет* (0), либо посредством логических или символьных переменных.

Если исследователю наряду с анализируемым свойством известны все возможные его градации вместе с правилом отнесения обследованного в ходе случайного эксперимента объекта к одной из этих градаций, то соответствующую номинальную величину называют *категоризованной.* В противном случае, она называется *некатегоризованной*.

Описание текущего состояния океана имеет ту особенность, что исчерпывающая характеристика какого-либо его процесса может быть дана набором нескольких переменных, выражаемых в различных шкалах. Например, полная характеристика волнения содержит информацию в номинальной шкале (вынужденная или свободная волна), порядковой шкале (балл состояния поверхности моря), метрических шкалах (направление распространения волны

в градусах, длина волны и ее высота в метрах, период волны в секундах). Естественно, это существенно затрудняет процедуру статистической обработки и последующего анализа океанологических процессов и явлений.

1.2. Понятие генеральной и выборочной совокупностей

Генеральная совокупность – это весь мыслимо возможный набор случайной величины. Генеральная совокупность может быть как конечного, так и бесконечного объема. Применительно к природной среде в качестве генеральной совокупности обычно используется совокупность бесконечного объема. Это связано с тем, что мы не имеем надежных сведений о начале образования и дальнейшей эволюции природной среды и тем более не можем предсказать ее конец. Впрочем, в некоторых случаях генеральная совокупность характеристик природной среды имеет конечный объем. Например, генеральная совокупность для температуры и влажности воздуха в конкретном здании всегда является конечной, поскольку началом ее служит дата его постройки, а концом – момент разрушения или перестройки.

Принято считать, что *все характеристики генеральной совокупности являются истинными.* Заметим, что хотя понятие генеральной совокупности представляет собой математическую абстракцию, оно является основным в теории вероятностей, а также широко используется в выводах при решении различных задач статистики. Далее в целях удобства истинные (теоретические) оценки при необходимости их сопоставления с выборочными аналогами будем обозначать полужирным шрифтом. Отметим, что в статистике под *оценкой* принято понимать любое числовое значение случайной величины или случайной функции.

Выборочная совокупность – любая последовательность значений случайной величины, извлеченная из генеральной совокупности. Другими словами – это любой статистический (в частности, временной) ряд, имеющий конечную длину. Следовательно, параметры такого ряда являются *выборочными параметрами.* Очевидно, что выборочная оценка параметра θ стремится к истинной оценке $\boldsymbol{\theta}$ при $n \to \infty$, где n – длина выборки. Например, выборочная оценка среднего арифметического стремится к истинной оценке

(математическому ожиданию) при неограниченном увеличении длины выборки, т. е. $\lim\limits_{n\to\infty}\overline{\theta}\to M[\theta]$.

Как было указано во введении, главный «водораздел» между теорией вероятности и математической статистикой состоит в том, что первая всегда имеет дело с истинными оценками случайных величин, а вторая – с их выборочными значениями.

Если выборка достаточно точно отражает основные закономерности, присущие генеральной совокупности, то она считается *представительной (репрезентативной)*. В этом случае выборочные параметры должны быть близкими к их истинным оценкам. Степень такой «близости» или, другими словами, степень «надежности» выборочных параметров обычно описывается с помощью следующих трех свойств статистических оценок: состоятельности, несмещенности и эффективности.

Несмещенность. Оценка параметра θ называется несмещенной, если ее *среднее значение, т. е. центр распределения выборочной совокупности случайной величины, совпадает с истинной величиной оцениваемого параметра.* В противном случае оценка является смещенной. Если это равенство не выполняется, то оценка θ, полученная по разным выборкам, будет либо завышать значение параметра **θ**, либо занижать его. Следовательно, требование несмещенности гарантирует отсутствие систематических ошибок при оценивании параметров. Итак, *требование несмещенности по существу означает, что выборочная средняя должна совпадать с ее истинной оценкой, т. е. с математическим ожиданием*. В результате имеем $\overline{\theta}=M\left[\theta\right]=m_{\theta}$.

Состоятельность. *Оценка параметра* θ *называется состоятельной, если она удовлетворяет закону больших чисел, т. е. при неограниченном возрастании объема выборки сходится по вероятности к оцениваемому параметру:*

$$\lim_{n\to\infty} p\left[\left|\theta-\hat{\theta}\right|<\varepsilon\right]=1,\ \varepsilon>0,$$

где $\hat{\theta}$ – истинная оценка параметра θ, ε – сколь угодно малое наперед заданное положительное число. Требование состоятельности означает, что с увеличением объема выборки рассеивание оценок θ относительно математического ожидания будет уменьшаться, и при достаточно большом значении *n* отклонение θ от $\hat{\theta}$ при доверительной

вероятности $p \to 1$ должно быть меньше любого наперед заданного числа. Отсюда следует асимптотический характер свойства состоятельности – проявляться лишь при неограниченном возрастании объема выборки. Таким образом, *оценка* θ *является состоятельной, если при неограниченном росте объема выборки она стремится к неизвестному истинному значению параметра, т. е.* $\theta \to \hat{\theta}$.

Эффективность. Несмещенная *оценка параметра* θ *называется эффективной, если она при заданном объеме выборки имеет наименьшую дисперсию среди всех возможных несмещенных оценок параметра* θ*, вычисленных по выборкам одного и того объема n, т. е.* $D[\theta] = D_{\min}$. Это означает, что эффективная оценка имеет меньшую вероятность появления грубой ошибки при определении параметров распределения.

Для эффективности оценки самым важным является задание закона распределения. Следует иметь в виду, что эффективная оценка параметра генеральной совокупности для одного закона распределения может не совпадать с эффективной оценкой параметра другого закона распределения.

Итак, если статистические параметры выборки отвечают указанным выше требованиям, то они считаются «хорошими» в статистическом смысле, а сама выборка является репрезентативной. Заметим, что оценка выборочной средней случайной величины обладает всеми тремя выше перечисленными свойствами: она является несмещенной, состоятельной и эффективной. Оценка дисперсии состоятельна и эффективна, но имеет малое отрицательное смещение по отношению к генеральной дисперсии, равное $\frac{n}{(n-1)}$. Поэтому для выборок малого объема ($n < 25$–30) в формуле оценки дисперсии вместо n целесообразно использовать $n - 1$, что позволяет устранить смещение.

Исследование свойств выборочных характеристик позволило установить, что в асимптотическом смысле, т. е. при неограниченном увеличении объема выборки, ее основные характеристики с ростом объема выборки стремятся к своим теоретическим аналогам и ведут себя при этом как нормально распределенные случайные величины.

1.3. Понятие о законе распределения случайной величины

С вероятностной точки зрения случайная величина может быть описана, если известны не только значения, какие она может принимать, но и как часто, т. е. с какой вероятностью она принимает эти значения. Другими словами, нужно задать закон распределения случайной величины.

В теории вероятностей под *законом распределения случайной величины понимается любое соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями.* Законы распределения могут быть выражены в табличной, графической и аналитической форме.

Графическая форма закона распределения состоит в том, что по оси абсцисс откладываются значения случайной величины, а по оси ординат – вероятности этих значений. Полученная таким образом фигура (рис. 1.2) называется многоугольником или *полигоном распределения.* При этом сумма ординат многоугольника, представляющая собой сумму вероятностей всех возможных значений случайной величины, всегда равна единице. В математической статистике *полигон распределения* представляет собой ломаную линию, соединяющую частоты вариационного ряда, т. е. выборки, построенной в порядке возрастания ее отдельных значений.

Табличная форма закона распределения аналогична графической форме, но только в этом случае возможные значения случайной величины X и соответствующие им вероятности p задаются в виде таблицы.

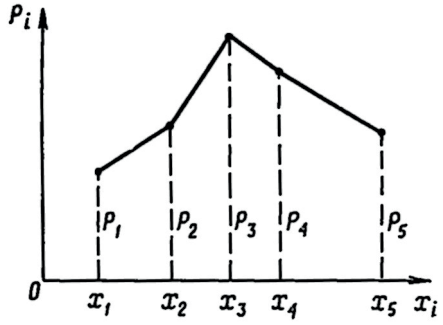

Рис. 1.2. Многоугольник (полигон) распределения

Аналитическая форма закона распределения описывается функцией распределения $\boldsymbol{F}(x)$, которая определяет вероятность того, что случайная величина X принимает значения меньше некоторого числа x, т. е.

$$\boldsymbol{F}(x) = p\left(X < x\right), \tag{1.1}$$

где p – вероятность, понимаемая применительно к выборочным данным как *частота события.* Геометрически это равенство можно истолковать так: $\boldsymbol{F}(x)$ *представляет вероятность того, что случайная величина примет значение, которое изображается на числовой оси точкой, лежащей левее точки x.*

Обычно функцию распределения (1.1) называют *интегральной функцией распределения* или интегральным законом. Она может быть использована как для дискретных, так и для непрерывных случайных величин. Еще раз подчеркнем, что функция распределения представляет наиболее общую форму описания случайной величины и полностью характеризует ее с вероятностной точки зрения.

Интегральная функция распределения обладает следующими основными свойствами.

Свойство 1. Значения функции распределения заключены в диапазоне [0,1], т. е. $0 \le \boldsymbol{F}(x) \le 1$. Это означает, что $\boldsymbol{F}(x) = 0$, если $x \to -\infty$ и $\boldsymbol{F}(x) = 1$, если $x \to \infty$.

Свойство 2. Функция $\boldsymbol{F}(x)$ является неубывающей, т. е. если $x_1 < x_2$, то $\boldsymbol{F}(x_1) \le \boldsymbol{F}(x_2)$.

Следствие 1. Вероятность того, что случайная величина примет значение, заключенное в интервале $[a,b]$, равна приращению функции распределения на этом интервале: $p(a \le X < b) = F(b) - F(a)$.

Следствие 2. Вероятность того, что случайная величина X примет одно определенное значение, равна нулю.

Свойство 3. Если возможные значения случайной величины принадлежат интервалу $[a,b]$, то $\boldsymbol{F}(x) = 0$ при $x \le a$ и $\boldsymbol{F}(x) = 1$ при $x \ge b$.

Следствие 3. Если возможные значения непрерывной случайной величины расположены на всей оси x, то справедливы следующие предельные соотношения: $\boldsymbol{F}(x) = 0$ при $x \to -\infty$ и $\boldsymbol{F}(x) = 1$ при $x \to \infty$.

Для непрерывной случайной величины график этой функции представляет непрерывную кривую, монотонно возрастающую от нуля до единицы (рис. 1.3-*а*). Для дискретной случайной величины функция распределения является ступенчатой функцией, непрерывной слева. При этом функция имеет разрыв в точках, совпадающих с возможными значениями случайной величины, а величины скачков

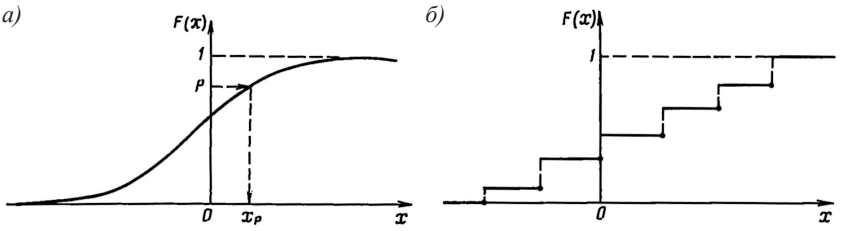

Рис. 1.3. Графики интегральной функции распределения для непрерывной (*а*) и дискретной (*б*) случайных величин

совпадают с соответствующими вероятностями, т. е. $p_i = p(X = x_i)$. График этой функции приводится на рис. 1.3-*б*.

Заметим, что в гидрометеорологических расчетах иногда используется функция обеспеченности, обратная функции распределения:

$$\boldsymbol{P}(x) = p\left(X \geq x\right) = 1 - \boldsymbol{F}(x). \qquad (1.2)$$

Естественно, что кривая обеспеченности симметрична кривой функции распределения и пересекается с ней при $\boldsymbol{F}(x) = P(x) = 0{,}5$.

Недостатком функции распределения является то, что она, являясь функцией «накопленной вероятности», не отражает распределения вероятностей по отдельным значениям случайной величины и не показывает, как часто появляются те или иные ее значения. Этого недостатка лишена *плотность распределения вероятностей,* называемая также *дифференциальной функцией распределения* (законом распределения), представляющая собой первую производную от $\boldsymbol{F}(x)$:

$$f(x) = F'(x) = \lim_{\Delta x \to 0} \frac{\Delta F}{\Delta x} = \lim_{\Delta x \to 0} \frac{p\left(x \leq X < x + \Delta x\right)}{\Delta x}. \qquad (1.3)$$

Отсюда видно, что плотность распределения есть предел отношения вероятности попадания случайной величины X в интервал $[x, x + \Delta x]$ к величине Δx при $\Delta x \to 0$. График плотности распределения (рис. 1.4) называется *кривой распределения.*

К основным свойствам плотности распределения относятся:

Свойство 1. Плотность распределения является неотрицательной функцией, т. е. $\boldsymbol{f}(x) \geq 0$;

Свойство 2. Интеграл от плотности распределения в бесконечных пределах равен 1, т. е. $\int\limits_{-\infty}^{\infty} f(x)dx = 1$.

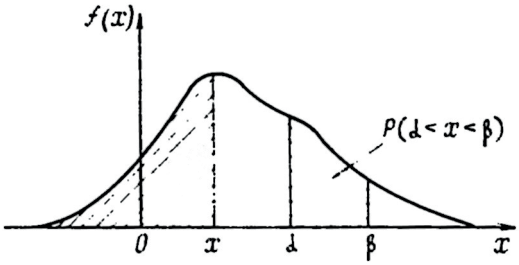

Рис. 1.4. График дифференциальной функции распределения

Это означает, что полная площадь, ограниченная кривой распределения и осью абсцисс, равна единице.

Распределение может быть теоретическим или эмпирическим (статистическим). Если известны истинные значения вероятностей случайной величины, то такое распределение является *теоретическим*. Несмотря на то, что во многих случаях истинные оценки вероятностей неизвестны, тем не менее, представляется возможным получить их приближенные оценки на основе опытных данных. Распределение вероятностей, полученных из опытных (эмпирических) данных достаточно большого объема, называется *эмпирическим распределением* случайной величины.

В этом случае под эмпирической функцией распределения понимается любое соотношение, устанавливающее связь между возможными значениями случайной величины X и соответствующими им относительными частотами события $X < x$. Отсюда следует:

$$F(x) = \frac{n_x}{n}, \tag{1.4}$$

где n_x – число вариант (значений), меньших x. Таким образом, различие между $\boldsymbol{F}(x)$ и $F(x)$ состоит в том, что первая определяет вероятность события $X < x$, а вторая определяет относительную частоту этого же события. Заметим также, что свойства эмпирической (статистической) функции полностью совпадают со свойствами теоретической функции распределения.

1.4. Статистические ряды распределения

В общем случае любая выборка может быть упорядочена, т. е. расположена в возрастающем (начиная с минимального значения) или убывающем (начиная с максимального значения) порядке.

Такая процедура называется ранжированием ряда, а сам ряд – *ранжированным* рядом. Если теперь этот ряд разбить на некоторое число интервалов (групп, градаций) и распределить отдельные значения по интервалам, то получим статистический ряд распределения. Другими словами, *статистический ряд распределения – это упорядоченное распределение единиц совокупности на группы по определенному варьирующему признаку.*

В зависимости от признака, положенного в основу образования такого ряда, различают атрибутивные и вариационные ряды распределения. *Атрибутивными* называют ряды распределения, построенные по качественным признакам. *Вариационными* называют ряды распределения, построенные по количественному признаку. Обычно вариационный ряд строится в порядке возрастания значений его членов и обозначается следующим образом: $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$. Каждый член этой последовательности ($x^{(i)}$) называется *порядковой статистикой.* Аппарат порядковых статистик широко используется при статистическом оценивании и проверке гипотез, непараметрическом анализе малых выборок и ряде других задач. Следует иметь в виду, что члены вариационного ряда в отличие от членов исходной выборки уже не являются взаимно независимыми (по причине своей предварительной упорядоченности). Соответственно, их частные распределения не являются одинаковыми, описываемыми одним и тем же законом распределения, как для исходной выборки.

Любой вариационный ряд состоит из двух элементов: вариантов и частот. *Вариантами* считаются отдельные значения признака, которые он принимает в данном ряду, т. е. конкретные значения варьирующего признака. *Частоты* − это численности отдельных вариантов или каждой группы вариационного ряда. Другими словами, это числа, показывающие, как часто встречаются те или иные варианты в ряду распределения. Сумма всех частот определяет объем выборки. *Частостями* называют частоты, выраженные в долях единицы или в процентах к итогу. Поэтому сумма частостей равна 1 или 100 %.

В зависимости от характера вариации признака различают дискретные и интервальные вариационные ряды. Дискретный вариационный ряд характеризует распределение единиц совокупности по дискретному признаку, а интервальный ряд – по непрерывному признаку, который может принимать на числовой оси любые значения.

Наглядное представление о характере изменения частот вариационного ряда дают полигон и гистограмма. *Полигон* используется

при изображении дискретного вариационного ряда. Для его построения в прямоугольной системе координат по оси абсцисс в одинаковом масштабе откладываются ранжированные значения признака, а по оси ординат – частоты. Соединив эти точки прямыми линиями, получим полигон распределения.

Гистограмма применяется для изображения интервального вариационного ряда. При построении гистограммы на оси абсцисс откладываются номера интервалов, а частоты изображаются прямоугольниками, опирающимися на соответствующие им интервалы. В результате получим гистограмму – *график, представляющий распределение частот по интервалам вариационного ряда.* Если середины интервалов соединить линией, то получим график плотности распределения, т. е. значения частот, приходящихся на единицу ширины интервала.

Довольно часто для изображения вариационных рядов используется *кумулятивная кривая.* При помощи кумуляты, т. е. кривой сумм, изображается ряд накопленных частот, который показывает, как быстро к 1 или 100 % приближается ряд распределения. Если на таком графике поменять местами оси ординат и абсцисс, то получим кривую, называемую *огивой*.

1.5. Основные этапы статистического анализа эмпирической информации

Слово «информация» в переводе с латинского означает «осведомление» и «доведение сведений о чем-либо». Очевидно, в общем случае под *информацией следует понимать любые сведения (в количественной и качественной форме) об исследуемом объекте.* Естественно, что со статистической точки зрения наибольший интерес вызывает количественная информация, частным случаем которой является гидрометеорологическая информация. Объектом гидрометеорологической информации служит, как известно, природная среда.

Всю совокупность информации целесообразно разделить на первичную и вторичную. *Первичная информация* – это результат непосредственного измерения метеорологических, гидрологических, океанологических и иных параметров со стационарной сети станций, постов, полученных во время экспедиций, натурных экспериментов, а также с помощью наземных, самолетных, спутниковых измерительных комплексов и т.д.

Вторичная информация уже представляет результаты расчетов, выполненных на основе первичной информации. Так, например, данные по испарению могут представлять собой первичную информацию, если оно измерено с помощью малоинерционной аппаратуры, или вторичную информацию, если испарение рассчитывается по первичным данным тем или иным методом. Отметим, что основной в этом случае является вторичная информация по испарению.

Логическая схема статистического анализа эмпирических данных какого-либо процесса или явления может быть представлена в виде следующих основных этапов (рис. 1.5):

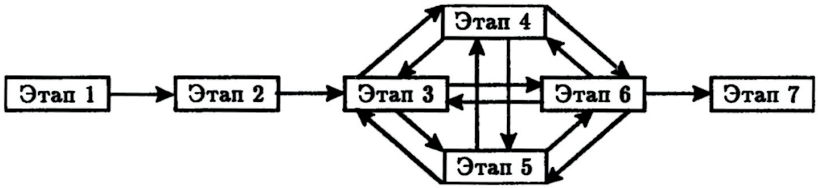

Рис. 1.5. Блок-схема структурных взаимосвязей основных этапов статистического анализа

Этап 1: постановка задачи. Сюда входит формулирование основных целей предполагаемого исследования и возможных результатов, которые могут быть получены с его помощью. Естественно, что в каждом конкретном случае формулирование целей носит произвольный характер и осуществляется на содержательном (физическом) уровне. При этом правильное физическое представление изучаемого процесса или явления является тем «базисом», на который затем ставится «надстройка», представляющая собой совокупность статистических методов. Следовательно, важной составной частью постановки задачи должен служить предварительный физический анализ исследуемого реального процесса или явления, отображением которого выступает система исходных наблюдений (измерений). После этого осуществляется формализованная постановка задачи, включающая по возможности вероятностную модель или совокупность статистических методов, которые могут быть применены к анализу изучаемой системы данных. Кроме того, в случае необходимости выполняется оценка общего времени и трудозатрат на выполнение исследования.

Этап 2: информационный. В том случае, если задача поставлена таким образом, что реальные экспериментальные данные для ее решения отсутствуют, то необходимым условием данного этапа

становится предварительное составление плана сбора исходной статистической информации. При этом желательно учитывать полную схему дальнейшего статистического анализа с тем, чтобы не возникали ситуации, когда становится очевидной невозможность проведения расчетов из-за отсутствия необходимых для этого данных. Если экспериментальные данные имеются в необходимом объеме, но только в табличном виде, то возникает задача их занесения на машинные носители, т. е. создание компьютерных архивов данных. Хотя эта задача носит технический характер, тем не менее, следует предварительно составить такую схему занесения данных в ЭВМ, чтобы в последующем их было удобно обрабатывать.

Этап 3: первичный анализ. В ходе первичной статистической обработки исходной информации обычно решаются следующие конкретные задачи:

– унификация типа переменных, т. е. перевод их в единую однородную систему;

– расчет и анализ первичных статистик;

– анализ резко выделяющихся наблюдений (выбросов);

– восстановление пропущенных наблюдений;

– проверка статистической независимости наблюдений, составляющих массив исходных данных;

– проверка свойств внутренней структуры временного ряда с помощью статистических гипотез;

– экспериментальный анализ закона распределения исходной совокупности и его параметризация.

Отметим, что задача унификации типа переменных возникает при автоматизированном анализе многомерного случайного процесса, когда одновременно могут встречаться переменные всех трех типов: количественные, ординальные и номинальные. В этом случае используются два альтернативных подхода. Первый связан с «оцифровкой» (шкалированием) неколичественных переменных, когда исследователь, руководствуясь дополнительными соображениями и допущениями, пытается преобразовать качественные данные в количественные. При другом подходе все наблюдения многомерной случайной величины смешанной природы делятся на определенное число градаций (интервалов, классов и т.п.), внутри которых все данные заменяются на нули или единицы. Естественно, при переходе от индивидуальных к сгруппированным значениям происходит потеря информативности исходных данных, но это неизбежная плата за подобные преобразования.

Этап 4: построение эмпирических зависимостей. К данному этапу относятся следующие конкретные задачи:

– определение вида связи между переменными;
– корреляционный анализ;
– построение и анализ линейной регрессии двух переменных;
– построение и анализ одномерной полиномиальной регрессии;
– подбор нелинейной эмпирической формулы;
– построение и анализ двухмерной полиномиальной регрессии;
– построение и анализ множественной линейной регрессии.

Этап 5: анализ временных рядов. При проведении временного анализа решаются следующие конкретные задачи:

– проверка стационарности временного ряда;
– построение и анализ трендов;
– гармонический анализ;
– автокорреляционный анализ;
– взаимнокорреляционный анализ;
– спектральный анализ;
– фильтрация временных рядов.

Этап 6: анализ пространственных полей. Данный этап включает в себя:

– определение числовых характеристик полей;
– оценка однородности и изотропности случайного поля;
– анализ статистической структуры полей;
– построение и анализ карт;
– объективный анализ случайных полей.

Сразу же отметим, что для этапов 3–6 перечислены в основном лишь те задачи, которые непосредственно рассмотрены в данном пособии. Безусловно, список их может быть существенно расширен. Следует также иметь в виду, что разделение на этапы во многом является условным. Прежде всего, оно не означает, что эти этапы осуществляются в строгой хронологической последовательности один за другим. Некоторые из них могут быть объединены вместе, другие, исходя из специфики исходного материала, вообще пропущены. Кроме того, ряд этапов (например, 3, 4, 5 и 6 этапы) находится в соотношении итерационного взаимодействия: результаты более поздних этапов могут содержать выводы о необходимости повторения предыдущих этапов (см. рис. 1.4).

Этап 7: интерпретация результатов и подведение итогов исследования. Очевидно, это самый неформальный этап. *Получение содержательных выводов – главный итог выполненного*

исследования. Однако прежде делается формальный статистический отчет, представляющий собой выводы из применения статистических процедур, результаты которых даются в виде таблиц, графиков, формул и т.п. Именно это и служит основой для формулирования содержательных выводов.

Отметим, что анализ получаемых результатов должен осуществляться не только в конце исследования, но и после каждого этапа, причем в зависимости от этого они могут подвергаться ревизии (пересмотру). Например, один статистический метод заменен другим, выдвигается новая вероятностная модель и т.д.

В заключение проверяется, в какой степени достигнуты намеченные на первом этапе исследования цели, и если не все они достигнуты, то объясняется, с чем это связано.

1.6. Общая характеристика океанологической информации

Под океанологической информацией понимают совокупность данных наблюдений и расчетов любых характеристик океана. К ним относятся, прежде всего, физические, химические, биологические и геологические характеристики. В этом заключается принципиальная сложность их статистического анализа, поскольку методы их измерения и тем более расчетов резко различаются по степени автоматизации, сложности и точности, а сами данные – по степени полноты охвата акватории Мирового океана. Очевидно, наиболее полно Мировой океан, особенно после внедрения в практику спутниковых систем измерения, освещен данными по температуре поверхности океана (ТПО) и уровню. Очень слабо известны многие биологические характеристики.

Обработка океанологической информации может быть разделена на несколько наиболее общих видов:

1) первичная обработка информации;
2) оперативный диагноз океанологических процессов;
3) подготовка выборочных массивов данных;
4) подготовка режимно-справочных обобщений;
5) количественный анализ в научных целях.

Принципиальное отличие первичной обработки и оперативного диагноза данных от других видов обработки состоит в том, что они осуществляются многократно в заданном ритме с дискретностью, равной дискретности поступления новой информации

об океане. Примером подобных видов обработки можно считать, например, оперативную обработку спутниковой карты распределения температуры поверхности океана. Для каждого из множества снимков, приходящих в течение суток, необходимо осуществить первичную обработку, т. е. выявить шумы (помехи), облачность и отделить сушу от водного пространства. Затем статистическими методами восстанавливаются недостающие данные и осуществляется собственно диагноз, который бы идентифицировал фронтальные зоны, вихри, распределение аномалий температуры и т.п. Последующие виды обработки данных производятся в основном эпизодически или даже в единичных случаях. Суть их достаточна очевидна.

Весь сложный процесс разнообразной обработки океанологической информации состоит из четырех основных этапов: сбор информации, накопление и хранение, собственно обработка информации, анализ результатов.

Практически любая исследовательская океанологическая работа начинается со сбора информации и заканчивается ее анализом. Достижение надежного результата в исследованиях океана возможно при выполнении условия взаимного соответствия друг другу и общим целям одновременно всех перечисленных этапов. Для этого сам процесс преобразования исходной информации должен быть системой, в которой планомерно, упорядоченно и закономерно выполняются все основные действия.

К сожалению, при обработке океанологической информации возникает целый ряд трудностей. Так, описание текущего состояния океана имеет ту особенность, что исчерпывающая характеристика его дается, как правило, набором нескольких переменных, выражаемых в различных шкалах. Как уже указывалось выше, полная характеристика волнения содержит информацию в номинальной шкале (вынужденная или свободная волна), порядковой шкале (балл состояния поверхности моря), метрических шкалах (направление распространения волны в градусах, длина волны и ее высота в метрах, период волны в секундах). Этой причиной объясняются многие трудности в процессе сбора, накопления, обмена и обработки океанологической информации. Трудно создать унифицированный код для передачи любой гидрометеорологической информации, поэтому сейчас используются различные коды. Еще более трудно вычислить и проанализировать статистические характеристики связи для переменных, задаваемых различными шкалами.

Традиционные виды работ, проводимые в открытом море (океане), можно разделить на четыре вида в зависимости от назначения.

1. *Наблюдения на вековых разрезах,* состоящие из стандартного комплекса измерений различных характеристик, систематически выполняемые ежегодно, один раз в сезон или в месяц. Наиболее уникальным представляется вековой разрез «Кольский меридиан», который вытянут от Мурманска на север вдоль 33° в.д. и включает порядка 10 гидрологических станций. Первые наблюдения на этом разрезе были выполнены еще в 20-е годы прошлого столетия. Наиболее полные систематические наблюдения начинаются с 1951 г. К настоящему времени количество выполнений данного разреза уже превысило 1000 раз. К сожалению, в последние годы из-за нехватки финансирования наблюдения на данном разрезе осуществляются все реже и реже.

2. *Комплексные океанографические съемки* по сетке стандартных разрезов и наблюдения на судах погоды, научно-исследовательских судах и буйковых станциях, регулярно выполняемые для оперативного обеспечения различных отраслей экономики и службы прогнозов гидрометеорологической и гидрохимической информацией о состоянии океанических акваторий и морей. Отметим важную роль судов погоды, которые начали функционировать с начала 1950-х годов. В Северной Атлантике начали работу девять судов погоды, в северной части Тихого океана – четыре. Многие потом, к сожалению, были закрыты по финансовым соображениям. Из действующих, на наш взгляд, громадное значение имеет судно погоды «М», расположенное почти в центре Норвежского моря (66° с.ш. и 2° в.д.). Дело в том, что акватория Норвежского моря является мощной энергоактивной зоной океана, имеющая исключительно важное значение в формировании и колебаниях гидрометеорологического режима сопредельных территорий, в том числе Европейской территории России. На судне «М» выполняется широкий комплекс глубоководных гидрологических, гидрохимических, а также метеорологических и даже аэрологических наблюдений. Общий период наблюдений, начатых в 1951 г., уже достигает 70 лет.

3. *Эпизодические океанографические наблюдения* и работы, выполняемые по специальным программам для обеспечения тематики научно-исследовательских работ, в том числе работы на полигонах. В качестве примера можно упомянуть научно-исследовательские программы «Полэкс-Север» и «Полэкс-Юг», которые были разработаны в начале 1970-х годов с целью изучения процессов

крупномасштабного взаимодействия океана и атмосферы и их пространственно-временной изменчивости в высоких широтах. За период порядка пятнадцати лет проведены комплексные натурные эксперименты, что позволило получить огромный объем уникальных экспериментальных данных, которые затем использовались при решении многих актуальных научных задач.

4. *Попутные гидрометеорологические наблюдения,* регулярно четыре раза в сутки осуществляемые штурманским составом коммерческих судов, которые предназначены для получения оперативной информации о состоянии погоды в районах плавания и составления режимно-справочных обобщений. Именно таким образом осуществлялся сбор, а затем последующая обработка данных о температуре поверхности океана, температуре воздуха и скорости ветра попутных (коммерческих) судов в Северной Атлантике от экватора до 70° с.ш. Дважды в сутки (0 и 12 ч по Гринвичу) радисты передавали в Центры погоды радиосводки с данными о температуре воды и воздуха, скорости ветра, атмосферном давлении. Во ВНИГМИ-МЦД бывшего СССР эти данные обрабатывались. При этом акватория Северной Атлантики была разделена на пятиградусные трапеции, границами которых являлись широты и долготы, кратные пяти. По координатам судна гидрометеорологические данные относились к той или иной трапеции. Накопленные за месяц наблюдения усреднялись и затем по истечении годового интервала времени публиковались атласы.

Кроме традиционных видов океанографических работ, в последние десятилетия все большое распространение получают дистанционные методы и прежде всего спутниковые наблюдения. С середины 70-х годов прошлого столетия выполняются измерения характеристик морского льда. С начала 80-х годов точность измерения температуры поверхности океана с ИСЗ становится достаточной для ее использования при решении многих научных и практических задач. Так, спутники NOAA-7, -10, -11 и -14 обеспечивают измерение ТПО с пространственным разрешением по широте и долготе 1/6° (примерно 18 км) и временным осреднением одна неделя. При этом точность измерения ТПО составляет 0,3 ºС. С 1993 г. стали доступными альтиметрические данные об уровне океана. Спутниковая альтиметрия осуществляет измерение расстояния между спутником и поверхностью отражения по времени прохождения сигнала бортового радарного высотомера, передающего со скоростью света высокочастотные радиосигналы и получающего отраженный от морской

поверхности сигнал. Независимое определение параметров орбиты спутника (широта, долгота, высота) относительно земного эллипсоида позволяет найти высоту уровня океана. При этом альтиметрические измерения, отсчитываемые от поверхности геоида, показывают возмущения относительно среднего стационарного состояния уровенной поверхности океана. Основным источником альтиметрических измерений являются спутники TOPEX/Poseidon, ERS-2, Jason-1. Эти данные имеют пространственное разрешение 1/3° в меркаторовской проекции, временное осреднение – одна неделя и точность расчета морского уровня – 4,2 см. Очевидно, именно за дистанционными методами наблюдений за различными гидрометеорологическими характеристиками со спутников – будущее.

Весьма важным источником натурных данных об океане является так называемый «реанализ» (ретроспективный анализ), представляющий собой синтезированные данные о состоянии атмосферы и океана, полученные путем обработки результатов предшествующих наблюдений с сети стационарных станций и данных спутникового дистанционного зондирования с ассимиляцией их в численные модели с целью корректировки прогнозов погоды. Многие системы «реанализа» носят глобальный характер, оперативно пополняются и находятся в свободном доступе в сети Интернет. В качестве примера обратимся к таблице 1.1, в которой приводятся сведения о некоторых глобальных архивах, содержащих данные о температуре поверхности океана.

Таблица 1.1

Общая характеристика некоторых архивов, содержащих данные о ТПО

Наименование архива	Пространственное разрешение (широта / долгота)	Пространственная протяженность	Временной период данных (гг.)	Временная дискретность данных
NOAA NCEP/ NCAR CDAS	1,875×1,875	Глобальное	с 1949	1 месяц
COADS	2×2	Глобальное	1854–1992	1 месяц
NOAA NCEP CMB GLOBAL (RESM)	1×1	Глобальное	с декабря 1981	5 дней 7 дней 1 месяц
UKMO	5×5	Глобальное	с 1856	1 месяц

Приведем теперь краткое описание этих архивов.

– *Архив CDAS* (Climate Data Assimilation System) системы NOAA NCEP/NCAR Reanalysis содержит глобальный архив ТПО,

а также разнообразных среднемесячных метеорологических данных и характеристик внешнего теплового баланса океана с 1949 г., оперативно пополняется с очень небольшим запаздыванием во времени и находится в свободном доступе на сайте (http://sgi62.wwb.noaa.gov:8080). Пространственное разрешение исходных данных – широтно-долготная сетка 1,875°×1,875°.

– *Система COADS* (Comprehensive Ocean-Atmosphere Data Set), содержащая среднемесячные данные по ТПО за период с 1854 по 1992 г. в двухградусных квадратах, доступ к которой находится на сайте http://iridl.ldeo.columbia.edu/SOURCES/.COADS/.

– *Архив Рейнольдса-Смита* (RESM) системы NOAA NCEP, содержащий данные по ТПО в одноградусной сетке с 1981 г. и оперативно пополняемый в реальном режиме времени, а свободный доступ к нему есть на сайте: http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP/.EMC/.CMB/.GLOBAL/.Reyn_SmithOIv2/.monthly/.dataset.

– *Архив метеорологической службы Англии UKMO* (United Kingdom Meteorological Office), содержащий данные о ТПО в пятиградусной сетке Мирового океана с 1856 г., имеющий свободный доступ на сайте http://www.cru.uea.ac.uk/ftpdata/hadcrut.nc.

Глава 2. Числовые характеристики случайной величины

2.1. Методы точечного оценивания

Параметры, назначение которых состоит в выражении в сжатой форме наиболее существенных особенностей распределения случайных величин, называются числовыми характеристиками. Определение числовых характеристик представляет собой суть статистического оценивания случайной величины. Отметим также, что поскольку числовые характеристики имеют чрезвычайно важное значение, то они очень часто используются в расчетах безотносительно законов распределения.

В настоящее время используются два основных метода статистического оценивания: точечное и интервальное (рис. 2.1). *Точечное оценивание* заключается в том, что с помощью статистических методов определяются конкретные (точечные) оценки выборочного параметра, около которого находятся его истинные значения.

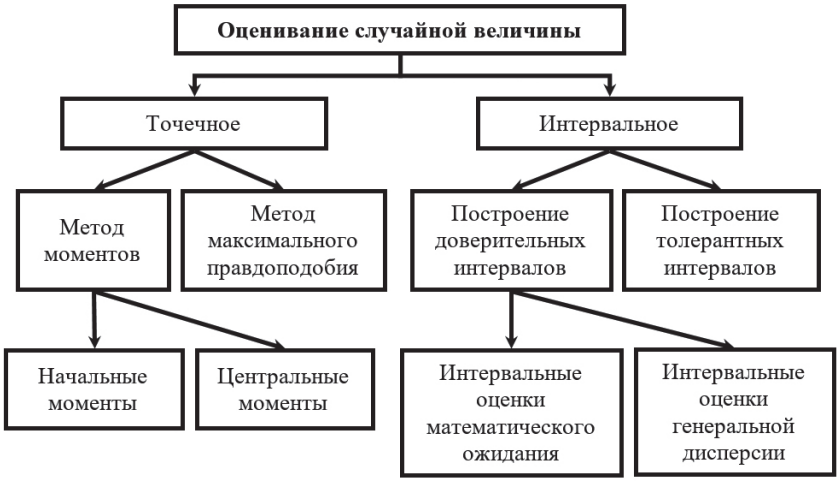

Рис. 2.1. Методы статистического оценивания случайной величины

Суть *интервального оценивания* состоит в том, что с помощью статистических методов мы получаем определенный диапазон (интервал) оценок выборочного параметра, внутри которого с большой заданной вероятностью находится его истинное неизвестное значение.

Точечное оценивание осуществляется с помощью следующих методов: моментов, максимального (наибольшего) правдоподобия, наименьших квадратов, наименьших абсолютных отклонений и др. Наиболее точным принято считать метод максимального правдоподобия, который рассматривается в виде некоторого эталона для других методов. Это связано с тем, что оценки максимального правдоподобия образуют класс оценок, имеющих наименьшую среднеквадратическую ошибку для выборок большого объема. Однако даже такой простой метод, как метод моментов для распределений, близких к нормальному закону, позволяет получить оценки параметров, хорошо согласующихся с методом максимального правдоподобия.

Суть *метода максимального правдоподобия,* предложенного Р. Фишером, состоит в следующем. Допустим, что вид функции распределения дискретной случайной величины X задан, но не известен параметр $\boldsymbol{\theta}$, которым определяется этот закон. Тогда функцией правдоподобия случайной величины X называют функцию аргумента $\boldsymbol{\theta}$:

$$L(x_1, x_2,\ldots, x_n; \boldsymbol{\theta}) = p(x_1; \boldsymbol{\theta})p(x_2; \boldsymbol{\theta})\ldots p(x_n; \boldsymbol{\theta}),$$

где $p(x_i; \boldsymbol{\theta})$ – вероятность того, что в результате испытаний величина X примет значение x_i.

В качестве точечной оценки параметра $\boldsymbol{\theta}$ принимают такое его значение $\theta = \theta(x_1, x_2, \ldots, x_n)$, при котором функция правдоподобия достигает максимума. При этом оценку θ называют оценкой максимального правдоподобия. В практических расчетах вместо функции L удобнее использовать логарифмическую функцию правдоподобия $\ln L$. Для нахождения точечной оценки θ нужно отыскать максимум функции $\ln L$ следующим образом:

1) вычислить производную $\dfrac{d \ln L}{d\theta}$;

2) приравнять производную нулю и найти критическую точку – корень полученного уравнения;

3) найти вторую производную $\dfrac{d^2 \ln L}{d\theta^2}$, и в том случае, если вторая производная при $\boldsymbol{\theta} = \theta$ отрицательна, то θ – точка максимума.

Найденную таким образом точку максимума θ принимают в качестве оценки максимального правдоподобия параметра $\boldsymbol{\theta}$. К достоинствам данного метода относится то, что он всегда дает состоятельные, эффективные и несмещенные оценки и использует всю информацию, содержащуюся в выборке. Существенный недостаток метода состоит в том, что полученные с его помощью оценки зависят от закона распределения, а также в том, что он часто требует довольно сложных вычислений.

Метод моментов, предложенный К. Пирсоном, опирается на понятие о моментах статистических совокупностей. При этом наиболее широкое распространение в вероятностных расчетах получили моменты двух видов: начальные и центральные.

Начальным моментом α_k порядка k называется математическое ожидание величины x^k:

$$\alpha_k = M\left[x^k\right]. \qquad (2.1)$$

Из формулы (2.1) следует, что при $k = 1$ имеем первый начальный момент, который соответствует математическому ожиданию случайной величины X.

Центральным моментом μ_k порядка k называется математическое ожидание центрированной величины $(x - m_x)^k$:

$$\mu_k = M\left[\left(x - m_x\right)^k\right], \tag{2.2}$$

где m_x – математическое ожидание случайной величины X. Отсюда видно, что центрированная случайная величина представляет отклонение от математического ожидания m_x. Из формулы (2.2) следует, что при $k = 2$ получаем генеральную (истинную) дисперсию случайной величины.

Таким образом, процесс центрирования, очень часто используемый в вероятностных расчетах, заключается в переносе начала координат в среднюю (центральную) точку, абсцисса которой совпадает с m_x. Так, из формулы (2.2) видно, что, например, второй центральный момент μ_2 соответствует дисперсии случайной величины X.

Между начальными и центральными моментами существует функциональная связь. Учитывая, что в теории вероятностей используются, в основном, первые четыре момента, связь между ними выражается следующими формулами:

$$\begin{aligned}
&\mu_1 = 0,\\
&\mu_2 = \alpha_2 - \alpha_1^2,\\
&\mu_3 = \alpha_3 - 3\alpha_1\alpha_2 + 2\alpha_1^3,\\
&\mu_4 = \alpha_4 - 4\alpha_1\alpha_3 + 6\alpha_1^2\alpha_2 - 3\alpha_1^4.
\end{aligned}$$

Суть метода моментов точечного оценивания неизвестных параметров заключается в приравнивании теоретических моментов соответствующим эмпирическим моментам того же порядка. Так, начальным эмпирическим моментом порядка k является выражение вида:

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k. \tag{2.3}$$

Теперь приравняем друг к другу начальный теоретический и эмпирический моменты первого порядка: $\boldsymbol{\alpha}_1 = \alpha_1$. Тогда с учетом формул (2.1) и (2.3) имеем:

$$M\left[x\right] = \bar{x}. \tag{2.4}$$

Отсюда следует, что точечной оценкой математического ожидания является среднее арифметическое, полученное по ограниченной выборке. Аналогичным образом, приравнивая центральный

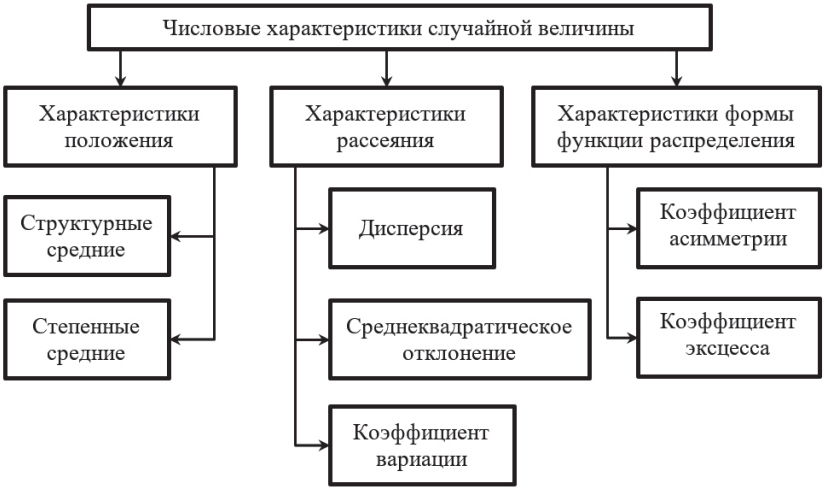

Рис. 2.2. Числовые характеристики случайной величины

теоретический и эмпирический моменты второго порядка друг к другу, можно получить точечную оценку дисперсии.

Вообще говоря, в зависимости от свойств случайных величин числовые характеристики могут быть разделены на несколько групп (рис. 2.2). Одна группа (математическое ожидание, медиана, мода и др.) определяет положение случайной величины на числовой оси и характеризует центр её группирования. Другая группа (дисперсия, амплитуда, интерквартильное расстояние и др.) показывает размах (масштаб) колебаний случайной величины и степень ее рассеяния от центра. Наконец, еще одна группа (коэффициенты асимметрии и эксцесса) является характеристикой формы функции распределения, определяя степень ее асимметрии и крутости.

2.2. Характеристики положения случайной величины

В общем случае различают 2 вида средних величин:

– степенные средние;
– структурные средние.

Степенные средние могут быть вычислены по следующей общей формуле:

$$\overline{x} = \frac{1}{n}\sqrt[m]{\sum x_i^m}, \qquad (2.5)$$

где m – показатель степени, x_i – текущее значение (вариант) усредняемого признака. В зависимости от величины m различают следующие виды степенных средних:

1) если $m = -1$, то имеем среднюю гармоническую $\left(\overline{x}_{\text{гар.}}\right)$;

2) если $m = 0$, то имеем среднюю геометрическую $\left(\overline{x}_{\text{геом.}}\right)$;

3) если $m = 1$ то имеем среднюю арифметическую $\left(\overline{x}_{\text{ар.}}\right)$;

4) если $m = 2$, то имеем среднюю квадратическую $\left(\overline{x}_{\text{кв.}}\right)$;

5) если $m = 3$, то имеем среднюю кубическую $\left(\overline{x}_{\text{куб.}}\right)$.

Из формулы (2.5) следует, что при использовании одних и тех же исходных данных, чем больше показатель степени m, тем больше значение средней величины, т. е.

$$\overline{x}_{\text{гар.}} \le \overline{x}_{\text{геом.}} \le \overline{x}_{\text{ар.}} \le \overline{x}_{\text{кв.}} \le \overline{x}_{\text{куб.}}. \qquad (2.6)$$

Это свойство называется *правилом мажорантности средних.* Естественно, что из указанных средних наибольшее распространение в статистике получила арифметическая средняя (выборочная средняя). Она применяется в форме простой средней и взвешенной средней.

Взвешенной выборочной средней дискретной случайной величины называется сумма произведений всех ее возможных значений (вариант) на их частоты (веса), т. е.:

$$\overline{x} = \frac{\left(x_1 f_1 + x_2 f_2 + \ldots + x_n f_n\right)}{\left(f_1 + \ldots + f_n\right)} = \frac{\sum x_i f_i}{\sum f_i}, \qquad (2.7)$$

где $f_i, \ldots, f_n$ – частоты. Если частоты случайной величины одинаковы ($f_1 = f_2 = \ldots = f_n$), что соответствует, например, проведению измерений в одинаковых условиях, то получаем *простую выборочную среднюю,* равную сумме отдельных значений выборки (вариант), деленной на общее число наблюдений:

$$\overline{x} = \frac{\left(x_1 + x_2 + \ldots + x_n\right)}{n} = n^{-1}\sum x_i. \qquad (2.8)$$

Среднее арифметическое значение характеризует *центр тяжести* (распределения) числового ряда.

Свойства выборочной средней

Свойство 1. Постоянный множитель *a* может быть вынесен за знак средней $\left(\overline{ax} = a\overline{x}\right)$.

Свойство 2. Средняя сумма равна сумме средних: $\overline{x+y} = \overline{x} + \overline{y}$.

Свойство 3. Сумма отклонений всех наблюденных данных от их средней равна нулю: $\sum_{i=1}^{n}\left(x_i - \overline{x}\right) = 0$.

Свойство 4. Сумма квадратов отклонений членов ряда от центра их тяжести достигает минимума по сравнению с аналогичной суммой, вычисленной относительно любого числа $a \neq \overline{x}$, т. е. $S = \sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \min$.

Свойство 5. Среднее арифметическое ряда, полученного путем объединения нескольких однородных статистических групп, образуется как среднее взвешенное значение частных средних, включенных в расчет с весами, равными объемам соединяемых совокупностей: $\overline{x} = \dfrac{\sum_{k=1}^{m} n_k \overline{x}_k}{\sum_{k=1}^{m} n_k}$.

Все перечисленные свойства среднего арифметического значения широко используются в выводах математической статистики и при решении различного рода практических задач.

Например, приведенное к длительному периоду значение средней арифметической по ряду многолетних наблюдений той или иной гидрометеорологической характеристики, называется *нормой*. Для вычисления норм по рекомендации Всемирной метеорологической организации необходимо, чтобы длина выборки составляла 30–40 лет.

Особым видом средних величин являются структурные средние. Они применяются для изучения внутреннего строения и структуры рядов распределения случайной величины. К структурным средним относятся мода и медиана.

Медианой называется значение случайной величины, которое соответствует среднему положению в ранжированном (вариационном) ряду. Если всем единицам ряда придать порядковые номера, то при нечетном числе членов ряда ($n = 2m + 1$) медиана определится

как $Me = x_{m+1}$. При четном числе членов ряда ($n = 2m$) за медиану условно принимается среднее значение между центральными значениями величин ранжированного ряда, т. е.

$$Me = \frac{1}{2}\left(x_m + x_{m+1}\right). \qquad (2.9)$$

Геометрически медиана – это абсцисса точки, в которой площадь, ограниченная кривой плотности вероятности, делится пополам (рис. 2.3). Сказанное означает, что справедливо следующее равенство $p(X < Me) = p(X > Me) = 0{,}5$.

Главное свойство медианы заключается в том, что сумма абсолютных отклонений членов ряда от медианы есть величина наименьшая:

$$\sum \left| x_i - Me \right| = \min .$$

Модой называется наиболее часто встречающаяся в данном статистическом ряду величина. Геометрически мода представляет собой наибольшую ординату кривой плотности вероятности в случае одновершинного распределения (рис. 2.3). Поэтому

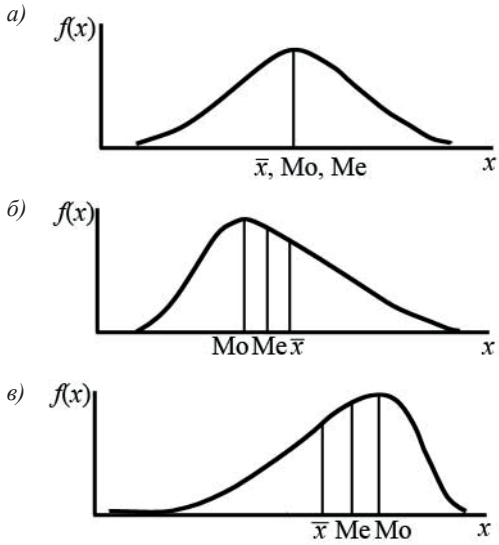

Рис. 2.3. Соотношение между средним арифметическим $(\bar{x})$, модой (*Mo*) и медианой (*Me*): *а* – симметричное распределение, *б* – положительно асимметричное распределение, *в* – отрицательно асимметричное распределение.

одновершинное распределение называют одномодальным. В тех случаях, когда распределение имеет несколько вершин, его называют многомодальным или полимодальным.

Для не очень асимметричных и одновершинных распределений мода может быть рассчитана по приближенному соотношению К. Пирсона:

$$Mo = \bar{x} + 3\left(Me - \bar{x}\right). \tag{2.10}$$

2.3. Характеристики рассеяния случайной величины

Простейшей мерой рассеяния (изменчивости) статистического ряда является размах (амплитуда) колебаний, определяемый как:

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

где $x_{\max}$, $x_{\min}$ – соответственно максимальный и минимальный члены ряда. Размах колебаний дает лишь самое общее представление об изменчивости, так как показывает, насколько отличаются друг от друга крайние значения, но не указывает, насколько велики отклонения отдельных значений внутри ряда.

Поэтому наиболее распространенными показателями изменчивости статистического ряда являются дисперсия, среднеквадратическое (стандартное) отклонение, коэффициент вариации, которые взаимосвязаны друг с другом.

Истинная оценка, т. е. полученная по генеральной совокупности, *дисперсии* дискретной случайной величины определяется по следующей формуле:

$$D = \frac{\sum_{i=1}^{n}\left(x_i - m_x\right)^2}{n} \tag{2.11}$$

и имеет размерность квадрата случайной величины. Выборочная оценка дисперсии иногда обозначается как σ^2 или s^2. Поскольку она имеет отрицательное смещение, то для выборок малого объема ($n < 25$–30), разделив (2.11) на поправку Бесселя $n/(n-1)$, получим:

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}. \tag{2.11а}$$

Это позволяет устранить смещение.

Среднеквадратическое отклонение случайной величины представляет собой корень квадратный из дисперсии и поэтому сохраняет размерность исходного ряда. В связи с этим при выполнении статистических расчетов данная величина обладает определенными преимуществами перед дисперсией.

Выборочный коэффициент вариации $C = \frac{\sigma}{\overline{x}}$ является безразмерной величиной, поэтому он очень удобен для анализа изменчивости рядов с различной размерностью.

Свойства дисперсии

Свойство 1. Дисперсия постоянной величины равна нулю $\sigma^2(a) = 0$.

Свойство 2. Дисперсия остается постоянной, если все члены ряда увеличить или уменьшить на одно и то же число $\sigma^2(a + x) = \sigma_x^2$.

Свойство 3. Постоянную величину можно выносить за знак дисперсии, возведя ее в квадрат $\sigma^2(ax) = a^2\sigma_x^2$.

Свойство 4. Дисперсия алгебраической суммы независимых случайных величин равна сумме их дисперсий:

$$\sigma^2\left(x_1 \pm x_2 \pm ... \pm x_n\right) = \sigma_{x_1}^2 + \sigma_{x_2}^2 + ... + \sigma_{x_n}^2 .$$

Свойство 5. Дисперсия суммы двух связанных между собой корреляционной зависимостью случайных величин определяется как:

$$\sigma^2\left(E + y\right) = \sigma_x^2 + \sigma_y^2 + 2\sigma_x\sigma_y r_{xy}, \qquad (2.12)$$

где r_{xy} – коэффициент корреляции между переменными x и y, получаемый по формуле (6.1).

Свойство 6. Дисперсия относительно среднего арифметического значения меньше, чем средний квадрат отклонений от любого значения в ряду x_i, на величину $\left(x_i - \overline{x}\right)^2$.

Свойство 7. Если некоторая величина y_i связана с x_i линейным уравнением $y_i = ax_i + b$, то $\sigma_y^2 = a^2\sigma_x^2$.

Перечисленные выше свойства дисперсии, как и свойства средней арифметической, широко используются в выводах математической статистики и при решении многих практических задач.

В некоторых случаях в качестве меры рассеяния используется среднее линейное отклонение, которое представляет собой среднюю арифметическую абсолютных значений отклонений отдельных

вариантов от их выборочной средней. Для несгруппированных данных она вычисляется как $d = \dfrac{\sum |x_i - \overline{x}|}{n}$, а для сгруппированных данных – по формуле $d = \dfrac{\sum |x_i - \overline{x}| f_i}{\sum f_i}$. Применение величины d оправданно в тех случаях, когда суммирование показателей без учета знаков имеет экономический смысл.

2.4. Характеристики формы кривой распределения случайной величины

Рассмотренные в предыдущих разделах характеристики положения и рассеяния случайной величины не дают представления о форме ее кривой распределения. Так, две случайные величины, имея одинаковые средние арифметические значения и дисперсии, могут обладать совершенно различными распределениями вероятностей. Это обстоятельство обусловливает необходимость введения для описания случайных величин таких характеристик, которые позволяют оценить степень асимметрии и крутости распределения.

Характеристикой асимметрии (скошенности) распределения случайной величины X является выборочный коэффициент асимметрии:

$$As = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^3}{n\sigma^3}. \qquad (2.13)$$

Нетрудно видеть, что коэффициент асимметрии величина безразмерная. Если члены ряда располагаются симметрично относительно среднего значения, то разные по величине положительные и отрицательные отклонения от среднего повторяются одинаково часто. В этом случае $As = 0$ и $\overline{x} = Mo = Me$ (рис. 2.3-*а*). В практических расчетах скошенность распределения обычно считается слабой, если $|As| < 0{,}25$, умеренной при $0{,}25 < |As| < 0{,}5$ и сильной, если $|As| > 0{,}5$.

При положительной асимметрии ($As > 0$) ряд будет включать немногочисленные, но большие по величине положительные отклонения, и более многочисленные, но менее значительные по величине отрицательные отклонения. В результате $\overline{x} > Mo$ и $\overline{x} > Me$ (рис. 2.3-*б*). При отрицательной асимметрии ($As < 0$) ряд будет

включать немногочисленные, но большие по величине отрицательные отклонения от среднего, и более многочисленные, но малые по величине положительные отклонения. Поэтому $\bar{x} < Mo$ и $\bar{x} < Me$ (рис. 2.3-*в*). Нетрудно запомнить следующее правило: при отрицательной (левой) асимметрии порядок следования слов *mean, median, mode* такой же, как в английском словаре. При правой асимметрии – наоборот.

Характеристикой крутости (островершинности или плосковершинности) кривой распределения служит выборочный коэффициент эксцесса:

$$Ex = \frac{1}{n\sigma^4}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4 - 3. \qquad (2.14)$$

Коэффициент эксцесса является безразмерной величиной и характеризует отклонение крутости эмпирической кривой от нормальной кривой распределения, причем при их совпадении принимается $Ex = 0$.

Если эмпирическая кривая распределения является более островершинной по сравнению с нормальной кривой, то $Ex > 0$ (рис. 2.4). В результате мода эмпирического распределения должна быть больше моды нормального распределения ($Mo > Mo_N$).

Если эмпирическая кривая распределения является более плосковершинной по сравнению с нормальной кривой, то $Ex < 0$ (рис. 2.4). В этом случае $Mo < Mo_N$.

Если предельная величина отрицательного эксцесса равна –2, то его положительная величина теоретически может быть сколь

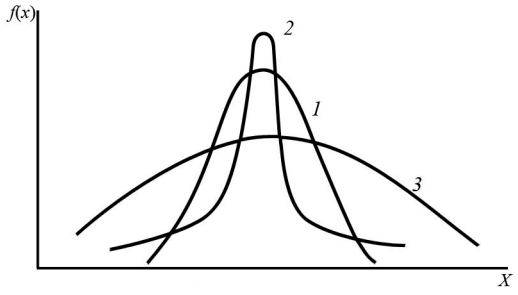

Рис. 2.4. Виды распределений с различной крутостью:
1) нормальное распределение; *2*) распределение с положительным эксцессом;
3) распределение с отрицательным эксцессом

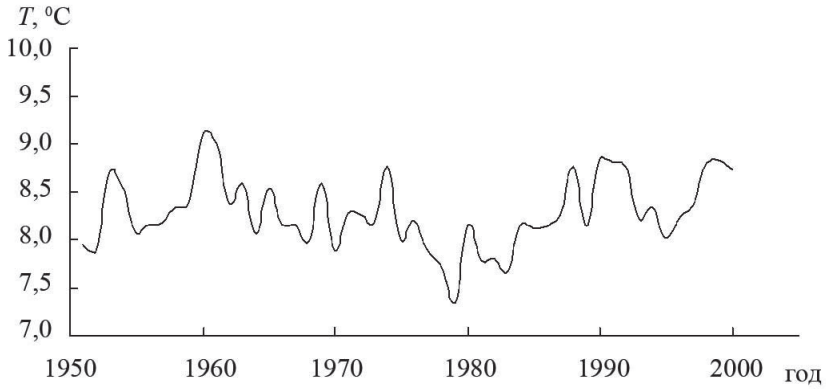

Рис. 2.5. Межгодовой ход температуры поверхности океана в районе судна погоды «М»

угодно большой. Обычно эксцесс считается незначительным, если выполняется условие $|Ex| < 0{,}5$.

Пример 2.1. Оценим числовые характеристики среднемесячных значений температуры поверхности океана (ТПО) в районе судна погоды «М» (66° с.ш. и 2° в.д.) за период 1951–2000 гг., которое расположено практически в центре Норвежского моря. Отметим, что непрерывные наблюдения за комплексом гидрологических и метеорологических характеристик в течение более полувека на судне погоды «М», безусловно, носят уникальный характер. Основные статистические характеристики ТПО для отдельных месяцев и года в целом представлены в таблице 2.1, а ее межгодовой ход дан на рис. 2.5.

Представленные в таблице 2.1 статистические характеристики позволяют анализировать одновременно внутригодовую и межгодовую изменчивость ТПО. Нетрудно видеть, что температура воды имеет довольно хорошо выраженный годовой ход, обусловленный годовым притоком солнечной радиации и запаздывающий от него на два месяца. Максимальные значения ее наблюдаются в августе, а минимальные – в феврале и марте.

Годовой ход размаха колебаний ТПО в общем повторяет годовой ход средней арифметической, но внутригодовая амплитуда его более чем в три раза меньше (соответственно 1,5 и 5,3 °С). Годовой ход среднеквадратического отклонения ТПО почти повторяет годовой ход величины R. Аналогичный характер годового хода

Таблица 2.1

Первичные статистические оценки среднемесячных значений температуры поверхности океана в районе судна погоды «М» за период 1951–2000 гг.

Месяц	Среднее, °С	*Me*, °С	σ, °С	*C*	x_{max}, °С	x_{min}, °С	*R*, °С	*As*	*Ex*
I	6,7	6,7	0,4	0,06	7,4	5,8	1,6	–0,21	–0,86
II	6,4	6,3	0,4	0,06	7,4	5,6	1,8	0,26	–0,49
III	6,4	6,4	0,4	0,06	7,3	5,5	1,8	–0,10	1,04
IV	6,5	6,5	0,4	0,06	7,2	5,6	1,6	–0,29	–0,45
V	7,4	7,4	0,4	0,05	8,2	6,6	1,6	0,03	–0,24
VI	9,1	8,9	0,7	0,08	10,9	7,9	3,0	0,68	0,04
VII	10,8	10,6	0,8	0,07	12,6	9,5	3,1	0,43	–0,73
VIII	11,7	11,5	0,8	0,07	13,6	10,2	3,4	0,54	–0,24
IX	10,7	10,8	0,7	0,07	11,9	9,0	2,9	–0,27	–0,58
X	9,0	8,9	0,6	0,07	10,3	8,1	2,2	0,46	–0,52
XI	7,8	7,7	0,5	0,06	9,1	6,9	2,2	0,72	0,29
XII	7,1	7,1	0,4	0,06	8,2	6,2	2,0	0,05	–0,48
Год	8,3	8,2	0,4	0,05	9,1	7,3	1,8	0,11	–0,19

в значениях $\bar{X}$ и σ обусловливает почти постоянство коэффициента вариации в течение всего года.

Межгодовая изменчивость ТПО невелика и практически одинакова для всех месяцев года. Действительно, различие между максимальным (июнь) и минимальным (май) коэффициентом вариации составляет лишь 0,03. Особенности распределения коэффициентов асимметрии и эксцесса и расхождение между средней и медианой позволяют выяснить особенности «поведения» эмпирической кривой плотности вероятности. Прежде всего, отметим, что только средние годовые значения ТПО имеют оценки *As* и *Ex* сравнительно мало отличающиеся от нуля, т. е. распределение средних годовых значений ТПО является близким к нормальному закону (см. гл. 3). Этого нельзя сказать в отношении всех месяцев года. Даже когда *As* мал, то *Ex* весьма велик (например, март) и наоборот (например, июнь).

В распределении асимметрии преобладают положительные значения *As*. Это означает, что в течение каждого из восьми месяцев временной ряд включает немногочисленные, но большие по величине положительные отклонения, и более многочисленные, но менее значительные по величине отрицательные отклонения. Отсюда следует, что должно выполняться неравенство $\bar{x} > Me$. Из таблице 2.1

видно, что при больших значениях *As* оценки среднего превышают оценки медианы на 0,1–0,2 °С. При $As < 0$ значения $\bar{x}$ должны быть меньше медианы. Но поскольку отрицательные оценки *As* невелики, то данное условие из четырех месяцев отмечается только в сентябре.

В распределении оценок *Ex* преобладают отрицательные значения. Это означает, что эмпирическая кривая распределения является более плосковершинной по сравнению с нормальной кривой. Только для трех месяцев выполняется условие $Ex > 0$, когда кривая распределения является более островершинной по сравнению с нормальной кривой.

2.5. Интервальное оценивание числовых характеристик

Естественно, что точечные оценки параметра θ в действительности являются приближенными значениями истинного неизвестного параметра **θ** даже в случае их несмещенности, эффективности и состоятельности. В связи с этим возникает вопрос: как сильно может отклоняться эта приближенная оценка от истинного значения? Другими словами, нельзя ли указать интервал вида $[\theta_1, \theta_2]$, который бы с заданной вероятностью, близкой к единице, накрывал неизвестную нам оценку истинного значения параметра **θ**? Такой интервал принято называть *доверительным интервалом* для параметра **θ**, а концы его называются *доверительными границами.* Поскольку доверительные границы находятся по выборочным данным, то они являются случайными величинами в отличие от оцениваемого параметра **θ** – величины неслучайной. Следовательно, *доверительный интервал – это область значений случайной величины внутри доверительных границ.*

Именно в построении доверительного интервала состоит суть интервальных оценок выборочных параметров, которые позволяют судить о степени разброса оценок выборочного параметра, внутри которого с высокой надежностью находятся его истинное неизвестное значение. Аналитически интервальная оценка произвольного выборочного параметра θ, имеющего некоторое теоретическое распределение, может быть записана в виде:

$$p\left(\theta_{н} < \boldsymbol{\theta} < \theta_{в}\right) = \gamma = 1 - \alpha, \qquad (2.15)$$

где **θ** – истинная оценка параметра θ, $\theta_{н}$ и $\theta_{в}$ – соответственно нижняя и верхняя доверительные границы, т. е. такие значения случайной

величины, выход за пределы которых имеет наперед заданную *доверительную вероятность* (надежность) γ, α – уровень значимости, представляющий собой вероятность события, которым решено пренебречь.

При симметричности доверительного интервала относительно оценки **θ** его нижняя и верхняя границы определяются как: $\theta_н = \boldsymbol{\theta} - \varepsilon$, $\theta_в = \boldsymbol{\theta} + \varepsilon$, где ε – половина длины доверительного интервала при заданном уровне значимости, означающим вероятность принятия ошибочного решения. Величину $|\theta - \boldsymbol{\theta}|$ можно рассматривать как возможную абсолютную ошибку оценки, полученной по данной выборке, а величина ε – это, по существу, предельная ошибка, которая может быть получена при оценке неизвестного параметра **θ** по данному ряду наблюдений. Иногда ошибка ε называется ошибкой репрезентативности выборки.

При установлении доверительных интервалов требуется знать закон распределения случайной величины. Особенно это касается малых объемов выборки (короткого временного ряда), поскольку для ее больших объемов можно условно принимать нормальность распределения, к которому асимптотически приближается случайная величина при $n \to \infty$.

Интервальной оценкой математического ожидания m_x нормально распределенной случайной величины X при известном среднем квадратическом отклонении генеральной совокупности (**σ**) служит доверительный интервал вида:

$$\bar{x} - z\left(\frac{\sigma}{n^{1/2}}\right) < m_x < \bar{x} + z\left(\frac{\sigma}{n^{1/2}}\right), \qquad (2.16)$$

где z – значение аргумента функции Лапласа $\Phi(z)$ (см. Приложение 1), при котором $\Phi(z) = \frac{(1-\alpha)}{2}$. Данный критерий на практике используется весьма редко, так как истинная (генеральная) оценка величины **σ** неизвестна.

Интервальная оценка математического ожидания нормально распределенной случайной величины при неизвестном генеральном стандартном отклонении определяется по формуле:

$$\bar{x} - t_\alpha\left[\frac{s}{(n)^{1/2}}\right] < m_x < \bar{x} + t_\alpha\left[\frac{s}{(n)^{1/2}}\right], \qquad (2.17)$$

где s – выборочная оценка генерального стандартного отклонения **σ**, рассчитываемая как: $s = \left[\frac{\sum (x_i - \bar{x})}{(n)} \right]^{1/2}$, t_α – критерий Стьюдента при заданном уровне значимости α и числе степеней свободы $\nu = n - 1$.

Из этих формул видно, что ширина доверительного интервала при заданном уровне значимости зависит от объема выборки. С ростом n она суживается и при $n \to \infty$ выборочная оценка параметра превращается в истинную оценку. Наоборот, с уменьшением n доверительный интервал расширяется, причем при $n \to 0$ $\bar{x} \to \infty$. Таким образом, смысл интервальной оценки состоит в том, что она *представляет собой статистическую ошибку оцениваемого параметра, обусловленную ограниченностью выборки.*

Отметим, что при достаточно больших значениях n доверительные границы, рассчитанные по формулам (2.16) и (2.17), почти не отличаются между собой. Это связано с тем, что исходя из центральной предельной теоремы среднее арифметическое $\bar{X}$ случайных величин $X_1, X_2, \ldots, X_n$ при увеличении n стремится к нормальному закону распределения. Однако при малых значениях n расхождения в доверительных границах могут быть заметными. На практике для построения интервальной оценки математического ожидания при любых n обычно ограничиваются формулой (2.17).

Рассмотрим теперь интервальные оценки для дисперсии. Поскольку ее величина распределена по закону χ^2, то доверительный интервал для генеральной дисперсии нормально распределенной случайной величины X при известном математическом ожидании рассчитывается по формуле:

$$p\left(\frac{ns^2}{\chi_2^2} < D_x < \frac{ns^2}{\chi_1^2} \right) = 1 - \alpha, \qquad (2.18)$$

где s^2 – выборочная оценка дисперсии, D_x – истинная (генеральная) оценка дисперсии, χ^2_2 и χ^2_1 – табличные значения статистики χ^2 при числе степеней $\nu = n - 1$, причем $\chi_2^2 = \chi^2_{\frac{\alpha}{2}}$, $\chi_1^2 = \chi^2_{1-\frac{\alpha}{2}}$. В том случае, если математическое ожидание неизвестно, то формула (2.18) преобразуется к виду:

$$p\left[\frac{ns^2}{\chi_2^2} < D_x < \frac{ns^2}{\chi_1^2}\right] = 1 - \alpha \qquad (2.19)$$

или

$$\frac{ns^2}{\chi^2_{1-\frac{\alpha}{2}}} < D_x < \frac{ns^2}{\chi^2_{\frac{\alpha}{2}}}. \qquad (2.19')$$

Нетрудно видеть, что расхождения в доверительных границах, рассчитанных по обеим формулам, становятся пренебрежимо малыми при возрастании величины n. Чтобы получить интервальную оценку для стандартного отклонения, достаточно в формуле (2.19′) извлечь квадратный корень.

Заметим, что интервальная оценка среднего квадратического отклонения нормально распределенной случайной величины X может быть получена также, исходя из следующих формул:

$$s(1 - q) < \boldsymbol{\sigma}_x < s(1+ q) \text{ (при } q < 1), \qquad (2.20)$$

$$0 < \boldsymbol{\sigma}_x < s(1+ q) \text{ (при } q > 1). \qquad (2.20')$$

Для определения величины q может быть использована специальная таблица 2.2, входными параметрами которой являются надежность γ и длина ряда n.

Таблица 2.2

Значения q в зависимости от длины ряда n и надежности γ

n	γ		n	γ	
	0,95	0,99		0,95	0,99
5	1,37	2,67	17	0,42	0,66
6	1,09	2,01	18	0,40	0,63
7	0,92	1,62	20	0,37	0,58
8	0,80	1,38	25	0,32	0,49
9	0,71	1,20	30	0,28	0,43
10	0,65	1,08	35	0,26	0,38
11	0,59	0,98	40	0,24	0,35
12	0,55	0,90	50	0,21	0,30
13	0,52	0,83	60	0,188	0,269
14	0,48	0,78	80	0,164	0,226
15	0,46	0,73	100	0,143	0,198
16	0,44	0,70	200	0,099	0,136

Пример 2.2. Рассмотрим построение доверительных интервалов для средних годовых значений солености поверхностного слоя воды на одной из прибрежных станций Баренцева моря, если известно, что среднее арифметическое солености $\overline{S}$ = 34,2 ‰, среднее квадратическое отклонение s = 21,8 ‰, период наблюдений n = 28 лет. Столь высокая изменчивость солености связана со значительными колебаниями морского льда в прибрежной зоне моря. В предположении нормального распределения исходных данных при построении доверительного интервала для математического ожидания солености воспользуемся формулой (2.17). Из таблицы распределения Стьюдента, принимая α = 0,05 и ν = n − 1 = 27, находим t_α = 2,05. После этого определяем нижнюю и верхнюю доверительные границы:

$$\overline{x} - t_\alpha \left[\frac{s}{(n)^{1/2}} \right] = 34,2 - 2,05 \times \frac{21,8}{(28)^{1/2}} = 25,6 \text{ ‰},$$

$$\overline{x} + t_\alpha \left[\frac{s}{(n)^{1/2}} \right] = 34,2 + 2,05 \times \frac{21,8}{(28)^{1/2}} = 42,8 \text{ ‰}.$$

Итак, получаем 25,6 < m_S < 42,8 ‰. Нетрудно видеть, что математическое ожидание солености находится в довольно широких доверительных границах. С одной стороны, это связано с относительно малой длиной выборки, а с другой – со значительной межгодовой изменчивостью, обусловленной колебаниями притока пресных вод и морского льда.

При построении доверительного интервала для среднего квадратического отклонения солености воспользуемся формулами (2.19) и (2.20). В первой из них неизвестными параметрами являются χ_{1*} и χ_{2*}. Из распределения χ^2 по значениям α = 0,05 и ν = n − 1 = 27 находим $\chi^2_{1-\frac{\alpha}{2}} = 14,6$ и $\chi^2_{\frac{\alpha}{2}} = 43,2.$ Теперь определяем нижнюю и верхнюю доверительные границы:

$$\frac{(n)^{1/2} s}{\chi_{2*}} = \frac{21,8(28)^{1/2}}{(43,2)^{1/2}} \approx 17,5 \text{ ‰},$$

$$\frac{(n)^{1/2} s}{\chi_{1*}} = \frac{21,8(28)^{1/2}}{(14,6)^{1/2}} \approx 30,2 \text{ ‰}.$$

Итак, на уровне значимости α = 0,05 генеральное значение среднего квадратического отклонения годовых значений солености находится в интервале $17{,}5 < \boldsymbol{\sigma}_S < 30{,}2$ ‰.

Неизвестным параметром в формуле (2.20) является величина q. По таблице 2.2 при $\gamma = 0{,}95$ и $n = 28$ находим, что $q = 0{,}30$. Отсюда получаем $15{,}26 < \boldsymbol{\sigma}_S < 28{,}34$ ‰. Нетрудно видеть, что оба доверительных интервала при заданных параметрах имеют почти одинаковую ширину, однако интервал, рассчитанный по формуле (2.20) несколько смещен в сторону более низких значений.

2.6. Понятие о толерантных интервалах

В отличие от доверительных интервалов, устанавливающих пределы изменчивости отдельных выборочных параметров случайной величины в зависимости от длины выборки, представляет интерес нахождение таких интервалов, которые показывают пределы случайной изменчивости всей рассматриваемой выборочной совокупности, т. е. определяют степень репрезентативности выборки. Это связано с тем, что сама генеральная совокупность нам, как правило, неизвестна. Данная задача может быть решена с помощью *толерантных* (допустимых) *интервалов*.

Примем, что случайная величина X распределена по нормальному закону с известными выборочными характеристиками $\bar{x}$ и σ. Нетрудно задать такие пределы:

$$u_1 = \bar{x} - k\sigma,\ u_2 = \bar{x} + k\sigma,$$

что с вероятностью p можно гарантировать попадание в них доли генеральной совокупности, не меньшей заданного предела Q. Эти пределы называются допустимыми (толерантными). Параметр k является функцией длины выборки n, Q и p:

$$k = k_\infty \left[1 + \frac{x_p}{(2n)^{1/2}} + \frac{\left(\sigma x^2{}_p + 10\right)}{12n} \right], \qquad (2.21)$$

где k_∞ − истинное значение k, соответствующее математическому ожиданию и истинной оценке дисперсии случайной величины X. Из свойств нормального распределения следует, что $2F_0(t) = Q$, а $0{,}5 - F_0(t) = 1 - p$. Таким образом, задавая величину Q можно определить для некоторой произвольной выборки пределы u_1 и u_2, в которых с вероятностью p заключена доля Q всей генеральной

совокупности. Однако толерантные интервалы не получили в статистике широкого распространения, ибо, как правило, априори величина Q неизвестна.

2.7. Понятие о малой выборке и квантильном анализе

Отметим, что рассмотренные выше числовые характеристики случайной величины имеют высокую надежность только при сравнительно большой длине выборки. С уменьшением длины выборки и, особенно под влиянием выбросов оценка первичных статистических характеристик довольно быстро теряет эффективность. Так, оценка медианы является более эффективной по сравнению со средним значением, так как она мало зависит от длины выборки. Еще менее устойчивой оказывается величина дисперсии, которая очень сильно зависит от длины ряда и возможных выбросов случайной величины. И уже совсем неустойчивыми оказываются оценки коэффициентов асимметрии и эксцесса. Таким образом, для коротких статистических рядов (малой выборки) желательны специальные методы оценивания, к которым относятся методы *непараметрической статистики.* Достоинством их является то, что они не привязаны к теоретическим законам распределения и наибольшую эффективность имеют как раз применительно к малым выборкам.

Однако, в статистике нет строгого определения малой выборки. Интуитивно понятно, что выборка длиной 10 значений является малой, а 100 значений – большой. Возникает вопрос, где провести границу малой выборки?

Учитывая, что закон распределения представляет собой важнейшую характеристику случайной величины, в качестве малой выборки можно считать такую, когда при обработке ее методами, основанными на группировке наблюдений, нельзя достичь заданной точности и достоверности. Однако данное определение вряд ли можно признать универсальным. В статистике есть множество задач, не связанных с оценкой функции распределения исходных данных. Например, как будет показано в разделе 6, при расчете коэффициентов корреляции вполне достаточно длины рядов $n = 30...35$.

Поэтому, возможно, более универсальным является следующее определение без привязки к закону распределения: *выборка является малой, если рассчитанные на ее основе стандартными методами статистические параметры не отвечают заданной точности*

и достоверности. Но и в данном случае присутствует некоторая доля субъективизма, ибо понятие заданной точности и достоверности являются неоднозначными и могут быть различными в зависимости от поставленной задачи даже для одной и той же выборки. Так, для выборки длиной $n = 40$ сложно построить надежную эмпирическую функцию распределения, но в то же время рассчитанные по ней коэффициенты корреляции оказываются достаточно точными. На практике довольно часто в качестве условной верхней границы малой выборки принимают $n < 25...30$ значений. Итак, *для малой выборки невозможно использовать методы группирования данных, нет смысла вычислять статистические моменты выше второго порядка и необходимо применять специальные методы анализа данных.*

Одним из таких специальных методов является *квантильный анализ,* который относится к методам *теории порядковых статистик.* Для вариационного ряда, расположенного в порядке возрастания его значений, *i*-й по порядку член называется *i*-й порядковой статистикой ряда объемом *n*. Любая порядковая статистика представляет собой функцию всех элементов выборки. При изменении ее объема порядковые статистики могут существенно измениться. Первой работой по математической теории порядковых статистик считается статья К. Пирсона, опубликованная в 1902 г. Однако наиболее интенсивное развитие она получила во второй половине XX столетия. Что касается квантильного нализа, то наибольшую известность он получил благодаря работам американского статистика Тьюки.

Квантилю, отвечающему заданному уровню вероятности *p*, соответствует такое значение $x = x_p$, при котором функция распределения принимает значение, равное *p*, т. е.

$$F(x_p) = p. \tag{2.22}$$

Отсюда следует, что выборочный квантиль x_p порядка *p* представляет собой элемент вариационного ряда $x^{(1)}, x^{(2)}, \dots, x^{(n)}$, полученного в результате преобразования выборки $x_1, x_2, \dots, x_n$. В статистической практике используют ряд квантилей, имеющих специальные названия:

– *персентили*: $P_1, P_2, \dots, P_{99}$ – квантили порядков 0,01; 0,02; … ; 0,99;

– *децили*: $D_1, D_2, \dots, D_9$ – квантили порядков 0,10; 0,20; … ; 0,90;

– *квартили*: Q_1, Q_2, Q_3 – квантили порядков 0,25, 0,50, 0,75.

Нетрудно видеть, что вариационный ряд делится тремя квартилями на четыре равные части: Q_1 или $x_{0,25}$ – значение, ниже которого лежит 25 % наблюдений, Q_2 или $x_{0,50}$ – 50 % наблюдений, Q_3 или $x_{0,75}$ – 75 % наблюдений. Указанные квартили имеют особые названия. Так, медианой называется квартиль, отвечающий доверительной вероятности p = 0,5, т. е. $x_{0,50}$. Вероятностям p = 0,25 и p = 0,75 соответствуют *нижний* $x_{0,25}$ и *верхний* $x_{0,75}$ *квартили.* Разность $Q = x_{0,75} - x_{0,25}$ называется *интерквартильным расстоянием.* Наиболее часто в вероятностных расчетах используются следующие порядковые статистики: $x_{\min}$, $x_{\max}$, $x_{0,25}$, $x_{0,75}$, $x_{0,50}$ и др.

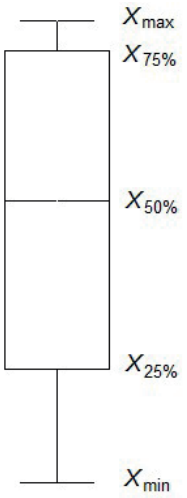

Рис. 2.6. Графическое представление «ящика с усами»

Наглядной формой представления результатов квантильного анализа является предложенный Тьюки так называемый «*ящик с усами*» (рис. 2.6). Для его построения чертится прямоугольник, верхняя и нижняя стороны которого соответствуют $x_{0,25}$ и $x_{0,75}$, а медиане соответствует поперечная черта. К ящику пристраиваются усы, т. е. отрезки, соединяющие каждый сгиб с соответствующим крайним ($x_{\min}$ или $x_{\max}$) значением выборки.

Несмотря на видимую простоту построения «ящика с усами», в нем содержится большое количество полезной информации о выборке. Действительно, медиана характеризует центр распределения. В некоторых случаях для придания центру распределения еще большей устойчивости используется так называемое *трехсреднее значение,* определяемое как:

$$\bar{X}_3 = \frac{\left(x_{0,25} + 2Me + x_{0,75}\right)}{4}. \qquad (2.23)$$

Основной характеристикой рассеяния служит интерквартильное расстояние, представляющее аналог среднеквадратического отклонения. Кроме того, другой характеристикой рассеяния служит размах колебаний $R = x_{\max} - x_{\min}$. Дополнительно более подробно изменчивость выборки может быть проанализирована при построении так называемых «барьеров», представляющих прямые линии, перпендикулярные к «усам». Внутренние барьеры отстоят от верхней

и нижней границ ящика на расстоянии 1,5Q, внешние барьеры – на расстоянии 3Q. Для случайной выборки, имеющей нормальное распределение, между внутренними барьерами содержится 99 % значений выборки, а между внешними – 99,9997 %. Отметим также, что при нормальном распределении данных между интерквартильным расстоянием и среднеквадратическим отклонением существует следующее соотношение:

$$Q = 1,34\sigma. \tag{2.24}$$

Кроме того, на основе квартилей может быть вычислен показатель асимметрии, формула для которого имеет вид:

$$As = \frac{\left(x_{0,75} + x_{0,25} - 2x_{0,50}\right)}{\left(x_{0,75} - x_{0,25}\right)}. \tag{2.25}$$

Пример 2.3. В течение 1979–1990 гг. (n = 12) в юго-восточной части Тихого океана, ограниченной по широте 30 и 45° ю.ш., а по долготе 80 и 105° з.д. судами бывшего Советского Союза осуществлялся круглогодичный промысел ставриды. В отдельные годы ее вылов превышал 1 млн т. Рассмотрим распределение «ящиков с усами» вылова рыбы для всех месяцев года (рис. 2.7), которые рассчитывались исключительно по фактическим данным, т. е. с учетом пропусков. В некоторые месяцы (сентябрь–ноябрь) число пропусков достигало 5 значений. В этих случаях длина ряда сокращалась до n = 7. Учитывая слишком короткую длину исходных рядов, барьеры не строились.

Из рис. 2.7 видно, что среднемесячные данные вылова ставриды имеют весьма сложную внутреннюю структуру, существенно неодинаковую для различных месяцев года. Прежде всего, следует отметить, что в статистических оценках вылова рыбы практически отсутствует годовой ход. Так, медиана достаточно случайно меняется в течение года. Ее максимальное значение отмечается в январе, а минимальное – в августе. Интерквартильное расстояние также испытывает хаотические изменения. Максимальное значение Q наблюдается в августе, а минимальное – в марте. Кроме того, заметно меняется при переходе от одного месяца к другому соотношение между медианой, интерквартильным расстоянием и размахом колебаний. Например, в октябре отмечается максимальный размах в оценках вылова рыбы, в то время как интерквартильное расстояние существенно меньше, чем в августе.

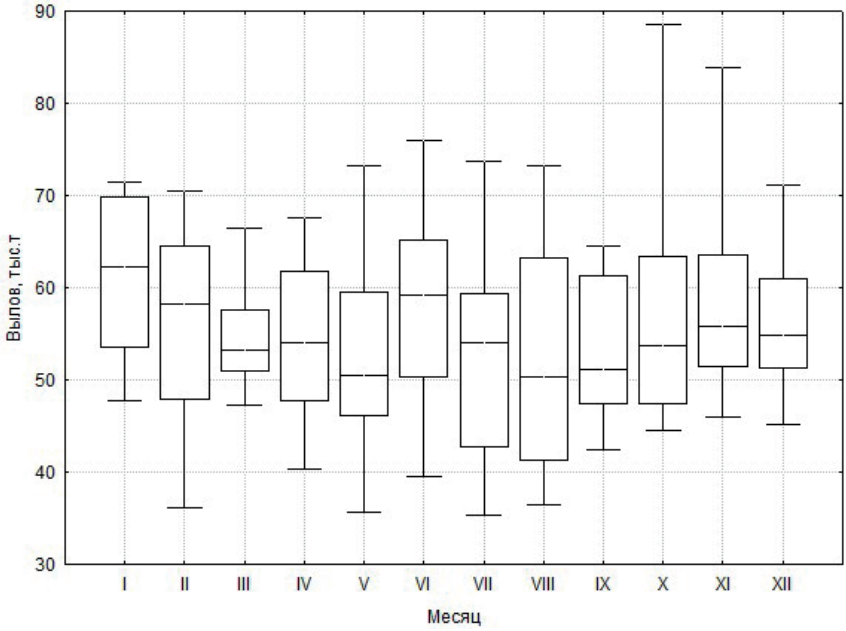

Рис. 2.7. Квантильный анализ вылова ставриды в юго-восточной части Тихого океана для отдельных месяцев года за период 1979–1990 гг.

Глава 3. Законы распределения случайной величины

Построение закона распределения – это один из наиболее простых и одновременно универсальных способов обобщения и анализа эмпирических данных, позволяющий в аналитическом виде представить их основные закономерности и внутреннюю структуру. Как известно, *в математической статистике случайная величина считается заданной, если известна ее функция распределения.* Этим обстоятельством определяется фундаментальное значение законов распределения. В настоящее время известно очень большое число самых разнообразных законов распределения.

Очевидно, основную их массу можно разделить на 2 группы: первая, наиболее многочисленная, включает законы распределения, которые непосредственно используются для обобщения эмпирических

данных. Вторая группа – это те законы, которые применяются в статистических расчетах (например, законы Фишера, Стьюдента и др.) при построении разного рода оценок, критериев и т.п. Особое место среди всех законов распределения принадлежит нормальному закону, выведенному немецким математиком Гауссом в результате изучения им ошибок при стрельбе артиллерийскими снарядами.

3.1. Нормальный закон распределения

Случайная величина X считается распределенной по нормальному закону (закону Гаусса), если ее плотность вероятности высчитывается по следующей формуле:

$$N(\bar{x}, s_x) = f(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \exp\left(-\frac{(x-\bar{x})^2}{2\sigma_x^2} \right), \qquad (3.1)$$

где $\bar{x}$ – среднее значение переменной X, σ^2_x – выборочная дисперсия. Таким образом, формула (3.1) отражает эмпирическую плотность вероятности, рассчитанную по выборочным данным. При замене среднего значения математическим ожиданием, а выборочной дисперсии – генеральной, получаем теоретическую (истинную) плотность вероятности.

Как видно из данной формулы, для выборочной совокупности достаточно знать всего два параметра, а именно – среднее значение и стандартное отклонение ($\bar{x}$ и σ_x), чтобы нормальный закон распределения считался заданным. Из формулы (3.1) следует, что нормальная кривая $f(x)$ располагается симметрично относительно максимальной ординаты, равной $f(x)_{\max} = \frac{1}{s_x (2\pi)^{1/2}}$ и проходящей через $\bar{x}$ (рис. 3.1). При $\bar{x} = 0$ нормальная кривая будет симметрична началу координат.

Если положить σ_x = const, но изменять параметр $\bar{x}$, то кривая нормального распределения будет смещаться параллельно оси абсцисс, не меняя своей формы. При изменении параметра σ_x $(\bar{x} = \text{const})$ происходит изменение формы кривой нормального закона. Если этот параметр увеличивается, то максимальное значение функции $f(x)$ убывает, и наоборот. Так как площадь, ограниченная кривой распределения и осью Ox, должна быть постоянной и равной 1, то с увеличением параметра кривая приближается к оси Ox и

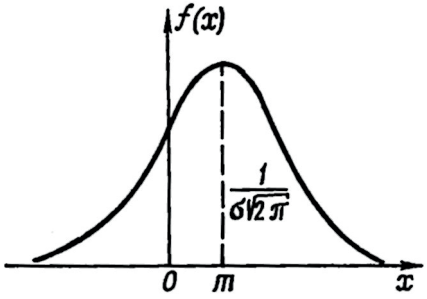

Рис. 3.1. Функция плотности вероятности нормального закона распределения.

растягивается вдоль нее, а с уменьшением σ кривая распределения сжимается с боков и вытягивается вдоль оси ординат.

Преобразуем выражение (3.1) к интегральному виду:

$$F(x) = \int_{-\infty}^{x} f(x)dx = \frac{1}{\sigma_x \sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{(x-\bar{x})^2}{2\sigma_x{}^2} \right] dx. \qquad (3.2)$$

Интеграл, входящий в эту формулу, аналитически определить нельзя, так как он через элементарные функции не выражается, но может быть вычислен путем замены переменной. Произведем замену переменной следующим образом:

$$t = \frac{(x-\bar{x})}{\sigma x},$$

где t – *стандартизированная случайная величина,* обладающая тем важным свойством, что при любом распределении случайной величины ее среднее значение равно нулю, а дисперсия – единице. С учетом данной формулы имеем:

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{0}^{t} \exp\left(\frac{-t^2}{2} \right) dt. \qquad (3.3)$$

Переход от переменной x к t по существу означает перенос начала координат в центр распределения и выражение абсциссы в долях от стандартного отклонения. Данный интеграл, выражающий площадь под нормальной кривой в интервале [0, t], носит название *функции Лапласа.* Численные значения его в пределах от $t = 0$ до $t = 5$ приводятся в Приложении 1. Заметим, что интегральная

теоретическая функция нормального распределения может быть представлена через функцию Лапласа по следующей формуле:

$$\boldsymbol{F}(x) = 0,5 + 0,5\Phi\left[\frac{(x - m_x)}{\sqrt{D_x}}\right], \qquad (3.4)$$

где D_x – генеральная дисперсия переменной X.

Далее перечислим **свойства функции Лапласа**.

Свойство 1. Функция Лапласа является нечетной, т. е. $\Phi(-t) = -\Phi(t)$.

Свойство 2. При $t = 0$ $\Phi(t) = 0$.

Свойство 3. При $t = \pm\infty$ $\Phi(t) = 0,5$.

Поскольку удвоенная функция Лапласа равна 1, то площадь, ограниченная интегральной кривой распределения, также равна единице. Поэтому, используя формулу (3.4), нетрудно вычислить площадь в пределах любого заданного интервала и таким образом рассчитать вероятность того, что нормально распределенная случайная величина X попадет в интервал $[\alpha, \beta]$.

Для симметричного относительно центра распределения интервала получим:

$$p\left(\left|X - m_x\right| < \varepsilon\right) = 2\Phi\left(\frac{\varepsilon}{\sqrt{D_x}}\right),$$

где ε – некоторая заранее заданная величина. Рассчитаем теперь по этой формуле вероятность попадания переменной X в интервалы $\pm\sigma = \pm\sqrt{D_x}$, $\pm 2\sqrt{D_x}$, $\pm 3\sqrt{D_x}$:

$$p\left(-\sigma_x \le x < \sigma_x\right) = p\left|X - m_x\right| < \sqrt{D_x} = 0,678,$$
$$p\left(-2\sigma_x \le x < 2\sigma_x\right) = p\left|X - m_x\right| < 2\sqrt{D_x} = 0,956,$$
$$p\left(-3\sigma_x \le x < 3\sigma_x\right) = p\left|X - m_x\right| < 3\sqrt{D_x} = 0,997.$$

Итак, с вероятностью 67,8 % возможное значение X находится в пределах $\pm\sigma$, 95,6 % – в пределах $\pm 2\sqrt{D_x}$, и 99,7 % – в пределах $\pm 3\sqrt{D_x}$. Поэтому можно сделать вывод, что основная часть наблюдений попадает уже в интервал $\pm 2\sqrt{D_x}$. И лишь три наблюдения из 1000 имеют числовое значение, выходящее из интервала $\pm 3\sqrt{D_x}$.

Естественно, что вероятность подобного события чрезвычайно мала. Это позволяет сформулировать «*правило трех сигм*»: если распределение случайной величины неизвестно, но в интервале $\pm 3\sigma = \pm 3\sqrt{D_x}$ содержится 99,7 % ее значений, то практически достоверно можно утверждать, что эта случайная величина распределена нормально.

Возможно также несколько иное толкование правила трех сигм. Если случайная величина распределена нормально, то есть основания считать, что в пределах $\pm 3\sigma$ содержатся практически все ее значения. Правило трех сигм довольно широко используется в практических расчетах, например, в теории ошибок (см. гл. 5).

Основные свойства нормального закона

Свойство 1. Плотность вероятности $f(x)$ всегда положительна, а область ее существования вся числовая ось.

Свойство 2. Функция $f(x)$ является четной, т. е. $f(x) = f(-x)$, ее график симметричен относительно прямой $x = \bar{x}$.

Свойство 3. Функция $f(x)$ имеет максимум в точке $x = \bar{x}$, равный $\dfrac{1}{\sigma\sqrt{2\pi}}$.

Свойство 4. Математическое ожидание, мода и медиана, а также их выборочные оценки совпадают, причем коэффициенты асимметрии и эксцесса равны нулю.

Свойство 5. Любое линейное преобразование исходной случайной величины X, имеющей нормальное распределение, сохраняет нормальность закона распределения.

Свойство 6. Если две независимые случайные величины X и Y распределены по нормальному закону с параметрами соответственно $\bar{x}$, σ_x и $\bar{y}$, σ_y, то их сумма $Z = X + Y$ будет также иметь нормальное распределение с параметрами $\bar{z} = \bar{x} + \bar{y}$ и $\sigma_z = \sqrt{\sigma_x^2 + \sigma_y^2}$.

Отметим, что все свойства и закономерности нормального закона распределения для генеральной и выборочной совокупности остаются неизменными.

Значение нормального закона. Главная особенность, выделяющая нормальный закон распределения среди многих других, состоит в том, что он является предельным, т. е. законом, к которому могут приближаться другие законы распределения при некоторых условиях. В частности, это вытекает из *центральной предельной теоремы*. Хотя существует несколько форм центральной

предельной теоремы, однако все они посвящены установлению условий, при которых сумма взаимно независимых случайных величин при неограниченном увеличении числа слагаемых стремится к нормальному закону распределения. Рассмотрим центральную предельную теорему в форме *теоремы Ляпунова.* Суть ее состоит в следующем.

Если взаимно независимые случайные величины $X_1, X_2, \ldots, X_n$ *имеют конечные абсолютные центральные моменты третьего порядка и если при* $n \to \infty$ *выполняется условие:*

$$\lim_{n\to\infty}\left(\frac{\sum_{i=1}^{n} M\left[\left|X_i - m_{x_i}\right|^3\right]}{D_x^{3/2}}\right) = 0, \qquad (3.5)$$

то распределение суммы случайных величин $X = \sum_{i=1}^{n} X_i$ *неограниченно (асимптотически) приближается к нормальному с параметрами* $m_x = M\left[\sum_{i=1}^{n} X_i\right]$ *и* $D_x = D\left[\sum_{i=1}^{n} X_i\right]$.

Условие (3.5) выражает тот факт, что вклад всех слагаемых в рассеяние величины X по отдельности ничтожно мал по сравнению с их суммарным эффектом.

В частном случае, когда все случайные величины $X_1, X_2, \ldots, X_n$ имеют одинаковые законы распределения с параметрами m_{xi} и $\boldsymbol{\sigma}_{xi}$, то при $n \to \infty$ условие (3.5) выполняется автоматически и, следовательно, может быть проигнорировано. Тогда в соответствии с центральной предельной теоремой распределение случайной величины $X = \sum_{i=1}^{n} X_i$ становится асимптотически нормальным с параметрами $m_x = nm$ и $\sigma_x = \sqrt{n\sigma^2}$. При этом среднее арифметическое $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$ будет иметь асимптотически нормальное распределение с параметрами $m_{\bar{x}} = m$ и $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$.

Следствие. Если случайная величина представляет собой результат взаимодействия большого числа сравнительно слабых и

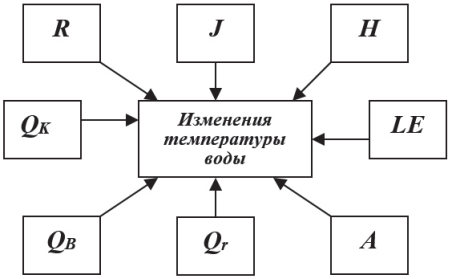

Рис. 3.2. Основные факторы изменения температуры поверхностного слоя воды в океане

примерно равноценных факторов, то, согласно центральной предельной теореме, можно ожидать, что эта случайная величина будет распределена по нормальному закону. Однако, если среди множества взаимодействующих факторов есть хотя бы один или два преобладающих, то у нас уже нет оснований утверждать, что случайная величина будет подчиняться нормальному закону.

Заметим, что априори значение n, при котором случайная величина X становится распределенной по закону, близкому к нормальному, вряд ли может быть установлено теоретически. Однако, как следует из результатов практических расчетов, для многих природных процессов достаточно четырех-пяти равноценных факторов, чтобы распределение случайной величины стало близким к нормальному закону.

Рассмотрим конкретный пример. В соответствии с уравнением теплового баланса океана изменение температуры поверхностного слоя воды (рис. 3.2) определяется следующими основными факторами:

– коротковолновым притоком солнечной радиации (Q_k);

– длинноволновым излучением радиации с поверхности океана ($Q_в$);

– длинноволновым потоком радиации из атмосферы в океан (Q_r);

– затратами тепла на испарение (LE);

– турбулентным теплообменом между океаном и атмосферой (H);

– адвекцией тепла течениями (A);

– горизонтальным турбулентным теплообменом (J);

– вертикальным обменом тепла с нижележащими слоями воды (R);

Если пренебречь рядом других факторов (например, диссипацией кинетической энергии в тепловую, тепловыми эффектами от замерзания или таяния морских льдов), то имеем восемь основных факторов, влияющих на изменения температуры воды в поверхностном слое. Очевидно, что значимость указанных факторов в значительной степени зависит как от масштабов временного осреднения процессов формирования теплового баланса, так и от географического района океана.

Примем, например, период осреднения равный 1 месяцу. В этом случае для большинства районов океана преобладающим фактором оказывается годовой ход коротковолнового притока солнечной радиации, который может значительно превышать вклад в изменения температуры воды других тепловых процессов. Именно вследствие преобладания этого фактора распределение среднемесячных значений температуры поверхности океана обычно не подчиняется нормальному закону.

Для исключения влияния годового хода солнечной радиации можно рассчитать аномалии температуры воды:

$$\Delta t_{ij} = t_{ij} - \overline{t}_j,\ i = 1,\ \ldots\ ,\ n,\ j = 1,\ \ldots\ ,\ m,$$

где n – количество лет, m – количество месяцев (m = 12), $\overline{t}_j$ – среднемноголетняя норма температуры для j-го месяца. В результате использования такой процедуры принято считать, что в аномалиях температуры воды годовой цикл солнечной радиации уже отсутствует. В этом случае вклад различных факторов в формирование температуры воды в большинстве районов океана становится более равноценным. Поэтому распределение аномалий среднемесячных величин температуры воды значительно чаще подчиняется нормальному закону.

Отметим, что если в качестве масштаба временного осреднения взять 1 год, то в этом случае радиационный фактор уже, как правило, не дает преобладающего вклада в колебания температуры поверхности океана. Поэтому распределение средних годовых значений температуры в отличие от среднемесячных величин носит значительно более симметричный характер.

Помимо центральной предельной теоремы, важное значение нормального закона состоит также в том, что он хорошо разработан теоретически, доступен и широко используется при решении многочисленных задач. В математической статистике нормальный закон играет роль некоторого стандарта, с которым сравниваются другие распределения. Кроме того, он широко используется во многих статистических методах анализа информации: методе наименьших

квадратов, корреляционном анализе, проверке статистических гипотез, методе ошибок и др.

В связи с этим проверка гипотезы нормальности распределения исходной выборки, т. е. степени соответствия эмпирического распределения нормальному, представляет собой один из важнейших этапов первичной обработки исходных данных.

Пример 3.1. На рис. 3.3 представлены гистограммы среднемесячных значений температуры поверхности океана (ТПО) и их аномалий для района Канарского апвеллинга, ограниченного по широте 20 и 24° с.ш. и по долготе 20° з.д. и берегом Африки. Нетрудно

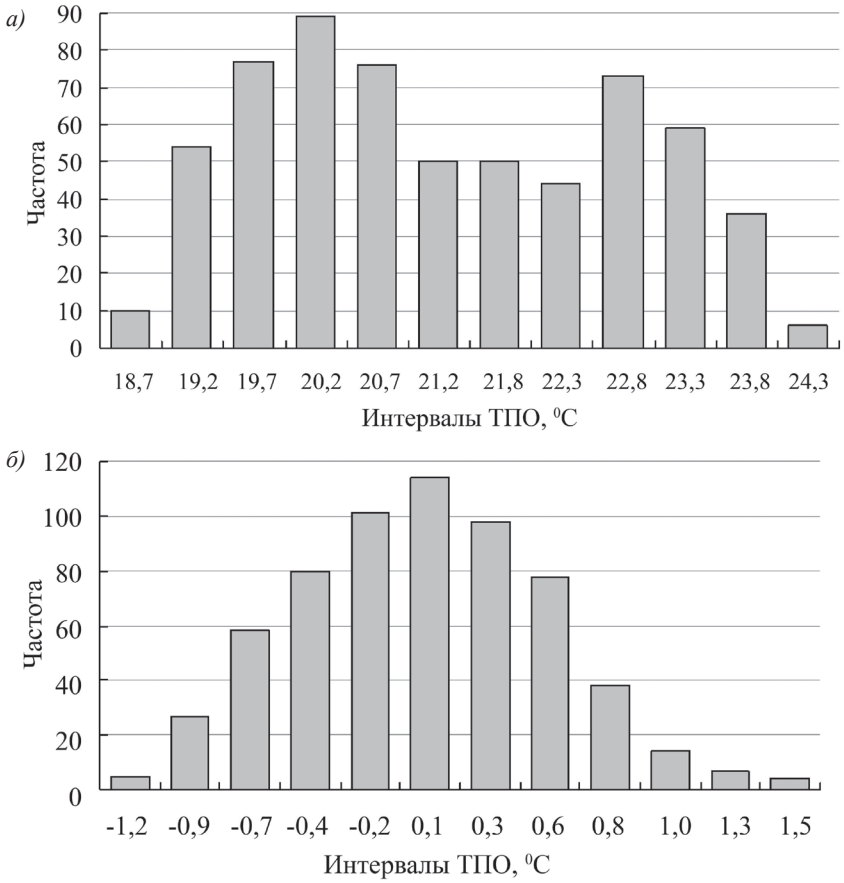

Рис. 3.3. Гистограммы среднемесячных (*а*) значений температуры поверхности океана и их аномалий (*б*) для района Канарского апвеллинга

видеть, что распределение среднемесячных значений ТПО является двухмодальным и, естественно, абсолютно не соответствует нормальному закону распределению. В то же время распределение аномалий ТПО кардинально отличается от распределения среднемесячных значений ТПО и уже носит симметричный характер, т. е. очень близко к нормальному распределению. Таким образом, можно считать установленным исключение превалирующего влияния потока суммарной радиации на годовой ход ТПО, вследствие чего вклад различных факторов в формирование температуры воды становится относительно равноценным.

3.2. Законы распределения, используемые в гидрометеорологии

Логарифмически нормальное распределение. Вообще говоря, на практике довольно часто встречается ситуация, когда случайная величина X не является нормально распределенной, однако путем ее несложного функционального преобразования можно получить случайную величину $Y = \varphi(X)$, распределенную по нормальному закону. При этом наибольшее распространение получило логарифмическое преобразование вида $Y = \log_a X$, которое допустимо лишь при $X > 0$.

Случайная величина X считается распределенной логарифмически нормально, если нормальному закону распределения подчиняется её логарифм $Y = \log_a X$. В соответствии с этим плотность вероятности выражается формулой:

$$f(y) = \frac{1}{\sqrt{2\pi D_y}} \exp\left[-\frac{\left(y - m_y\right)^2}{2D_y} \right], \tag{3.6}$$

где $m_y = M[\log_a X]$ – математическое ожидание, $D_y = \boldsymbol{\sigma}_y^{\,2} = D[\log_a X]$ – генеральная дисперсия, a – основание логарифма, причем наиболее часто принимается, что $a = e$, т. е. $Y = \ln X$.

Выполнив несложные преобразования, можно от формулы (3.6) перейти к плотности распределения исходной случайной величины, которая будет иметь следующий вид:

$$f\left(x\right) = f\left(y\right) = \frac{1}{x\sigma_y \sqrt{2\pi}} \exp\left[-\frac{\left(y - m_y\right)^2}{2\sigma_y^{\,2}} \right], \tag{3.7}$$

где $m_y = M[\ln X]$, $\boldsymbol{\sigma}_y^{\,2} = D[\ln X]$.

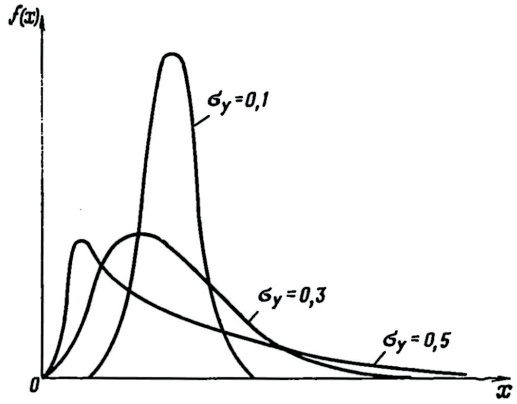

Рис. 3.4. Функция плотности вероятности логарифмически нормального распределения при различных значениях параметра σ_y и $m_y = 1$

График плотности вероятности данного распределения приведен на рис. 3.4. Как следует из формулы (3.7) и рис. 3.4, логарифмически нормальное распределение характеризуется положительной асимметрией, возрастающей с увеличением $\boldsymbol{\sigma}_y$. Естественно, чем меньше $\boldsymbol{\sigma}_y$, тем ближе друг к другу значения моды, медианы и математического ожидания и тем ближе кривая распределения к нормальному закону. Такое распределение свойственно случайным величинам, формирование которых происходит в результате умножения большого числа влияющих на них независимых равнозначных факторов.

Между параметрами нормального и логарифмически нормального распределения существуют следующие соотношения:

$$m_x = \exp\left(m_y + 0,5\boldsymbol{\sigma}_y^2\right),$$

$$\boldsymbol{\sigma}_x = m_x\left[\exp\left(\boldsymbol{\sigma}_y^2 - 1\right)\right]^{1/2},$$

или

$$m_y = \ln\left[\frac{m_x}{\left(1 + C_x^2\right)^{1/2}}\right],$$

$$\boldsymbol{\sigma}_y = \left[\ln\left(1 + C_x^2\right)\right]^{1/2},$$

где C_x – коэффициент вариации величины X.

Мода логарифмически нормального распределения функционально связана с математическим ожиданием и коэффициентом вариации:

$$Mo = \frac{m_x}{\left(1 + C_x^2\right)^{3/2}}.$$

С увеличением коэффициента вариации различия между *Mo* и m_x возрастают.

Распределение Вейбулла. *Непрерывная случайная величина X считается распределенной по закону Вейбулла, если её плотность вероятности определяется следующей формулой:*

$$f(x) = \begin{cases} 0, & при\ x < 0, \\ mcx^{m-1} \exp\left(-cx^m\right), & при\ x \geq 0, \end{cases} \qquad (3.8)$$

где m и c – параметры распределения, которые могут принимать только положительные значения.

Кривая распределения Вейбулла имеет различный вид в зависимости от значения параметра m. В связи с этим параметр m является характеристикой формы, а параметр c – характеристикой масштаба. При $m > 1$ распределение Вейбулла одномодально.

Интегральная функция распределения данного закона выражается формулой:

$$F(x) = \begin{cases} 0, & при\ x < 0, \\ 1 - \exp\left(-cx^m\right), & при\ x \geq 0. \end{cases} \qquad (3.9)$$

Заметим, что некоторые виды распределений являются частными случаями распределения Вейбулла. Так, например, при $m = 1$ получим показательное распределение, плотность вероятности которого определяется как:

$$f(x) = \begin{cases} 0, & при\ x < 0, \\ c\exp\left(-cx\right), & при\ x \geq 0, \end{cases} \qquad (3.10)$$

а функция распределения показательного (экспоненциального) закона имеет вид:

$$F(x) = \begin{cases} 0, & при\ x < 0, \\ 1 - \exp\left(-cx\right), & при\ x \geq 0. \end{cases} \qquad (3.11)$$

График плотности вероятности показательного распределения приводится на рис. 3.5. Важным свойством показательного

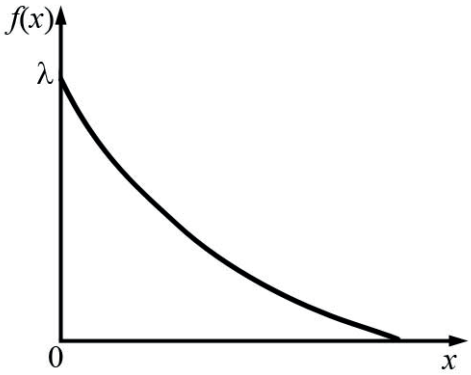

Рис. 3.5. Функция плотности вероятности показательного распределения

закона является то, что математическое ожидание и стандартное отклонение равны и функционально связаны с параметром c, т. е. $m_x = \sigma_x = 1/c$.

Кроме того, для показательного распределения характерно и то, что его коэффициенты вариации, асимметрии и эксцесса не зависят от параметра c и имеют следующие значения: $C_v = 1$, $As = 2$, $Ex = 9$.

Другим частным случаем распределения Вейбулла является *распределение Релея,* которое может быть получено, если принять $c = 1/2\sigma_x^2$ и $m = 2$, т. е.

$$f(x) = \begin{cases} 0, & \text{при } x < 0, \\ \left(\dfrac{x}{\sigma_x^2}\right)\exp\left(\dfrac{-x^2}{2\sigma_x^2}\right), & \text{при } x \geq 0, \end{cases} \qquad (3.12)$$

где σ_x – единственный определяемый параметр. График плотности вероятности данного распределения при различных значениях σ_x представлен на рис. 3.6. Нетрудно видеть, что кривая распределения, особенно при малых значениях σ_x, является резко асимметричной.

Интегральная функция распределения закона Релея выражается формулой:

$$F(x) = \begin{cases} 0, & \text{при } x < 0, \\ 1 - \exp\left(\dfrac{-x^2}{2\sigma_x^2}\right), & \text{при } x \geq 0. \end{cases} \qquad (3.13)$$

При этом основные числовые характеристики имеют вид:

$$m_x = \sigma\sqrt{\frac{\pi}{2}},\ D_x = \left(2 - \frac{\pi}{2}\right)\sigma^2,\ As \approx 0,63,\ Ex \approx -0,3.$$

Следовательно, кривая распределения Релея имеет большую положительную асимметрию и является более плосковершинной по сравнению с кривой нормального закона.

Равномерное распределение. *Непрерывная случайная величина X распределена равномерно, если плотность вероятности во всем интервале её возможных значений постоянна, а за его пределами равна нулю.* В соответствии с этим плотность вероятности в интервале $[a, b]$ может быть представлена в виде:

$$f(x) = \begin{cases} 0, & при\ x \le a, \\ \dfrac{1}{b-a}, & при\ a < x < b, \\ 0, & при\ x \ge b, \end{cases} \qquad (3.14)$$

а интегральная функция распределения записана как:

$$F(x) = \int_{-\infty}^{\infty} f(x)dx = \begin{cases} 0, & при\ x \le a, \\ \dfrac{x-a}{b-a}, & при\ a < x < b, \\ 1, & при\ x \ge b. \end{cases} \qquad (3.15)$$

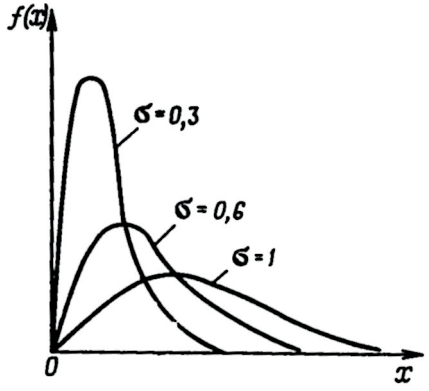

Рис. 3.6. Функция плотности вероятности распределения Релея при различных значениях параметра σ

Итак, равномерное распределение определяется двумя параметрами: a и b. При этом основные числовые характеристики равномерного закона могут быть выражены следующим образом:

$$m_x = \frac{(a+b)}{2},\ D_x = \frac{(b-a)^2}{12},\ As = 0,\ Ex = -1,2.$$

Биномиальное распределение. *Дискретная случайная величина X с возможными исходами x = m = 0, 1, 2, ..., n имеет биномиальное распределение, если вероятность того, что X = m определяется формулой:*

$$p(X = m) = P_{m,n} = c_n^m p^m q^{n-m}. \qquad (3.16)$$

При этом функция биномиального распределения выражается следующим образом:

$$F(m) = p(X < m) = \begin{cases} 0, & при\ \ m < 0, \\ \sum\limits_{m_i < m} P_{m_i,n} & при\ \ 0 < m < n, \\ 1, & при\ \ m > n. \end{cases} \qquad (3.17)$$

Биномиальное распределение определяется двумя параметрами: p и n, причем основные числовые характеристики связаны с этими параметрами как:

$$m_x = np,\ D_x = npq,\ A_S = \frac{q-p}{\sqrt{npq}},\ E_x = \frac{1-6pq}{npq}.$$

Отсюда следует, что с увеличением n коэффициенты асимметрии и эксцесса стремятся к нулю. При $n \to \infty$ и $np \to \infty$ биномиальное распределение становится асимптотически нормальным. На практике биномиальное распределение считают асимптотически нормальным уже при $npq \geq 9$.

3.3. Законы распределения, используемые в статистических расчетах

Как уже указывалось выше, при решении многих задач (статистическое оценивание, проверка гипотез, дисперсионный анализ, регрессионный анализ и др.) в качестве некоторых стандартов используется ряд теоретических законов распределения. Прежде

всего, к ним относятся непараметрическое распределение Пирсона χ^2 и параметрические законы Стьюдента и Фишера.

Распределение χ^2. Пусть имеется n независимых случайных величин $X_1, X_2, ..., X_n$, каждая из которых распределена по нормальному закону с нулевым средним значением и единичной дисперсией. Тогда *распределением χ^2 (хи-квадрат) с ν степенями свободы называется распределение суммы квадратов независимых случайных величин $\chi^2 = X^2_1 + X^2_2 + ... + X^2_n$, распределенных по стандартному нормальному закону.*

При этом *число степеней свободы – это количество значений, функционально не связанных между собой,* или, другими словами, *число независимых параметров.*

Если, например, мы имеем ряд наблюдений из четырех членов (4 + 6 + 8 + 3), то последний член является зависимой величиной. Действительно, сумма первых трех членов равна 18. Сумма же всего ряда равна 21. Поэтому на четвертый член остается величина 3, ибо никакая другая величина не даст нам требуемую сумму. Таким образом, для статистического ряда число степеней свободы равно $\nu = n - 1$.

Плотность вероятности распределения χ^2 имеет вид:

$$f(x) = \left\{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)\right\}^{-1} x^{\left(\frac{\nu}{2}-1\right)} \exp\left(-0,5x^2\right), \qquad (3.18)$$

где $\Gamma\left(\frac{\nu}{2}\right)$ – гамма-функция Эйлера, определяемая как $\Gamma\left(\frac{\nu}{2}\right) = \int_0^\infty t^{\left(\frac{\nu}{2}-1\right)} \exp(-t)dt$. Итак, распределение χ^2 зависит лишь от одного параметра − числа степеней свободы, которое определяется как $\nu = k - 1 - l$, где l – число параметров распределения. Поскольку $l = 1$, то $\nu = k - 2$.

Значения распределения χ^2 затабулированы для различных степеней свободы и уровней значимости (Приложение 2). На графике плотности вероятности распределения χ^2 для различных степеней свободы (рис. 3.7) видно, что оно резко несимметрично при малом числе ν. Однако с возрастанием ν плотность вероятности $f(x)$ становится все более симметричной и похожей на кривую нормального распределения, что вытекает из центральной предельной теоремы. Практически при ν = 13–15 случайная величина χ^2 уже подчиняется нормальному закону.

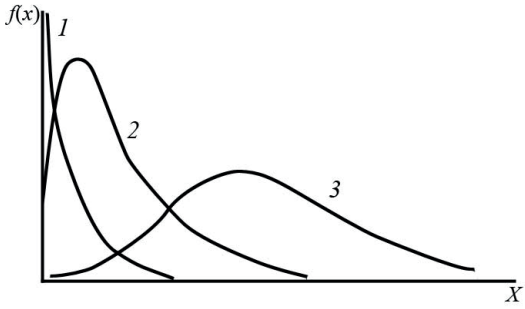

Рис. 3.7. Функция плотности вероятности χ^2-распределения Пирсона для различных степеней свободы (ν): *1)* ν = 1, *2)* ν = 4, *3)* ν = 20

Распределение Стьюдента. Пусть Z и V — независимые случайные величины, причем величина Z является нормально распределенной с параметрами $M(Z) = 0$, $D(Z) = 1$, а V — распределенной по закону χ^2 с ν степенями свободы. Тогда *случайная величина* $t = \dfrac{Z}{\sqrt{V/\nu}}$ *имеет распределение, которое называется распределением Стьюдента с ν степенями свободы.* Плотность вероятности величины t выражается следующей формулой:

$$f(x) = \frac{\Gamma\left(\dfrac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\dfrac{\nu}{2}\right)}\left[1+\frac{x}{\nu}\right]^{\frac{-(\nu+1)}{2}}. \qquad (3.19)$$

Из графика плотности вероятности $f(x)$ видно, что она симметрична относительно начала координат (рис. 3.8). По мере увеличения числа степеней свободы t-распределение приближается к нормальному закону, причем скорость этого приближения выше, чем у распределения χ^2.

Значения t-распределения затабулированы для различных степеней свободы и уровней значимости (Приложение 3). В таблице приведены значения t-статистики для двухстороннего и одностороннего критерия значимости.

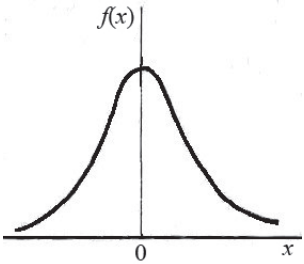

Рис. 3.8. Функция плотности вероятности t-распределения Стьюдента для степеней свободы ν = 9

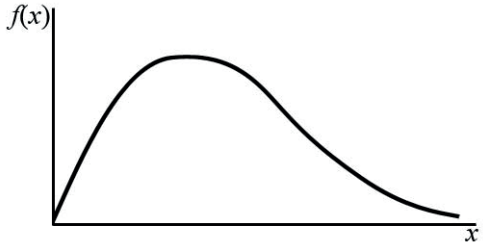

Рис. 3.9. Функция плотности вероятности *F*-распределения Фишера для степеней свободы: $\nu_1 = 10$, $\nu_2 = 25$

Распределение Фишера. Пусть мы имеем две случайные величины, дисперсии которых известны, причем $D_1 > D_2$. Тогда *дисперсионное отношение* $F = D_1 / D_2$ *имеет распределение, называемое распределением Фишера* или иногда распределением Фишера–Снедекора. Плотность вероятности этого распределения выражается следующей формулой:

$$f(x) = \frac{\nu_1^{\frac{1}{2}\nu_1}\nu_2^{\frac{1}{2}\nu_2}\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{1}{2}\nu_1\right)\Gamma\left(\frac{1}{2}\nu_2\right)} x^{\frac{1}{2}\nu_1-1}\left(\nu_1 x+\nu_2\right)^{-\frac{\nu_1+\nu_2}{2}}, \qquad (3.20)$$

где ν_1 и ν_2 – числа степеней свободы первой и второй выборки, причем $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 2$.

Как следует из формулы (3.20), распределение Фишера (*F*-распределение) не зависит от дисперсий входных выборок, а зависит лишь от числа степеней свободы. График плотности вероятности $f(x)$ приведен на рис. 3.9.

Для *F*-распределения составлены таблицы значений для различных степеней свободы и уровня значимости $\alpha = 0{,}05$ (Приложение 4).

Заметим, что эти таблицы даны для двухстороннего критерия значимости, т. е. когда проверяется, например, условие $D_1 = D_2$. В том случае, если необходимо проверить, например, неравенство дисперсий по двум выборкам, т. е. $D_1 > D_2$ или $D_1 < D_2$, то используется односторонний критерий.

3.4. Особенности построения эмпирической функции распределения

Как уже отмечалось выше, *эмпирической* (статистической) *функцией распределения* $F(x)$ случайной величины X называется закон изменения частоты события $X < x$ в данной статистической выборке, т. е.

$$F(x) = p(X < x),$$

где $p = m/n$ – относительная частота события $X < x$; m – число событий (эмпирическая повторяемость) в данном интервале (классе) k; n – длина выборки.

При $n \to \infty$ $p \to \boldsymbol{p}$, где $\boldsymbol{p}$ – теоретическая вероятность события $X < x$ и $F(x) \to \boldsymbol{F}(x)$, где $\boldsymbol{F}(x)$ – теоретическая функция распределения.

В гидрометеорологических расчетах в некоторых случаях используется соотношение, имеющее следующий вид: $G(x) = p(X \geq x)$, которое называется *эмпирической функцией обеспеченности.* Графическое изображение эмпирической функции обеспеченности называется *эмпирической кривой обеспеченности* (рис. 3.10).

Если объем выборки n весьма велик, то построение надежных в статистическом смысле эмпирических функций распределения и обеспеченности не представляет затруднений. Однако, если

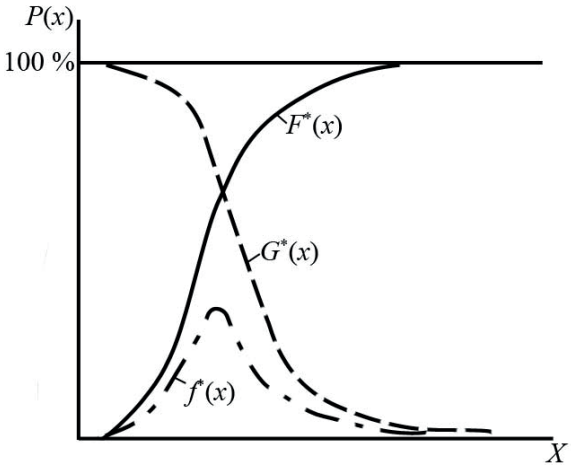

Рис. 3.10. Распределение эмпирических функций распределения $F^*(x)$, функции плотности распределения $f^*(x)$ и функции обеспеченности $G^*(x)$

в каждом интервале $m < 7–8$ значений, то для повышения надежности результатов обычно используется ряд приемов.

В общем случае процесс построения эмпирической функции распределения можно разбить на несколько этапов. Рассмотрим вкратце каждый из них.

Этап 1. Ранжирование исходного ряда наблюдений, т. е. расположение его в убывающем порядке от максимального значения до минимального ($x_n \geq x_{n-1} \geq x_{n-2} \geq \ldots \geq x_1$) или, наоборот, в возрастающем порядке. Очевидно, данный этап комментариев не требует.

Этап 2. Оценка оптимального числа интервалов (градаций). Данный вопрос представляется весьма важным, поскольку имеют место две противоречивые тенденции. С одной стороны, увеличивая число интервалов, мы получаем более детальную картину распределения. Однако из-за ограниченности объема выборки в каждый интервал попадает малое число наблюдений, в результате чего групповые частоты p начинают обнаруживать существенные случайные колебания. С другой стороны, при уменьшении числа интервалов случайные колебания значений p сглаживаются, но одновременно с этим сглаживаются и характерные черты распределения.

По-видимому, наиболее приемлемым будет некоторый компромисс, обеспечивающий достаточно четкое выявление основных особенностей изучаемого распределения. К сожалению, не существует строгого решения данной задачи. Обычно для выбора числа градаций используются те или иные эмпирические формулы. В качестве примера укажем две из них:

$$k \approx 1 + 3{,}32 \lg n, \quad k \approx 5 \lg n.$$

Первая может считаться излишне «жесткой». Поэтому для не очень больших выборок лучше ориентироваться на вторую формулу. Необходимо также помнить, что количество градаций k может быть только целым числом.

Этап 3. Нахождение ширины градаций. В первом приближении ширина градаций находится по следующей формуле:

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

Так как рассчитанное значение Δc может не совсем удачно характеризовать исходную выборку, то оно изменяется (обычно в большую сторону) до приемлемого для нас значения. Заметим, что на практике обычно принимается соответствие в числе значащих цифр Δc и x.

Этап 4. Определение границ градаций. Границы (c_1, c_2), (c_2, c_3), ... , (c_k, c_{k+1}) находятся с учетом найденной ширины Δc, причем для крайних границ c_1 и c_{k+1} и крайних членов выборки x_{max} и x_{min} должны выполняться условия $c_1 \le x_{min}$, $c_{k+1} \ge x_{max}$. В некоторых случаях за начало первой градации рекомендуется брать $c_1 = \dfrac{x_{min} - \Delta c}{2}$.

В процессе группирования выборки могут быть случаи точного совпадения отдельных наблюдений с границами градаций. Если число точно совпадающих членов выборки четное, то их распределяют пополам в смежные градации. При нечетном числе таких членов остаточное от деления пополам наблюдение относят в меньшую из смежных градаций.

Этап 5. Оценка числа событий m в каждом интервале и построение гистограммы. Число событий определяется путем суммирования значений случайной величины X для каждой градации. Отметим, что вследствие неравнозначности эмпирических повторяемостей m (в средней части распределения значений m представлено, как правило, значительно больше, чем в его крайних участках) могут возникать существенные погрешности при определении крайних частей кривых распределения и обеспеченности. С целью уменьшения искажения между эмпирической и истинной кривыми распределения предложен ряд эмпирических формул. Например:

$$p = \frac{m - 0,5}{n};\ p = \frac{m}{n+1};\ p = \frac{m - 0,3}{(n + 0,4)}.$$

Все эти формулы в какой-то степени учитывают выборочность имеющихся наблюдений, что выражается в асимптотическом приближении $F(x)$ к $\boldsymbol{F}(x)$ при $n \to \infty$. В средних частях кривой распределения данные формулы дают практически одинаковые результаты и различаются лишь для нижней и верхней частей кривой распределения. Построение гистограммы обычно осуществляется в декартовой системе координат. По оси абсцисс откладываются интервалы, по оси ординат – соответствующие им эмпирические повторяемости.

Пример 3.2. Покажем построение эмпирической функции распределения для гидрологической станции в Белом море, где в летний период в течение месяца выполнены четырехразовые наблюдения за поверхностной температурой воды (ПТВ). Общая длина выборки составила n = 100 значений температуры воды. Используя

формулу Стерджесса $k \approx 1 + 3{,}32 \lg n$, имеем $k = 8$ градаций (интервалов). Далее определяем ширину градации $\Delta c = \frac{(14{,}1 - 9{,}7)}{8} =$ $= 0{,}57$ °C. Так как рассчитанное значение Δc не очень удачно характеризует ширину градации, то округляем его в большую сторону до $\Delta c = 0{,}6$ °C. За начальное значение первого интервала примем величину $c_1 = \frac{x_{\min} - \Delta c}{2} = \frac{9{,}7 - 0{,}6}{2} = 9{,}4$ °C.

Распределение значений температуры воды по градациям, т. е. оценки эмпирической частоты, приведено в таблице 3.1. Кроме того, в данной таблице представлены оценки относительной частоты ПТВ, называемой частостью. Частость выражается в долях единицы или в процентах. Накопленная частота показывает, сколько наблюдалось вариантов со значением признака, меньше x. Из таблицы 3.1 видно, что эмпирическое распределение ПТВ является близким к симметричному.

Таблица 3.1

Распределение данных поверхностной температуры воды на гидрологической станции в Белом море по градациям

Градация	Ширина градации, °C	Эмпирическая частота, m_i	Частость, m_i/n	Накопленная частота, Σm_i	Накопленная частость, $\Sigma m_i/n$
1	9,4–10,0	3	0,03	3	0,03
2	10,0–10,6	7	0,07	10	0,10
3	10,6–11,2	11	0,11	21	0,21
4	11,2–11,8	20	0,20	41	0,41
5	11,8–12,4	28	0,28	69	0,69
6	12,4–13,0	19	0,19	88	0,88
7	13,0–13,6	10	0,10	98	0,98
8	13,6–14,2	2	0,02	100	1,00
	Σ	100	1,00	–	–

3.5. Понятие о нормализации исходных данных

Учитывая исключительно большое значение нормального закона распределения в статистических расчетах, целесообразно исходные данные приводить к «нормальному» виду в тех случаях, когда

их распределение носит явно выраженный асимметричный характер. Основанием для этого может послужить анализ эмпирической гистограммы.

Действительно, если на графике члены ряда располагаются несимметрично относительно среднего значения, то это означает скошенность распределения, причем в зависимости от характера скошенности нормализация осуществляется различным образом.

Для положительной асимметрии ($As > 0$), как уже указывалось выше, левая ветвь гистограммы является более крутой, а правая – более пологой. В этом случае обычно используется логарифмическое преобразование вида:

$$x' = \lg\left(x \times 10^{a}\right).$$

Множитель 10^a вводится сюда для того, чтобы исключить появление отрицательных значений параметров. Кроме того, для приведения распределения к симметричному виду иногда применяются и другие преобразования:

$$x' = \frac{1}{x},\ x' = \frac{1}{\left(x\right)^{1/2}}.$$

Отметим, что обратная величина $1/x$ является наиболее «сильным» преобразованием, нормализующим выборки с существенной положительной асимметрией.

Для отрицательной асимметрии ($As < 0$) левая ветвь гистограммы, наоборот, является более пологой, а правая – более крутой. Нормализация исходной выборки в этом случае осуществляется преобразованием $x' = x^{\xi}$, где показатель степени может принимать различные положительные значения, большие единицы ($\xi > 1$). При умеренно отрицательной асимметрии обычно принимается $\xi = 1{,}5$, при более сильной асимметрии $\xi = 2$.

Поскольку существуют различные варианты приведения исходных данных к нормальному виду, то возникает вопрос их оценки. Другими словами, необходимо определить, какое преобразование наилучшим образом нормализует исходную выборку. На наш взгляд, для этой цели целесообразно воспользоваться критерием Пирсона χ^2 (см. разд. 4.3), который характеризует соответствие эмпирической и теоретической функций распределения. Тот вариант нормализации исходной выборки, при котором критерий χ^2 достигает минимума, следует считать наилучшим.

Глава 4.
Статистическая проверка гипотез

Раздел математической статистики, устанавливающий на основе различных критериев наличие (отсутствие) тех или иных предположений относительно свойств случайной величины, называется *статистической проверкой гипотез.*

В общем случае различают параметрическое и непараметрическое оценивание гипотез (рис. 4.1). При *параметрическом оценивании* предполагаются известными вид функции распределения генеральной совокупности (как правило, принимается нормальный закон) и отдельные параметры. Проверка гипотез относится к неизвестному параметру θ_0 о принадлежности его некоторому подмножеству $\theta_0 \subset \theta$. К параметрическим критериям относятся статистики Фишера, Стьюдента и др.

Непараметрические критерии не требуют знания законов распределения изучаемой случайной величины, поэтому они являются более общими по сравнению с параметрическими критериями. Заметим также, что для проверки гипотез с помощью непараметрических критериев обычно требуется меньший объем вычислений. Однако существенным недостатком непараметрических критериев является их меньшая мощность (эффективность). Это приводит к тому, что какие-либо имеющиеся различия в свойствах изучаемого процесса являются значимыми реже, чем при использовании соответствующих параметрических критериев. К непараметрическим критериям относятся критерии согласия, критерий Уилкоксона, серий, знаков и др.

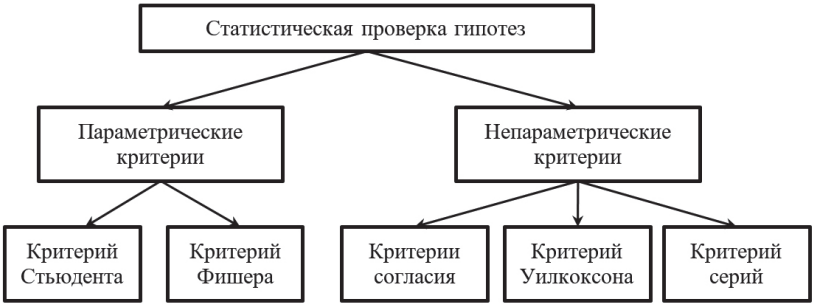

Рис. 4.1. Статистические критерии проверки гипотез

4.1. Общие положения проверки гипотез

В общем случае гипотеза – это сформулированное предположение относительно объективных свойств изучаемого явления. В математической статистике основной является так называемая *нулевая гипотеза,* т. е. *предположение об отсутствии различий в тех или иных свойствах случайного процесса.*

Нулевая гипотеза обозначается как H_0. Тогда, например, запись нулевой гипотезы в виде:

$$H_0 : \overline{\theta}_1 = \overline{\theta}_2$$

означает, что среднее арифметическое первой выборки равно среднему арифметическому второй выборки.

Если имеется нулевая гипотеза, то обязательно должны существовать *альтернативные* (противоположные) *гипотезы*, являющиеся логическим отрицанием нулевой гипотезы. Вообще говоря, их может быть бесчисленное множество, однако в некоторых простых случаях они могут быть представлены в виде единственной альтернативы. Например, в рассматриваемом примере альтернативная гипотеза имеет вид:

$$H_1 : \overline{\theta}_1 \neq \overline{\theta}_2.$$

Гипотеза может быть простой или сложной. *Простой* называется такая гипотеза, в которой проверяемый параметр может принять только одно значение. Так, приведенная выше нулевая гипотеза является простой. Если же проверяемый параметр может принимать некоторое множество (два и более) значений, то такая гипотеза называется *сложной*. В общем случае сложная гипотеза может быть записана как:

$$H_0 : \theta \in C,$$

где C – некоторое множество значений параметра θ. Например, запись сложной гипотезы $H_0 : \overline{x}_1 = a_1 < x < a_2$ означает, что среднее арифметическое случайной величины X должно принимать значение в диапазоне $[a_1, a_2]$. В дальнейшем мы будем рассматривать только простые гипотезы.

Естественно, что нулевая гипотеза как предположение должна подлежать проверке (испытанию). Задача проверки состоит в том, чтобы установить, противоречит ли выдвинутая гипотеза результатам наблюдений над исследуемой величиной или нет. Для этого используются *статистические критерии* (параметрические и

непараметрические), *которые представляют собой определенный свод правил, указывающих, при каких результатах наблюдений рассматриваемая гипотеза отклоняется, а при каких – нет.* Однако нулевая гипотеза может быть как истинной, так и ложной. Это приводит к тому, что возникает четыре комбинации исходов, две из которых приводят к правильному, а две – к неправильному выводу. Возможные комбинации принятия (отвержения) нулевой гипотезы представлены в таблице 4.1.

Таблица 4.1

Возможные комбинации принятия (отвержения) нулевой гипотезы

Нулевая гипотеза	Гипотеза верна	Гипотеза неверна
Принимается	Правильное решение	Ошибка второго рода
Отвергается	Ошибка первого рода	Правильное решение

Только принятие правильной или отклонение неправильной гипотезы можно считать верным решением. Если нулевая гипотеза отвергается, в то время как на самом деле она верна, то возникает ошибка, называемая ошибкой *первого* рода. Наоборот, если ошибочная гипотеза принимается, то совершается ошибка *второго* рода.

Вероятность появления ошибки первого рода называется уровнем значимости критерия и обозначается как α. Если величина α всегда задается заранее, то вероятность появления ошибки второго рода, обозначаемой обычно β, остается неизвестной. Если, например, в рассматриваемом выше примере нулевая гипотеза отвергается, то можно сделать вывод о том, что обе изучаемые выборки имеют различные средние значения, и вероятность того, что принято ошибочное решение, равна α. С другой стороны, если H_0 не отвергается, то утверждение того, что средние значения двух выборок совпадают, может оказаться ложным с неизвестной вероятностью β.

Итак, вероятность события, которым решено пренебречь в данном исследовании, и представляет уровень значимости α. Практический смысл уровней значимости заключается в следующем. Пусть α = 5 %. Тогда в предположении, что нулевая гипотеза верна, разность средних двух выборок можно ожидать не менее чем пять раз на каждые 100 испытаний, проведенных в неизменных условиях. Если частота появления исследуемой статистики окажется меньше указанной разности, то гипотеза отвергается.

Вообще говоря, выбор уровня значимости является произвольным. Действительно, на практике всегда приходится выбирать

между двумя противоположными тенденциями. С одной стороны, с увеличением вероятности того, что некоторая статистика принимает какое-либо значение, увеличивается вероятность ошибочного отбрасывания верной гипотезы, а с другой – с уменьшением вероятности возрастает число испытаний, необходимое для эффективного применения критерия значимости. Поэтому он устанавливается на основе опыта как уровень, дающий практическую уверенность, что ошибочные заключения будут сделаны только в очень редких случаях. Наиболее часто в гидрометеорологических расчетах используются уровни значимости 1, 5 и 10 %.

По аналогии с уровнем значимости *ошибка второго рода – это вероятность отвергнуть верную конкурирующую (альтернативную) гипотезу.* Очевидно, при фиксированной ошибке первого рода, чем меньше вероятность ошибки второго рода, тем эффективнее будет критерий. Другими словами, вероятность сделать правильный выбор в этом случае становится максимальной. Отсюда приходим к понятию мощности критерия, под которым понимается вероятность попадания заданной статистики в критическую область, когда верна альтернативная гипотеза. Другими словами, *мощность критерия – это вероятность не допустить ошибку второго рода, т. е. принять нулевую гипотезу, когда она неверна.* Итак, мощность критерия функционально связана с β, т. е. $\gamma = 1 - \beta$. Используя юридическую терминологию, можно сказать, что α – вероятность вынесения судом обвинительного приговора, когда обвиняемый на самом деле невиновен, а β – вероятность вынесения судом оправдательного приговора, в то время как обвиняемый виновен в преступлении.

Значения статистики, при которых гипотеза опровергается, т. е. вероятность которых меньше заданного уровня значимости, образуют *критическую область* проверяемой гипотезы. Естественно, если значения этой статистики имеют вероятность больше уровня значимости, то получаем *область допустимых значений* или доверительную область (рис. 4.2). В связи с этим задача проверки гипотезы сводится к построению критической области для выбранного уровня значимости. Если статистика попадет в критическую область, то это указывает на несоответствие гипотезы наблюденным данным, и нулевая гипотеза опровергается.

Кроме того, как следует из рис. 4.2, с увеличением уровня значимости увеличивается критическая область, что влечет за собой и увеличение вероятности попадания исследуемой статистики в критическую область. Однако вместе с тем возрастает вероятность

Рис. 4.2. Формирование области допустимых значений и критической области гипотезы

ошибочного отбрасывания гипотезы. Таким образом, в выборе уровня значимости присутствует известное противоречие: *с одной стороны, этот уровень должен быть достаточно велик для отбрасывания ложных гипотез, а с другой – он должен быть достаточно мал, чтобы приводить к отбрасыванию лишь немногих верных гипотез*. В общем случае критическую область нужно задавать такой, чтобы при заданном уровне значимости мощность критерия γ была максимальной. Задача построения такой критической области при проверке гипотез решается с помощью теоремы Неймана–Пирсона. Однако в связи со сложностью построения оценок мощности статистических критериев на практике обычно ограничиваются проверкой нулевой гипотезы по уровню значимости.

При проверке гипотез следует различать двусторонний и односторонний уровни значимости. *Двусторонний уровень значимости* применяется в тех случаях, когда требуется, например, оценить расхождение между двумя случайными величинами, причем для нас одинаково представляют интерес как положительные, так и отрицательные разности между изучаемыми величинами. В тех случаях, когда нужно убедиться, что одна случайная величина в среднем строго больше (меньше) другой, применяется *односторонний критерий значимости.* Поскольку двусторонний уровень значимости на практике используется значительно чаще, то в статистических таблицах, как правило, приводятся именно его оценки. Поэтому, если надо применить, например, 5%-ный уровень значимости при одностороннем критерии, мы должны взять в соответствующей таблице для двустороннего критерия 10%-ный уровень значимости.

При выбранном уровне значимости критическую область следует строить так, чтобы мощность критерия была бы максимальной. Выполнение данного требования должно обеспечить минимальную ошибку второго рода. Ясно, что критическая область тем лучше, чем меньше вероятности ошибок первого и второго рода. Однако при заданном объеме выборки уменьшить одновременно α и β невозможно. Если уменьшить α, то β будет возрастать. Единственный способ одновременного уменьшения вероятностей ошибок первого и второго рода состоит в увеличении объема выборки.

Заметим также, что *уровень значимости – величина, функционально связанная с доверительной вероятностью* ($\alpha = 1 - p$). Наконец, следует помнить одно из основных положений математической статистики: *при помощи критерия значимости нулевая гипотеза может быть отвергнута, но не может быть доказана.* На примере рассмотренного выше случая о равенстве средних двух выборок это означает, что мы вправе утверждать об их неравенстве, но не вправе сделать вывод о том, что они равны. Мы можем лишь полагать, что данные наблюдений согласуются с нулевой гипотезой и, следовательно, не дают оснований ее отвергнуть. Другими словами, рассматриваемая гипотеза не находится в противоречии с данными наблюдений.

На практике для большей уверенности принятия гипотезы ее проверяют другими способами или повторяют ее проверку, увеличив объем выборки. Отметим, что при изменении объема выборки данная гипотеза может приобрести даже противоположный смысл. Поэтому следует иметь в виду, что принцип проверки статистической гипотезы не дает абсолютного доказательства ее верности или неверности.

Общая схема проверки нулевой гипотезы

1. Исходя из постановки задачи, записывается в том или ином виде нулевая гипотеза.

2. Выбирается альтернативная гипотеза, от вида которой строится критическая область. Например, если альтернативную гипотезу задать как $H_1 : \overline{\theta}_1 \neq \overline{\theta}_2$, то в этом случае строится двусторонняя критическая область. Если же альтернативная гипотеза принимается в виде неравенств $H_1 : \overline{\theta}_1 > \overline{\theta}_2$ или $H_1 : \overline{\theta}_1 < \overline{\theta}_2$, то соответственно строится правосторонняя (левосторонняя) критическая область.

3. Выбирается какой-либо статистический критерий $\mathbf{\theta}$, наилучшим образом отвечающий, по мнению исследователя, проверке нулевой гипотезы.

4. Рассчитывается по экспериментальным данным выборочное значение параметра θ.

5. Осуществляется проверка неравенства $\theta > \boldsymbol{\theta}_{кр}(\alpha,\nu)$, где $\boldsymbol{\theta}_{кр}(\alpha,\nu)$ – критическое (пороговое) значение статистики $\boldsymbol{\theta}$, выбираемое из соответствующей таблицы теоретического распределения по заданному уровню значимости α и числу степеней свободы ν.

6. При проверке неравенства возможно три исхода. Если данное неравенство выполняется, то нулевая гипотеза всегда отвергается. Если данное неравенство не выполняется, то из-за невозможности доказать нулевую гипотезу мы можем лишь предположить альтернативный вывод. Если же получаем $\theta = \boldsymbol{\theta}_{кр}(\alpha,\nu)$, то следует изменить уровень значимости для получения однозначного вывода.

Произвольность выбора уровня значимости представляет, вероятно, самое неприятное условие проверки гипотезы. Хорошо, если при задании разных вариантов уровня значимости (например, 0,1, 0,05 и 0,01) удается получить однозначные результаты, т. е. во всех вариантах нулевая гипотеза отвергается или, наоборот, нет оснований для ее отвержения. Значительно сложнее принять решение при противоположных исходах проверки нулевой гипотезы. Поэтому, чтобы избежать такой неопределенности, целесообразно рассчитывать минимальный уровень значимости, при котором отвергается нулевая гипотеза. Польза его оценки состоит уже в том, что он показывает, насколько сильно наблюденное значение противоречит гипотезе H_0.

Отметим, что задаваемые оценки уровня значимости трактуются различным образом. Обычно, если $\alpha \geq 0{,}1$, то принято считать, что данные согласуются с H_0, при $\alpha = 0{,}05$ возможна значимость, но есть некоторые сомнения в истинности H_0 и при $\alpha = 0{,}01$ существует высокая значимость, гипотеза H_0 почти наверняка не подтверждается. Наконец, следует помнить, что чем меньше уровень значимости, тем сложнее отвергнуть нулевую гипотезу. На практике целесообразно задавать разные оценки α. Как уже указывалось выше, наиболее часто используются уровни 10, 5 и 1 %.

4.2. Проверка гипотез о равенстве выборочных средних и дисперсий

Одним из важнейших понятий случайного процесса является стационарность, под которой, как будет указано в разд. 9, приближенно можно понимать постоянство во времени выборочных средних и дисперсии. Понятие стационарности относится к числу

ключевых при анализе случайных процессов. Одним из простейших способов проверки стационарности является использование статистических гипотез о равенстве выборочных средних и дисперсий. При этом не обязательно выборку делить пополам или на несколько равных частей. Впрочем, проверка этих гипотез широко применяется при решении многих других задач. Критериями для проверки гипотез служат параметрические критерии Стьюдента и Фишера.

Гипотеза о равенстве выборочных средних при неизвестных генеральных дисперсиях

Рассмотрим две независимые выборки переменных X и Y, объемы которых равны m и n соответственно, причем известно, что они извлечены из нормальных генеральных совокупностей, имеющих равные дисперсии ($D_x = D_y = D$). Но при этом сами генеральные (истинные) дисперсии, а также математические ожидания m_x и m_y неизвестны. Прежде всего сформулируем нулевую гипотезу о равенстве выборочных средних значений этих выборок, т. е. $H_0 : \overline{x} = \overline{y}$. Альтернативную гипотезу примем в виде $H_1 : \overline{x}_1 \neq \overline{y}$.

Поскольку указанные выборочные средние $\overline{x}$ и $\overline{y}$ имеют нормальное распределение, то естественно считать, что их разность также будет распределена по нормальному закону. В этом случае для проверки нулевой гипотезы может быть использована статистика Стьюдента, рассчитываемая по следующей формуле:

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{(m-1)s_x^2 + (n-1)s_y^2}} \sqrt{\frac{mn(m+n-2)}{m+n}}, \qquad (4.1)$$

где s^2_x и s^2_y – выборочные оценки дисперсий первой и второй совокупностей, m и n – соответственно длина первой и второй выборки. Как известно, статистика t распределена по закону Стьюдента с $\nu = n + m - 2$ степенями свободы (Приложение 3).

После этого осуществляется проверка неравенства $|t| > t_{кр}(\alpha,\nu)$, где $t_{кр}(\alpha,\nu)$ – критическое значение статистики Стьюдента, соответствующее уровню значимости α и числу степеней свободы $\nu = n + m - 2$. *Если данное неравенство выполняется, то нулевая гипотеза отвергается и делается вывод, что выборочные средние, извлеченные из нормальных генеральных совокупностей, имеют значимые расхождения (не равны друг другу) при заданном уровне значимости.* В противоположном случае, т. е. при $|t| < t_{кр}(\alpha,\nu)$, у нас есть основания считать, что расхождения между выборочными средними не являются значимыми.

Гипотеза о равенстве выборочных средних при известных генеральных дисперсиях

Нулевая гипотеза формулируется аналогичным образом, причем, если известны дисперсии генеральных совокупностей, то проверить ее гораздо легче. Для этого необходимо вычислить критерий:

$$Z = \frac{\left|\overline{x} - \overline{y}\right|}{\left(\frac{D_x}{m} + \frac{D_y}{n}\right)^{1/2}}, \qquad (4.2)$$

где D_x и D_y – генеральные дисперсии двух выборок. Затем по таблице функции Лапласа находится критическая точка $\boldsymbol{Z}_{\text{кр}}$ из равенства:

$$\Phi\left(\boldsymbol{Z}_{\text{кр}}\right) = \frac{\left(1-\alpha\right)}{2}.$$

Если выполняется неравенство $Z > \boldsymbol{Z}_{\text{кр}}$, то нулевая гипотеза о равенстве средних отвергается, если $Z < \boldsymbol{Z}_{\text{кр}}$, то у нас нет оснований отвергать нулевую гипотезу.

Заметим, что указанные критерии являются точными и могут быть использованы как для больших, так и для малых выборок, извлеченных из нормальных генеральных совокупностей. С известной долей осторожности формула (4.1) может быть использована в тех случаях, когда $D_x \neq D_y$, а также для больших выборок с неизвестным законом распределения, ибо в соответствии с центральной предельной теоремой величины $\overline{x}$ и $\overline{y}$ распределены асимптотически нормально. Отметим, что поскольку генеральные дисперсии известны редко, то Z-критерий не нашел широкого применения.

Гипотеза о равенстве выборочных дисперсий при неизвестных значениях математического ожидания

Рассмотрим две независимые выборки переменных X и Y, объемы которых равны m и n соответственно. Эти выборки извлечены из нормальных генеральных совокупностей, причем математические ожидания их неизвестны. Требуется проверить равенство выборочных дисперсий. Для этого составляем нулевую гипотезу вида $H_0 : s_x^2 = s_y^2$ при альтернативе $H_1 : s_x^2 \neq s_y^2$. Наиболее точным критерием ее проверки, как известно, является статистика Фишера (дисперсионное отношение), определяемая по формуле:

$$F = \frac{s_x^2}{s_y^2}, \qquad (4.3)$$

причем принимается, что $s^2_x > s^2_y$. Выборочные оценки s^2_x и s^2_y рассчитываются как:

$$s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2,$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2.$$

Далее осуществляется проверка неравенства $F > \boldsymbol{F}_{\text{кр}}(\alpha;\nu_1,\nu_2)$, где $\nu_1 = n - 1$, $\nu_2 = m - 1$ (приложение 4). *Если данное неравенство выполняется, то нулевая гипотеза отвергается и делается вывод, что дисперсии выборок, извлеченных из нормальных генеральных совокупностей, имеют значимые расхождения (не равны друг другу) при заданном уровне значимости.* Если это неравенство не выполняется, то у нас нет оснований отвергать нулевую гипотезу. Поэтому можно предположить, что расхождения между выборочными дисперсиями малы и носят незначимый характер.

Гипотеза о равенстве выборочных дисперсий при известных значениях математического ожидания

Данная гипотеза проверяется аналогично предыдущей. Различие состоит в том, что при оценке выборочных дисперсий используются значения математических ожиданий m_x и m_y, т. е.

$$s_x^2 = m^{-1}\sum_{i=1}^{m}\left(x_i - m_x\right)^2,$$

$$s_y^2 = n^{-1}\sum_{i=1}^{n}\left(y_i - m_y\right)^2.$$

Заметим, что данная гипотеза проверяется очень редко, поскольку математические ожидания генеральных совокупностей за редким исключением неизвестны.

Гипотеза о равенстве нескольких выборочных дисперсий

Для сравнения нескольких дисперсий по выборкам одинакового объема, взятых из нормальных генеральных совокупностей, может быть использован критерий Кочрена, а различного объема – критерий Бартлетта. Однако оба критерия обладают недостатками. Критерий Бартлетта является весьма приближенным, а распределение критерия Кочрена хотя известно, но оно имеет существенно меньшую мощность, чем критерий Фишера. Поэтому для сравнения нескольких дисперсий все же следует использовать критерий Фишера. С этой целью осуществляется ранжирование величин дисперсий и

затем производится сравнение наибольшей и наименьшей дисперсий. Если окажется, что различие между ними незначимо, то различие между остальными дисперсиями тем более будет незначимо. В противном случае выбирается следующая пара дисперсий, имеющая максимальную разность, и процедура их сравнения повторяется.

Пример 4.1. В первой декаде июля осуществлена съемка физических и химических характеристик воды Финского залива. При этом 8 гидрологических станций были выполнены в пределах акватории Невской губы до о. Котлин, а другие 9 станций – сразу же за о. Котлин. Средняя поверхностная температура воды до о. Котлин составила $\bar{x}$ = 16,2 °С, а ее стандартное отклонение s_x = 3,2 °С. Средняя температура воды за о. Котлин оказалась заметно ниже $\bar{y}$ = 13,9 °С при стандартном отклонении s_y = 2,1 °С. На уровне значимости α = 0,05 выяснить насколько существенно влияние острова Котлин на распределение средней температуры воды и дисперсии за период проведения гидрологической съемки.

Вначале рассмотрим равенство выборочных дисперсий. Нулевая гипотеза имеет вид $H_0 : s_x^2 = s_y^2$, альтернативная гипотеза – $H_1 : s_x^2 \neq s_y^2$. В этом случае критическая область является двусторонней. Рассчитываем фактическое значение критерия Фишера по формуле (4.3), которое равно F = 2,32. После этого определяем критическое значение статистики Фишера при числе степеней свободы $\nu_1 = n - 1 = 8$, $\nu_2 = m - 1 = 7$ и уровню значимости α = 0,05. Из Приложения 4 находим, что $\boldsymbol{F}_{кр}(\alpha;\nu_1,\nu_2) = 3{,}73$. Так как $F < \boldsymbol{F}_{кр}$, то мы можем полагать, что расхождения между выборочными дисперсиями не являются значимыми и, следовательно, влияние о. Котлин не сказывается на дисперсии температуры воды.

Рассмотрим теперь равенство выборочных средних. В соответствии с общей схемой проверки гипотез записываем нулевую гипотезу как $H_0 : \bar{x} = \bar{y}$, т. е. средние значения температуры воды для обоих участков гидрологической съемки равны. В качестве альтернативной гипотезы возьмем гипотезу $H_1 : \bar{x} > \bar{y}$, принятие которой означает существенное влияние о. Котлин на среднюю температуру воды. Проверке гипотезы отвечает критерий Стьюдента. Рассчитывая его фактическое значение по формуле (4.1), получаем t = 1,62. Теперь определяем критическое значение статистики Стьюдента при числе степеней свободы ν = 9 + 8 – 2 = 15 для односторонней области, соответствующей удвоенному уровню значимости, т. е. 2α.

Из приложения 3 находим $\boldsymbol{t}_{кр}(2\alpha = 0{,}10, \nu = 15) = 1.75$. Поскольку $t < \boldsymbol{t}_{кр}$, то у нас есть основания считать, что расхождения между выборочными средними не являются значимыми. Другими словами, влияние о. Котлин не сказывается существенно на среднем значении температуры воды западнее острова.

Пример 4.2. Как известно, для подавляющего большинства районов Мирового океана характерно очень плохое покрытие его гидрометеорологическими данными вообще и температурой поверхности океана в частности. В связи с этим постоянно возникает вопрос о степени репрезентативности тех или иных архивов «реанализа», содержащих гидрологические характеристики и представляющих собой по существу некие «черные ящики». Естественно, для этого необходимы реперные данные. К их числу, безусловно, относятся уникальные гидрологические наблюдения, измеренные на судне погоды «М», расположенном почти в центре Норвежского моря.

Рассмотрим степень соответствия температуры поверхности океана в районе судна «М» (66° с.ш. и 2° в.д.) и полученной из глобального архива «реанализа» CDAS (Climate Data Assimilation System), сведения о котором приведены в разделе 1. Значения температуры из архива CDAS брались для двухградусного квадрата, центр которого (65,7° с.ш. и 1,9° в.д.) почти совпадает с местоположением судна «М».

В таблице 4.2 приведены первичные статистические характеристики ТПО (выборочные средние и дисперсии) для отдельных месяцев за период 1951−2001 гг. ($N = 51$), а также вычисленные критерии Стьюдента и Фишера. Из сравнения средних видно систематическое занижение данных CDAS в течение всего года, которое колеблется в пределах 0,2−0,5 °С. В среднем за год оно равно 0,3 °С. Кроме того, в большинстве месяцев года проявляется занижение дисперсии данных CDAS, особенно заметное летом. Возникает вопрос – насколько существенны расхождения в оценках средних и дисперсий. Отметим, что критическое значение критерия Стьюдента при $\alpha = 0{,}05$ и $\nu = 101$ равно $\boldsymbol{t}_{кр} = 1{,}98$, а критерия Фишера при $\alpha = 0{,}05$ и $\nu_1 = 50$, $\nu_2 = 50$ равно $\boldsymbol{F}_{кр} = 1{,}60$.

Как видно из таблицы 4.2, для всех 12 месяцев $t > \boldsymbol{t}_{кр}$, т. е. различия между средними значениями значимы. Что касается сравнения величин дисперсий, то расхождения значимы в летне-осенний (июнь–декабрь) период, когда $F > \boldsymbol{F}_{кр}$. В течение января–мая изменчивость ТПО по натурным данным и архива CDAS можно полагать близкой.

Таблица 4.2

Проверка соответствия средних значений и дисперсий ТПО в районе судна погоды «М» и точке с координатами 65,7° с.ш. и 1,9° в.д. для отдельных месяцев периода 1951–2001 гг.

Месяц	Среднее значение, °С		Дисперсия, °С		Критерий Стьюдента	Критерий Фишера
	«М»	CDAS	«М»	CDAS		
Январь	6,65	6,38	0,19	0,12	3,40	1,58
Февраль	6,38	6,11	0,18	0,14	3,41	1,29
Март	6,38	5,99	0,14	0,16	3,06	1,17
Апрель	6,46	6,24	0,13	0,12	3,14	1,12
Май	7,39	7,18	0,15	0,11	2,91	1,40
Июнь	9,10	8,75	0,45	0,18	3,11	2,43
Июль	10,80	10,40	0,66	0,27	2,94	2,43
Август	11,70	11,20	0,69	0,25	3,65	2,76
Сентябрь	10,70	10,40	0,52	0,20	2,50	2,56
Октябрь	9,03	8,80	0,32	0,11	2,47	2,98
Ноябрь	7,78	7,58	0,25	0,09	2,42	2,78
Декабрь	7,10	6,83	0,20	0,10	3,46	1,98
Год	8,28	7,99	0,14	0,07	4,45	2,04

Очевидно, что главной причиной этих расхождений является наличие систематической ошибки в данных архива CDAS. Для ее устранения достаточно к значениям температуры за весь рассматриваемый период времени прибавить 0,3 °С. Действительно, пересчет после этого критерия Стьюдента показал, что для всех месяцев года уже выполняется условие $t < t_{кр}$. В то же время в соответствии со вторым свойством дисперсии ее величина остается постоянной для всех месяцев года, поэтому оценки критерия Фишера в таблице 4.2 не изменятся.

Итак, использование критериев Стьюдента и Фишера позволило выявить не только существенную нерепрезентативность среднемесячных значений ТПО в Норвежском море полученным из архива CDAS, но и в значительной степени устранить ее простым способом.

4.3. Проверка гипотезы соответствия эмпирической и теоретической функций распределения

В настоящее время известно большое число самых разнообразных тестов на проверку соответствия экспериментальных данных заданной теоретической функции распределения. В общем случае такая проверка может быть выполнена с помощью как упрощенных, так и более строгих методов. Приближенные способы позволяют производить быструю проверку по относительно простым тестам (критериям). Более строго это можно осуществить на основе критериев согласия. *Критериями согласия принято называть статистические критерии, предназначенные для проверки соответствия между гипотетической теоретической моделью и реальными данными.* Другими словами, эти критерии показывают, насколько предположения о распределении случайных величин соответствуют экспериментальным данным, т. е. не вступает ли принятая теоретическая модель в противоречие с исходными данными. Учитывая, что такой теоретической моделью для случайной выборки служит закон распределения, то критерии согласия чаще всего применяются для проверки соответствия эмпирической и теоретической функций распределения.

Критериями согласия являются статистики Пирсона χ^2, Колмогорова–Смирнова, Мизеса–Крамера ω^2. Рассмотрим наиболее широко используемые в практических расчетах первые два критерия.

Критерий Пирсона χ^2. Данный критерий является непараметрическим и используется для выборок достаточно большого объема при проверке любых теоретических функций распределения, которые должны быть заданы в дифференциальном виде. Предварительно осуществляется ранжирование ряда и разбиение его на градации. Считается, что длина выборки должна быть $n \geq 50$, причем число градаций должно быть не меньше 5, и желательно, чтобы в каждой из градаций число должно быть минимум 5–7 наблюдений. Последнее требование на практике обычно очень сложно выполнить, так как «хвосты» (края) эмпирического распределения имеют, как правило, более низкую повторяемость по сравнению с его центральной частью.

Прежде всего, формулируется нулевая гипотеза. Например, соответствие эмпирической функции распределения с параметрами $\overline{x}$, s^2 нормальному закону с параметрами m_x, D_x может быть записано

как $H_0 : f\left(\bar{x}, s^2\right) = f\left(m_x, D_x\right)$. Альтернативную гипотезу зададим в обычном виде $H_1 : f\left(\bar{x}, s^2\right) \neq f\left(m_x, D_x\right)$.

В качестве меры расхождения между эмпирическими данными и теоретической функцией распределения используется выражение:

$$\chi^2 = n\sum_{i=1}^{k}\frac{\left(p_i - \boldsymbol{p}_i\right)^2}{\boldsymbol{p}_i} = \sum_{i=1}^{k}\frac{\left(m_i - n\boldsymbol{p}_i\right)^2}{n\boldsymbol{p}_i}, \qquad (4.4)$$

где p_i – эмпирическая вероятность в *i*-й градации; $\boldsymbol{p}_i$ – теоретическая вероятность; k – число градаций в выборке объемом n; m_i – абсолютная эмпирическая частота (число событий) в *i*-й градации.

После того как на основе эмпирических данных по формуле (4.4) вычисляется величина χ^2, осуществляется проверка неравенства $\chi^2 > \boldsymbol{\chi}^2_{\text{кр}}(\alpha,\nu)$ (Приложение 2). При этом число степеней свободы определяется как $\nu = k - \xi - 1$, где ξ – число параметров теоретического распределения. Поскольку для нормального закона $\xi = 2$, то имеем $\nu = k - 3$. *Если данное неравенство выполняется, то нулевая гипотеза отвергается и делается вывод, что эмпирическая функция распределения не соответствует нормальному закону*. Если $\chi^2 < \boldsymbol{\chi}^2_{\text{кр}}(\alpha,\nu)$, то у нас нет оснований отвергать нулевую гипотезу о нормальном распределении случайной величины X. В связи с этим можно полагать, что расхождения между эмпирическими и теоретическими частотами являются незначимыми, т. е. носят случайный характер.

Следует иметь в виду, что градации с малым числом событий ($m < 5$) целесообразно объединять вместе. Естественно, в этом случае величина k определяется по числу окончательных градаций.

Критерий Колмогорова–Смирнова. Данный критерий также является непараметрическим и может быть использован для проверки любой теоретической функции распределения. В отличие от критерия Пирсона он используется для интегральных функций распределения. Поэтому нулевая гипотеза на соответствие эмпирической функции распределения нормальному закону записывается как $H_0 : F\left(\bar{x}, s^2\right) = \boldsymbol{F}\left(m_x, D_x\right)$. Проверка ее осуществляется с помощью статистики D, представляющей собой модуль максимального уклонения между эмпирической $F(x)$ и теоретической $\boldsymbol{F}(x)$ функциями распределения, т. е.

$$D_{x\in(-\infty,\infty)} = \max\left|F(x) - \boldsymbol{F}(x)\right|. \qquad (4.5)$$

Рис. 4.3. Оценка максимального уклонения
между эмпирической $F(x)$ и теоретической $\boldsymbol{F(x)}$ функциями распределения

Статистика D является случайной величиной, предельное распределение которой было установлено Колмогоровым. Оно выражает вероятность того, что при неограниченном возрастании объема выборки значение D не будет превосходить заданного числа λ_0:

$$\lim_{n\to\infty} p\left|D\sqrt{n} > \lambda\right| = \sum (1)^k \exp\left(-2k^2t^2\right) = p(\lambda_0).$$

В практических расчетах более удобно пользоваться величиной λ, которая может быть вычислена как:

$$\lambda = D\sqrt{n}. \qquad (4.6)$$

Оценка величины D как максимального уклонения между $F(x)$ и $\boldsymbol{F}(x)$ демонстрируется на рис. 4.3. Значения статистики $\boldsymbol{\lambda_{кр}}$, зависящие лишь от уровня значимости, затабулированы и приводятся в таблице 4.3.

Таблица 4.3

Распределение статистики $\lambda_{кр}$ в зависимости от уровня значимости α

Уровень значимости	0,30	0,20	0,10	0,05	0,025	0,01	0,005	0,001
Критическое значение $\lambda_{кр}$	0,97	1,07	1,22	1,36	1,48	1,63	1,73	1,95

Главное условие к исходной информации – непрерывность. Поскольку на практике мы имеем дело обычно с дискретными данными, то вариационный (ранжированный) ряд должен быть предварительно сгруппирован по очень малым градациям, чтобы различия между ними были как можно меньше. В принципе статистика

λ может быть вычислена и непосредственно по индивидуальным (несгруппированным) значениям, однако в этом случае к выводам, получаемым с помощью критерия Колмогорова–Смирнова, следует относиться с максимальной осторожностью.

Итак, общая последовательность проверки гипотезы о законе распределения с помощью критерия Колмогорова–Смирнова заключается в следующих пунктах.

4. Строятся эмпирическая функция распределения $F(x)$ и предполагаемая теоретическая функция $\boldsymbol{F}(x)$.

5. Определяется статистика D и вычисляется величина λ.

6. Если выполняется неравенство $\lambda > \boldsymbol{\lambda}_{кр}(\alpha)$, то нулевая гипотеза отвергается и делается вывод, что случайная величина X не соответствует заданному теоретическому закону распределения. В противном случае у нас нет оснований отвергать нулевую гипотезу и, следовательно, она не противоречит тому, что исходные данные распределяются по заданному закону распределения.

Следует иметь в виду, что при использовании данного критерия учитывается лишь наибольшее уклонение эмпирических данных от принятой теоретической функции распределения. Поэтому он использует не всю информацию, заключенную в исходной выборке. Действительно, нетрудно представить, что эмпирические данные систематически уклоняются от принятой теоретической кривой в разные стороны, но не настолько, чтобы сделать значительным максимальное уклонение, т. е. величину D. В этих случаях критерий Колмогорова будет показывать на хорошее соответствие теоретической и эмпирической функций распределения.

Если к этой же выборке применить критерий Пирсона, то в данном случае будет осуществляться суммирование квадратов уклонений для каждой из градаций. Поскольку сумма может оказаться весьма значительной и превысить критическое значение критерия, то эмпирическая функция распределения будет уже не соответствовать теоретической.

Итак, при использовании критериев согласия можно получить противоположные выводы. Какой же из них более верный? На наш взгляд, более точным при проверке данной нулевой гипотезы следует считать критерий χ^2, так как он использует практически всю информацию, содержащуюся в исходной выборке.

Пример 4.3. Как было показано в примере 3.2, эмпирическое распределение поверхностной температуры воды на гидрологической станции в Белом море в летний период является близким

к симметричному. Учитывая важность нормального закона распределения для статистического анализа, выполним оценку степени соответствия исходных данных указанному теоретическому закону на основе критериев согласия Пирсона и Колмогорова. Предварительный анализ значений температуры воды, разбитых на 8 градаций (интервалов), был ранее представлен в таблице 3.1, поэтому воспользуемся оценками эмпирической частоты, которые перенесем в таблице 4.4.

Далее по формуле $\boldsymbol{f}(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \exp\left(-\frac{(x - m_x)^2}{2D_x} \right)$ рассчитываем теоретические оценки вероятности нормальной функции распределения для середин интервалов (таблице 4.4). Но при этом в качестве m_x и D_x берутся выборочные оценки среднего арифметического и стандартного отклонения $(\bar{x} = 11{,}9\ ^\circ\text{C},\ s = 0{,}9\ ^\circ\text{C})$.

Так как эмпирические частоты первого и последнего интервалов малы (меньше 5), то для получения более достоверных результатов целесообразно при использовании критерия Пирсона объединить указанные градации с соседними. Эти оценки приведены в таблице 4.4 в скобках. Итак, теперь уже нетрудно рассчитать статистику χ^2, которая дана в последней графе: $\chi^2 = 2{,}27$. Далее осуществляется проверка неравенства $\chi^2 > \boldsymbol{\chi^2}_{\text{кр}}(\alpha,\nu)$, причем число степеней свободы $v = k - 3 = 6 - 3 = 3$. Принимая уровень значимости $\alpha = 0{,}05$, находим по распределению Пирсона $\boldsymbol{\chi^2}_{\text{кр}} = 7{,}82$. Нетрудно

Таблица 4.4

Проверка соответствия эмпирических данных нормальному закону распределения по критерию Пирсона

Номер градации	Градация, °С	Эмпирическая частота, m_i	Вероятность, $\boldsymbol{p}_i$	Теоретическая частота, $n\boldsymbol{p}_i$	$(m_i - n\boldsymbol{p}_i)^2$	$\frac{(m_i - n\boldsymbol{p}_i)^2}{n\boldsymbol{p}_i}$
1	9,4–10,0	3 (10)	0,017	1,7 (7,6)	5,76	0,758
2	10,0–10,6	7	0,059	5,9		
3	10,6–11,2	11	0,141	14,1	9,61	0,682
4	11,2–11,8	20	0,228	22,8	7,84	0,344
5	11,8–12,4	28	0,247	24,7	10,89	0,441
6	12,4–13,0	19	0,182	18,2	0,64	0,035
7	13,0–13,6	10	0,087	8,7	0,16	0,014
8	13,6–14,2	2 (12)	0,029	2,9 (11,6)		
	Σ	100	0,990	99,0		2,27

видеть, что $\chi^2 < \boldsymbol{\chi}^2_{кр}$. Следовательно, у нас нет оснований отвергать нулевую гипотезу о нормальном распределении генеральной совокупности. Можно полагать, что гипотеза о выбранном нормальном распределении согласуется с опытными данными, а расхождения между эмпирическими и теоретическими частотами носят случайный характер.

Теперь подвергнем проверке нулевую гипотезу о соответствии эмпирической функции распределения нормальному закону с помощью критерия Колмогорова. С этой целью пересчитаем эмпирическую дифференциальную функцию $f(x)$ в интегральную функцию $F(x)$. Эмпирические оценки функции $F(x)$, которые соответствуют накопленной частости (см. таблицу 3.1), приведены в таблице 4.5.

Далее следует рассчитать теоретические оценки $\boldsymbol{F}(x)$. Для этого воспользуемся формулой (3.4):

$$\boldsymbol{F}(x) = 0,5 + 0,5\Phi\left[\frac{(x - m_x)}{D_x}\right].$$

Исходя из этой формулы, для первого значения поверхностной температуры воды получим:

$$\boldsymbol{F}(9,4) = 0,5 + 0,5\Phi\left[\frac{(9,4 - 11,9)}{0,93}\right] = 0,5 + 0,5\Phi(-2,69) =$$

$$= 0,5 - 0,5 \times 0,9928 \approx 0,004.$$

Таблица 4.5

Сравнение эмпирической и теоретической (нормальной) функций распределения для температуры воды на гидрологической станции в Белом море

x	9,4	10,0	10,6	11,2	11,8	12,4	13,0	13,6	14,2
$F(x)$	0,010	0,030	0,100	0,210	0,410	0,690	0,880	0,980	1,000
$\boldsymbol{F}(x)$	0,004	0,021	0,080	0,221	0.449	0,695	0,878	0,964	0,993

Аналогичным образом рассчитываются все остальные оценки функции $\boldsymbol{F}(x)$. Сравнение значений эмпирической и теоретической функций распределения, приведенных в таблице 4.5, показывает, что максимальное расхождение между ними отмечается при температуре T = 11,8 °С. Величина D = |0,410–0,449| = 0,039. Вычислим $\lambda = D(n)^{1/2} = 0,039(100)^{1/2} = 0,39$. Так как $\lambda < \boldsymbol{\lambda}_{кр}$ при любом числе степеней свободы, то можно полагать, что нулевая гипотеза о выбранном нормальном распределении согласуется с опытными данными.

4.4. Приближенные способы проверки нормальности распределения выборки

В некоторых случаях для проверки нормальности исходной выборки достаточным оказывается использование упрощенных схем. Вспомним, что исходя из свойств нормального закона, должны выполняться следующие условия: среднее значение, мода и медиана совпадают, а коэффициенты асимметрии и эксцесса равны нулю. Отсюда следует, что необходима проверка на равенство среднего, моды и медианы друг другу, а коэффициентов асимметрии и эксцесса на равенство нулю.

Вначале осуществляется сравнение характеристик положения. Естественно, при их малых расхождениях относительно друг друга исходная выборка может считаться приближенно нормальной. В отличие от среднего арифметического, ошибки единичных значений моды и медианы оценить довольно сложно, поэтому в качестве критерия точности можно принять условие $\bar{x} \pm k\sigma$, где k – доля стандартной ошибки рассматриваемой характеристики. Если оценки моды и медианы будут отличаться от среднего значения на величину менее $\pm k\sigma$, то такое распределение можно считать близким к нормальному.

Более точно можно оценить степень отклонения от нуля коэффициентов асимметрии и эксцесса на основе критерия Стьюдента. Для этого записывается нулевая гипотеза в виде $H_0 : |As| = 0$ и $H_0 : |Ex| = 0$. Эмпирические оценки статистики Стьюдента вычисляются по следующим формулам:

$$t_{As} = \frac{|As|}{\sigma_A},$$

$$t_{Ex} = \frac{|Ex|}{\sigma_E},$$

где значения σ_A и σ_E определяются соответственно по формулам (5.7) и (5.8). После этого осуществляется проверка неравенств $t_{As} > \boldsymbol{t}_{\text{кр}}(\alpha,\nu = n - 2)$ и $t_{Ex} > \boldsymbol{t}_{\text{кр}}(\alpha,\nu = n - 2)$. Если эти неравенства выполняются, то нулевая гипотеза отвергается и делается вывод, что распределение не соответствует нормальному. В противном случае у нас есть основания полагать, что распределение может носить нормальный характер.

4.5. Проверка гипотезы об однородности выборки

Предположим, что мы имеем две независимые выборки случайных величин X_1 и X_2, описывающих один и тот же процесс (явление). Требуется установить, являются ли они выборками одного и того же неизвестного теоретического распределения или нет. Если статистические параметры случайных величин X_1 и X_2 (среднее выборочное, стандартное отклонение и др.) отличаются друг от друга, то возникает вопрос, являются ли наблюдаемые расхождения следствием объективного различия законов эмпирического распределения $F_1(x)$ и $F_2(x)$, принадлежащих общему теоретическому распределению $\boldsymbol{F}(x)$, или они могут быть объяснены случайностью выборки. Другими словами, нужно проверить нулевую гипотезу вида $H_0 : F_1(x) = F_2(x)$ при альтернативе $H_1 : F_1(x) \neq F_2(x)$. *Если расхождения между этими законами распределения не значимы, то есть основания считать, что выборки принадлежат одной и той же генеральной совокупности и, следовательно, являются однородными.*

Для проверки нулевой гипотезы может быть использован ряд критериев.

Критерий Колмогорова–Смирнова. Он основан уже на рассмотренной выше статистике D, которая в отличие от критерия согласия сравнивает две эмпирические функции распределения, т. е.

$$D = \max \left| F_1(x) - F_2(x) \right|.$$

Затем вычисляется величина:

$$\lambda' = \left[\frac{(n_1 n_2)}{(n_1 + n_2)} \right]^{1/2} \max \left| F_1(x) - F_2(x) \right|, \qquad (4.7)$$

где n_1 и n_2 – объемы выборок, причем необязательно $n_1 = n_2$. Далее проверяется неравенство $\lambda' > \boldsymbol{\lambda}'_{кр}(\alpha)$. Установлено, что для довольно длинных выборок ($n_1 \geq 50$, $n_2 \geq 50$) распределение статистики $\boldsymbol{\lambda}'$ сходится к распределению статистики $\boldsymbol{\lambda}$. Поэтому в данном случае можно воспользоваться распределением $\lambda_{кр}$ (см. таблицу 4.2). Если неравенство $\lambda' > \boldsymbol{\lambda}_{кр}(\alpha)$ выполняется, то нулевая гипотеза отвергается и делается вывод, что выборки не принадлежат одной генеральной совокупности, т. е. не являются однородными. В противоположном случае мы можем предполагать, что выборки являются однородными. Для более коротких выборок используются специальные таблицы. В этом случае приближенную оценку однородности можно получить, если разность между значениями λ' и $\boldsymbol{\lambda}_{кр}(\alpha)$ значительна.

Пример 4.4. Как известно, при измерении осадков на метеостанциях неоднократно происходила смена приборов. В частности, в России в течение довольно длительного периода времени систематические измерения осадков осуществлялись дождемером с защитой Нифера. Затем была произведена замена этого дождемера на осадкомер Третьякова, обладающего улучшенными аэродинамическими качествами благодаря специальной планочной защите. Именно этот осадкомер до настоящего времени остается основным сетевым прибором измерения осадков в России. Требуется проверить, является ли однородной выборка среднемесячных значений осадков после замены дождемера на осадкомер. Объем первой части выборки составил $n_1 = 110$, а второй – $n_2 = 100$. Результаты распределения значений осадков по девяти градациям представлены в таблице 4.6.

Таблица 4.6

Оценка эмпирической повторяемости среднемесячных значений осадков для обеих частей выборки

Градация	Ширина градации, мм/мес	Первая выборка	Вторая выборка
1	25–30	3	5
2	30–35	10	12
3	35–40	15	8
4	40–45	20	25
5	45–50	12	10
6	50–55	5	8
7	55–60	25	20
8	60–65	15	7
9	65–70	5	5
	Σ	110	100

Прежде всего, рассчитываем накопленные частоты для обеих частей выборок Σm_i, используемых для оценок эмпирических функций распределения: $F_1(x) = \dfrac{\sum m_i}{n_1}$ и $F_2(x) = \dfrac{\sum m_i}{n_2}$, распределение которых дается в таблице 4.7. После этого определяем максимальное уклонение между ними, которое отмечается для шестой градации и составляет $D = 0{,}089$.

По формуле (4.7) рассчитываем величину $\lambda' = 0{,}644$. По таблице 4.3 находим, что при уровне значимости $\alpha = 0{,}05$ $\boldsymbol{\lambda}_{кр} = 1{,}36$. Поскольку $\lambda' < \boldsymbol{\lambda}_{кр}$, то у нас есть основание считать, что различия между

Таблица 4.7

Сравнение эмпирических распределений $F_1(x)$ и $F_2(x)$ среднемесячных значений осадков для обеих частей выборки

Градация	Накопленная частота, Σm_{1i}	Накопленная частота, Σm_{2i}	$F_1(x)$	$F_2(x)$	$\lvert F_1(x) - F_2(x)\rvert$
30–35	3	5	0,027	0,050	0,023
35–40	13	17	0,118	0,170	0,052
40–45	28	25	0,254	0,250	0,004
45–50	48	50	0,436	0,500	0,064
50–55	60	60	0,545	0,600	0,550
55–60	65	68	0,591	0,680	0,089
60–65	90	88	0,818	0,880	0,072
65–70	105	95	0.955	0,950	0,005
70–75	110	100	1,000	1,000	0,000

этими законами распределения незначимы, т. е. выборки принадлежат одной и той же генеральной совокупности и, следовательно, общая выборка является однородной.

Критерий Уилкоксона. Данный критерий был предложен Уилкоксоном в 1945 г. для выборок одинакового объема, а затем обобщен в 1947 г. Манном и Уитни для выборок произвольных объемов. Критерий Уилкоксона является непараметрическим и ранговым. **Ранг** – номер места, которое занимает наблюдение в вариационном ряду. Тогда статистики, зависящие только от рангов, называются ранговыми, а критерии, основанные на этих статистиках – *ранговыми критериями.*

Суть этого критерия заключается в следующем. Расположим выборки $x_1, x_2, \ldots, x_m$ и $y_1, y_2, \ldots, y_n$ в общую последовательность в порядке возрастания их значений. Отметим, что m и n могут иметь различную длину, причем примем условие $m \le n$. Если это не так, то выборки следует перенумеровать. Затем каждому значению объединенного ряда присвоим свой ранг (порядковый номер). Пусть, например, общий вариационный ряд имеет вид:

x_1	y_1	x_2	x_3	y_2	x_4	y_3	y_4	x_5	x_6	y_5	y_6
1	2	3	4	5	6	7	8	9	10	11	12

Теперь подсчитаем сумму рангов для каждой выборки (w_x и w_y):

сумма рангов по x: $w_x = 1 + 3 + 4 + 6 + 9 + 10 = 33$;

сумма рангов по y: $w_y = 2 + 5 + 7 + 8 + 11 + 12 = 45$.

Нужно иметь в виду, что условием правильного определения числа ранговых сумм является выполнение следующего равенства:

$$w_x + w_y = \frac{(m+n)(m+n+1)}{2}.$$

В рассматриваемом нами случае имеем 33 + 45 = 78, $\frac{12(12+1)}{2} = 78.$

Заметим, что если несколько значений одной выборки одинаковы, то в общем вариационном ряду им приписываются различные порядковые номера; если же совпадают значения разных выборок, то всем им присваивают один и тот же порядковый номер, равный среднему арифметическому рангов, которые они могли бы иметь до совпадения.

Критерием, лежащим в основе проверки гипотезы однородности, может служить сумма рангов w, в качестве которой при $m < n$ принимается w_x, а при $n = m$ принимается меньшее ее значение. Очевидно, чем меньше отличаются друг от друга суммы рангов по x_i и по y_i, тем выше должна быть степень однородности выборок. Естественно, при $n \approx m$ это возможно в том случае, когда суммы рангов близки к среднему значению:

$$\overline{w} = \frac{(m+n)(m+n+1)}{4}.$$

Проверка нулевой гипотезы $H_0 : f(x) = f(y)$ при альтернативе $H_1 : f(x) \neq f(y)$ осуществляется путем построения доверительных интервалов $\boldsymbol{w}_{\text{ниж}} < w < \boldsymbol{w}_{\text{вер}}$. Если окажется, что сумма рангов $w < \boldsymbol{w}_{\text{ниж}}$ или $w > \boldsymbol{w}_{\text{вер}}$, т. е. она выходит за пределы доверительного интервала, то *нулевая гипотеза об однородности выборок отвергается* и, наоборот, если w попадает внутрь доверительного интервала, то у нас нет оснований отвергать нулевую гипотезу.

При этом проверка гипотезы зависит от длины выборки. Если длина хотя бы одной из выборок превышает 25 значений, то в этом случае нижняя критическая точка $\boldsymbol{w}_{\text{ниж}}(q = \alpha/2, m, n)$ определяется по формуле:

$$w_{\text{ниж}} = \frac{\left[(m+n+1)m - 1\right]}{2} - \boldsymbol{z}_{\text{кр}}\psi, \qquad (4.8)$$

где $\boldsymbol{z}_{\text{кр}}$ – квантиль функции Лапласа, определяемый по Приложению 1 в соответствии с равенством $\Phi\left(\boldsymbol{z}_{\text{кр}}\right) = \frac{(1-\alpha)}{2}$, а величина ψ, имеющая смысл среднего квадратического отклонения суммы рангов, равна:

$$\psi = \left[\frac{mn(m+n+1)}{12} \right]^{1/2}.$$

После этого находится верхняя критическая точка $\boldsymbol{w}_{\text{вер}}$ как:

$$\boldsymbol{w}_{\text{вер}} = \left[(m+n+1)m - 1 \right] - \boldsymbol{w}_{\text{ниж}}. \qquad (4.9)$$

В том случае, если объем обеих выборок не превышает 25 значений, то для нахождения нижней критической точки $\boldsymbol{w}_{\text{ниж}}$ используется специальная таблица Уилкоксона, входными параметрами для которой служат значения m, n и уровень значимости $\alpha/2$. Далее по формуле (4.9) определяется величина верхней критической точки $\boldsymbol{w}_{\text{вер}}$. В зависимости от того, попадает или не попадает величина w в доверительный интервал, делается соответствующий вывод.

В рассматриваемом нами примере, учитывая, что $n = m$, доверительный интервал составляется для $w_x = 33$. Находим нижнюю критическую точку $\boldsymbol{w}_{\text{ниж}}$ при $q = \alpha/2 = 0{,}025$, которая равна $\boldsymbol{w}_{\text{ниж}} = 26$. Верхняя критическая точка равна $\boldsymbol{w}_{\text{вер}} = 52$. Нетрудно видеть, что $26 < 33 < 52$, т. е. w_x попадает в доверительный интервал. Итак, у нас нет оснований отвергнуть нулевую гипотезу.

Следует иметь в виду, что данный критерий наиболее чувствителен к различию выборок по характеристикам положения и довольно слабо реагирует на различие в значениях дисперсий.

Пример 4.5. Воспользуемся данными по осадкам из предшествующего примера. Оценим степень однородности выборки с помощью критерия Уилкоксона. Сначала рассчитаем сумму рангов по меньшей выборке (обозначим ее через x), а затем по второй (обозначим через y). Получим $w_x = 10\,504$, $w_y = 11\,651$. Общая сумма рангов равна $w_x + w_y = \dfrac{(m+n)(m+n+1)}{2} = \dfrac{(210 \times 211)}{2} = 22\,155$. Нетрудно видеть, что сумма рангов подсчитана правильно. Теперь определяем $\boldsymbol{z}_{\text{кр}}$ по равенству $\Phi(\boldsymbol{z}_{\text{кр}}) = \dfrac{(1-\alpha)}{2} = \dfrac{(1-0{,}05)}{2} = 0{,}4975$. По таблице функции Лапласа находим $\boldsymbol{z}_{\text{кр}} = 2{,}81$. После этого вычисляем нижнюю критическую точку $\boldsymbol{w}_{\text{ниж}}$ при $q = \alpha/2 = 0{,}025$. Величина ψ равна $\psi = \left[\dfrac{mn(m+n+1)}{12} \right]^{1/2} = 60{,}8$. В результате имеем:

$$\boldsymbol{w}_{\text{ниж}} = \frac{\left[\left(m+n+1\right)m-1\right]}{2} - \boldsymbol{z}_{\text{кр}}\psi = \frac{\left(211\times 100-1\right)}{2} - 2{,}81\times 60{,}8 = 10\ 379.$$

Осталось найти верхнюю критическую точку $\boldsymbol{w}_{\text{вер}}$:

$$\boldsymbol{w}_{\text{вер}} = \left[\left(m+n+1\right)m-1\right] - \boldsymbol{w}_{\text{ниж}} = 21\ 099 - 10\ 379 = 10\ 720.$$

Составляем доверительный интервал: 10 379 < 10 504 <10 720. Следовательно, у нас нет оснований отвергать нулевую гипотезу. Поэтому мы можем полагать, что выборка среднемесячных значений осадков после замены дождемера на осадкомер остается однородной, т. е. принадлежит одной и той же генеральной совокупности.

Критерий серий. Данный критерий также является непараметрическим, но заметно более простым по сравнению с критерием Уилкоксона. Он был предложен в 1940 г. Вальдом и Вольфовитцем и состоит в следующем. Две выборки случайных величин X_1 и X_2 объемом $n_1 + n_2$ соединяются вместе и строится объединенный вариационный ряд. В этом ряду принадлежность данных к выборкам X_1 и X_2 определяется с помощью кодирующей переменной, принимающей два значения (0 и 1, А и В и т.п.). Полученная таким образом последовательность называется последовательностью кодов. *Серией принято называть участок последовательности, состоящий из идущих подряд одинаковых кодов и ограниченный с обеих сторон противоположными кодами, либо находящийся в начале или конце исходной последовательности.*

Например, в последовательности кодов 0 1 0 0 0 1 1 1 1 1 0 0 имеется пять серий: (0), (1), (0 0 0), (1 1 1 1 1), (0 0). Статистикой критерия является число серий N в последовательности кодов. Понятно, что чем больше число серий и чем меньше их длина, тем выше вероятность однородности двух выборок. Если же эмпирические распределения $F_1(x)$ и $F_2(x)$ несимметричны относительно друг друга, т. е. одно сдвинуто по отношению к другому, то число серий будет мало, но они будут весьма длинными. Следовательно, если нулевая гипотеза верна, то обе выборки будут хорошо перемешаны в вариационном ряду. В противном случае выборки получены из разных генеральных закономерностей.

При достаточно больших объемах выборок ($n_1 \geq 20$ и $n_2 \geq 20$) для проверки нулевой гипотезы используется статистика:

$$Z = \frac{\left[\left|N - \left(T_1 + 1\right)\right| - 0{,}5\right]}{\left(\frac{T_2}{T_3}\right)^{1/2}}, \qquad (4.10)$$

где

$$T_1 = \frac{\left(2n_1n_2\right)}{\left(n_1 + n_2\right)},\ T_2 = 2n_1n_2\left(2n_1n_2 - n_1 - n_2\right),\ T_3 = \left(n_1 + n_2\right)^2\left(n_1 + n_2 - 1\right).$$

Если нулевая гипотеза верна, то статистика Z имеет нормальное распределение. Поэтому для ее проверки используется $z_{кр}$ – квантиль функции Лапласа при уровне доверительной вероятности $p = 1 - \alpha$. *Если $Z > z_{кр}$, то нулевая гипотеза о принадлежности двух выборок одной генеральной совокупности отклоняется.* Если $Z < z_{кр}$, то у нас нет оснований отвергать нулевую гипотезу.

В том случае, когда объемы выборок несущественно меньше 20 значений, то принимается, что статистика Z приближенно подчиняется нормальному закону и соответственно используется $z_{кр}$. Для очень малых выборок построена специальная таблица, в которой критическая область задается неравенствами $N \le N_1$ и $N \le N_2$, где значения N_1 и N_2 определяются объемами выборок n_1, n_2 и уровнем значимости α.

Пример 4.6. В наблюдениях на прибрежных станциях, расположенных на побережье Северного Ледовитого океана, всегда присутствует довольно много пропусков, особенно в солености воды. Поэтому для одной из прибрежных станций были выбраны две непрерывные группы среднегодовых значений солености, одна продолжительностью 15 лет, а другая – 21 год. Задаем нулевую гипотезу в виде $H_0 : F_1(S) = F_2(S)$, т. е. обе выборки взяты из одной генеральной совокупности. Альтернативная гипотеза $H_1 : F_1(S) \neq F_2(S)$ означает, что выборки получены из разных генеральных совокупностей. Присвоим элементам первой группы код 1, а элементам второй группы код 0. Затем объединим выборки, запишем вариационный ряд и составим последовательность кодов:

1 1 0 0 1 0 0 0 1 0 1 0 1 0 0 1 1 1 0 1 1 0 1 0 0 1 0 0 0 0 1 0 1 0 0 0.

Число серий в данной последовательности равно $N = 22$. Теперь вычисляем статистку $Z = 1{,}044$. Далее обратимся к таблице функции Лапласа и получаем, что доверительной вероятности $p = 0{,}95$

соответствует $z_{кр} = 1{,}65$. Нетрудно видеть, что выполняется условие $Z < z_{кр}$, т. е. у нас нет оснований отвергать нулевую гипотезу. Очевидно, рассматриваемые выборки среднегодовых значений солености принадлежат одной генеральной совокупности.

Глава 5. Анализ погрешностей измерений и расчетов

5.1. Основные положения

В общем случае практически любое уравнение (балансовое, гидродинамическое и т.д.) может быть представлено следующим образом:

$$\sum_{i=1}^{k} x_{i,\text{ист}} = 0, \tag{5.1}$$

где k – число членов исходного уравнения; $x_{\text{ист}.i}$ – «истинная» оценка каждой компоненты исходного уравнения. Данное уравнение предполагает отсутствие погрешностей измерений и расчетов, что в действительности никогда не выполняется. Вследствие этого алгебраическая сумма всех членов (5.1) обычно не равна нулю. С учетом сказанного «истинное» уравнение (5.1) приобретает вид:

$$\sum_{i=1}^{k} x_i = \eta, \tag{5.2}$$

где x_i – наблюдаемая компонента исходного уравнения; η – суммарная погрешность определения всех компонент уравнения (5.2), называемая невязкой (дисбалансом).

Величина невязки может быть представлена следующим образом:

$$\eta = \sum_{i=1}^{k} \xi_i + \sum_{i=1}^{k} \delta_i + \sum_{j=1}^{l} \mu_j, \tag{5.3}$$

где ξ_i и δ_i – систематическая и случайная погрешности i-й компоненты исходного уравнения, μ_j – величина не учитываемой в исходном уравнении j-й компоненты; l – число не учитываемых членов (рис. 5.1).

Рис. 5.1. Виды погрешностей и способы их оценивания

Запишем, например, уравнение пресноводного баланса Мирового океана в виде:

$$\int_{\Gamma}(P-E)\,d\Gamma+Q=0,$$

где P – осадки; E – испарение; Г – площадь Мирового океана; Q – глобальный речной сток. В этом уравнении не учитываются подземный приток в океан, не дренируемый русловой сетью рек, и айсберговый сток (в основном с Антарктиды и частично с Гренландии). Величина этих составляющих значительно меньше других компонент уравнения пресноводного баланса, поэтому их не учет не может приводить к существенному влиянию на оценку величины невязки.

В настоящее время для оценки погрешностей существующих методов определения составляющих уравнения (5.2) используются четыре основных метода:

1) сравнение результатов различных независимых методов расчета отдельных составляющих уравнения (5.2);

2) сравнение расчетов составляющих уравнения (5.2) с их измерениями при помощи специальной аппаратуры;

3) оценка вероятных ошибок расчетов путем анализа примененных формул;

4) оценка погрешности расчета всех составляющих путем замыкания уравнения (5.2) при независимом определении всех его членов.

Если первые три способа позволяют оценить лишь ошибки отдельных составляющих уравнения (5.2), то последний способ дает возможность определить непосредственно величину невязки. Как показывает опыт, даже в сравнительно простых ситуациях выделить в «чистом» виде все погрешности чрезвычайно сложно. В связи с этим последний способ, являющийся наиболее объективным, следует рассматривать как основной в общей схеме анализа точности расчетов. Тогда первые три способа, которые можно рассматривать как составные части этой общей схемы, служат для оценки систематических и случайных погрешностей отдельных компонент уравнения (5.2).

Пример 5.1. Рассмотрим анализ погрешностей применительно к уравнению водного баланса Каспийского моря, которое для средних годовых интервалов времени можно записать следующим образом:

$$Q + P - E - \Delta V = \eta, \tag{5.4}$$

где Q – приток речных вод к морю; P – осадки, выпадающие на поверхность моря; E – испарение с акватории моря; ΔV − изменения полезного объема моря. Компоненты водного баланса могут быть выражены либо в единицах объема (км^3/год), либо в единицах слоя (мм/год). Отметим, что сток в залив Кара-Богаз-Гол рассматривается в (5.4) как составляющая испарения.

В уравнении (5.4) не учитываются, прежде всего, приток подземных (не дренируемых русловой сетью) вод (U) и плотностные (стерические) изменения объема (ΔV_ρ) за счет колебаний во времени температуры и солености деятельного слоя моря, вызывающие изменения плотности морской воды, т. е.

$$\sum_{j=1}^{2} \mu_j = U + \nabla V_\rho .$$

Оценить подземный приток вод к морю и особенно его межгодовую изменчивость весьма сложно. Хотя в оценках величины U, полученных разными авторами, отмечаются заметные расхождения, достаточно уверенно можно принять ее норму близкой к 3–4 км^3/год. Учитывая, что величина притока речных вод за разные многолетние периоды времени составляет Q = 275–305 км^3/год, то вклад U в суммарный приток составляет менее 2 %. Поскольку есть основания

полагать, что межгодовые колебания подземного притока, по крайней мере, не превышают самой величины U, то они вряд ли могут заметно сказаться на точности оценок Q. Поэтому в первом приближении достаточно ограничиться учетом нормы U, прибавляя ее к значениям речного стока.

Что касается межгодовых изменений плотностной компоненты ΔV_ρ, то она определяется главным образом колебаниями температуры воды в верхнем слое моря (0–100 м). Межгодовая изменчивость ее носит случайный характер и не превышает нескольких десятых градуса. Характерная величина межгодовых колебаний уровня моря составляет 0,4 см, что в пересчете на изменения объема моря дает $\Delta V_\rho = 1{,}5$ км^3/год. Но поскольку в рассматриваемый период в межгодовом ходе температуры воды отсутствовала трендовая компонента, то плотностные колебания уровня, очевидно, можно не принимать во внимание при расчетах межгодовых изменений водного баланса. Несколько по-иному обстоит дело, если рассматривать внутригодовые изменения составляющих водного баланса. В этом случае, вследствие отчетливо выраженного сезонного хода температуры воды, амплитуда ΔV_ρ уже будет составлять десятки км^3/год.

Выполненный расчет компонент уравнения (5.4) независимыми методами за 60-летний период (1930–1989 гг.) по разным системам исходных данных позволил оценить значения невязок, межгодовой ход которых приводится на рис. 5.2. Ее положительным значениям соответствует превышение притока речных вод и осадков

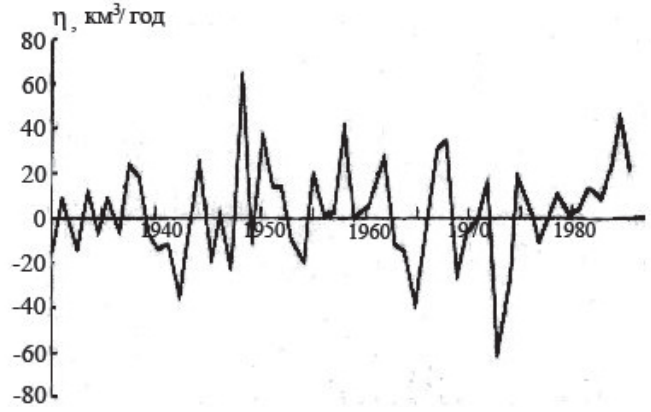

Рис. 5.2. Межгодовой ход значений невязки η уравнения водного баланса Каспийского моря за 1930–1989 гг.

над испарением (с учетом изменений полезного объема ΔV). Наоборот, при отрицательных значениях невязки испарение превышает осадки и приток речных вод. Нетрудно видеть, что в отдельные годы невязки по абсолютной величине являются весьма значительными. Максимальное значение наблюдается в 1948 г. и, судя по всему, обусловлено ненадежной оценкой эффективного испарения ($E-P$) с акватории моря. Также очевидно, что распределение значений невязок носит преимущественно случайный характер, имеет нормальное распределение и при осреднении за 60 лет невязка становится весьма малой. Действительно, осреднение компонент баланса за указанный период времени показало (таблице 5.1), что невязка (дисбаланс) уравнения водного баланса составляет лишь −3,0 км^3/год. Если теперь учесть приток подземных вод к морю, который примем 3,5 км^3/год, то оказывается, что невязка уменьшается до 0,5 км^3/год.

Отсюда следует, что межгодовые колебания уровня обусловлены главным образом соответствующими изменениями водного баланса, т. е. климатическими факторами. Все другие факторы, воздействующие на уровень в рассматриваемом диапазоне времени (тектонические движения земной коры, водообмен через дно моря, стерические колебания уровня, донное осадконакопление и т.п.) являются либо малыми, либо в крайнем случае имеют разнонаправленный характер, вследствие чего их суммарный эффект близок к нулю.

Таблица 5.1

Первичные статистические характеристики составляющих водного баланса Каспийского моря за 1930–1989 гг., км^3/год

Характеристика	Q	P	E	ΔV	η
$\bar{x}$	290,0	76,8	379,7	–9,9	−3,0
σ	45,5	13,4	19,3	54,1	11,2
x_{max}	385,0	109,7	410	102,6	65,0
x_{min}	213,0	49,1	303	–114,5	−98

Естественно, представляет интерес выявление степени связи невязки с отдельными компонентами водного баланса. Из таблицы 5.2 следует, что наиболее высокая корреляция значений невязки отмечается с испарением и изменениями полезного объема моря. Очевидно, что эти составляющие водного баланса определяются с наибольшей случайной погрешностью.

Таблица 5.2

Корреляционная матрица межгодовых колебаний составляющих водного баланса Каспийского моря

Составляющие	η	*P*	*Q*	*E*	Δ*V*
P	–0,11	1,00			
Q	0,18	–0,03	1,00		
E	–0,26	0,09	–0,13	1,00	
Δ*V*	–0,25	0,23	0,80	–0,37	1,00

5.2. Случайные погрешности

Как известно, *случайной погрешностью величины x_i называется погрешность, которая при испытаниях в одинаковых условиях меняется произвольным образом.* Причина ее возникновения заключается в совокупном воздействии на переменную *X* множества различных факторов, каждый из которых обладает собственной погрешностью, причем учесть их в отдельности обычно не представляется возможным.

Мерой случайной погрешности единичного значения любого члена x_i может служить выборочная оценка среднеквадратического отклонения, определяемая по формуле:

$$\sigma_x = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}.$$

Помимо стандартной ошибки существуют и другие меры случайной погрешности. Например, абсолютная случайная погрешность – это погрешность, выражаемая в единицах измеряемой (рассчитываемой) величины, т. е.

$$A_\delta = \left| x - x_{\text{ист}} \right|.$$

Относительная случайная ошибка выражается в долях единицы или в процентах, т. е.

$$\varepsilon = \frac{A_\delta}{x_{\text{ист}}}.$$

Иногда абсолютную ошибку соотносят со стандартным отклонением случайной величины *X* как:

$$A_\delta' = 0,8\sigma_x.$$

При долгосрочном прогнозировании гидрометеорологических процессов величина A_δ' называется допустимой ошибкой прогноза. Наконец, в некоторых случаях используется вероятная случайная ошибка, определяемая по формуле:

$$\sigma_{\text{В}} \approx \left(\frac{2}{3}\right)\sigma_x.$$

При условии распределения исходных данных по нормальному закону и отсутствия внутрирядной связи, что соответствует модели временного ряда «белый шум» (см. гл. 9), случайные погрешности разных характеристик первых четырех статистических моментов определяются следующим образом:

– ошибка выборочного среднего:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{(n-1)^{1/2}}, \tag{5.5}$$

– ошибка среднего квадратического отклонения:

$$\sigma_\sigma \approx \frac{\sigma_x}{(2n-1)^{1/2}}, \tag{5.6}$$

– ошибка коэффициента асимметрии:

$$\sigma_A = \sqrt{\frac{6(n-1)}{(n+1)(n+3)}} \approx \sqrt{\frac{6}{n}}, \tag{5.7}$$

– ошибка коэффициента эксцесса:

$$\sigma_E = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)^2}} \approx \sqrt{\frac{24}{n}}. \tag{5.8}$$

Из формулы (5.5) вытекает важное следствие. С увеличением n случайная погрешность уменьшается и при достаточно больших значениях n $\sigma_{\bar{x}} \to 0$, в результате чего ею можно пренебречь.

В результате автокоррелированности (связности) ряда случайная ошибка занижается. Поскольку в действительности гидрометеорологические ряды очень редко соответствуют модели «белый шум», то возникает необходимость учета их связности. В общем случае это может быть осуществлено путем введения так называемой *эквивалентно-независимой длины временного ряда n**. В результате формула ошибки выборочного среднего приобретает вид:

$$\sigma_{X_{ср}} = \frac{\sigma_x}{\left(n^*_{X_{ср}}\right)^{1/2}}, \qquad (5.9)$$

где $n^*_{X_{ср}} = \dfrac{n}{1+2n^{-1}\left[(n-1)r_1+(n-2)r_2+...+1r_{n-1}\right]}$.

Здесь r_1 – коэффициент автокорреляции на первом сдвиге (лаге) $\tau = 1$, r_2 – коэффициент автокорреляции на втором сдвиге $\tau = 2$ и т.д. Из этой формулы следует, что при отсутствии автокорреляции, т. е. бессвязности ряда, $n^* = n$. Если временной ряд представляет собой простую цепь Маркова, которая характеризуется наличием автокорреляции только между смежными значениями ряда, т. е. при $\tau = 1$, то имеем:

$$\sigma_{x_{ср}} = \frac{\sigma_x}{\sqrt{n^*_{X_{ср}}}} \approx \frac{\sigma_x}{\sqrt{n}}\sqrt{\frac{1+r_1}{1-r_1}}. \qquad (5.10)$$

Аналогичным образом осуществляется учет связности при оценке ошибки стандартного отклонения. Величина n^*_σ вычисляется как:

$$n^*_\sigma = \frac{1}{1+2n^{-1}\left[(n-1)r_1^2+(n-2)r_1^2+...+1r_{n-1}^2\right]}.$$

Тогда случайная ошибка среднего стандартного отклонения для временного ряда, имеющего автокорреляцию между смежными значениями, выражается формулой:

$$\sigma_\sigma = \frac{\sigma_x}{\sqrt{n^*_\sigma}} \approx \frac{\sigma_x}{\sqrt{n}}\sqrt{\frac{1+r_1^2}{1-r_1^2}}. \qquad (5.11)$$

Итак, как следует из приведенных формул, с увеличением степени связности ряда величина случайной погрешности возрастает. Так, при сильной инерционности процесса (например, $r_1 = 0{,}60$), величина ошибки выборочного среднего увеличивается в 2 раза, а ошибка стандартного отклонения возрастает почти в 1,5 раза.

5.3. Систематические погрешности

Систематической погрешностью величины x_i называется погрешность, изменяющаяся по определенному закону. Но поскольку зачастую, исключая инструментальные ошибки, такой закон

не известен, то в первом приближении принимается равенство ее постоянной величине.

В общем случае систематические погрешности можно разделить на четыре вида:

– инструментальные ошибки, возникающие из-за дефектов приборов измерений;

– личные ошибки, связанные с ограниченностью органов чувств самого наблюдателя;

– внешние ошибки, обусловленные недоучетом факторов или изменениями внешней среды (например, влияние корпуса корабля на приборы);

– теоретические ошибки, связанные с методами измерений и расчетов.

В свою очередь, каждый вид систематической ошибки может состоять из отдельных элементарных ошибок. Например, суммарная инструментальная погрешность складывается из многих ошибок отдельных приборов. Очень сложно разделить на элементарные ошибки внешние и теоретические погрешности.

В качестве примера рассмотрим ошибки определения такой легко измеряемой на первый взгляд характеристики, как величина осадков. Специальные экспериментальные исследования показали, что измеренные осадкомером Третьякова на открытых местах жидкие осадки преуменьшены на 5–20 %, а твердые – на 30–50 %. Причиной этого служат систематические ошибки, обусловленные действием ряда факторов. Так, суммарная систематическая погрешность определения осадков может быть выражена как:

$$\sum \xi_{р} = \Delta P_{в} + \Delta P_{и} + \Delta P_{с} - \Delta P_{м},$$

где $\Delta P_{в}$ – ветровая поправка, обусловленная искажением попадания осадков в ведро при ветре; $\Delta P_{и}$ – поправка на испарение части выпавших осадков; $\Delta P_{с}$ – поправка на смачивание; $\Delta P_{м}$ – метелевая поправка, связанная с надуванием поднятых ветром с поверхности земли снежинок. Установлено, что наибольший вклад в суммарную систематическую погрешность дает ветровой фактор. Заметим, что при суммировании нескольких видов систематической погрешности их сумма уже может рассматриваться как случайная погрешность.

Еще более сложной задачей является определение осадков над открытой водной поверхностью, точность которых практически не поддается количественной оценке именно по причине

большого числа систематических ошибок. Помимо уже рассмотренных выше систематических ошибок (исключая метелевый фактор) при измерении осадков на палубе научно-исследовательских судов дополнительно следует учитывать ошибки за счет попадания в приемное отверстие прибора брызг морской воды, капель и брызг с судовых надстроек и мачт, а также отклонения плоскости приемного отверстия от горизонтали из-за качки. Возможное сочетание погрешностей, обусловленных всеми факторами, в реальных условиях весьма разнообразно и практически не поддается строгому количественному учету. Именно поэтому осадки над морем считаются наиболее плохо определяемой составляющей водного баланса.

Следует иметь в виду, что поскольку в уравнении (5.2) присутствуют обычно все виды систематических ошибок, то отделить их друг от друга, как правило, не представляется возможным. Поэтому, очевидно, имеет смысл определять лишь суммарную систематическую погрешность каждой компоненты $x_i(\xi)$ или даже их общую сумму $\sum\xi_i$.

Для выявления систематических погрешностей можно, например, использовать сравнение измеренных (рассчитанных) компонент уравнения (5.2) с их «эталонными» измерениями (расчетами). При этом должны соблюдаться следующие условия: систематическая погрешность «эталонных» измерений должна быть мала, а случайная погрешность «эталонных» и обычных (сетевых) измерений – примерно одинакова. Так, в результате тщательных экспериментальных исследований установлено, что очень близкие к действительным суммам жидких, смешанных и твердых осадков можно получить, если стандартный осадкомер разместить в массиве густого лиственного кустарника высотой 2–4 м.

5.4. Понятие о косвенных погрешностях

Под *косвенной погрешностью случайной величины Y понимается такая погрешность, которая непосредственно не определяется, но может быть вычислена через измеряемые параметры* x_1, x_2, ... , x_m, *используемые для оценки Y, т. е.* $y_i = f(x_1, x_2, \ldots, x_m)$.

Предположим, что отдельные погрешности параметров x_j подчиняются нормальному закону распределения, по абсолютной величине значительно уступают средним значениям $\left(\sigma_1 \ll \overline{x}_1, \ \ldots\ , \sigma_m \ll \overline{x}_m\right)$ и некоррелированы друг с другом. В этом случае для оценки косвенных

Таблица 5.3

Формула для оценки случайных погрешностей различных зависимостей

Вид зависимости	Абсолютная погрешность	Относительная погрешность
$y = x_1 + x_2$	$\sigma_y = \sqrt{\sigma_1^2 + \sigma_2^2}$	$\varepsilon_y = \dfrac{\sqrt{\sigma_1^2 + \sigma_2^2}}{\bar{x}_1 + \bar{x}_2}$
$y = x_1 - x_2$	$\sigma_y = \sqrt{\sigma_1^2 + \sigma_2^2}$	$\varepsilon_y = \dfrac{\sqrt{\sigma_1^2 + \sigma_2^2}}{\bar{x}_1 - \bar{x}_2}$
$y = x_1 x_2$	$\sigma_y = \sqrt{x_2\sigma_1^2 + x_1\sigma_2^2}$	$\varepsilon_y = \sqrt{\dfrac{\sigma_1^2}{\bar{x}_1^2} - \dfrac{\sigma_2^2}{\bar{x}_2^2}}$
$y = x_1 / x_2$	$\sigma_y = \sqrt{\dfrac{\left(x_2\sigma_1^2 + x_1\sigma_2^2\right)}{x_2^2}}$	$\varepsilon_y = \sqrt{\dfrac{\sigma_1^2}{\bar{x}_1^2} + \dfrac{\sigma_2^2}{\bar{x}_2^2}}$
$y = x_1 x_2 x_3$	$\sigma_y = \sqrt{\bar{x}_2^2\bar{x}_3^2\sigma_1^2 + \bar{x}_3^2\bar{x}_1^2\sigma_2^2 + \bar{x}_1^2\bar{x}_2^2\sigma_3^2}$	$\varepsilon_y = \sqrt{\dfrac{\sigma_1^2}{\bar{x}_1^2} + \dfrac{\sigma_2^2}{\bar{x}_2^2} + \dfrac{\sigma_3^2}{\bar{x}_3^2}}$
$y = ax_1 + bx_2 + cx_3$	$\sigma_y = \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2 + c^2\sigma_3^2}$	$\varepsilon_y = \sqrt{\dfrac{a^2\sigma_1^2 + b^2\sigma_2^2 + c^2\sigma_3^2}{a\bar{x}_1 + b\bar{x}_2 + c\bar{x}_3}}$

погрешностей может быть использован так называемый *метод частных погрешностей:*

$$\sigma_y = \sqrt{\left(\frac{\partial f}{\partial x_1}\right)^2 \sigma_1^2 + \left(\frac{\partial f}{\partial x_2}\right)^2 \sigma_2^2 + \ldots + \left(\frac{\partial f}{\partial x_n}\right)^2 \sigma_m^2}. \qquad (5.12)$$

Придавая функции f конкретный вид, можно получить аналитические формулы для оценки случайных погрешностей. Формулы для некоторых простейших зависимостей приводятся в таблице 5.3. Заметим, что особенностью данного метода является то, что он оказывается корректным только для абсолютных погрешностей. Их относительные значения должны находиться соответствующим пересчетом. Конкретные примеры такого пересчета также указаны в таблице 5.3. Если ошибки коррелированы друг с другом, что, вообще говоря, нельзя упускать из виду, то формулы для оценки случайных погрешностей усложняются. Так, например, если $y = x_1 + x_2$, то

$$\sigma_y = \sqrt{\sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2 r_{1,2}},$$

где $r_{1,2}$ – коэффициент корреляции между погрешностями x_1 и x_2. Чем больше величина $r_{1,2}$, тем существеннее роль последнего слагаемого в формировании суммарной погрешности σ_y. Однако в зависимости от знака $r_{1,2}$ суммарная погрешность σ_y может либо увеличиться, либо уменьшиться. При $r_{1,2} > 0$ σ_y увеличивается, при $r_{1,2} < 0$ σ_y уменьшается.

Если $y = ax_1 + bx_2 + cx_3$, то в этом случае:

$$\sigma_y = \sqrt{a\sigma_1^2 + b\sigma_2^2 + c\sigma_3^2 + 2\sigma_1\sigma_2 r_{1,2} + 2\sigma_1\sigma_3 r_{1,3} + 2\sigma_2\sigma_3 r_{2,3}}.$$

Таким образом, наличие корреляции между погрешностями существенно усложняет анализ и оценку определения суммарной погрешности σ_y.

Итак, расчет погрешности косвенных измерений состоит из двух этапов. Первый этап – это вывод формулы для абсолютной или относительной погрешности результата косвенных измерений исходя из вида функции $y_i = f(x_1, x_2, \ldots, x_m)$. Второй этап – расчет погрешности в соответствии с полученной формулой путем суммирования ее составляющих по правилам суммирования случайных погрешностей с учетом корреляционных связей.

Пример 5.2. Соленость в автоматизированных системах зондирования морских вод, как правило, непосредственно не измеряется. К непосредственно измеряемым с необходимой точностью параметрам относятся температура t, удельная электропроводность χ, гидростатическое давление P и скорость распространения звука c. Для расчета солености достаточно измерить три из указанных четырех параметров. Если, например, принять $S = f(t,\chi,P)$, то в этом случае средняя квадратическая погрешность вычисления солености будет определяться выражением:

$$\sigma_S = \left[\left(\frac{\partial S}{\partial t} \right)_{x,P}^2 \sigma_t^2 + \left(\frac{\partial S}{\partial \chi} \right)_{t,P}^2 \sigma_\chi^2 + \left(\frac{\partial S}{\partial P} \right)_{t,x}^2 \sigma_P^2 \right]^{\frac{1}{2}}.$$

Средние квадратические погрешности первичных параметров, а также оценки частных производных, приводятся ниже:

$$\sigma_t = 0{,}05\ ^\circ\text{C};\ \ \sigma_\chi = 0{,}05\ \text{мСм}\cdot\text{см}^{-1};\ \ \sigma_p = 10^{-2}\ \text{МПа};$$

$$(\partial S/\partial t)_{\chi,p} = -1{,}145\ ‰/^\circ\text{C};\ \ (\partial S/\partial(\chi)_{t,p} = 1{,}34\ \%\ /\ (\text{мСм}\cdot\text{см}^{-1});$$

$$(\partial S/\partial P)_{t,\chi} = -0{,}06\ ‰/\text{МПа}.$$

Используя эти данные, нетрудно оценить стандартную погрешность вычисления солености, которая оказывается равной σ_S = 0,0087 ‰. Аналогичным образом могут быть определены погрешности вычисления солености для других комбинаций первичных параметров. Опуская промежуточные цифры, приводим сразу окончательные результаты:

$$S = f(t,\chi,c), \quad \sigma_S = 0{,}0096\ ‰,$$
$$S = f(t,c,\mathrm{P}), \quad \sigma_S = 0{,}075\ ‰,$$
$$S = f(\chi,c,\mathrm{P}), \quad \sigma_S = 0{,}088\ ‰.$$

Нетрудно видеть, что наиболее точные результаты при расчете солености могут быть получены, если в качестве первичных параметров используются температура, удельная электропроводность и гидростатическое давление. Лишь немного по точности уступает оценка солености по данным о температуре, удельной электропроводности и скорости звука.

5.5. Выявление и устранение грубых погрешностей

В статистических рядах довольно часто можно обнаружить *выбросы – резко выделяющиеся наблюдения, которые существенно отклоняются от распределения остальных выборочных данных.* Эти данные могут отражать как экстремальные свойства изучаемого явления (переменной), так и быть обусловлены ошибками измерений, расчетов, возникающих в результате ручной или машинной обработки. В первом случае выбросы представляют особый интерес, поскольку они связаны обычно со стихийными природными процессами. Так, например, мощнейшее цунами в Индийском океане в декабре 2004 г. вызвало нагонную волну высотой 5–10 м, обрушившуюся на побережье Таиланда, Индонезии и других стран, которая привела к катастрофическим разрушениям и гибели только по официальным данным 223 тыс. человек. Поэтому экстремальная аномалия уровня обязательно должна учитываться в статистических расчетах, ибо отражает реально произошедшее катастрофическое событие.

В то же время, если, например, в данных футшточных наблюдений на постах Таиланда за ноябрь 2004 г. будет присутствовать аналогичная величина уровня, то она должна быть исключена из анализа, поскольку является грубой ошибкой и ничем более.

Грубые ошибки могут приводить к существенному искажению получаемых результатов и соответственно к их неправильной интерпретации. Поэтому выявление и исключение грубых ошибок (промахов), относящихся по характеру своего происхождения к случайным погрешностям, является важной задачей первичного анализа информации.

Грубые ошибки в статистических данных должны выявляться, прежде всего, путем физического анализа и, желательно, в реальном режиме времени. Если они отличаются от основной массы данных на порядок и более, то их выявление и устранение не представляет особых затруднений и может быть осуществлено визуально. Значительно более сложной является задача нахождения промахов при ретроспективном анализе данных, особенно в тех случаях, когда они не слишком сильно отличаются от других результатов, а физический анализ процессов, приводящих к формированию сомнительных оценок в данных, оказывается невозможным. Очевидно, в этом случае без использования специальных статистических приемов, которые представлены на рис. 5.3, не обойтись.

Рассмотрим наиболее простые методы. Пусть мы имеем вариационный ряд $x_1, x_2, \ldots, x_n$, причем величина x_n резко выделяется. Необходимо решить вопрос о принадлежности x_n остальным наблюдениям исследуемой выборки. С этой целью составляется нулевая

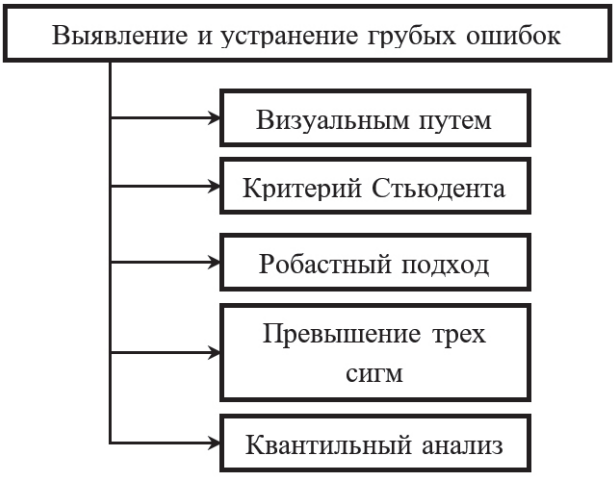

Рис. 5.3. Способы выявления и устранения грубых погрешностей

гипотеза вида $H_0 : |\overline{x} = x_n|$, проверка которой при условии нормальности исходных данных осуществляется с помощью критерия Стьюдента:

$$t = \frac{|\overline{x} - x_n|}{\sigma}. \qquad (5.13)$$

Здесь выборочные характеристики $\overline{x}$ и σ вычисляются без учета величины x_n. Затем проверяется выполнение неравенства $t > t_{кр.}$ (α, ν = n – 1). Если это неравенство выполняется, то нулевая гипотеза отвергается и делается вывод, что резко отличающееся наблюдение x_n входит в противоречие с данной выборкой и поэтому может быть из нее исключено. Если это неравенство не выполняется, то мы можем полагать, что крайнее наблюдение x_n нецелесообразно исключать. После исключения крайнего значения, данную процедуру можно повторить и для следующего по абсолютной величине максимального отклонения, но предварительно необходимо пересчитать $\overline{x}$ и σ для выборки нового объема n – 1.

В некоторых случаях для малых выборок вместо (5.13) вычисляется отношение:

$$t = \frac{|\overline{x} - x_n|}{\sigma\sqrt{\frac{(n-1)}{n}}}. \qquad (5.14)$$

Далее процедура обнаружения грубых погрешностей аналогична изложенной выше. Введение множителя $\sqrt{\frac{(n-1)}{n}}$ учитывает смещенность оценок при малых объемах выборки. Данный способ выявления грубых ошибок весьма прост и легко применим на практике, однако он имеет существенные недостатки. В частности, он оказывается нечувствительным, если в исследуемой выборке выбросы группируются вместе, но они отстоят довольно далеко от основной массы наблюдений. Кроме того, далеко не всегда исходная выборка имеет нормальное распределение.

Более точным, по сравнению со статистикой Стьюдента, способом оценки грубых ошибок представляется робастный подход. В переводе с английского *robust – крепкий*, *здоровый*. Однако в статистике под термином «робастный» понимается «устойчивый».

Строго говоря, термин *«робастность» означает нечувствительность к малым отклонениям от предположений.* Применительно к анализу выбросов робастное оценивание заключается в получении надежных статистических критериев случайной величины с учетом неявности ее закона распределения и наличия существенных отклонений в значениях данных.

Согласно П. Хьюберу, все робастные методы можно разделить на подходы, базирующиеся на применении *L*-оценок, *R*-оценок и *M*-оценок. Первые представляют линейные комбинации порядковых статистик (медиана, винзорированное среднее и т.п.), вторые определяются на основе ранговых статистик, а третьи вычисляются аналогично методу максимального правдоподобия с помощью различных весовых функций. При этом статистическое оценивание выборки в случае ее «засорения» данными, резко отличающимися от основной совокупности, осуществляется с помощью *L* и *E*-критериев. Для верхней части ранжированного ряда *L*-критерий имеет вид:

$$L = \frac{\sum_{i=1}^{n-k} \left(x_i - \overline{x}_k \right)^2}{\sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2}, \qquad (5.15)$$

где x_i – исходные данные в выборке, распределенной по какому-либо одному признаку; n − объем выборки; k − число наблюдений с резко отклоняющимися значениями признака; $\overline{x}$ – общая для всей совокупности средняя величина; $\overline{x}_k$ – средняя величина совокупности из $n - k$ наблюдений.

Для нижней части ранжированного ряда используется L'-критерий, по смыслу идентичный *L*-критерию:

$$L' = \frac{\sum_{i=k+1}^{n-k'} \left(x_i - \overline{x}_{k'} \right)^2}{\sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2}, \qquad (5.16)$$

где $\overline{x}$ – средняя величина, рассчитанная по $n - k$ наблюдениям, остающимся после отбрасывания k–грубых ошибок «снизу».

E-критерий используется, когда предположительно в выборке присутствуют грубые ошибки с максимальным или минимальным значениями ряда. Данный критерий имеет вид:

$$E = \frac{\sum_{i=k+1}^{n-k'} \left(x_i - \overline{x}_{k'}\right)^2}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}, \qquad (5.17)$$

где $\overline{x}_{k'}$ – средняя величина, вычисленная по «истинным» данным после отбрасывания из выборки наименьших (k) и наибольших (k') значений, засоряющих исходную совокупность.

Данные критерии имеют табулированные критические значения для заданного уровня значимости α при известном объеме выборки и предполагаемом числе ошибок. Если наблюденные значения критериев оказываются меньше критических (пороговых) оценок, то ошибки в данных, подвергаемые проверке, признаются грубыми, существенно отклоняющимися от основного массива данных. При обратном соотношении оценок указанных критериев данные гипотетически предполагаются типичными для изучаемой совокупности.

Очевидным недостатком указанных критериев является то, что на практике величина k, как правило, заранее неизвестна. Последнее обстоятельство существенно влияет на оценку их критического уровня и, следовательно, на получаемые результаты.

Иногда для исключения грубых погрешностей используется так называемое «правило трех сигм». Если предположить, что случайные ошибки подчиняются нормальному закону распределения, то в этом случае плотность вероятности случайных ошибок описывается формулой:

$$f(\delta) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-\delta_i^2}{2\sigma^2}}. \qquad (5.18)$$

Исходя из свойств нормального распределения $f(|\delta| < 3\sigma) = 0{,}9973$. Тогда $f(|\delta| > 3\sigma) = 0{,}0027$, т. е. вероятность превышения какой-либо величины $|x_m - \overline{x}|$ значения 3σ ничтожно мала и составляет менее 0,3 %. Поэтому, если выполняется условие:

$$\left|\overline{x} - x_n\right| > 3\sigma, \qquad (5.19)$$

где x_n – крайнее значение в ряду, то его следует исключить. Необходимо отметить, что данный критерий при решении некоторых задач считается излишне жестким. В этих случаях используется условие:

$$\left|\bar{x} - x_n\right| > 2\sigma, \qquad (5.20)$$

которому соответствует вероятность появления случайной ошибки в исследуемом ряду, равная ≈ 5 %.

Кроме того, для выявления грубых ошибок можно воспользоваться квантильным анализом, основы которого рассмотрены в разделе 2. В квантильном анализе точка, резко выделяющаяся от остальной совокупности, называется выбросом, если для нее выполняются следующие условия:

– для положительных аномалий: $x_{выб} > x0{,}75 + kQ$,

– для отрицательных аномалий: $x_{выб} < x0{,}25 - kQ$,

где k – подгоночный коэффициент, называемый коэффициентом выбросов. В пакете «Статистика» по умолчанию он принимается равным $k = 1{,}5$. Для случайной выборки, имеющей нормальное распределение, в диапазоне $\pm kQ$ содержится 99 % значений выборки. Если принять величину $k = 3$, то в диапазоне $\pm 3Q$ содержится – 99,9997 % значений выборки.

Следует иметь в виду, что применение указанных выше статистических критериев для выявления и устранения грубых погрешностей должно осуществляться с максимальной осторожностью, опираясь на всесторонний физический анализ условий, формирующих исследуемый процесс. В противном случае возможен вариант, при котором должны быть отвергнуты в соответствии с рассмотренными выше критериями крайние значения статистического ряда, реальность существования которых сомнений не вызывает.

Пример 5.3. В январе 1989 г. температура воздуха на одной из метеостанций Ленинградской области составила 0 ºС. При этом средняя многолетняя величина и стандартное отклонение температуры воздуха за 50 лет соответственно равны –8 ºС и 2,5 ºС. По формуле (5.13) получаем $t = 3{,}5$, а критерий Стьюдента при уровне значимости $\alpha = 10$ % составляет $\boldsymbol{t}_{кр} = 1{,}96$. Следовательно, проверяемое значение температуры воздуха должно быть исключено из выборки. Аналогичный результат получается и при использовании правила трех сигм. Естественно, что такой вывод не соответствует действительности и поэтому не должен быть принят во внимание.

Пример 5.4. Визуальный анализ среднемесячных наблюдений за температурой воды на прибрежной гидрометеорологической станции Болванский Нос ($\varphi = 70°27.6'$ с.ш., $\lambda = 59°07.5'$ в.д.) за период с 1963 по 2004 г. показал наличие сомнительных данных, причем в некоторые годы наблюдения отсутствовали. Поскольку

ретроспективный физический анализ сопутствующих этим данным гидрометеорологических условий оказался невозможен, то для выявления выбросов мы воспользовались изложенными выше статистическими приемами.

В таблице 5.4 представлены первичные статистические характеристики температуры воды для станции Болванский Нос. Нетрудно видеть, что длина наблюдений в разные месяцы различна, причем она минимальна в зимний период. Однако, учитывая, что именно зимой температура практически постоянна, то это обстоятельство не должно нас беспокоить. Здесь же даны критические значения статистики Стьюдента, причем от уровня значимости α в значительной степени будет зависеть отнесение данных к выбросам. Например, при α = 0,025 и $n \geq 30$ имеем $\boldsymbol{t}_{кр} = 2$, т. е. критерий Стьюдента совпадает с критерием «двух сигм». Увеличение критерия значимости существенно занижает величину $\boldsymbol{t}_{кр}$ и тем самым будет завышать число выбросов. Уменьшение его до α ≥ 0,005 или 0,001, наоборот, приближает $t_{кр}$ к критерию «трех сигм», что является малоинформативным. Поэтому мы приняли промежуточный вариант: α = 0,01.

При анализе значений температуры воды обращает на себя внимание наличие ее больших отрицательных значений в течение

Таблица 5.4

Первичные статистические характеристики температуры воды на ГМС Болванский Нос

Месяц	Статистическая характеристика					
	Длина ряда	Среднее арифметическое	Стандартное отклонение	X_{max}	X_{min}	Критерий Стьюдента (α = 0,01)
Январь	29	–1,96	0,55	–1,60	–3,60	2,46
Февраль	29	–2,00	0,57	–1,70	–3,70	2,46
Март	31	–1,98	0,55	–1,70	–3,70	2,46
Апрель	30	–1,77	0,07	–1,60	–1,90	2,46
Май	29	–1,65	0,20	–1,00	–1,80	2,46
Июнь	30	–0,69	0,98	1,70	–1,70	2,46
Июль	38	2,48	2,22	7,30	–0,90	2,43
Август	41	4,31	2,40	8,40	–0,30	2,42
Сентябрь	39	3,57	1,75	7,60	–0,70	2,43
Октябрь	40	1,63	1,32	4,80	–1,20	2,42
Ноябрь	35	–0,80	0,79	1,20	–1,80	2,44
Декабрь	32	–1,57	0,31	–0,30	–1,80	2,45

периода с января по март (–3,5, –3,7 ºС). Всего таких значений зарегистрировано 9. Оценить реальность существования подобных выбросов нетрудно, поскольку температура замерзания морской воды в зависимости от солености приближенно описывается формулой:

$$T_3 = -0{,}003 - 0{,}0527S - 0{,}00004S^2 - 0{,}0000004S^3.$$

Подставляя в эту формулу среднее значение солености, получим, что температура замерзания составляет $T_3 = -1{,}9$ ºС. Таким образом, совершенно очевидно, что значения температуры воды ниже –1,9 ºС являются грубыми ошибками, которые следует исключить из исходной выборки. Для последующих месяцев такие резко выделяющиеся значения не отмечаются, поэтому нами для выявления выбросов использованы описанные выше приемы. В таблице 5.5 приводится число выбросов для температуры воды на ст. Болванский Нос, полученных с помощью разных критериев.

Таблица 5.5

Число выбросов в межгодовом ходе температуры воды на ГМС Болванский Нос

Месяц	Критерий			
	Квантильный анализ	Стьюдента	Три сигма	Робастная оценка
Апрель	0	0	0	0
Май	3	0	0	0
Июнь	0	1	0	0
Июль	0	0	0	0
Август	5	0	0	0
Сентябрь	0	1	1	0
Октябрь	0	0	0	0
Ноябрь	0	1	0	0
Декабрь	3	4	1	0

При построении «ящика с усами» использовалась величина $k = 1{,}5$. Робастное оценивание осуществлялось по E-критерию для максимального и минимального значений ряда. Как оказалось, E-критерий является чересчур мягким, то есть выбросов он не выделяет. Из таблицы 5.5 следует, что полученные оценки выбросов неплохо согласуются друг с другом. Пожалуй, только в августе число выбросов, полученное по квантильному анализу, сильно отличается от других оценок. Поэтому обратимся к рис. 5.4, на котором приведены оценки выбросов, определенные квантильным анализом. При этом использовались два варианта коэффициента k ($k = 1{,}5$

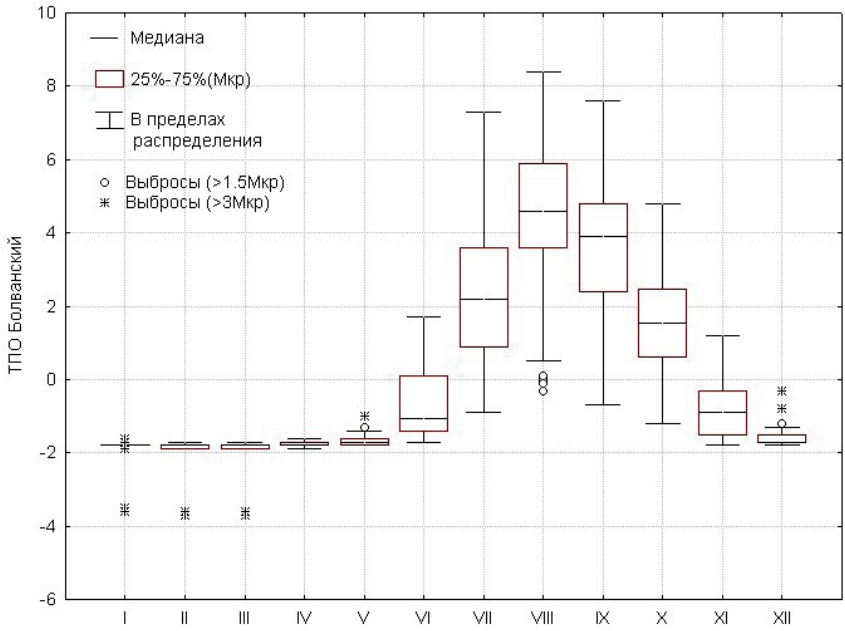

Рис. 5.4. Квантильный анализ наблюдений за температурой воды на ГМС Болванский Нос

и k = 3,0). Отметим, что все выбросы в августе регистрируются по величине k = 1,5.

Дополнительный анализ градиентов $\Delta x = x_{i+1} - x_i$, где i – номер месяца, показал, что градиент для выбросов, приведенных в таблице 5.5, незначимо по критерию Стьюдента отличается от других градиентов. Фактически это означает, что найденные нами выбросы следует считать скорее экстремальными оценками температуры воды, чем ее грубыми ошибками.

5.6. Понятие о теории выбросов

Поскольку аномальные условия формирования гидрометеорологических процессов имеют важное практическое значение, то для их изучения используется специальный раздел теории случайных функций – *теория выбросов*. Заметим, что нахождение закономерностей выбросов является весьма сложной математической задачей. Кроме того, использование теории выбросов предполагает наличие

весьма длинных статистических рядов, что далеко не всегда оказывается возможным. Поэтому многие исследования направлены на изыскание пригодных для практического использования приближенных формул и на развитие методов изучения характеристик выбросов путем статистического моделирования на ЭВМ.

Как уже отмечалось выше, значительные аномалии во временном ряду далеко не во всех случаях относятся к грубым погрешностям, а могут характеризовать экстремальные свойства случайного процесса. Например, если взять временной ряд срочных значений уровня Невы (ст. Горный институт) в осенний период года, то на фоне мало меняющихся средних значений будут обнаруживаться серии повышений уровня, обусловленных подпором воды в устье Невы при прохождении интенсивных циклонов над Финским заливом. Временной интервал времени, в течение которого уровень будет превышать отметки 160 см по Кронштадскому футштоку, интерпретируется как наводнение. Со статистической точки зрения серии повышенных значений уровня означают наличие во временном ряду уровня выбросов. В общем случае *выброс можно трактовать как участок реализации временного ряда, лежащий выше или ниже некоторого заданного уровня C.* Соответственно этому имеем положительные и отрицательные выбросы.

Непосредственный расчет характеристик выбросов является весьма трудоемкой задачей и требует наличия очень длинных реализаций случайного процесса. В то же время, если он имеет нормальное распределение, то для отдельных характеристик выбросов могут быть получены сравнительно простые расчетные формулы. Так, среднее число выбросов выходящее за уровень C на интервале T может быть вычислено по следующей приближенной формуле:

$$\bar{N}_c = \bar{N}_a \exp\left[-\frac{(C-a)(C+a-2m_x)}{2\sigma_x^2}\right], \qquad (5.21)$$

где $\bar{N}_a$ – среднее число выбросов за уровень a. Итак, чтобы рассчитать число выбросов за уровень C, необходимо предварительно определить среднее число выбросов за уровень a.

Другой важной характеристикой выбросов является среднее время пребывания случайного процесса выше уровня C (продолжительность выброса). Для его оценки может быть использована формула вида:

$$\overline{T}_A = T\left[0{,}5 - \Phi(t)\right], \tag{5.22}$$

где $\Phi(t)$ – функция Лапласа, определяемая по формуле (3.3); T – продолжительность интервала, на котором оцениваются выбросы. Зная значения $\overline{N}_c$ и $\overline{T}_c$, можно оценить среднюю длительность единичного выброса:

$$\overline{\theta} = \frac{\overline{N}_A}{\overline{T}_c}. \tag{5.23}$$

Во многих случаях важно знать мощность выброса S_c, которая определяется совокупным действием превышения и продолжительности выброса. Для нормально распределенных процессов приближенная формула оценки мощности выбросов за уровень C выражается формулой:

$$g\left(S_c\right) = \left(\frac{1}{3}\right)\lambda^{2/3} S_c^{\,-1/3} \exp\left[-0{,}5\left(\lambda S_c\right)^{2/3}\right], \tag{5.24}$$

где $\lambda = \dfrac{3\left(C - m_x\right)^2\left[-r''(\tau)\right]^{1/2}}{2\sigma_x^3}$, $r''(\tau)$ – вторая производная нормированной автокорреляционной функции (10.22). Еще более сложный вид имеет формула оценки среднего числа максимумов, амплитуда которых превышает заданный уровень C. Заметим также, что формулы (5.21)–(5.24) относятся к непрерывным процессам. При переходе к дискретному случайному процессу дополнительно принимаются условия его стационарности и эргодичности, а сами формулы для оценки отдельных характеристик выбросов еще более усложняются. Это является одной из причин того, что теория выбросов пока не получила широкого распространения на практике.

Часть 2. Построение эмпирических зависимостей

Глава 6. Корреляционный анализ

6.1. Виды связей между двумя переменными

При анализе гидрометеорологических явлений или процессов очень часто возникает необходимость в установлении связи между ними. В общем случае эта связь может быть трех типов: функциональной (детерминированной), стохастической (вероятностной) и случайной, характеризующей полное отсутствие связи.

Функциональная связь. Если каждому значению одной переменной соответствует единственное значение другой переменной, то такая зависимость носит название функциональной (рис. 6.1-*а*). Функциональные зависимости обычно могут быть доказаны теоретическим путем (например, с помощью логических рассуждений) и не нуждаются в опытной проверке. Естественно, что доверительная вероятность таких связей равна единице ($p = 1$).

Случайная связь. Если любому значению одной переменной может соответствовать практически любое значение другой переменной, то это означает, что связь отсутствует, и переменные X и Y являются независимыми по отношению друг к другу (рис. 6.1-*б*). В общем случае условие независимости выражается следующим образом: $F(x,y) = F_1(x)F_2(x)$, т. е. функция распределения системы независимых случайных величин равна произведению функций распределения отдельных случайных величин. Доверительная вероятность такой связи равна нулю ($p = 0$).

Стохастическая (вероятностная) связь. Если каждому значению одной переменной с определенной вероятностью ($0 < p < 1$) соответствует значение другой переменной, то такая зависимость называется стохастической. Естественно, что она носит промежуточный характер между функциональным и случайным типами связи (рис. 6.1-*в*). Стохастическая зависимость характеризуется теснотой связи. Чем выше теснота связи, тем ближе стохастическая зависимость приближается к функциональной, а доверительная вероятность – к единице. Наоборот, по мере уменьшения тесноты связи сопоставляемые переменные становятся все более независимыми.

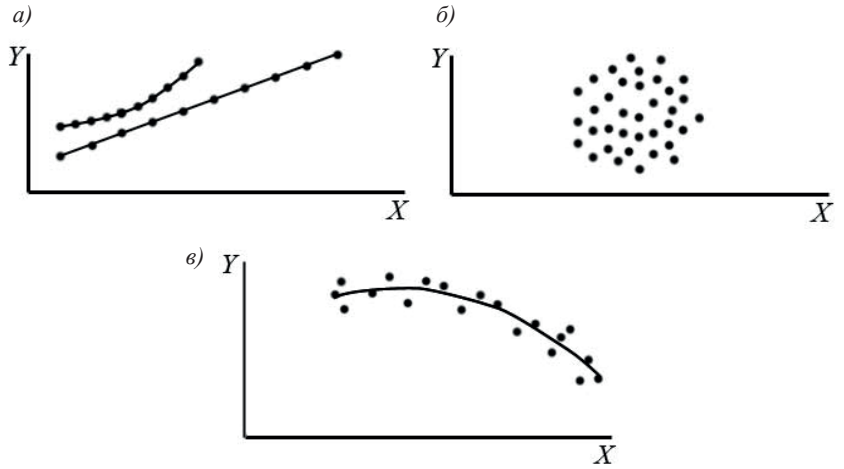

Рис. 6.1. Виды связей между переменными *X* и *Y*:
а) функциональная; *б)* случайная; *в*) стохастическая

Для описания стохастических связей обычно используется аппарат корреляционного и регрессионного анализа. При этом основной задачей корреляционного анализа является выявление связи между переменными и оценка ее тесноты, а основной задачей регрессионного анализа − установление формы и анализ зависимости между переменными.

Заметим, что связь между исследуемыми процессами в общем случае может носить как *линейный*, так и *нелинейный* характер. Линейные зависимости в свою очередь могут быть *прямо пропорциональными* и *обратно пропорциональными.* Что касается построения нелинейных зависимостей, то это более сложная задача. Способам ее решения будет посвящен раздел 8.

6.2. Коэффициент корреляции и его свойства

В настоящее время известно много различных показателей тесноты стохастических связей двух рядов. Обычно они разделяются на параметрические, применение которых предполагает знание теоретического (как правило, нормального) закона распределения, и непараметрические, не требующие выполнения данного условия.

К непараметрическим критериям связи относятся коэффициент знаков Фехнера, ранговые коэффициенты связи Спирмена и

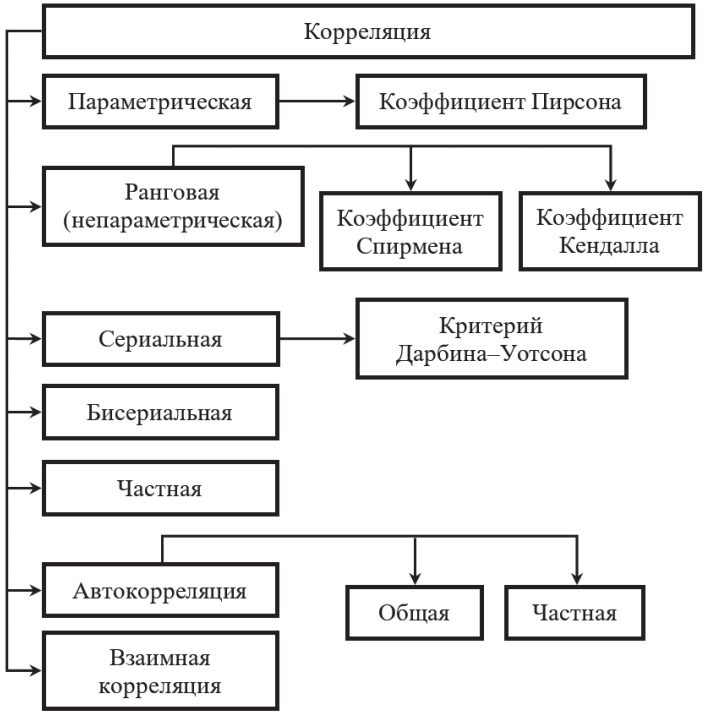

Рис. 6.2. Различные виды корреляции, используемые в статистических расчетах

Кендалла, коэффициенты сопряженности Пирсона и Чупрова и ряд других показателей.

К параметрическим критериям относятся парный коэффициент корреляции Пирсона, коэффициенты регрессии и др. *Параметрический характер коэффициента корреляции следует из того, что он является параметром двумерного нормального распределения.* Следовательно, прежде чем рассчитывать коэффициенты корреляции, надо убедиться в том, что система двух случайных величин имеет нормальное распределение. Из показателей связи в практических расчетах именно коэффициент корреляции Пирсона получил наибольшее распространение (рис. 6.2).

Выборочный коэффициент корреляции представляет собой безразмерную параметрическую характеристику линейной взаимосвязи двух случайных величин X и Y, т. е.

$$r = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}. \qquad (6.1)$$

или

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}. \qquad (6.1')$$

Истинная оценка коэффициента корреляции может быть получена при замене средних арифметических и выборочных дисперсий переменных X и Y на их математические ожидания и генеральные дисперсии. Отметим, что формула (6.1) справедлива для длинных статистических рядов. Вследствие того, что выборочный коэффициент корреляции имеет отрицательное смещение, для коротких рядов в знаменатель формулы (6.1) вводится множитель $(n - 1)$ вместо n.

Формула (6.1) удобна для физического анализа, однако в численных расчетах на ЭВМ обычно используется формула:

$$r = \frac{\sum xy - \left(n^{-1}\right)(\Sigma x)(\Sigma y)}{\sqrt{\left[\sum x^2 - \left(n^{-1}\right)(\Sigma x)^2\right]\left[\sum y^2 - \left(n^{-1}\right)(\Sigma y)^2\right]}}. \qquad (6.2)$$

Перечислим некоторые важные **свойства коэффициента корреляции.**

Свойство 1. Коэффициент корреляции не изменится, если:

а) прибавить (вычесть) к переменным X и Y какие-либо постоянные слагаемые;

б) умножить (разделить) переменные X и Y на произвольные положительные числа.

Свойство 2. Коэффициент корреляции изменяется в пределах $-1 \le r \le 1$.

Свойство 3. При линейной функциональной связи между переменными X и Y величина $r = \pm 1$. В этом случае облако точек на графике связи вырождается в прямую линию, наклоненную под некоторым углом к оси абсцисс.

Свойство 4. Если $r > 0$, то связь между X и Y прямая, т. е. обе переменные одновременно возрастают или убывают. Если $r < 0$, то

связь между X и Y обратная, т. е. с возрастанием одной величины другая убывает.

Свойство 5. Если переменные X и Y являются независимыми в статистическом смысле, то $r = 0$, вследствие чего проведение линии связи между переменными равновероятно в любом направлении (см. рис. 6.1-*б*).

Заметим также, что если одна из переменных является постоянной, то коэффициент корреляции не может быть определен, так как происходит деление на нуль. Облако точек на графике связи в этом случае превращается в прямую линию, параллельную одной из осей координат.

Анализ стохастической связи между переменными удобно осуществлять в так называемом *корреляционном поле.* Суть его заключается в том, что в декартовой системе координат по оси абсцисс откладывают значения одной переменной, а по оси ординат − другой переменной. Затем полученные точки соединяют друг с другом ломаной линией, которая называется *эмпирической линией связи.* По ее виду можно судить не только о наличии, но и о форме зависимости между рассматриваемыми переменными.

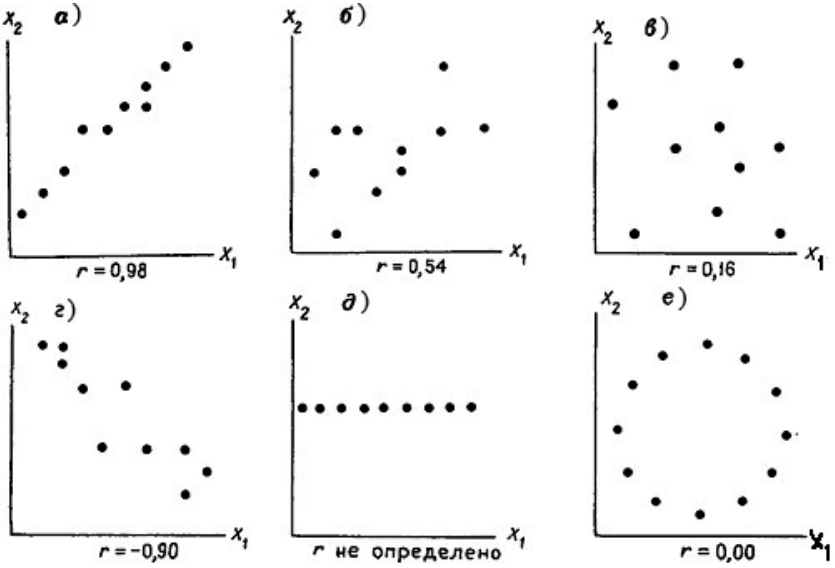

Рис. 6.3. Графики корреляционных полей
с различными коэффициентами корреляции между переменными X_1 и X_2

На рис. 6.3 приводится несколько корреляционных графиков. В том случае, когда между переменными X_1 и X_2 отмечается ярко выраженная прямо пропорциональная зависимость (рис. 6.3-*а*), коэффициент корреляции близок к единице. Менее отчетливая положительная корреляция ($r = 0{,}54$) наблюдается между переменными на рис. 6.3-*б*. Практически случайная связь на рис. 6.3-*в*, когда коэффициент корреляции близок к нулю, обусловлен тем, что в качестве переменных X_1 и X_2 использовались значения из таблицы случайных чисел. Очень сильная отрицательная корреляция ($r = -0{,}90$) между переменными изображена на рисунке 6.3-*г*. Определение коэффициента корреляции на рисунке 6.3-*д* невозможно, поскольку переменная X_1 постоянна и, следовательно, в соответствии с формулой (6.1) необходимо осуществить деление на нуль. На рис. 6.3-*е* наблюдения X_1 и X_2 расположены на окружности, поэтому между ними существует функциональная зависимость вида $X_2 = (a^2 - X^2_1)^{1/2}$, где a – радиус окружности. Поэтому, несмотря на наличие между переменными X_1 и X_2 нелинейной зависимости, коэффициент корреляции равен нулю, ибо он является характеристикой линейной связи.

6.3. Оценка достоверности и значимости коэффициента корреляции

Распределение выборочных коэффициентов корреляции очень сильно зависит от длины сравниваемых рядов и от самой величины r. При малых значениях r и достаточно больших объемах выборки распределение коэффициентов корреляции подчиняется нормальному закону. В этом случае для оценки случайных погрешностей применимы обычные параметрические методы проверки гипотез.

С увеличением r и уменьшением длины рядов n распределение коэффициентов корреляции приобретает все более несимметричный характер. Поэтому необходимы уже специальные методы оценки достоверности величин r.

Рассмотрим способы оценки достоверности выборочных коэффициентов корреляции при различных значениях n и r.

а) Оценка коэффициентов корреляции при $|r| < 0{,}3–0{,}4$ и $n > 30–40$.

Исходя из нормального закона распределения, для оценки среднеквадратических погрешностей коэффициентов корреляции используется следующая формула:

$$\sigma_r = \frac{1 - r^2}{n^{1/2}}. \tag{6.3}$$

Отсюда видно, что чем больше значения r и n, тем меньше ошибка коэффициента корреляции. После расчета σ_r находят отношение $|r|/\sigma_r$. Если $|r|/\sigma_r > 3$, то можно уверенно утверждать, что искомый коэффициент корреляции надежен и достоверно отражает связь между переменными. Для оценки генерального коэффициента корреляции строятся доверительные интервалы на основе t-статистики Стьюдента:

$$r - \boldsymbol{t}_{\text{кр}}\sigma_r < \boldsymbol{r} < r + \boldsymbol{t}_{\text{кр}}\sigma_r, \tag{6.4}$$

где $\boldsymbol{t}_{\text{кр}}$ – критерий Стьюдента при уровне значимости α и числе степеней свободы $\nu = n - 2$.

Оценка значимости коэффициента корреляции осуществляется на основе нулевой гипотезы, которая выбирается относительно проверки r на равенство нулю, т. е. $H_0 : |r| = 0$ при $H_1 : |r| \neq 0$. Коэффициент корреляции считается значимым, если он отличается от нуля неслучайным образом, т. е. его величина существенно выше (прямая связь) или ниже (обратная связь) нуля. Для проверки нулевой гипотезы используется критерий Стьюдента в виде:

$$t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}. \tag{6.5}$$

Затем осуществляется проверка неравенства $t > \boldsymbol{t}_{\text{кр}}(\alpha, \nu = n - 2)$.

Если неравенство $t > \mathbf{t}_{\text{кр}}(\alpha, \nu = n - 2)$ выполняется, то нулевая гипотеза отвергается и делается вывод, что коэффициент корреляции значим, т. е. отклоняется от нуля неслучайным образом. Если же оно не выполняется, то у нас есть основания полагать, что коэффициент корреляции не значим, т. е. отклоняется от нуля случайным образом.

Существенный недостаток формулы (6.5) состоит в том, что при оценке значимости коэффициента корреляции нужно постоянно пересчитывать значения t. Отметим, что этого можно избежать, если оценивать непосредственно критические значения коэффициента корреляции. Поставим в соответствие в формуле (6.5) критическому значению статистики Стьюдента критическое значение коэффициента корреляции, т. е.

$$\frac{\boldsymbol{t}_{\text{кр}}\left(1 - \boldsymbol{r}^2_{\text{кр}}\right)^{1/2}}{\left(n-2\right)^{1/2}} = \boldsymbol{r}_{\text{кр}}. \tag{6.6}$$

Возведем левую и правую части уравнения в квадрат и выполним несложные преобразования:

$$t^2_{кр} - t^2_{кр} r^2_{кр} = r^2_{кр}(n-2)$$

или

$$t^2_{кр} = r^2_{кр}(n - 2 + t^2_{кр}).$$

Отсюда

$$r_{кр} = \frac{t_{кр}}{\left(n - 2 + t^2_{кр}\right)^{1/2}}. \qquad (6.7)$$

Итак, подставляя в формулу (6.7) значения $t_{кр}$ и n, нетрудно вычислить критические значения коэффициента корреляции. Поскольку для длинных выборок (n > 30–40) при $\alpha = 0{,}05$ $t_{кр} \approx 2{,}0$, то формулу (6.7) можно упростить:

$$r_{кр} \approx \frac{2}{\sqrt{n+2}}. \qquad (6.8)$$

Преимущество формул (6.7) и (6.8) перед (6.5) состоит в том, что величину $r_{кр}$ не нужно пересчитывать как величину $t_{кр}$ для каждого выборочного значения r. Возведем теперь формулу (6.8) в квадрат и получим приближенную оценку критической величины коэффициента детерминации:

$$r^2_{кр} \approx \frac{4}{n+2}, \qquad (6.8')$$

которая может быть использована для приближенной оценки адекватности линейных регрессионных моделей.

б) Оценка коэффициентов корреляции при $|r| > 0{,}3-0{,}4$ и $n < 30–40$.

В этом случае, как уже отмечалось выше, распределение выборочных коэффициентов корреляции является резко асимметричным. Поэтому точность r обычно оценивается с помощью преобразования Фишера, основанного на использовании специальной переменной z, функционально связанной с r следующим выражением:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r} = \operatorname{arcth}(r), \qquad (6.9)$$

где *th* – гиперболический тангенс.

Распределение величины z (Приложение 5) почти не зависит от n и r, причем с возрастанием n оно быстро приближается к нормальному с математическим ожиданием:

$$M(z) = \frac{1}{2}\ln\frac{1+r}{1-r} + \frac{r}{2(n-1)}$$

и дисперсией $D(z) = \frac{1}{(n-3)}$.

Отсюда нетрудно видеть, что стандартная погрешность величины z зависит лишь от длины выборки и определяется как:

$$\sigma_z = \frac{1}{\sqrt{n-3}}. \tag{6.10}$$

Доверительные границы для z записываются следующим образом:

$$z - t_{кр}\sigma_z < z < z + t_{кр}\sigma_z. \tag{6.11}$$

Построив доверительные границы для z, нетрудно от них перейти к доверительным границам для r, используя для этого обратное преобразование $r = f(z)$. Для этого можно воспользоваться специальной таблицей (Приложение 5). Входим в нее с величиной z и получаем на выходе значение r. Возможен и аналитический вариант определения r. Исходя из формулы (6.9), можно получить:

$$r = \frac{\exp(2z-1)}{\exp(2z+1)}. \tag{6.12}$$

Тогда доверительный интервал для величины r примет вид:

$$\frac{\exp(2z_1-1)}{\exp(2z_1+1)} < r < \frac{\exp(2z_2-1)}{\exp(2z_2+1)}, \tag{6.13}$$

где $z_1 = z - t_{кр}\sigma_z$; $z_2 = z + t_{кр}\sigma_z$.

Отметим, что рассчитанные по формуле (6.13) доверительные границы могут быть несимметричными относительно величины r.

в) Оценка коэффициентов корреляции при $|r| < 0{,}3–0{,}4$, $n < 30–40$ и при $|r| > 0{,}3–0{,}4$, $n > 30–40$.

В этих случаях приближенно можно считать, что распределение выборочных коэффициентов корреляции не очень заметно отличается от нормального закона, поэтому для оценки точности величин r можно воспользоваться вариантом «*а*».

Пример 6.1. Найти с помощью преобразования Фишера интервальную оценку коэффициента корреляции, если $r = 0{,}74$, $n = 50$, $\alpha = 0{,}05$.

1) $z = \dfrac{0,5\ln(1+0,74)}{(1-0,74)} = 0,95.$

2) $\sigma_z = \dfrac{1}{(50-3)^{1/2}} = \dfrac{1}{6,86} = 0,146.$

3) $(0,95 - 2,01\times 0,146) < z < (0,95 + 2,01\times 0,146) \rightarrow 0,66 < z < 1,24.$

4) Вычислив по формуле (6.9) оценки *r*, окончательно получим $0,58 < \boldsymbol{r} < 0,84$. Это совпадает с оценками, полученными по приложению 5.

Поскольку рассмотренный пример совпадает с вариантом «в», то найдем интервальную оценку непосредственно по формуле (6.4). Имеем:

$$0,74 - 2,01\times 0,066 < \boldsymbol{r} < 0,74 + 2,01\times 0,066 \rightarrow 0,61 < \boldsymbol{r} < 0,87.$$

Сравнение интервальных оценок показывает, что оба доверительных интервала имеют одинаковую ширину, но интервал, вычисленный с помощью преобразования Фишера, является несимметричным, смещенным в сторону более высоких оценок *r*. Очевидно, симметричный доверительный интервал заслуживает большего доверия.

Коэффициенты корреляции, как мера линейной связи между процессами в силу простоты и доступности вычисления, получили самое широкое распространение в статистических расчетах. Однако следует помнить, что корреляция показывает только *силу* (тесноту) *связи* и ни в коей мере не может указывать на существование *зависимости* между переменными. Дело в том, что понятие «зависимость» подразумевает, что изменения одной переменной обусловлены изменением другой переменной. Другими словами, связь между ними носит причинно-следственный характер.

Очевидно, чтобы выяснить, какая из переменных влияет на другую переменную, необходим, прежде всего, содержательный физический анализ связи между ними. Иногда, сделать это весьма просто. Например, если в качестве переменных используются скорость ветра и высота волнения, то очевидно, что именно ветер влияет на волнение, а не наоборот. В других случаях это сделать принципиально невозможно. Так, всем известен философский спор: что первично – курица или яйцо. И, наконец, возможен вариант, когда причинно-следственная связь между переменными может существовать, но

физический анализ по каким-либо причинам затруднителен. В этом случае возникает необходимость в дополнительном использовании статистических методов.

В частности, существует тест Гранжера на причинность. Суть этого теста довольно проста. Если переменная X влияет на переменную Y, то изменения X должны предшествовать изменениям Y, но никак не наоборот. Очевидно, в этом случае должно выполняться следующее условие:

$$y_t = \alpha_0 + \sum_{j=1}^{m} \alpha_j x_{t-j} + \varepsilon_t, \qquad (6.14)$$

где m – интервал запаздывания Y по отношению к X, ε_t – случайная компонента.

Как следует из формулы (6.14), при $j = 0$ связь между X и Y синхронная, а при $j \geq 1X$ предшествует изменениям Y. Если изменения X значимо влияют на Y, то, соответственно, коэффициенты α_j будут значимо отличаться от нуля. Следовательно, мы можем записать нулевую гипотезу о том, что X влияет на Y в виде $H_0 : \alpha_1 = \ldots = \alpha_m = 0$. Как будет показано в разделе 7, это означает проверку регрессионной модели на адекватность по критерию Фишера. Если данная модель окажется адекватной, то можно уверенно сделать вывод о статистическом влиянии X на Y. Если переменная X не влияет на переменную Y, то аналогичным образом можно осуществить проверку возможного влияния Y на X.

Весьма важно, что корреляция является основой многих статистических методов и, в частности, многих методов многомерной статистики, основанных на анализе корреляционных матриц.

Если мы имеем значения какого-либо параметра (например, температуры воды), измеренного в M точках за промежуток времени N, причем $N > M$, то нетрудно составить матрицу исходных значений температуры размером $M \times N$, в которой столбцы представляют гидрологические станции, а строки – время измерения на них температуры, т. е.

$$T = \begin{vmatrix} T_{11} & T_{12} & \ldots & T_{1M} \\ T_{21} & T_{22} & \ldots & T_{2M} \\ \ldots & \ldots & \ldots & \ldots \\ T_{N1} & T_{N2} & \ldots & T_{NM} \end{vmatrix}.$$

В результате вычисления коэффициентов корреляции между рядами для отдельных точек данная матрица превращается в квадратную симметрическую матрицу следующего вида:

$$R = \begin{vmatrix} r_{11} & r_{12} & \dots & r_{1M} \\ r_{21} & r_{22} & \dots & r_{2M} \\ \dots & \dots & \dots & \dots \\ r_{M1} & r_{M2} & \dots & r_{MM} \end{vmatrix}.$$

Диагональные элементы этой матрицы всегда равны $r_{11} = r_{22} = \dots = r_{MM} = 1$, поскольку отражают корреляцию исходного ряда с самим собой. Индексы при r_{ij} указывают номера точек, между которыми рассчитываются коэффициенты корреляции.

Высокие положительные значения r связывают с синфазностью колебаний, высокие отрицательные значения r обычно интерпретируются как противофазность колебаний, наконец, $r \approx 0$ – отсутствие статистической связи между точками пространственного поля.

Пример 6.2. При решении многих гидрометеорологических задач большое значение имеет наличие длительных наблюдений за основными характеристиками среды. В этом плане уникальным представляется разрез «Кольский меридиан», первые наблюдения на котором были выполнены еще в 20-е годы прошлого столетия. Наиболее полные систематические наблюдения начинаются с 1951 г. К настоящему времени количество выполнений данного разреза уже превысило 1000 раз. Отметим, что Кольский разрез вытянут от Мурманска на север вдоль 33° в.д. и включает порядка 10 гидрологических станций.

Воспользуемся среднемесячными данными по температуре воды и солености на локальном разрезе между станциями 3−7, осредненными в слое от 0 до 50 м за период с 1951 по 1998 г. Рассчитаем коэффициенты корреляции для межгодовых изменений температуры воды и солености, которые приведены в корреляционной матрице (таблица 6.1). Коэффициенты корреляции для температуры воды составляют верхний треугольник матрицы, а солености – нижний треугольник.

Прежде всего, оценим значимость коэффициентов корреляции в таблице 6.1. Воспользуемся для этого формулой (6.7) $\boldsymbol{r}_{\text{кр}} = \dfrac{\boldsymbol{t}_{\text{кр}}}{\left(n - 2 + \boldsymbol{t}^2_{\text{кр}}\right)^{1/2}}$. Величина критерия Стьюдента при α = 0,05,

Таблица 6.1

Корреляционная матрица межгодовых изменений температуры воды (верхний треугольник) и солености (нижний треугольник) на разрезе Кольский меридиан, станции 3–7, слой 0–50 м за период с 1951 по 1998 г.

месяц	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
I	1	0,92	0,80	0,75	0,63	0,52	0,33	0,26	0,30	0,32	0,15	0,12
II	0,86	1	0,91	0,85	0,74	0,65	0,40	0,35	0,40	0,39	0,24	0,21
III	0,72	0,88	1	0,95	0,86	0,75	0,53	0,49	0,51	0,47	0,35	0,35
IV	0,71	0,84	0,95	1	0,93	0,78	0,60	0,54	0,60	0,51	0,39	0,37
V	0,69	0,80	0,88	0,92	1	0,88	0,71	0,64	0,67	0,60	0,51	0,44
VI	0,58	0,75	0,85	0,89	0,89	1	0,82	0,74	0,69	0,62	0,56	0,47
VII	0,53	0,68	0,77	0,79	0,81	0,92	1	0,85	0,72	0,55	0,53	0,51
VIII	0,48	0,65	0,74	0,77	0,75	0,79	0,85	1	0,82	0,59	0,52	0,50
IX	0,35	0,52	0,61	0,64	0,62	0,68	0,77	0,88	1	0,80	0,63	0.48
X	0,30	0,44	0,50	0,55	0,52	0,59	0,69	0,78	0,90	1	0,82	0,53
XI	0,17	0,36	0,41	0,48	0,41	0,54	0,57	0,66	0,74	0,84	1	0,82
XII	0,11	0,29	0,42	0,47	0,41	0,54	0,52	0,52	0,58	0,62	0,87	1

(Незначимые оценки коэффициентов корреляции отмечены чертой снизу.)

$\nu = 46$ равна $t_{кр} = 2{,}02$. Тогда $r_{кр} = \dfrac{2}{(50)^{1/2}} = 0{,}28$. Далее осуществляется проверка неравенства $|r| > r_{кр}$. Если данное неравенство выполняется, то коэффициент корреляции считается значимым. Незначимые оценки коэффициентов корреляции отмечены в таблице 6.1 чертой снизу.

Как видно из таблицы 6.1, отмечается высокая внутригодовая связность значений температуры и солености. Действительно, значимая корреляция наблюдается почти на протяжении всех двенадцати месяцев, причем отсутствует переход ее к отрицательным значениям. Столь высокая инерционность обусловлена, с одной стороны, адвекцией тепла течениями, и прежде всего Нордкапским течением, приносящим сравнительно теплые воды из Норвежского моря, а с другой – крупномасштабными метеорологическими процессами, имеющими значительную пространственно-временную сопряженность. Отметим, что высокая внутригодовая связность колебаний основных океанологических параметров отмечается сравнительно редко и является важной характеристикой рассматриваемого района моря.

Пример 6.3. При решении многих статистических задач (например, анализе пространственно-временной изменчивости, долгосрочном прогнозе различных характеристик и др.) важное значение имеет построение и последующий анализ карты изолиний равной корреляции, называемой полем изокоррелят, между рассматриваемыми переменными (y_i и x_{ij}), причем одна из них (x_j) задана в n точках пространства ($j = 1,n$). Вначале для каждой точки этого пространства рассчитывается парный коэффициент корреляции r_{yxj}, а затем строятся изолинии равной корреляции. Довольно просто это сделать в статистическом пакете «Серфер».

Как известно, важнейшей характеристикой атмосферной циркуляции в Северной Атлантике является Северо-Атлантическое колебание (САК), представляющее собой разность атмосферного давления между центрами Азорского максимума и Исландского минимума. САК характеризует интенсивность геострофического зонального переноса воздушных масс в умеренных широтах. Чем выше значения САК, тем интенсивнее зональный перенос. На рис. 6.4 приводится карта изокоррелят между средними годовыми значениями САК и индексом обшей циклоничности C, представляющим собой произведение интенсивности циклонов на их

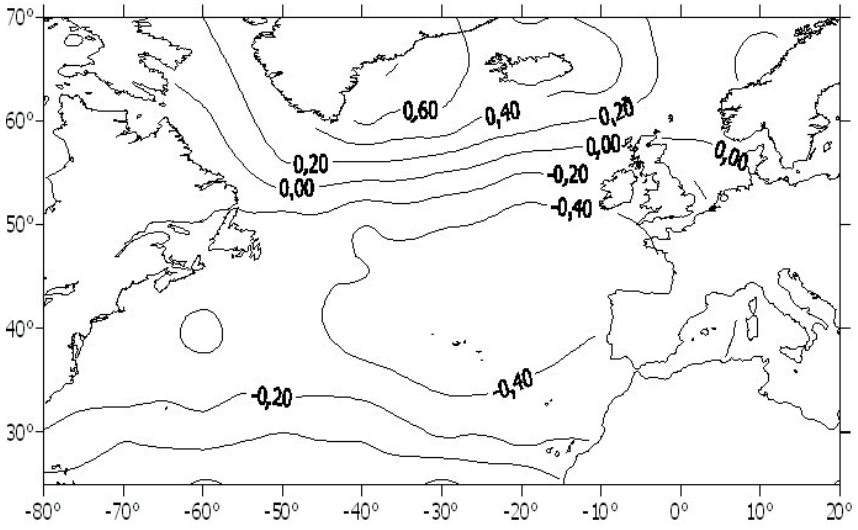

Рис. 6.4. Пространственное распределение коэффициентов корреляции между индексом общей циклоничности C и интенсивностью Северо-Атлантического колебания

повторяемость и таким образом являющимся интегральным показателем циклонических синоптических вихрей. Значения индексов *C* были заданы в 63 узлах географической сетки с шагом 5° по широте и 10° по долготе. Объем выборки составил 51 год (с 1946 по 1996 г.). При анализе примем для простоты критическое значение коэффициента корреляции $r_{кр} = 0{,}30$.

Из рис. 6.4 видно, что большая часть акватории Северной Атлантики занята значимой корреляцией САК с индексом *C*. При усилении (ослаблении) САК в области Исландской депрессии отмечается возрастание (уменьшение) интенсивности циклонической активности ($r > 0{,}40$) и ее ослабление (возрастание) в зоне влияния Азорского максимума давления. Действительно, вследствие существенно большей изменчивости атмосферного давления в зоне Исландской депрессии по сравнению с Азорским центром действия интенсивность САК зависит преимущественно от межгодовых колебаний давления в Исландской депрессии. Следует иметь в виду, что мы не можем только по данному полю изокоррелят установить причинно-следственные связи между рассматриваемыми переменными. Однако, учитывая характер изменчивости давления в зоне Исландской депрессии, можно предположить, что именно возрастание здесь циклонической активности должно усиливать общий зональный перенос воздушных масс над Северной Атлантикой.

6.4. Понятие ранговой корреляции

Для коротких статистических рядов, а также при изучении качественных или количественных признаков, распределенных по неизвестному закону, классические подходы корреляционного анализа оказываются неэффективными. В этом случае для изучения тесноты связи между переменными используются методы непараметрической статистики, среди которых наибольшее распространение получили ранговые коэффициенты связи. *Под ранговой корреляцией понимается линейная стохастическая связь между порядковыми переменными.*

Ранг – это порядковый номер значений признаков, расположенных в порядке возрастания или убывания их величины. Если значения признаков одинаковы, то их ранги равны среднему арифметическому от соответствующих номеров мест этих признаков. Такие ранги называются связными.

В качестве непараметрических коэффициентов связи наибольшее распространение получили ранговые коэффициенты Спирмена (ρ) и Кендалла (τ). Эти коэффициенты могут быть использованы для определения линейной тесноты связи как между количественными, так и между качественными признаками.

Если нет связных рангов, то ранговый коэффициент корреляции Спирмена рассчитывается по формуле:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}, \tag{6.15}$$

где d_i^2 – квадрат разности рангов, т. е. $d_i = R(x_i) - R(y_i)$; n – число пар рангов (число наблюдений).

Ранговый коэффициент корреляции Спирмэна, как и парный коэффициент корреляции изменяется в пределах $-1 \le \rho \le 1$. Если $\rho = 1$, то ранги переменных X и Y полностью совпадают, если $\rho = -1$, то ранги X и Y полностью противоположны. При $\rho = 0$ линейная связь между исходными переменными отсутствует.

Значимость ρ проверяется на основе t-критерия Стьюдента, расчетное значение которого определяется по формуле:

$$t = \rho\sqrt{\frac{n-2}{1-\rho^2}}. \tag{6.16}$$

Значение коэффициента Спирмена ρ считается значимым, если $|t| > \boldsymbol{t}_{\text{кр}}(\alpha, \nu = n - 2)$.

При наличии связных рангов определение коэффициента корреляции Спирмэна существенно усложняется. Однако в связи с тем, что учет связных рангов очень мало сказывается на точности оценки коэффициента ρ, то эти громоздкие формулы редко используются на практике.

Коэффициент ранговой корреляции Кендалла вычисляется как:

$$\tau = \frac{4Q}{n\left(n-1\right)} - 1, \tag{6.17}$$

где Q – сумма рангов по ряду y_i, т. е. $Q = Q_1 + Q_2 + \ldots + Q_{n-1}$. Ранги Q_i определяются следующим образом. Ранжированному ряду x_i ставится в соответствие ряд y_i. Другими словами, каждому значению члена ряда x_i будет соответствовать значение ряда y_i со своим порядковым номером (рангом). Далее берем первое значение y_1 и считаем,

сколько в ряде y_i правее находится рангов, больших y_1. После этого аналогичным образом считается, сколько рангов имеется правее y_2 и т.д. Последний ранг будет определяться для $n-1$ значения ряда y_i.

Ранговый коэффициент корреляции Кендалла также изменяется в пределах $-1 \le \tau \le 1$. Если $\tau = 1$, то ранги переменных X и Y полностью совпадают, если $\tau = -1$, то ранги X и Y полностью противоположны. При $\tau = 0$ линейная связь между исходными переменными отсутствует.

При проверке значимости τ исходят из того, что при отсутствии корреляционной связи между переменными имеет место приближенный нормальный закон распределения с математическим ожиданием равным нулю и стандартным отклонением:

$$\sigma = \sqrt{\frac{2(2n+5)}{9n(n-1)}}. \tag{6.18}$$

В этом случае t-критерий приобретает вид:

$$t = \frac{|\tau|}{\sigma} = |\tau| \sqrt{\frac{9n(n-1)}{2(2n+5)}}. \tag{6.19}$$

Отметим, что хотя вычисление коэффициента Кендалла более трудоемко по сравнению с вычислением коэффициента ρ, однако он имеет определенные преимущества перед ним. Это связано с большей исследованностью его статистических свойств и возможности использования в частной корреляции рангов.

Между коэффициентами корреляции Спирмена и Кендалла при достаточно большом объеме исходной выборки существует почти функциональная связь:

$$\tau \approx \left(\frac{2}{3}\right)\rho, \tag{6.20}$$

т. е. величина τ всегда меньше ρ.

Дополнительно отметим, что для определения тесноты связи двух качественных признаков, каждый из которых состоит только из двух групп (да и нет), могут применяться коэффициенты ассоциации и контингенции.

Пример 6.4. Требуется сравнить степень взаимосвязи межгодовых значений максимальной ледовитости для двух районов Балтийского моря за десять лет. Оценки ледовитости, выраженные в процентах от общей площади района, приведены в таблице 6.2.

Таблица 6.2

Порядок расчета коэффициентов ранговой корреляции Спирмена и Кендалла

1 район (x_i)	86	75	95	70	90	84	60	50	62	57
2 район (y_i)	83	55	92	60	93	80	72	70	45	62
Ранжирование x_i	95	90	86	84	75	70	62	60	57	50
Ранги x_i	1	2	3	4	5	6	7	8	9	10
Ряд y_i	92	93	83	80	55	60	45	72	62	70
Ранги y_i	2	1	3	4	9	8	10	5	7	6
Разность рангов d_i	–1	1	0	0	–4	–2	–3	3	2	4
Число рангов Q_i	8	8	7	6	1	1	0	2	0	–

Сначала выполним ранжирование значений ледовитости 1-го района в порядке убывания и присвоим им ранг от 1 до 10 (см. таблицу 6.2). Теперь определим ранг значений ледовитости 2-го района. Максимальному значению ледовитости 1 района (95 %) соответствует второе по величине значение ледовитости 2 района (92 %). Следовательно, ранг $y_1 = 2$. Рангу $x_2 = 2$ соответствует ранг $y_2 = 1$. Аналогичным образом мы можем определить ранги для всех значений y_i. После этого находим разность рангов как $d_i = x_i - y_i$. Вычислим сумму квадратов разностей рангов:

$$\sum d_i^2 = 1+1+16+4+9+9+4+16 = 60.$$

Теперь нетрудно рассчитать коэффициент корреляции Спирмена:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)} = \frac{1-\left(6\times 60\right)}{\left(10^3-10\right)} = 0,64.$$

После этого оценим значимость коэффициента корреляции. Выдвигаем нулевую гипотезу $H_0 : |\rho| = 0$ при альтернативе $H_1 : |\rho| \neq 0$. Проверку нулевой гипотезы осуществляем с помощью статистики Стьюдента:

$$t = |\rho|\left[\frac{(n-2)}{\left(1-\rho^2\right)}\right]^{1/2} = 0,64\left(\frac{8}{0,59}\right)^{1/2} = 2,36.$$

Критическое значение $\boldsymbol{t}_{\text{кр}}(\alpha = 0,05,\ \nu = 8) = 2,31$. Отсюда видно, что выборочное значение коэффициента корреляции Спирмена является *значимым*.

Теперь оценим степень линейной взаимосвязи между ледовитостью двух районов с помощью рангового коэффициента корреляции Кендалла. С этой целью воспользуемся уже полученными оценками рангов для x_i и y_i, приведенными в таблица 6.2. В соответствии с формулой (6.13) требуется проанализировать ранги для y_i. Справа от ранга $y_1 = 2$ имеется $Q_1 = 8$ рангов (3, 4, 9, 8, 10, 5, 7, 6), больших $y_1 = 2$. Справа от $y_2 = 1$ имеется $Q_2 = 8$ рангов, больших $y_2 = 1$. Аналогичным образом получим все остальные оценки Q_i. Для десятого члена ряда ранги отсутствуют. Вычислим сумму:

$$Q = Q_1 + \ldots + Q_9 = 8 + 8 + 7 + 6 + 1 + 1 + 2 = 33.$$

Подставляя величину Q в формулу (6.17), находим $\tau = \left(\frac{4 \times 33}{90}\right) - 1 = 0,47$. Для оценки его значимости рассчитываем статистику Стьюдента $t = 0,47\left[\left(\frac{90 \times 9}{2 \times 25}\right)\right]^{1/2} = 1,88$. Сравнивая эту оценку с критической величиной $\boldsymbol{t}_{\text{кр}} = 2,31$ видим, что вычисленный коэффициент ранговой корреляции Кендалла оказывается *незначимым*.

Итак, использование коэффициентов корреляции Спирмена и Кендалла для одной и той же выборки дает противоположные результаты. Возникает вопрос: какой вывод можно сделать в данной противоречивой ситуации? Безусловно, учитывая малую длину выборки желательно продление рядов максимальной ледовитости. Только в этом случае можно будет получить более объективные оценки взаимосвязи колебаний ледовитости этих районов. Если же вывод необходимо сделать только на основании полученных результатов, то тогда, на наш взгляд, целесообразно ориентироваться на более жесткие оценки. В данном случае – это коэффициент корреляции Кендалла. Поэтому представляется разумным полагать отсутствие значимой связи между межгодовыми значениями максимальной ледовитости для двух районов Балтийского моря за данный промежуток времени.

Пример 6.5. Оценим взаимосвязь вылова ставриды в юго-восточной части Тихого океана (ЮВТО) с крупномасштабными метеорологическими параметрами за период работы в этом весьма важном рыбопромысловым районе советских судов в течение 1979–1990 гг. В качестве метеорологических параметров использовались: индексы Антарктического Колебания Тихоокеанского (ААО) и Южного Колебания (SOI), параметры Южнотихоокеанского антициклона (ЮТА) (давление P, смещение по широте φ и долготе λ). Оценки

Таблица 6.3

Оценка линейной связи между выловом ставриды и метеорологическими параметрами при различных сдвигах (годы) относительно вылова рыбы за период с 1979 по 1990 г.

Параметр	Коэффициент Спирмена			Коэффициент корреляции Пирсона		
	0	1 год	2 года	0	1 год	2 года
AAO	0,18	–0,38	0,16	0,27	–0,48	0,12
SOI	**0,59**	–0,15	**–0,71**	0,37	–0,29	**–0,78**
P	**0,64**	0,30	–0,52	0,57	0,27	–0,53
λ	**–0,70**	–0,24	**0,61**	**–0,65**	–0,24	0,53
φ	–0,42	–0,10	0,26	–0,20	0,05	0,32

коэффициента Спирмена приведены в таблице 6.3. Одновременно для сравнения в ней также даны и коэффициенты корреляции Пирсона. При этом кроме нулевого сдвига, соответствующего синхронной связи, указанные коэффициенты рассчитывались для сдвигов 1 и 2 года, имеющих прогностический смысл для вылова рыбы. Значимые по критерию Стьюдента коэффициенты выделены полужирным шрифтом.

Как видно из таблицы 6.3, наиболее высокая теснота синхронной связи ($\rho = –0,70$) отмечается между выловом ставриды и смещением ЮТА по долготе. При смещении на восток ЮТА происходит увеличение вылова рыбы. К сожалению, комментировать физический смысл между выловом рыбы и смещением ЮТА по долготе весьма сложно, учитывая короткую длину рядов. Но со статистической точки зрения данная связь, несомненно, является значимой и существенной. Кроме того, значимая синхронная связь отмечается для вылова рыбы с межгодовой изменчивостью давления в центре ЮТА ($\rho = 0,64$) и с индексом Южного Колебания ($\rho = 0,59$). Обращает на себя внимание высокая корреляция вылова рыбы с метеопараметрами при прогностическом сдвиге $\tau = 2$ года. Весьма высокая отрицательная корреляция наблюдается с индексом Южного Колебания ($\rho = –0,71$), а также с другими параметрами. Это означает, что с заблаговременностью в два года можно построить прогностическую модель вылова ставриды.

Сравнение коэффициентов Спирмена с коэффициентами корреляции Пирсона свидетельствует о том, что расхождения между ними носят преимущественно случайный характер и, как правило, не превышают по абсолютной величине 0,1. Однако, примерно в одной трети случаев расхождения между ними все же превышают 0,1, т. е. являются уже существенными.

6.5. Понятие бисериальной корреляции

Для оценки связи между качественной альтернативной (да, нет) и количественной переменными используется *бисериальный коэффициент корреляции.* При этом качественная альтернативная переменная получила название *дихотомической*. В расчетах она обычно принимается в виде 1 и 0. В качестве показателя линейной связи между дихотомической и непрерывной количественной переменными используется бисериальный коэффициент корреляции:

$$r_{cx} = \frac{(\bar{x}_1 - \bar{x}_0)\, p_c q_c}{s_x z_p}, \qquad (6.21)$$

где $\bar{x}_1$ и $\bar{x}_0$ – выборочные средние количественной переменной, соответствующие наличию (1) и отсутствию (0) явления C; p_c и q_c – относительные частоты наличия и отсутствия явления С при рассматриваемых условиях; s_x – выборочное среднее квадратическое отклонение переменной X для всей выборки; z_p – величина стандартного нормального z-распределения. Отметим, что случайная величина X, используемая при расчете бисериального коэффициента корреляции, должна быть распределена нормально.

Вычисление r_{cx} по формуле (6.21) осуществляется следующим образом. Вначале по всему исходному ряду рассчитывается величина s_x. Затем этот ряд делится на две совокупности: в одну включаются все значения X, когда имело место явление C, а в другую – когда оно отсутствовало. Далее вычисляются средние обеих совокупностей: $\bar{x}_1$ и $\bar{x}_0$, а также частоты $p_c = n_1/n$, $q_c = 1 - p_c$, где n_1 – число наблюдений при наличии явления C, n – общая длина исходного ряда. По найденной частоте p_c определяется значение величины z_p. Для этого вначале под кривой нормального распределения ищется точка, разделяющая площадь, ограниченная кривой, на части, пропорциональные p и q. Такая точка находится по значению p из таблицы функции Лапласа (Приложение 1) как $x_p = \Phi^{-1}(|p - 0{,}5|)$. Затем по величине x_p из таблицы плотности нормального распределения находится значение $z_p = f(x_p)$. Подставляя найденное значение z_p в формулу (6.21), вычисляем окончательно бисериальный коэффициент корреляции. Отметим, что бисериальный коэффициент не получил широкого распространения в практических расчетах, ибо чаще всего переменные одновременно являются количественными или качественными.

6.6. Понятие ложной корреляции

Если две переменные X_1 и X_2 не содержат в себе какой-либо информации о третьей переменной, то корреляция между X_1 и X_2 является истинной. В том случае, если переменные X_1 и X_2 связаны каким-либо образом с третьей переменной X_3, то возникает ложная (автоматическая) корреляция.

В этом нетрудно убедиться, если обратиться к следующему примеру. Пусть мы имеем переменные X_1 и X_2, корреляция между которыми отсутствует, т. е. $r_{X_1,X_2} = 0$. Сформируем два ряда отношений $Z_1 = X_1/X_3$ и $Z_2 = X_2/X_3$, где X_3 – некоторая третья переменная. В этом случае между рядами Z_1 и Z_2 возникает корреляция, величина которой зависит от изменчивости исходных выборок.

Как было установлено в результате численных расчетов, при малых коэффициентах вариации всех трех переменных и их приблизительном равенстве ($C_1 \approx C_2 \approx C_3$) коэффициент корреляции между Z_1 и Z_2 будет составлять $r_{Z_1,Z_2} = 0,5$. Если изменчивость третьего ряда в 2 раза больше первых двух, т. е. $C_3 \approx 2C_1 \approx 2C_2$, то $r_{Z_1,Z_2} = 0,8$. Если $C_3 \approx 3C_1 \approx 3C_2$, то уже $r_{Z_1,Z_2} = 0,9$.

Таким образом, из некоррелированных рядов мы получили почти функциональную связь. Естественно, что такой результат стал возможным потому, что на величину ложной корреляции существенное влияние оказывает изменчивость статистических рядов.

Наглядный пример ложной корреляции – корреляция годовых или суточных реализаций гидрометеорологических характеристик. Годовой ход, как известно, обусловлен солнечной радиацией, а суточный ход – вращением Земли вокруг собственной оси. Поэтому без исключения годового и суточного хода гидрометеорологических характеристик, корреляция между ними заведомо завышается.

Для исключения годового и суточного хода используют обычно ряд приемов, простейший из которых заключается в вычислении аномалий данной величины. Корреляция между аномалиями гидрометеорологических характеристик в значительной степени уменьшает ложную корреляцию. Поэтому в общем случае, когда третья переменная неизвестна, эффект ложной корреляции приближенно может быть определен как $r_{\text{ложн}} = |r_{\text{набл}}| - |r_{\text{ист}}|$, где $r_{\text{ист}}$ – корреляция между аномалиями рассматриваемых рядов.

Разумеется, приведенным выше примером не исчерпываются возможности появления эффекта ложной корреляции при изучении гидрометеорологических процессов или явлений. Так, она возникает при использовании одного и того же математического преобразования одновременно к обеим переменным. Например, применение операторов фильтрации, особенно полосовых фильтров, неминуемо приводит к появлению эффекта ложной корреляции.

Поэтому, прежде всего, нужно начинать с содержательной постановки задачи, после этого выполняются численные расчеты, а затем должна следовать обязательная физическая интерпретация полученных результатов. Если физическая связь рассматриваемых процессов не поддается расшифровке, то вряд ли стоит переоценивать значение даже высоких коэффициентов корреляции. Возможно, это является эффектом ложной корреляции.

Пример 6.6. Для оценки эффекта ложной корреляции воспользуемся ежемесячными картами температуры поверхности океана, составляемыми Гидрометцентром с 1977 г. Для двух пятиградусных квадратов с центрами (первый $\varphi = 60°$ с.ш., $\lambda = 10°$ з.д.; второй $\varphi = 65°$ с.ш., $\lambda = 20°$ з.д.) составлены выборки среднемесячных значений температуры воды за десятилетний (1977–1986 гг.) период. Таким образом, длина каждой выборки равна $n = 120$.

Коэффициент корреляции между статистическими рядами составил $r = 0{,}94$. Затем для каждого календарного месяца были рассчитаны аномалии температуры воды, как $\Delta x = x_i - \overline{x}$, где $\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i$. В результате получены две новые выборки, корреляция между которыми оказалась равной $r_{\Delta T} = 0{,}15$. Следовательно, ложная корреляция, обусловленная эффектом годового хода солнечной радиации, составляет $r_{\text{лож}} = 0{,}79$.

Пример 6.7. Известно, что в течение XX столетия уровень Мирового океана почти монотонно повышался. Средняя скорость его роста составляла примерно 1,5–1,8 мм/год. Главной причиной повышения уровня послужило глобальное потепление климата. В течение прошлого столетия глобальная температура воздуха повысилась на 0,6 °С. Естественно, это вызвало интенсивное таяние горных ледников, уменьшение массы шельфовых ледников в Антарктиде, термическое расширение объема вод океана и др., что в результате и привело к росту уровня Мирового океана.

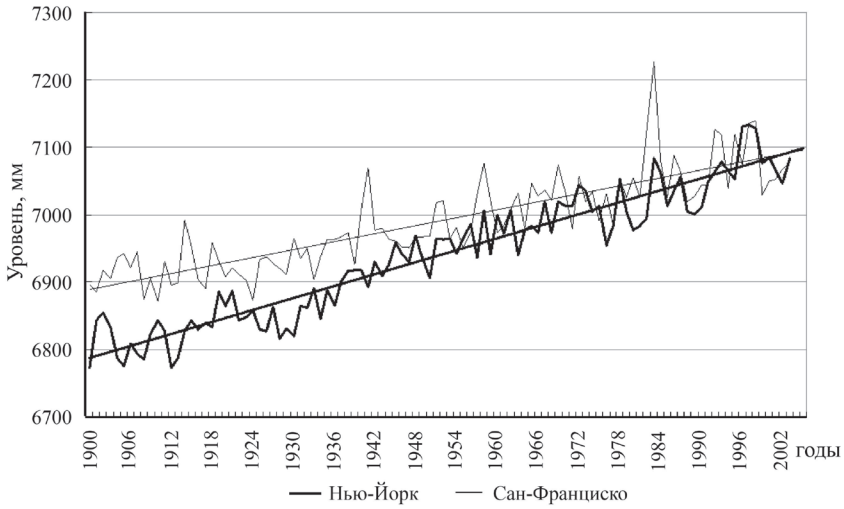

Рис. 6.5. Межгодовой ход морского уровня
на станциях Сан-Франциско и Нью-Йорк с 1901 г.

На рис. 6.5 представлен межгодовой ход уровня в XX столетии для двух станций, расположенных на противоположных берегах американского континента: Нью-Йорк (побережье Атлантики) и Сан-Франциско (побережье Тихого океана). Нетрудно видеть, что в межгодовых изменениях уровня на обеих станциях отмечается ярко выраженная тенденция к его повышению. Однако если средняя скорость роста уровня в Сан-Франциско составляет около 2,0 мм/год, то в Нью-Йорке он достигает 3,0 мм/год, что практически в 2 раза превышает рост глобального уровня океана. Причиной этого являются вертикальные движения земной коры. В районе Нью-Йорка происходит ее опускание со скоростью около 1,5 мм/год.

Корреляция между указанными рядами морского уровня равна $r = 0{,}85$. Естественно полагать, что наличие в каждом из рассматриваемых рядов трендовой компоненты (см. разд. 10.2) приводит к появлению эффекта ложной корреляции. Поэтому рассчитаем уравнение линейного тренда (10.11) и вычтем тренд из рядов уровня (рис. 6.6). Как видно из рис. 6.6, межгодовой ход уровня на этих станциях существенно изменился. В результате имеем $r = 0{,}24$. Хотя корреляция на уровне $\alpha = 0{,}05$ является значимой, но она

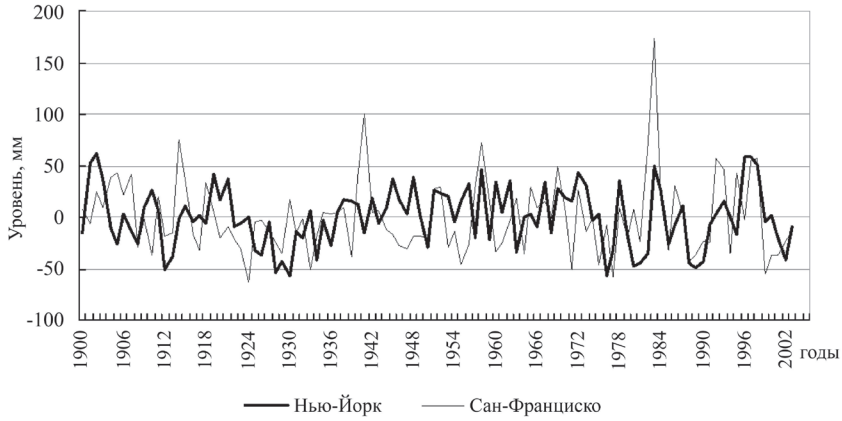

Рис. 6.6. Межгодовой ход морского уровня
на станциях Сан-Франциско и Нью-Йорк после исключения линейного тренда

существенно ниже по сравнению с корреляцией между исходными рядами. Итак, получаем, что ложная корреляция оказывается равной $r_{лож} = 0{,}85 - 0{,}24 = 0{,}61$, т. е. является весьма высокой.

Глава 7.
Линейный регрессионный анализ

7.1. Понятие о методе наименьших квадратов

Метод наименьших квадратов (МНК), без преувеличения, является классическим методом анализа данных и лежит в основе многих других методов статистического анализа. Метод наименьших квадратов впервые был сформулирован в 1805 г. Лежандром, поэтому он иногда называется принципом Лежандра. Теоретические основы метода изложены немецким математиком Карлом Фридрихом Гауссом в 1809 г., который затем неоднократно возвращался к нему в течение всей своей жизни.

Пусть мы имеем некоторую совокупность наблюдений $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$), причем между ними существует некоторое приближенное соотношение $y = f(x; a_1, a_2, \ldots, a_m)$, где $a_1, a_2, \ldots, a_m$ – неизвестные параметры данной зависимости. В этом случае для отыскания

неизвестных коэффициентов может быть использован метод наименьших квадратов, суть которого заключается в том, чтобы *сумма квадратов отклонений точек от линии связи должна быть наименьшей,* т. е.

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - y_{(x)i} \right)^2 = \sum_{i=1}^{n} \left[y_i - f\left(x; a_1, a_2, ..., a_m\right) \right]^2 = \min, \qquad (7.1)$$

где ε_i – остатки (ошибки), представляющие собой разность между фактическими (y_i) и вычисленными ($y_{(x)i}$) значениями случайной величины Y, n – длина выборки, причем $n > m$. Итак, в соответствии с формулой (7.1) требуется минимизировать сумму квадратов значений ε_i. Как видно из рис. 7.1, остатки ε_i представляют расстояние по оси ординат между фактическими y_i и вычисленными $y_{(x)i}$ значениями функции отклика.

Так как в уравнении (7.1) x_i и y_i известны (результаты наблюдений), то при заданном общем виде сглаживающей функции величину $\sum_{i=1}^{n} \varepsilon_i^2 = S$ можно рассматривать как функцию от неизвестных коэффициентов a_1, a_2, ... , a_m. Для их нахождения в соответствии с общим правилом определения экстремума дифференцируемой функции необходимо найти частные производные от нее по этим коэффициентам и приравнять их к нулю:

$$\frac{\partial S}{\partial a_1} = -2 \sum_{i=1}^{n} \left[y_i - f\left(x; a_1, a_2, ..., a_m\right)_i \right] \left(\frac{\partial f}{\partial a_1} \right) = 0,$$

$$\frac{\partial S}{\partial a_2} = -2 \sum_{i=1}^{n} \left[y_i - f\left(x; a_1, a_2, ..., a_m\right)_i \right] \left(\frac{\partial f}{\partial a_2} \right) = 0,$$

.....,

$$\frac{\partial S}{\partial a_m} = -2 \sum_{i=1}^{n} \left[y_i - f\left(x; a_1, a_2, ..., a_m\right)_i \right] \left(\frac{\partial f}{\partial a_m} \right) = 0.$$

Данная система уравнений называется системой нормальных уравнений. Число уравнений в системе всегда равно числу неизвестных, поэтому задача вычисления коэффициентов a_1, … , a_m является определенной и имеет единственное решение, которое может быть найдено, например, с помощью методов Крамера, Гаусса или обратных матриц.

Пример 7.1. Предположим, две выборки экспериментальных данных x_i и y_i образуют в координатной плоскости $x0y$ некоторую криволинейную статистическую связь (рис. 7.1). Наша задача состоит в наилучшем подборе вида аппроксимирующей функции и определении ее коэффициентов. В соответствии с методом наименьших квадратов необходимо минимизировать отклонения ε_i, т. е. подобрать такую кривую, которая дает наилучшее приближение к исходным точкам в смысле (7.1).

Очевидно, зависимость на рис. 7.1 может быть представлена в виде полинома второй степени $y = a_0 + a_1x + a_2x^2$. Тогда имеем систему из 3-х линейных уравнений:

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = -2\sum\limits_{i=1}^{n}\left[y_i - a_0 - a_1x_i - a_2x_i^2 \right] = 0, \\ \dfrac{\partial S}{\partial a_1} = -2\sum\limits_{i=1}^{n}\left[y_i - a_0 - a_1x_i - a_2x_i^2 \right] x_i = 0, \\ \dfrac{\partial S}{\partial a_2} = -2\sum\limits_{i=1}^{n}\left[y_i - a_0 - a_1x_i - a_2x_i^2 \right] x_i^2 = 0. \end{cases}$$

В результате несложных преобразований этой системы получим:

$$\begin{cases} a_0 n + a_1\sum\limits_{i=1}^{n} x_i + a_2\sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} y_i, \\ a_0\sum\limits_{i=1}^{n} x_i + a_1\sum\limits_{i=1}^{n} x_i^2 + a_2\sum\limits_{i=1}^{n} x_i^3 = \sum\limits_{i=1}^{n} x_i y_i, \\ a_0\sum\limits_{i=1}^{n} x_i^2 + a_1\sum\limits_{i=1}^{n} x_i^3 + a_2\sum\limits_{i=1}^{n} x_i^4 = \sum\limits_{i=1}^{n} x_i^2 y_i. \end{cases} \qquad (7.2)$$

Решение данной системы линейных нормальных уравнений с использованием современных ЭВМ не представляет каких-либо затруднений. Однако следует иметь в виду, что в систему (7.2) входят высокие степени переменной x_i, причем самая высокая степень равна удвоенной степени полинома. Это обстоятельство является главным источником вычислительных ошибок.

К очевидным достоинствам МНК можно отнести то, что при нормальном распределении исходных данных МНК дает оценки параметров, совпадающие с методом максимального правдоподобия, признающимся в статистике наиболее точным. Однако в общем

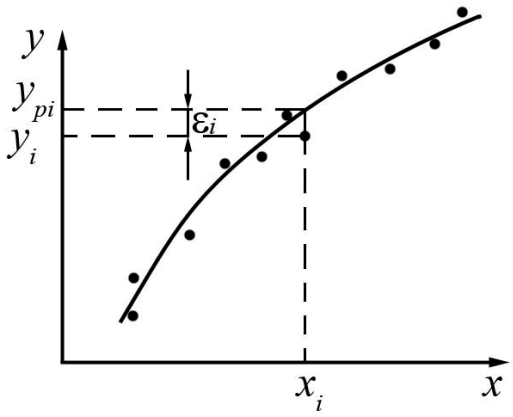

Рис. 7.1. Условие минимизации суммы квадратов отклонений переменной *Y* в стохастической зависимости между переменными *X* и *Y*

случае требование нормальности не входит в условия теоремы Гаусса–Маркова, объявляющей оценки МНК оптимальными среди всех линейных оценок. Другими словами, оценки МНК являются наилучшими линейными оценками, т. е. состоятельными, несмещенными оценками, которые обладают минимальными дисперсиями среди множества всех линейных несмещенных оценок. Но, как было показано Фишером, это связано не столько с «хорошими» свойствами МНК, сколько с «плохими» свойствами линейных оценок, проявляющимися почти всюду, за исключением очень малой окрестности нормального распределения остатков.

Кроме того, другими важными достоинствами МНК являются те, что он весьма прост, хорошо теоретически разработан, легко алгоритмизируется и служит основой многих других методов статистического анализа.

Одновременно с этим необходимо отметить и недостатки МНК:

1) желательность нормального распределения исходных данных;

2) линейность по параметрам $a_1, \ldots, a_m$;

3) чувствительность к выбросам.

Поскольку первые два условия достаточно очевидны, то рассмотрим третье условие. Как было указано в гл. 5, к выбросам относятся резко выделяющиеся наблюдения, которые существенно отклоняются от распределения остальных выборочных данных. Если в выборке имеются выбросы, то они несут с собой опасность искажения в интерпретации результатов при использовании

классических статистических процедур. На рис. 7.2 представлен тестовый пример регрессионного анализа по шести точкам, значения которых приведены далее в таблице 7.4. Нетрудно видеть, что крайняя правая точка резко отличается от остальной совокупности. Поэтому два уравнения прямых, построенные с помощью МНК, одно из которых по всем исходным точкам (рис. 7.2-*а*), а другое – без учета выпадающей точки (рис. 7.2-*б*), резко различаются друг от друга. Такое различие обусловлено тем, что в соответствии с условием (7.1) осуществляется минимизация квадратов расстояний от линии связи до каждой точки, т. е. остатков ε_i, которое именно до крайней правой точки является максимальным.

Таким образом, даже одно единственное, резко выделяющееся наблюдение может полностью изменить наклон регрессионной

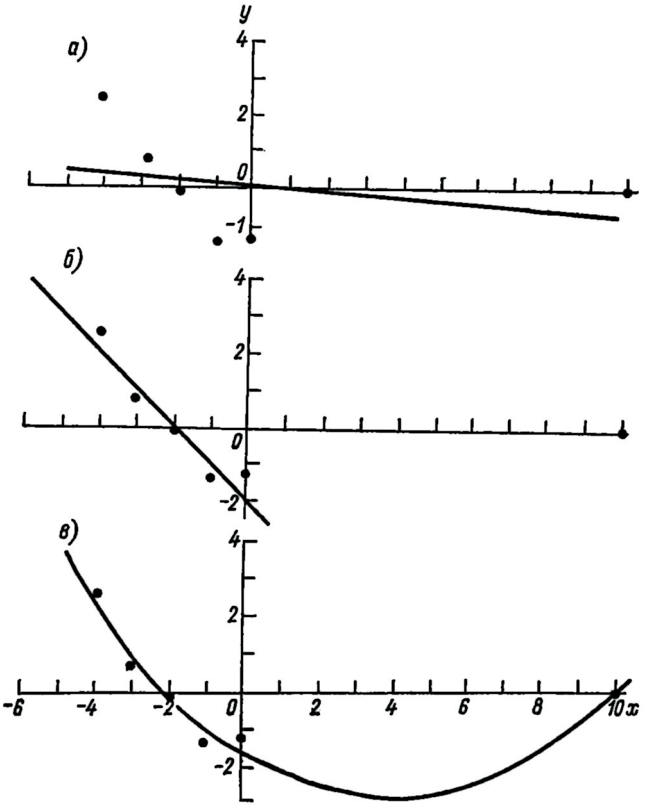

Рис. 7.2. Тестовый пример робастной регрессии

линии и, следовательно, привести к значительному искажению оценок регрессионных параметров. Резко выделяющиеся наблюдения нарушают независимость дисперсии остатков от его математического ожидания и тем самым противоречат методу наименьших квадратов. Действительно, резко выделяющееся наблюдение приводит к несимметричности распределения остатков, вследствие чего условие независимости дисперсии от математического ожидания сразу нарушается. Обработка таких данных методом обычного регрессионного анализа может привести к ошибкам настолько большим, что полученная модель не будет иметь смысла. Поэтому регрессионному анализу, построенному на МНК, должен предшествовать тщательный анализ на содержание в исходной выборке аномальных наблюдений.

7.2. Основы метода линейной регрессии двух переменных

Термин «регрессия» (*regression – отступление, движение назад*) впервые был введен в научную литературу Френсисом Гальтоном в 1886 г., причем непосредственного отношения к статистике данный термин не имел. Гальтон изучал зависимость между ростом родителей и их детей. Он обнаружил, что рост детей у высоких (низких) родителей обычно выше (ниже) среднего, но не совпадает с ростом родителей. Линия, показывающая, в какой мере рост детей отклоняется в среднем от роста родителей, была интерпретирована Гальтоном как «регрессия (отступление) к посредственности», т. е. к среднему. В дальнейшем под регрессией стали понимать стохастические связи между переменными.

В общем случае любую регрессионную зависимость можно представить следующим образом:

$$y = f(b_j, x^i_j),\ i = (1,k),\ j = (1,m),$$

где y – зависимая переменная, b_j – коэффициент регрессии, x_j – независимая переменная, k – степень регрессии, m – число независимых переменных. Если данную зависимость представить в координатах $m0k$ (рис. 7.3), то получим совокупность возможных видов уравнений регрессии. Из рис. 7.3 следует:

1) при $m = 1, k = 1$ имеем парную линейную регрессию (точка 1);

2) при $m \geq 2, k = 1$ имеем множественную линейную регрессию (прямая 2);

3) при $m = 1, k \geq 2$ имеем одномерную нелинейную регрессию (прямая 3);

4) при $m \geq 2$, $k \geq 2$ имеем нелинейную множественную регрессию (система прямых 4).

Итак, рис. 7.3. позволяет наглядно представить всю совокупность методов регрессионного анализа.

Запишем теперь два линейных уравнения в следующем виде:

$$y_i = c_0 + c_1 x_i, \tag{7.3}$$

$$y_i = a_0 + a_1 x_i + \varepsilon_i, \tag{7.4}$$

где ε_i – остатки, не описываемые уравнением (7.3). Появление в формуле (7.4) остатков связано со следующими объективными предпосылками. Во-первых, изменчивость переменной y_i носит более сложный характер и не может быть описана всего лишь одним фактором x_i, входящим в уравнение (7.4). Во-вторых, переменные y_i и x_i, как правило, измерены или рассчитаны с некоторыми ошибками, имеющими случайный характер, ибо систематические ошибки не входят в остатки ε_i.

Итак, первое уравнение представляет *математическое уравнение прямой линии,* в то время как второе – это уже *статистическое уравнение линейной регрессии двух переменных.* Принципиальное различие между ними состоит в том, что если первое уравнение является чисто теоретическим и не требует никаких предположений, то во втором на остатки ε_i уже накладывается целый ряд допущений. К ним относятся:

1) ошибки (остатки) модели регрессии должны иметь нулевое среднее значение $(\overline{\varepsilon} = 0)$;

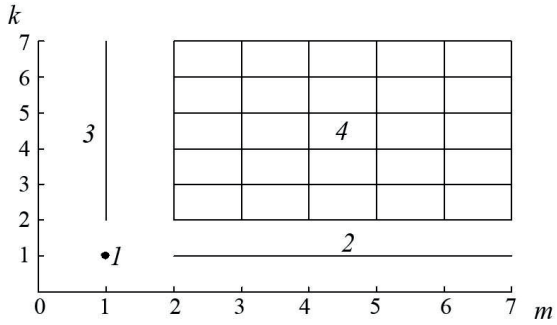

Рис. 7.3. Классификация методов регрессионного анализа:
1) линейная регрессия 2-х переменных; *2*) множественная линейная регрессия;
3) нелинейная одномерная регрессия; *4*) множественная нелинейная регрессия

2) дисперсия остатков должна быть постоянной $\left(\sigma_\varepsilon^2 = \text{const}\right)$, т. е. выполняется условие гомоскедастичности регрессионных остатков;

3) ошибки должны быть независимы (некоррелированы) с переменными X и Y;

4) независимая переменная X носит неслучайный характер;

5) желательно, но не обязательно, нормальное распределение остатков.

Заметим, что в уравнении (7.4) переменная X может называться независимой, факторной, регрессором, предиктором, а переменная Y – зависимой, результативной, функцией отклика, предиктантом. Иллюстрация второго условия регрессионных остатков приводится на рис. 7.4. Нетрудно видеть, что в первом случае (рис. 7.4-*а*) остатки (отклонения точек от прямой) имеют равномерное рассеяние (условие гомоскедастичности), а на другом графике (рис. 7.4-*б*) рассеяние остатков последовательно увеличивается, т. е. их дисперсия не остается постоянной (условие гетероскедастичности).

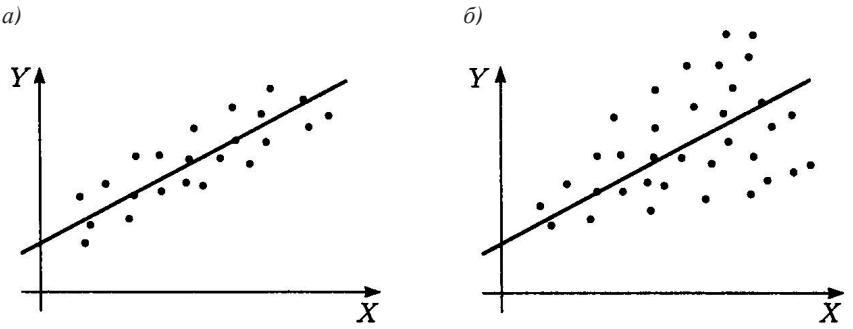

Рис. 7.4. Характер распределения остатков в регрессионной модели:
а) условие гомоскедастичности; *б)* условие гетероскедастичности

Первые три предположения являются необходимыми условиями использования метода наименьших квадратов и, очевидно, не нуждаются в комментариях. В соответствии с четвертым предположением, если переменная X не случайна, то это означает, что ее элементами служат известные числа, точно задаваемые исследователем. Отсюда следует, что единственным источником случайных возмущений значений y_i являются случайные возмущения регрессионных остатков ε_i. Но поскольку по определению ε_i − случайная

величина, то и Y тоже является случайной величиной, причем ее закон распределения соответствует закону распределения ε_i.

В результате сделанных предположений появляется возможность корректного использования метода наименьших квадратов (МНК) для определения неизвестных коэффициентов: a_0 – свободного члена, a_1 – коэффициента регрессии.

В соответствии с методом наименьших квадратов требуется минимизировать разность квадратов фактических и рассчитанных значений y_i, т. е.

$$S = \sum_1^n \left[y_i - \left(a_1 x_i + a_0 \right) \right]^2 = \min.$$

Отсюда нетрудно получить систему из двух нормальных линейных уравнений:

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = -2\sum\limits_{i=1}^{n} \left[y_i - a_0 - a_1 x_i \right] = 0, \\ \dfrac{\partial S}{\partial a_1} = -2\sum\limits_{i=1}^{n} \left[y_i - a_0 - a_1 x_i \right] x_i = 0 \end{cases}$$

или, после несложных преобразований:

$$\begin{cases} a_0 n + a_1 \sum x_i = \sum y_i, \\ a_0 \sum x_i + a_1 \sum x_i^2 = \sum x_i y_i. \end{cases} \qquad (7.5)$$

Решая (7.5) относительно параметров a_0 и a_1, имеем:

$$a_1 = \frac{\sum^n \left(x_i y_i - n\overline{xy} \right)}{\sum^n x_i^2 - n\overline{x}^2}, \qquad (7.6)$$

$$a_0 = \overline{y} - a_1 \overline{x}. \qquad (7.7)$$

Заметим, что параметры a_0 и a_1 могут быть представлены в несколько ином виде. Так, из (7.6) можно получить:

$$a_1 = r \frac{\sigma_y}{\sigma_x}. \qquad (7.8)$$

Отсюда видно, что коэффициент регрессии прямо пропорционален коэффициенту корреляции. Естественно, при отсутствии корреляции $a_1 = 0$.

Подставляя (7.8) в (7.7), имеем:

$$a_0 = \overline{y} - r\frac{\sigma_y}{\sigma_x}\overline{x}. \qquad (7.9)$$

С учетом выражений (7.8) и (7.9) классическое уравнение регрессии может быть переписано в следующем виде:

$$y - \overline{y} = r\frac{\sigma_y}{\sigma_x}\left(x - \overline{x}\right) + \varepsilon. \qquad (7.10)$$

Аналогичным образом может быть представлено уравнение регрессии x по y:

$$x - \overline{x} = r\frac{\sigma_x}{\sigma_y}\left(y - \overline{y}\right) + \varepsilon. \qquad (7.11)$$

Следует иметь в виду, что уравнения (7.10) и (7.11) являются различными самостоятельными зависимостями, взаимно не получаемыми одно из другого. Модель (7.10) условно называют *прямой регрессией,* а модель (7.11) – *обратной регрессией.* Прямые регрессии пересекаются в точке с координатами $\overline{x}$ и $\overline{y}$ и образуют «ножницы». При $|r| = 1$ прямые совпадают.

Физический смысл параметров a_0 и a_1 становится очевидным, если обратиться к рис. 7.5.

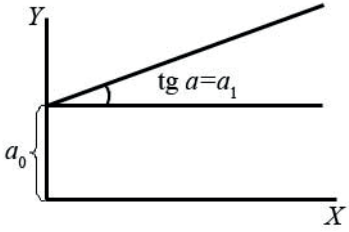

Рис. 7.5. Интерпретация коэффициентов в уравнении линейной регрессии

Коэффициент регрессии a_1 *характеризует тангенс угла наклона линии регрессии к оси абсцисс* ($a_1 = tg\alpha$), а свободный член a_0 *представляет собой расстояние от начала координат до точки пересечения оси ординат с линией регрессии.* Величина a_1 показывает насколько в среднем изменится зависимая переменная Y при изменении факторной переменной на единицу своего измерения.

В зависимости от знака при параметрах a_0 и a_1 линия регрессии занимает различное положение в декартовой системе координат (рис. 7.6). Если $a_0 = 0$, то линия регрессии проходит через начало координат.

Нетрудно показать, что линия регрессии должна обязательно проходить через точку с координатами $x_i = \overline{x}$ и $y_i = \overline{y}$. Другой такой

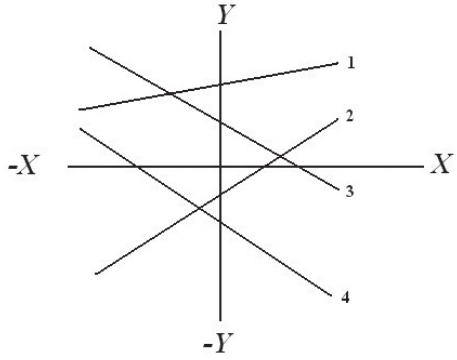

Рис. 7.6. Вид графика линейной регрессии в зависимости от значений параметров: *1*) $a_0 > 0$, $a_1 > 0$; *2*) $a_0 < 0$, $a_1 > 0$; *3*) $a_0 > 0$, $a_1 < 0$; *4*) $a_0 < 0$, $a_1 < 0$.

же «выдающейся» точкой является свободный член – точка пересечения прямой линии с осью *Y*.

Заметим, что в большинстве пакетов прикладных программ одновременно с обычными коэффициентами регрессии вычисляются их стандартизированные аналоги. Для этого предварительно производится расчет стандартизированных переменных *X* и *Y* по формуле $z_{ij} = \dfrac{\left(x_{ij} - \overline{x}_j\right)}{\sigma_j}$, т. е. из каждого наблюдения переменной вычитается средняя арифметическая, и результат делится на ее стандартное отклонение. Напомним, стандартизированная переменная z_i обладает тем свойством, что ее среднее значение равно нулю, а дисперсия равна единице. В результате применения метода наименьших квадратов к новым переменным, получаем следующее стандартизированное уравнение линейной регрессии:

$$z_y = \beta z_x. \qquad (7.12)$$

Здесь z_y, β, z_x – стандартизированные значения функции отклика, коэффициента регрессии и независимой переменной. Нетрудно видеть, что свободный член в уравнении (7.12) равен нулю. Физический смысл стандартизованного коэффициента регрессии состоит в том, что он показывает относительную роль переменной *X* в описании изменчивости функции отклика. Между коэффициентами в уравнениях (7.4) и (7.12) существует функциональная взаимосвязь:

$$\beta = a_1 \left(\frac{\sigma_x}{\sigma_y} \right), \tag{7.13}$$

где σ_x – стандартное отклонение переменной *X*. Отсюда следует, что чем больше изменчивость *X*, тем больше величина β.

7.3. Оценивание параметров линейной регрессии двух переменных

Рассмотрим основные критерии качества линейной модели регрессии.

Линейный коэффициент детерминации:

$$r^2 = \frac{\sigma^2_{y(x)}}{\sigma^2_y} = 1 - \left(\frac{\sigma^2_\varepsilon}{\sigma^2_y} \right), \tag{7.14}$$

где $\sigma^2_{y(x)}$ – дисперсия вычисленных по уравнению регрессии значений функции отклика, σ^2_y – выборочная дисперсия фактических значений переменной *Y*, σ^2_ε – дисперсия остатков. Отсюда видно, что *коэффициент детерминации показывает долю объясненной дисперсии функции отклика*. Если, например, $r^2 = 0{,}80$, то это означает, что 80 % изменчивости функции отклика описывается с помощью модели регрессии. Коэффициент детерминации изменяется в пределах от 0 до 1. При $r^2 = 1$ $\sigma^2_\varepsilon = 0$, при $r^2 = 0$ $\sigma^2_\varepsilon = \sigma^2_y$. В первом случае остатки отсутствуют, вычисленные по модели и фактические значения переменной *Y* совпадают, во втором случае вся дисперсия переменной *Y* уходит в остатки.

Среднеквадратическое (стандартное) отклонение модели:

$$\sigma_{y(x)} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - y_{(x)i} \right)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-1}}. \tag{7.15}$$

Можно показать, что данная величина для нормально распределенных совокупностей функционально связана с линейным коэффициентом детерминации формулой:

$$\sigma_{y(x)} = \sigma_y \sqrt{1 - r^2}. \tag{7.16}$$

Отсюда видно, что чем выше коэффициент корреляции, тем меньшей оказывается стандартная погрешность уравнения регрессии.

Стандартные ошибки коэффициента корреляции и коэффициентов регрессии:

$$\sigma_r = \frac{1-r^2}{\sqrt{n-1}}; \tag{7.17}$$

$$\sigma_{a_1} = \frac{\sigma_y}{\sigma_x}\sqrt{\frac{1-r^2}{n-1}}; \tag{7.18}$$

$$\sigma_{a_0} = \frac{\sigma_{y(x)}}{\sqrt{n-1}} = \frac{\sigma_y\sqrt{1-r^2}}{\sqrt{n-1}}. \tag{7.19}$$

Заметим, что в некоторых статистических изданиях для более корректного учета смещенности рекомендуется $n - 1$ заменять на $n - 2$. Впрочем, это может сказываться на оценках ошибок только для очень малых выборок. Напомним также, что использование формулы (7.17) правомерно только при условии, что выборочные значения r подчиняются нормальному закону, т. е. при сравнительно малых значениях r и большой длине исходных рядов n. При больших значениях r и малых значениях n следует применять *z-преобразование Фишера*.

Итак, формулы (7.17) – (7.19) имеют очень близкую структуру. Стандартные ошибки коэффициента корреляции и коэффициентов регрессии обратно пропорциональны коэффициенту детерминации и длине выборки. Чем они больше, тем меньше оценки стандартных ошибок. Кроме того, стандартные ошибки коэффициентов регрессии прямо пропорциональны оценке стандартного отклонения функции отклика.

Формулы (7.17) – (7.19) используются при проверке коэффициента корреляции и коэффициентов регрессии на значимость и построении доверительных интервалов. Для этой цели применяется *t-статистика Стьюдента*. Методика их построения для коэффициента корреляции приводится в главе 6. Аналогичным образом осуществляется оценка значимости и построение доверительных интервалов для коэффициентов регрессии. Вначале записывается нулевая гипотеза вида $H_0 : |b_j| = 0$ при альтернативе $H_1 : |b_j| \neq 0$, для проверки которой вычисляется $t = \dfrac{|b_j|}{\sigma_{bj}} = 0$. Далее проверяется неравенство: $t > \boldsymbol{t}_{\text{кр}}(\alpha, \nu = n - 2)$.

Если нулевая гипотеза отвергается, то соответствующий коэффициент регрессии считается значимым, т. е. отклоняющимся от нуля неслучайным образом. Заметим, что в большинстве современных пакетах прикладных статистических программ процедура проверки значений b_j на значимость реализуется через *p-критерий* (*p-level*), представляющий собой отношение коэффициента регрессии к его стандартному отклонению, который затем с учетом числа степеней свободы по распределению Стьюдента переводится в уровень значимости. Например, *p-level* = 0,03 означает, что рассматриваемый коэффициент регрессии значим на уровне 0,05 и не значим на уровне 0,01. По существу, *p-level представляет минимальный уровень значимости, при котором отвергается нулевая гипотеза.* Заметим, что проверка на значимость коэффициента корреляции эквивалентна проверке на значимость коэффициента регрессии. Если *r* значим, то коэффициент регрессии тоже является значимым и наоборот.

7.4. Оценка адекватности регрессионной модели

Адекватность в переводе на русский язык означает «соответствие», «тождественность». Поэтому *под адекватностью регрессионной модели понимается, насколько хорошо она соответствует исходным данным*. Оценка адекватности регрессионной модели основывается на положениях дисперсионного анализа. Прежде всего, вспомним важное свойство дисперсий двух рядов:

$$D(x+y) = \sigma_x^2 + \sigma_y^2 + 2r\sigma_x\sigma_y,$$

т. е. дисперсия суммы двух переменных равна сумме дисперсий плюс удвоенное произведение средних квадратических отклонений переменных X и Y на коэффициент корреляции между ними. При $r = 0$ переменные являются некоррелированными.

Представим зависимую переменную y_i в виде $y_i = y_{(x)i} + \varepsilon_i$, где $y_{(x)i}$ – вычисленные по уравнению регрессии значения y_i; ε_i – ошибки регрессии, для которых предполагается нормальное распределение с нулевым средним $(\bar{\varepsilon} = 0)$ и отсутствие корреляции с переменными x_i и y_i. В этом случае дисперсия переменной y_i будет равна:

$$D_y = D_{y(x)} + D_\varepsilon.$$

Или при переходе к выборочным параметрам:

$$\sigma^2_y = \sigma^2_{y(x)} + \sigma^2_\varepsilon.$$

Но так как:

$$D_{y(x)} = D\left(a_0 + a_1 x\right) = a_1^2 D_x = r^2 \left(\frac{D_y}{D_x}\right) D_x = r^2 D_y,$$

то получим:

$$r^2 = \frac{D_{y(x)}}{D_y}, \quad \frac{D_\varepsilon}{D_y} = 1 - r^2.$$

Итак, мы имеем три характеристики:

– дисперсию исходной переменной y_i, характеризующую ее общую изменчивость (D_y);

– дисперсию вычисленных по модели значений $y_{(x)i}$, характеризующую изменчивость линии регрессии относительно среднего значения ($D_{y(x)}$);

– дисперсию остатков ε_i, характеризующую отклонение прямой, построенной по методу наименьших квадратов, от результатов наблюдений (D_ε).

Исходя из дисперсионного анализа, первый вид дисперсии интерпретируется как *общая дисперсия,* второй – *межгрупповая дисперсия,* третий – *внутригрупповая дисперсия.*

Отношение $r^2 = \frac{D_{y(x)}}{D_y}$ показывает степень (качество) приближения вычисленных значений $y_{(x)i}$ по уравнению регрессии к фактическим значениям y_i или, другими словами, долю дисперсии функции отклика, описываемой моделью регрессии. Квадрат коэффициента корреляции r^2, как было указано выше, – это *линейный коэффициент детерминации.*

Отношение D_ε / D_y показывает степень неопределенности уравнения регрессии, т. е. долю изменчивости переменной y_i, которая не может быть объяснена переменной x_i Фактически отношение D_ε / D_y – это некоторая шумовая составляющая уравнения регрессии.

Естественно, чем выше коэффициент детерминации, тем выше качество приближения вычисленных значений $y_{(x)i}$ к фактическим значениям y_i и тем меньше уровень шума (зашумленность) уравнения регрессии.

На практике для оценки адекватности регрессионной модели обычно используется дисперсионный анализ. Согласно его основной идее мы можем записать:

$$\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2=\sum_{i=1}^{n}\left(y_{(x)i}-\overline{y}\right)^2+\sum_{i=1}^{n}\left(y_{(x)i}-y_i\right)^2$$

или

$$Q=Q_R+Q_\varepsilon,$$

где Q – сумма квадратов фактических отклонений зависимой переменной от арифметической средней; Q_R − сумма квадратов вычисленных по уравнению регрессии отклонений зависимой переменной от средней; Q_ε − остаточная сумма квадратов, характеризующая случайную ошибку (влияние неучтенных факторов).

В результате оценки указанных сумм квадратов составляется таблица по специальной форме (таблица 7.1). Число степеней свободы зависит от длины ряда (n) и числа оцениваемых в модели параметров (q), причем $q = m + 1 = 2$, где m – число независимых переменных в модели. Нетрудно показать, что суммы квадратов связаны со значениями дисперсий как:

$$\sigma^2_{y(x)}=\frac{Q_R}{(q-1)},\ \ \sigma^2_\varepsilon=\frac{Q_e}{(n-q)},\ \ \sigma^2_y=\frac{Q}{(n-1)}.$$

Таблица 7.1

Проверка адекватности линейной регрессионной модели $y = a_0 + a_1x + \varepsilon$

Источник изменчивости	Сумма квадратов	Число степеней свободы	Критерий Фишера
Линейная регрессия	$Q_R=\sum\left(y_{(x)i}-\overline{y}\right)^2$	$q-1$	
Отклонение от регрессии	$Q_\varepsilon=\sum\left(y_{(x)i}-y_i\right)^2$	$n-q$	
Общая изменчивость	$Q=\sum\left(y_i-\overline{y}\right)^2$	$n-1$	$F=\frac{\sigma^2_{y(x)}}{\sigma^2_\varepsilon}$

При оценке адекватности (значимости) модели составляется нулевая гипотеза о равенстве дисперсий, т. е. $H_0:\sigma^2_{y(x)}=\sigma^2_\varepsilon$ при альтернативе $H_0:\sigma^2_{y(x)}\neq\sigma^2_\varepsilon$. Проверка нулевой гипотезы осуществляется с помощью F-критерия, который в соответствии с таблицей 7.1 имеет вид:

$$F=\frac{\sigma^2_{y(x)}}{\sigma^2_\varepsilon}=\frac{Q_R(n-q)}{Q_\varepsilon(q-1)}=Q_R\frac{(n-2)}{Q_\varepsilon}. \qquad (7.20)$$

После этого проверяется неравенство:

$$F > \boldsymbol{F}_{\text{кр}}(\alpha, \nu_1 = 1, \nu_2 = n - 2).$$

Если $F > \boldsymbol{F}_{\text{кр}}(\alpha, \nu_1, \nu_2)$, то нулевая гипотеза о равенстве дисперсий отвергается и, следовательно, справедливой является альтернативная гипотеза $H_1 : \sigma^2_{y(x)} \neq \sigma^2_{\varepsilon}$. Отсюда следует, что дисперсия переменной Y, вычисленная по линейной *регрессии, неслучайным образом отличается от дисперсии шума.* Поэтому регрессионная модель является адекватной (значимой), т. е. она хорошо описывает исходные данные. В противном случае у нас есть основания полагать, что она плохо описывает исходные данные.

Пример 7.2. Средний годовой уровень Каспийского моря зависит от внутригодовых изменений объема его вод ΔV. Если $\Delta V > 0$, то уровень моря повышается, если $\Delta V < 0$, то уровень понижается. Из уравнения (5.4) видно, что изменения объема вод в свою очередь складываются из притока речных вод, осадков и испарения с акватории моря, причем определяющим фактором является приток речных вод. Было установлено, что последний фактор практически линейно зависит от годового стока Волги ($Q_{\text{В}}$), изменчивость которого в основном обусловлена зоной формирования стока выше г. Самары. Поэтому представляет интерес выявление степени связи межгодовых колебаний стока Волги и изменений объема вод моря. Отметим, что в этом случае нет смысла в физическом анализе связи, ибо априори ясно, что сток Волги влияет на колебания объема вод моря, а не наоборот.

В нашем распоряжении имелись данные по внутригодовым изменениям объема моря (км^3/год) и годовому стоку Волги в г. Самара (м^3/с) за период с 1890 по 1990 г., т. е. за 101 год. Прежде всего, было построено корреляционное поле данных характеристик, которое приводится на рис. 7.7. Нетрудно видеть, что между ними отмечается хорошо выраженная линейная связь. Первичные статистические характеристики ΔV и $Q_{\text{В}}$ приведены в таблице 7.2.

В результате анализа эмпирической гистограммы было установлено, что исходные данные не в полной мере соответствуют нормальному закону распределения. Однако, как известно, степень надежности рассчитываемых статистических характеристик возрастает с увеличением длины выборки. В данном случае объем выборки весьма внушителен. Поэтому целесообразность построения линейной регрессионной модели между значениями изменений объема моря и годовым стоком Волги ($Q_{\text{В}}$) является очевидной.

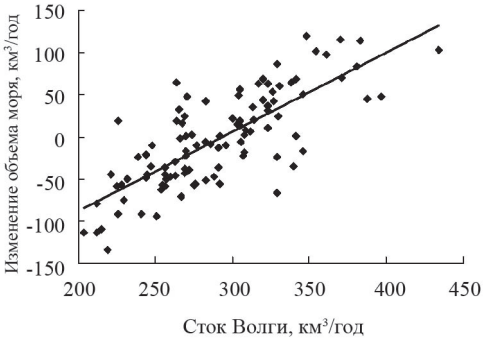

Рис. 7.7. Корреляционное поле между внутригодовыми изменениями объема Каспийского моря (км³/год) и годовым стоком Волги в г. Самара (км³/год) за период с 1890 по 1990 г.

Таблица 7.2.

Первичные статистические характеристики изменений объема моря (км³/год) и стока Волги (м³/с)

Параметр	Среднее	Стандартное отклонение	X_{max}	X_{min}	R
ΔV	–3,3	56,4	120	–135	255
$Q_В$	7531	1430	11 600	4680	6920

При использовании МНК получено следующее уравнение регрессии:

$$\Delta V = -246{,}7 + 0{,}0323 Q_В. \qquad (7.21)$$

Оценки параметров этого уравнения даются ниже:

– коэффициент корреляции $r = 0{,}79$,

– коэффициент детерминации $r^2 = 0{,}63$,

– среднеквадратическое (стандартное) отклонение модели $\sigma_{y(x)} = 34{,}6$ км³/год,

– стандартная ошибка коэффициента корреляции $\sigma_r = 0{,}04$,

– стандартная ошибка коэффициента регрессии $\sigma_{a1} = 0{,}0025$,

– стандартная ошибка свободного члена регрессии $\sigma_{a0} = 19{,}1$.

После этого осуществляется проверка параметров регрессии на значимость. С этой целью воспользуемся их оценками *p-level.* Для свободного члена *p-level* $= 6{,}8 \times 10^{-23}$, для коэффициента регрессии *p-level* $= 5{,}7 \times 10^{-23}$, т. е. значимость их настолько велика, что они очень близки к истинным оценкам, которые могли бы быть получены по генеральной совокупности. Очевидно, поэтому в построении доверительных интервалов нет необходимости.

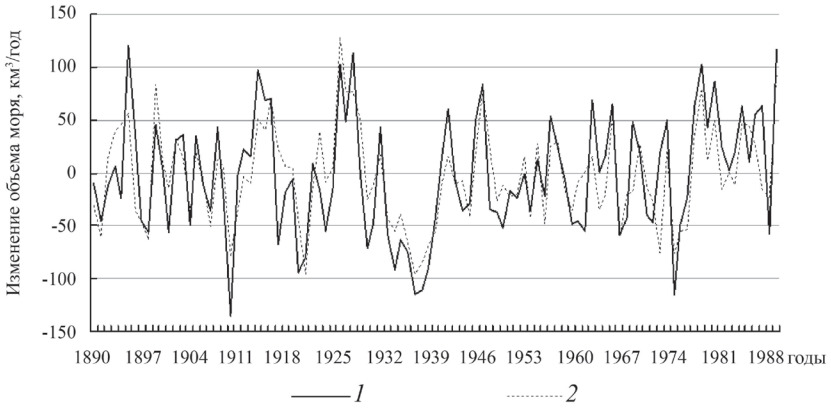

Рис. 7.8. Сопоставление фактических (*1*) и вычисленных по модели (*2*) значений изменений объема вод Каспийского моря за период с 1890 по 1990 г.

Адекватность регрессионной модели проверяем по критерию Фишера: $F = \dfrac{\sigma^2_{y(x)}}{\sigma^2_{\varepsilon}} = 166{,}9$. Критическое значение статистики Фишера при $\alpha = 0{,}05, \nu_1 = 1, \nu_2 = 99$ равно $\boldsymbol{F}_{\text{кр}} = 3{,}93$. Отсюда следует, что модель (7.21) полностью адекватна. Сопоставление фактических и вычисленных по модели значений ΔV представлено на рис. 7.8. Нетрудно видеть, что в целом отмечается неплохое соответствие, хотя в отдельные годы расхождения между этими характеристиками весьма значительны. Так, максимальное завышение наблюдается в 1917 г. и составляет −90 км³/год, а максимальное занижение – в 1973 г., достигающее 93,8 км³/год.

7.5. Анализ остатков регрессионной модели

Анализ остатков является необходимым условием проверки оптимальности регрессионной модели. Особенно он полезен в следующих ситуациях:

– когда возможны грубые ошибки наблюдений, выбросы, ошибки при записи на машинные носители;

– когда линейная форма модели может быть не пригодна для описания данных;

– когда основные гипотезы модели не выполняются и оценки ее параметров являются мало надежными;

– когда необходимо изменить масштаб координатных осей или провести преобразование исходных данных.

Анализ остатков, в отличие от параметров модели, обычно проводится визуально. С этой целью строится ряд графиков:

– общий график остатков в координатах нормального распределения (гистограмма);

– зависимость остатков от времени;

– зависимость остатков от переменной Y;

– зависимость остатков от переменной X.

Графики включены в состав большинства пакетов прикладных программ, поэтому их построение и анализ не вызывает каких-либо затруднений. Отметим только, что первый график строится в том случае, когда длина выборки достаточно большая и позволяет произвести разбиение остатков на градации. Проверку соответствия остатков нормальному закону можно осуществить на основе критерия Пирсона χ^2. Если остатки подчиняются нормальному закону, а на перечисленных графиках их распределение оказывается независимым, т. е. наблюдается горизонтальная полоса рассеяния, параллельная оси абсцисс (рис. 7.9-*а*), то это означает адекватность модели. Если полоса рассеяния расширяется (сужается), когда значения x_i или y_i возрастают (рис. 7.9-*б*), то это указывает на непостоянство дисперсии остатков, называемое *гетероскедастичностью*. Если остатки зависят линейно или нелинейно от времени (рис. 7.9-*в*), то это свидетельствует о наличии в значениях функции отклика отчетливо выраженного тренда, который должен быть исключен из исходных данных. Наличие криволинейной полосы рассеяния в зависимости остатков от переменной X (рис. 7.9-*г*) означает, что линейная модель неудовлетворительно описывает связь этой переменной с функцией отклика. Поэтому необходимо перейти от линейной модели к нелинейной.

Наличие в модели гетероскедастичности является весьма неприятным фактом, ибо постоянство дисперсии остатков относится к числу ключевых предпосылок использования МНК. При невыполнимости этой предпосылки оценки коэффициентов регрессии не будут эффективными, причем их дисперсии оказываются смещенными. В результате все выводы, полученные с использованием статистик Стьюдента и Фишера, станут ненадежными. Поэтому возрастает возможность сделать неверные статистические

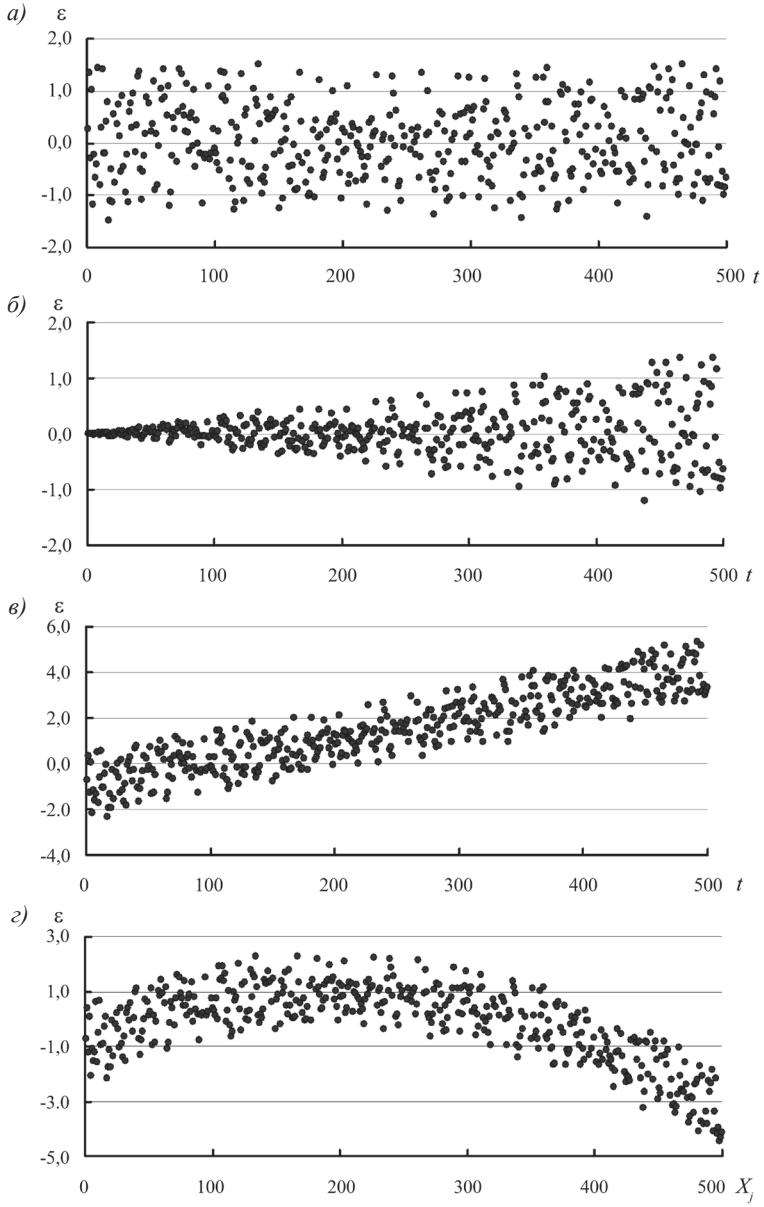

Рис. 7.9. Характеристика модели в зависимости от распределения остатков: *а)* модель адекватна, *б)* гетероскедастичность модели, *в)* наличие в модели линейного тренда, *г)* наличие нелинейной связи функции отклика с переменной X

выводы. Один из способов «борьбы» с гетероскедастичностью заключается в использовании так называемого *метода взвешенных наименьших квадратов,* рассмотрение которого выходит за рамки данной книги.

Наконец, необходима дополнительная проверка на наличие корреляции остатков между собой. Такое явление получило название сериальной корреляции. Если регрессионная модель представляет совокупность временных рядов, то в этом случае *сериальная корреляция* превращается в автокорреляционную функцию остатков. Широко распространенным критерием выявления сериальной корреляции является *критерий Дарбина–Уотсона,* который приводится практически во всех ППСП. Данный критерий состоит в вычислении статистики d:

$$d = \frac{\sum\left(\varepsilon_{i+1} - \varepsilon_i\right)^2}{\sum \varepsilon_i^{\,2}}, \qquad (7.22)$$

которая учитывает наличие взаимосвязи только между смежными значениями остатков. Можно показать, что для достаточно длинных рядов должно выполняться следующее соотношение:

$$d \approx 2(1 - r),$$

где r – коэффициент корреляции между остатками ε_{i+1} и ε_i. Отсюда видно, что при наличии высокой положительной корреляции между остатками ($r \to 1$) величина d становится близкой к нулю ($d \to 0$), при высокой отрицательной корреляции ($r \to -1$) $d \to 4$. Отсутствие корреляции означает, что $d \to 2$.

Для оценки значимости коэффициента сериальной корреляции составляется нулевая гипотеза вида $H_0 : d = 0$ при альтернативе $H_0 : d \neq 0$. Для ее проверки можно воспользоваться специальными таблицами, позволяющими определять критические величины данной статистики (d_1, d_2 – нижняя и верхняя границы) по уровню значимости и числу степеней свободы, соответствующих числу переменных в модели. При этом величина d может принимать значения в интервале $0 \leq d \leq 4$. Чтобы проверить значимость отрицательной корреляции, нужно вычислить величину $4 - d$. Далее проверка осуществляется по указанной выше схеме.

В результате проверки нулевой гипотезы возможно несколько исходов, которые представлены в таблице 7.3.

Из таблицы 7.3 видно, что для двух диапазонов значений статистики d возникает неопределенность в интерпретации результатов,

Таблица 7.3

Оценка значимости коэффициента сериальной корреляции

Значение статистики d	Вывод
$4 - d_1 < d < 4$	Гипотеза H_0 отвергается, есть отрицательная корреляция
$4 - d_2 < d < 4 - d_1$	Неопределенность
$d_2 < d < 4 - d_2$	Гипотеза H_0 не отвергается
$d_1 < d < d_2$	Неопределенность
$0 < d < d_1$	Гипотеза H_0 отвергается, есть положительная корреляция

причем интервал $[d_1, d_2]$ особенно при малой длине выборки n довольно широк. Следовательно, значительной оказывается неопределенность в интерпретации результатов. Кроме того, другой существенный недостаток таблиц критических значений d состоит в том, что число переменных, входящих в модель, ограничено до $m = 5$.

Если же не прибегать к помощи таблиц, то можно отметить следующее. Чем меньше величина d, тем сильнее отмечается положительная корреляция между остатками, а чем ближе величина d приближается к 4, тем сильнее отрицательная корреляция. Отсутствие сериальной корреляции для линейной регрессии проявляется в некотором диапазоне значений d. Например, при уровне значимости $\alpha = 0{,}05$ и $n = 20$ сериальная корреляция отсутствует, если $1{,}41 < d < 2{,}59$, а при $n = 50$ она отсутствует при $1{,}59 < d < 2{,}41$, т. е. величина d находится вблизи 2. Очевидно, приближенно можно принять выполнимость условия $1{,}50 < d < 2{,}50$.

Пример 7.3. Оценим остатки регрессионной модели (7.21) по расчету изменений объема воды Каспийского моря. Распределение остатков в зависимости от стока Волги приводится на рис. 7.10. Нетрудно видеть, что остатки представляют собой довольно хорошо выраженную горизонтальную полосу рассеяния, параллельную оси абсцисс. Критерий Дарбина–Уотсона равен $d = 1{,}84$. Следовательно, можно сделать вполне определенный вывод, что распределение остатков носит случайный характер.

7.6. Понятие о робастной регрессии

Как отмечалось выше, термин «робастный» означает устойчивость к сравнительно малым отклонениям от принятых предположений. Очевидно, робастная регрессия может служить дополнением к классическому методу наименьших квадратов, если исходная

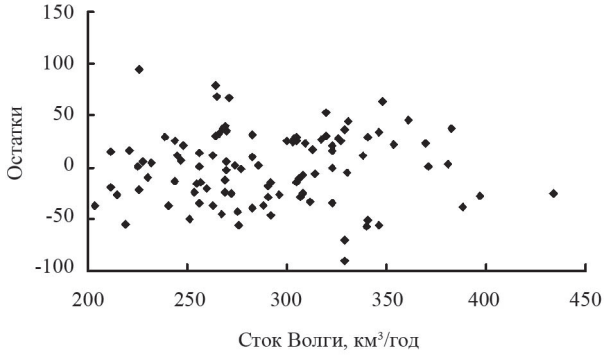

Рис. 7.10. Распределение остатков регрессионной модели (7.21) в зависимости от стока Волги

выборка является «засоренной», т. е. содержит резко отличающиеся от исходной совокупности данные.

Чтобы понять трудности классического регрессионного анализа на основе МНК, рассмотрим следующий пример. Пусть мы имеем по 6 значений переменных X и Y (таблица 7.4), для которых необходимо построить линейную регрессию. На рис. 7.2-*а* дана прямая, построенная с помощью МНК, уравнение которой имеет вид:

$$y_{i1} = 0{,}068 - 0{,}081x_i. \qquad (7.23)$$

Коэффициент детерминации этой модели мал, а сама модель незначима по критерию Фишера. Тем не менее, на первый взгляд, ошибки этой модели ($\varepsilon_i = y_i - y_{(x)i}$) сравнительно невелики. При этом аномально больших ошибок не отмечается. Сомнение может вызвать точка 1, где ошибка максимальна и точка 6, которая находится довольно далеко от остальной совокупности. Однако именно точка 6 далеко «увела» линию регрессии от остальных точек.

Таблица 7.4

Тестовый пример построения регрессионной модели

Точка	X	Y	Ошибки регрессионной модели		
			Модель (7.23)	Модель (7.24)	Модель (7.25)
1	–4	2,48	2,09	0,44	0,25
2	–3	0,73	0,42	–0,33	0,26
3	–2	–0,04	–0,27	–0,12	–0,13
4	–1	–1,44	–1,59	–0,54	–0,44
5	0	–1,32	–1,39	0,55	0,42
6	10	0	0,75	11,64	–0,01

Без учета последней точки уравнение линейной регрессии примет уже другой вид (рис. 7.2-*б*):

$$y_{i2} = -1{,}87 - 0{,}977x_i. \qquad (7.24)$$

При этом коэффициент детерминации вырос и модель стала значимой даже при самом жестком критерии Фишера. Естественно, стандартная ошибка модели тоже резко уменьшается (таблица 7.5).

Таблица 7.5

Оценки параметров регрессионных моделей

Модель	Параметры регрессионной модели		
	R^2	F	$\sigma_{(x)}$
7.23	0,10	0,4	1,55
7.24	0,91	31,1	0,95
7.25	0,95	29,7	0,40

Впрочем, исходя из полученных результатов, можно сделать и другой вывод: в данном конкретном случае линейная модель просто не применима и поэтому следует воспользоваться более сложной моделью в виде параболы (рис. 7.2-*в*):

$$y_{i3} = -1{,}74 - 0{,}66x + 0{,}08x_i^2. \qquad (7.25)$$

Точность данной модели стала лишь чуть-чуть выше (таблица 7.5), ибо коэффициент детерминации увеличился только на величину 0,04. В то же время стандартная ошибка нелинейной модели уменьшилась более чем в два раза. Если судить по ошибкам, то последний вариант, безусловно, заслуживает предпочтения, ибо ему соответствует самая малая ошибка. Однако в действительности данный пример является искусственным. К шести точкам, лежащим на прямой $y = -2 - x$, были добавлены случайные ошибки, причем к первым пяти – ошибки с нулевым средним и стандартным отклонением $\sigma = 0{,}6$, а к 6-й – большая ошибка, равная $\delta = 12$.

Итак, классический метод наименьших квадратов весьма чувствителен к выбросам и при их наличии может сильно искажать истинное уравнение. Поэтому желательно иметь такой метод, который бы «распознавал» выбросы и автоматически исключал их из расчетов. Таким методом как раз и является робастная регрессия. Для ее решения используются, в основном, *M*-оценки максимального правдоподобия. При этом вместо минимизации непосредственно суммы квадратов остатков осуществляется минимизация некоторой функции от остатков:

$$M = \sum_{i=1}^{n} \rho(\varepsilon_i) \to \min. \qquad (7.26)$$

Значение, обращающее условие (7.26) в минимум для некоторой функции, называют *M*-оценкой. Эта оценка рассматривается как оценка максимума правдоподобия. Заметим, что поскольку выбор функции ρ довольно произволен, то в зависимости от ее вида могут быть получены разные значения коэффициентов регрессии. На функцию ρ накладывается условие существования первой и второй производной, т. е. $\rho'(\varepsilon) = \psi(\varepsilon)$, $\rho''(\varepsilon) = \gamma(\varepsilon)$. Это означает, что ρ является выпуклой непрерывной функцией. Дифференцируя (7.26), получаем систему линейных уравнений:

$$\sum_{i=1}^{n} \Psi(\varepsilon_i) x_{ik} = 0 \quad (k = 1,\ 2,\ \ldots,\ m). \qquad (7.27)$$

Кроме того, при построении робастных оценок обычно вводят параметр масштаба остатков *s*, что приводит к решению системы уравнений:

$$\sum_{i=1}^{n} \frac{\Psi(\varepsilon_i)}{sk^*} x_{ik} = 0, \qquad (7.28)$$

где k^* − некоторая константа, выбираемая из соображений неформального характера. Так как параметр масштаба *s* неизвестен, то на практике его оценка в простейшем случае может быть найдена как медиана абсолютных отклонений медианы от оценки остатка, т. е.

$$MAD = med\{|\varepsilon_i - med(\varepsilon_i)|\}.$$

При этом остатки ε_i предварительно находятся с помощью классического МНК.

Для «засоренных» нормально распределенных выборок П. Хьюбер предложил семейство оценок, определяемых функцией ρ:

$$\rho(\varepsilon) = 0{,}5\ \varepsilon^2 \qquad \text{при } |\varepsilon| < k^*s = h,$$

$$\rho(\varepsilon) = k^*s(|\varepsilon| - 0{,}5\ k^*s) \qquad \text{при } |\varepsilon| \geq k^*s = h.$$

Здесь *h* – параметр робастности. Заметим, что выбор ρ, ψ и *k* представляет собой весьма сложную задачу и не является однозначным.

Полученные П. Хьюбером оценки оказываются робастными в том случае, когда при больших ошибках ε_i скорость роста функции ρ(ε) становится меньше скорости роста принятой в МНК суммы

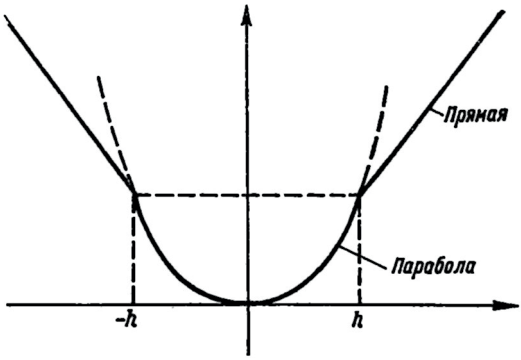

Рис. 7.11. График функции Хьюбера на отрезке [–*h*, *h*]

квадратов остатков. Как следует из графика функции Хьюбера (рис. 7.11), парабола на отрезке [–*h*, *h*] продолжается далее двумя прямыми линиями. Значение *h*, определяющее порог, после которого происходит уменьшение скорости роста ρ, по существу является параметром, определяющим степень робастности оценок метода.

Если значение параметра *h* достаточно велико, то полученные регрессионные оценки совпадают с обычными оценками метода наименьших квадратов. Значения *h* обычно подбирают исходя из конкретных свойств исследуемых статистических совокупностей и, в частности, из представлений о «критических» отклонениях от расчетной модели.

Для вычисления коэффициентов робастной регрессии используются методы модифицированных весов, модифицированных остатков и псевдонаблюдений. Наиболее простым представляется метод псевдонаблюдений.

Пример 7.3. Как было показано в примере 6.5, между выловом ставриды в юго-восточной части Тихого океана (ЮВТО) и смещением Южнотихоокеанского антициклона (ЮТА) по долготе отмечается довольно высокая статистическая связь (коэффициент корреляции Спирмена ρ = –0,70). Корреляционное поле между этими переменными приводится на рис. 7.12, из которого отчетливо видно, что, вообще говоря, из общей совокупности точек выделяются две, которые явно лежат вне линии связи. Поэтому рассмотрим возможный эффект использования робастной регрессии применительно к данным переменным. С этой целью запишем линейную регрессионную модель в стандартном виде:

$$V = b_0 + b_1\lambda_{\text{ЮТА}} + \varepsilon, \qquad (7.29)$$

где V – вылов рыбы в 10^3 т.

Коэффициенты регрессии будем определять классическим методом наименьших квадратов и робастным вариантом – методом модифицированных весов. Статистические параметры уравнения регрессии даны в таблице 7.6, а сами графики уравнений – на рис. 7.12.

Нетрудно видеть, что робастная регрессия не учитывает не две точки, как это казалось визуально, а три точки, вылов в которых составлял меньше 50×10^3 т/год. Именно поэтому стандартное отклонение для модели робастной регрессии оказалось несколько больше, чем для классической регрессии. Впрочем, это расхождение несущественно. В то же время достаточно очевидным является, что точность описания всех других точек рассматриваемой совокупности заметно выше по сравнению с классическим МНК. В частности, коэффициент ранговой корреляции Спирмена без трех точек возрос до $\rho = -0{,}93$.

Таблица 7.6

Статистические параметры уравнения регрессии (7.29)

Регрессия	Свободный член	Коэффициент регрессии	Коэффициент корреляции	Стандартное отклонение, 10^3 т/год
Робастная	–9,58	–0,71	0,65	3,56
Классическая	–18,55	–0,79	0,65	3,13

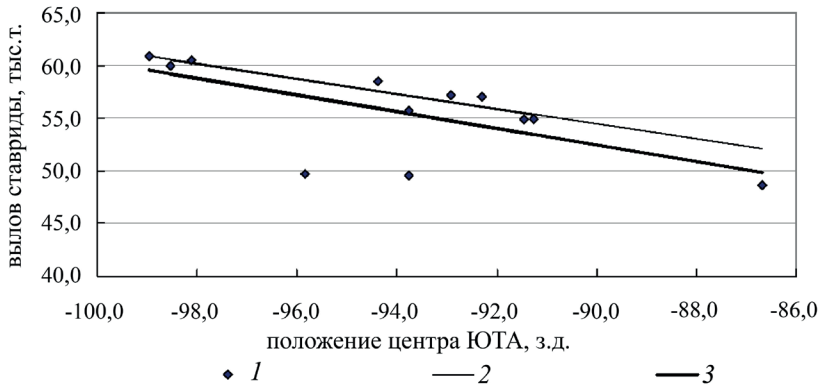

Рис. 7.12. График связи вылова ставриды и смещения ЮТА по долготе:
1) фактические значения, *2)* робастная регрессия, *3)* линейная регрессия

Если же исключить выпадающие точки из выборки и осуществить расчет уравнения регрессии классическим МНК, то оценки стандартного отклонения для обоих видов регрессии будут различаться еще меньше, чем в таблице 7.6. Таким образом, с помощью робастной регрессии можно установить точки, далеко отстоящие от линии связи между переменными. Если же такие точки установлены другим (например, экспертным) путем, то в этом случае использование робастной регрессии теряет свою эффективность. Однако далеко не всегда определение отскакивающих точек (выбросов) является очевидным. Так, крайняя точка справа на рис. 7.12 визуально вряд ли может быть признана отскакивающей. И только с помощью робастного подхода удалось это установить.

7.7. К построению кусочно-линейных моделей регрессии

В некоторых случаях анализ структуры связи между переменными в корреляционном поле позволяет предположить, что разбиение исходной выборки на две части может существенно повысить точность описания функции отклика. Действительно, как следует из рис. 7.13, на левом графике корреляция между переменными X и Y является довольно высокой и равна $r = 0{,}80$. В тоже время отчетливо видно, что если данную выборку разбить на две части, то корреляция между переменными заметно увеличивается. Так, для первой подвыборки $r = 0{,}91$, а для второй подвыборки $r = 0{,}92$.

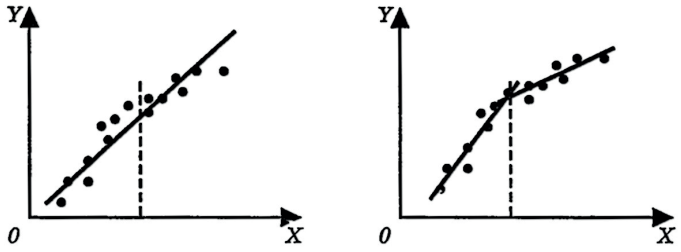

Рис. 7.13. Применение теста Чоу при построении кусочно-линейной регрессии

В связи с этим возникает задача нахождения таких условий, при которых целесообразен переход от общей регрессии, построенной по полной выборке, к кусочно-линейным моделям, построенным

для двух или более подвыборок. Решение данной задачи возможно на основе критерия Чоу, суть которого заключается в следующем.

Пусть выборка имеет объем n. Обозначим через S_0 сумму квадратов отклонений y_i от общего уравнения регрессии. С помощью визуального анализа разбиваем всю выборку на две части объемами n_1 и n_2 соответственно. Для каждой из них необходимо построить свое собственное уравнение линейной регрессии. Обозначим теперь через S_1 и S_2 суммы квадратов отклонений значений y_i каждой из подвыборок от соответствующих уравнений регрессии. Очевидно, равенство $S_0 = S_1 + S_2$ возможно лишь при совпадении коэффициентов регрессии для всех трех уравнений.

Естественно, чем сильнее различие в поведении Y для двух подвыборок, тем больше значение S_0 будет превосходить сумму $S_1 + S_2$. Тогда разность $S_0 - (S_1 + S_2)$ может быть интерпретирована как улучшение качества модели при разбиении объема выборки n на две части. Отсюда следует, что дробь $\frac{\left[S_0 - \left(S_1 + S_2\right)\right]}{\left(m+1\right)}$ определяет оценку уменьшения дисперсии за счет построения двух уравнений регрессии вместо одного. При этом число степеней свободы сократится на $(m + 1)$, поскольку вместо $(m + 1)$ параметров объединенного уравнения теперь необходимо оценивать $(2m + 2)$ параметров двух регрессий. Следовательно, дробь $\frac{\left(S_1 + S_2\right)}{\left(n - 2m + 2\right)}$ представляет собой необъясненную дисперсию зависимой переменной при использовании двух регрессий. Отсюда можно сделать вывод о том, что общую выборку целесообразно разбить на две ее части только в том случае, если уменьшение дисперсии будет значимо больше оставшейся необъясненной дисперсии. Это означает, что нулевая гипотеза может быть записана как $H_0 : (S_0 - S_1 - S_2) = S_1 + S_2$ при альтернативе $H_1 : (S_0 - S_1 - S_2) \neq S_1 + S_2$.

Проверка нулевой гипотезы осуществляется по стандартной процедуре сравнения дисперсий с помощью критерия Фишера. При этом F-статистика имеет вид:

$$F = \frac{S_0 - S_1 - S_2}{S_1 + S_2} \times \frac{n - 2m - 2}{m + 1}. \qquad (7.30)$$

Далее осуществляется проверка неравенства $F > \boldsymbol{F}_{\text{кр}}(\alpha,\nu_1,\nu_2)$, где $\nu_1 = m + 1$, $\nu_2 = n - 2m - 2$, m – число переменных в модели. Если

при выбранном уровне значимости α данное неравенство выполняется, то нулевая гипотеза о равенстве дисперсий отвергается и делается вывод о целесообразности разбиения выборки на две подвыборки. В противном случае у нас есть основания полагать, что разбивать выборку на отдельные части не имеет смысла.

Некоторая неопределенность использования критерия Чоу заключается в том, что визуальный анализ не всегда позволяет однозначно определять границу между подвыборками с максимальной корреляцией между переменными внутри каждой из них. Очевидно, прежде чем применять данный критерий, необходимо использовать несколько различных вариантов разделения общей выборки на отдельные части и выбрать тот из них, для которого корреляция между переменными является максимальной.

7.8. Множественная линейная регрессия

Известно, что зависимости между характеристиками природной среды носят, как правило, многофакторный характер, т. е. когда рассматриваемая переменная зависит не от одной, а уже от многих других переменных. Естественно, что в этом случае построение парной линейной регрессии теряет смысл. В результате мы приходим к необходимости построения модели множественной линейной регрессии (МЛР), уравнение которой можно представить в следующем виде:

$$y_i = b_0 + \sum_{j=1}^{m} b_j x_{ij} + \varepsilon_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im} + \varepsilon_i, \quad (7.31)$$

где ε_i – вектор остатков (ошибок), которые не описываются уравнением регрессии, m – число независимых переменных. Нетрудно видеть, что МЛР представляет собой обобщение линейной регрессии двух переменных на многомерный случай. Однако, если парная регрессия имеет четкую геометрическую интерпретацию, то для МЛР сделать это практически невозможно, так как для многомерного пространства не существует аналогичной интерпретации. Например, если мы имеем две независимые переменные, то в этом случае решением уравнения (7.31) служит плоскость (сечение), проходящая в трехмерном (кубическом) пространстве таким образом, что суммарный квадрат разброса исходных точек относительно нее минимален.

Естественно, с увеличением размерности пространства представлять такую плоскость становится все сложнее. Поэтому (m + 1)-мерное пространство – это лишь удобный математический

прием, позволяющий экстраполировать свойства двухмерного пространства на многомерное. Отсюда следует, что *уравнение (7.31) можно интерпретировать как некую условную гиперплоскость в (m + 1)-мерном пространстве, которая обладает тем свойством, что сумма квадратов отклонений точек (y_i, x_{ij}, … , x_{im}) от нее меньше, чем до любой другой поверхности.*

В уравнении (7.31) Y – зависимая переменная (функция отклика, предиктант и т.п.), X_j – независимая переменная (фактор, предиктор и т.д.), b_j – коэффициент регрессии.

Основные предположения, накладываемые на регрессионную модель, состоят в следующем:

1) ошибки (остатки) модели МЛР должны иметь нулевое среднее значение $\left(\overline{\varepsilon} = 0\right)$;

2) дисперсия остатков должна быть постоянной ($\sigma^2_{\varepsilon} = \text{const}$), т. е. выполняется условие гомоскедастичности регрессионных остатков;

3) ошибки должны быть независимы (некоррелированы) по отношению к факторам и функции отклика;

4) исходные факторы x_1, x_2, … , x_m являются неслучайными переменными;

5) ранг матрицы исходных данных $\boldsymbol{X}$ должен быть максимальным, но при этом меньше n, т. е. $\text{rank}\boldsymbol{X} = (m + 1) < n$;

6) желательно, но не обязательно, нормальное распределение остатков.

Первые три предположения являются необходимыми условиями использования метода наименьших квадратов и уже были рассмотрены выше. В соответствии с четвертым предположением, если матрица $\boldsymbol{X}$ не случайна, то это означает, что ее элементами служат известные числа, точно задаваемые исследователем. Отсюда следует, что единственным источником случайных возмущений значений y_i являются случайные возмущения регрессионных остатков ε_i. Но поскольку по определению ε_i – случайная величина, то переменная y_i тоже является случайной величиной, причем ее закон распределения соответствует закону распределения ε_i.

Очевидным следствием данного предположения является то, что данная модель является совершенно точной только для конкретного числа переменных, входящих в модель. Если исследователь хочет распространить полученные выводы на более широкий класс факторов, непосредственно не участвующих в построении модели МЛР, то переменные x_1, … , x_m будут уже носить случайный

характер. Тем самым возникает неопределенность в статусе функции отклика y_i.

Обсуждая пятое условие, прежде всего, напомним, что *ранг матрицы может быть определен как наибольший порядок ее отличного от нуля минора, который совпадает с максимальным числом линейно независимых столбцов.* В свою очередь оно должно быть меньше числа строк в матрице, поскольку в противоположном случае становится невозможной оценка коэффициентов регрессионной модели с помощью метода наименьших квадратов. Итак, если требование к рангу матрицы $\boldsymbol{X}$ не выполняется, т. е. он не является максимальным, то возникает линейная зависимость хотя бы между двумя столбцами. Это означает существование функциональной линейной взаимосвязи между исходными факторами. В результате происходит вырождение матрицы $\boldsymbol{X'X}$ и, следовательно, ее детерминант (главный определитель) становится равным нулю, т. е. $\det(\boldsymbol{X'X}) = 0$, что приводит к возникновению проблемы *мультиколлинеарности*.

При выполнении первых пяти условий получаем классическую модель МЛР. Если дополнительно постулируется нормальный характер распределения регрессионных остатков, то имеем нормальную классическую модель МЛР. В том случае, когда наблюдается гетероскедастичность, т. е. дисперсия регрессионных остатков меняется во времени, получаем *обобщенную модель МЛР.*

7.9. Вычисление и оценивание параметров множественной линейной регрессии

Коэффициенты регрессии в модели МЛР определяются методом наименьших квадратов, в соответствии с которым требуется минимизировать сумму разности квадратов фактических и вычисленных по уравнению (7.31) значений функции отклика, т. е.

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \tilde{y}_i \right)^2 = \sum_{i=1}^{n} \left[y_i - \left(b_0 + b_1 x_{i1} + \ldots + b_m x_{im} \right) \right]^2 = \min,$$

где $\tilde{y}_i$ – вычисленные по уравнению МЛР значения функции отклика. Для отыскания минимума данного выражения необходимо найти частные производные по всем неизвестным коэффициентам и затем, приравняв их к нулю, получить систему линейных нормальных уравнений. В матричном виде она может быть записана как:

$$(\boldsymbol{X'X})\boldsymbol{B} = \boldsymbol{X'Y}, \quad (7.32)$$

где $\boldsymbol{X}$ – матрица исходных данных; $\boldsymbol{B}$ и $\boldsymbol{Y}$ – диагональные матрицы коэффициентов регрессии и функции отклика. Для решения этой системы, умножим (7.32) на матрицу, обратную ($\boldsymbol{X'X}$), т. е.

$$(\boldsymbol{X'X})^{-1}(\boldsymbol{X'X})\boldsymbol{B} = (\boldsymbol{X'X})^{-1}(\boldsymbol{X'Y}). \quad (7.33)$$

Но поскольку произведение в левой части выражения (7.33) представляет единичную матрицу $(\boldsymbol{X'X})^{-1}(\boldsymbol{X'X}) = \boldsymbol{I}$, то решение системы нормальных уравнений в матричной форме запишется следующим образом:

$$\boldsymbol{B} = (\boldsymbol{X'X})^{-1}(\boldsymbol{X'Y}). \quad (7.34)$$

Стандартизированное уравнение МЛР по аналогии с (7.12) примет вид:

$$z_y = \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_m z_m + e = \sum_j^m \beta_j z_j + e. \quad (7.35)$$

Здесь z_y, β_j, z_j – стандартизированные значения функции отклика, коэффициентов регрессии и предикторов соответственно, e – безразмерный вектор остатков. Нетрудно видеть, что свободный член в уравнении (7.35) равен нулю. Напомним, что *физический смысл стандартизированных коэффициентов регрессии состоит в том, что они показывают относительную роль каждого предиктора в описании изменчивости функции отклика.*

Приступая к оцениванию параметров модели МЛР (рис. 7.14), прежде всего, отметим, что оно осуществляется точно так же, как для уравнения линейной регрессии двух переменных. Различия связаны главным образом с оценкой числа степеней свободы. При этом основным требованием к исходным данным, как уже упоминалось выше, является выполнение многомерного нормального закона распределения. Но поскольку проверить такой закон на практике вряд ли возможно, то обычно постулируется более «мягкое» требование, состоящее в необходимости выполнения одномерного нормального закона распределения для каждой переменной.

Основные параметры модели МЛР

– *Множественный коэффициент линейной корреляции,* представляющий собой аналог обычного парного коэффициента корреляции. Он характеризует меру линейной связи между фактическими и вычисленными по уравнению МЛР значениями функции отклика, т. е.

$$R = \frac{1}{n\sigma_y \sigma_{y(x)}} \sum_{i=1}^{n} \left(y_i - \bar{y}\right)\left(\tilde{y}_i - \bar{y}\right), \qquad (7.36)$$

где $\tilde{y}_i$ – вычисленные по модели МЛР значения отклика, $\sigma_{y(x)}$ – стандартное отклонение значений $\tilde{y}_i$. Величина R изменяется в пределах $0 \le R \le 1$. При $R = 1$ имеем функциональную линейную модель, когда факторы полностью описывают дисперсию функции отклика, вследствие чего остатки равны нулю ($\varepsilon_i = 0$). При $R = 0$, напротив, изменчивость функции отклика полностью уже определяется остатками ε_i. Это означает, что все коэффициенты парной корреляции вектора столбца $\boldsymbol{X'Y}$, характеризующего меру связи переменной Y с факторами X_j, равны нулю.

Следует иметь в виду, что во многих ППСП одновременно с величиной R приводится также *скорректированный множественный коэффициент корреляции* $R_{ск}$. Дело в том, что, как будет показано ниже, величина R имеет положительное смещение, которое устраняется с помощью следующей формулы:

$$R_{CK} = \sqrt{1 - \frac{\sigma_\varepsilon^2 \left(n-1\right)}{\sigma_y^2 \left(n-m-1\right)}} = \sqrt{1 - \frac{\left(1-R^2\right)\left(n-1\right)}{n-m-1}}. \qquad (7.37)$$

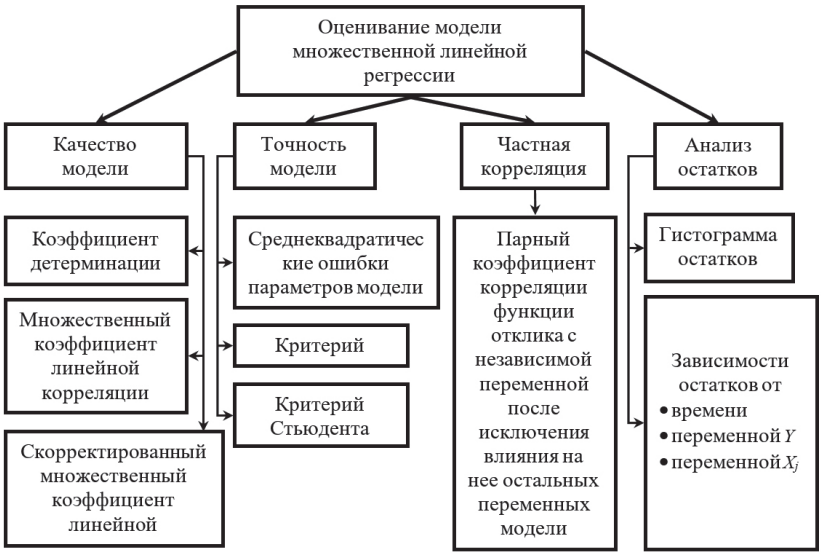

Рис. 7.14. Оценивание модели множественной линейной регрессии

Итак, *разность* $R - R_{ск}$ *– это поправка на положительное смещение величины R.* Из анализа формулы (7.37) видно, что для множественной регрессии ($m > 1$) $R_{ск} < R$ и только для случая одномерной регрессии ($m = 1$) парный коэффициент корреляции является уже несмещенной оценкой, т. е. $R_{ск} = R$. Важным свойством $R_{ск}$ является то, что он увеличивается при добавлении новой переменной тогда и только тогда, когда *t*-статистика этой переменной по модулю больше единицы. Поэтому добавление в модель новых переменных может осуществляться до тех пор, пока он растет.

Отметим, что несмотря на статистические достоинства, величина $R_{ск}$ не получила на практике широкого распространения. Это связано с тем, что при включении в модель новых предикторов величина R уменьшаться не может, в то время как для $R_{ск}$ такое уменьшение возможно, ибо с ростом m величина $(1 - R^2)$ обычно уменьшается медленнее, чем $n - m$. Кроме того, если разность $n - m$ мала, то коэффициент $R_{ск}$ может принимать даже отрицательные значения. Например, при $n = 20$, $m = 18$ и $R^2 = 0{,}5$ по формуле (7.37) получим $R^2_{ск} = -3{,}75$, т. е. $R_{ск}$ становится мнимой величиной, чего не может быть в действительности.

– *Линейный коэффициент детерминации,* представляющий собой квадрат множественного коэффициента линейной корреляции:

$$R^2 = \frac{\sigma^2_{y(x)}}{\sigma^2_y} = 1 - \left(\frac{\sigma^2_\varepsilon}{\sigma^2_x} \right). \qquad (7.38)$$

Отсюда следует, что коэффициент детерминации показывает долю объясненной дисперсии функции отклика. Он функционально связан со стандартизированными коэффициентами регрессии формулой:

$$R^2 = \sum \beta_j r_{yj} = \beta_1 r_{y1} + \beta_2 r_{y2} + \ldots + \beta_m r_{ym}, \qquad (7.39)$$

где r_{yj} – парный коэффициент корреляции между предиктантом и j-м предиктором. Отсюда следует, что произведение $\beta_j r_{yj}$ представляет собой вклад каждого из предикторов X_j в описание изменчивости функции отклика. При этом влияние факторов X_j на изменчивость Y зависит не только от коэффициента корреляции между ними, но и от величины стандартизированного коэффициента регрессии.

Кроме того, можно отметить еще одно важное свойство: при включении в состав предикторов дополнительной $m + 1$ переменной величина R^2 возрастает или, в крайнем случае, остается на том же уровне, т. е. $R^2_{m+1} \geq R^2_m$. Эти величины равны только в том случае,

когда дисперсия новой $m + 1$ переменной полностью описывается набором из m предикторов и, следовательно, ее вклад в описание дисперсии функции отклика будет равен нулю.

Заметим, что математическое ожидание R^2 при $b_1 = b_2 = \ldots = b_m = 0$, т. е. когда отклик полностью описывается остатками ($\sigma^2_y = \sigma^2_\varepsilon$) и, следовательно, величина R^2 тоже должна быть равна нулю, определяется по следующей формуле:

$$M\left(R^2\right) = \frac{m}{n-1}.$$

Из этой формулы видно, что чем меньше разность $n - m$, тем больше величина R^2 отличается от нуля. По существу это означает, что коэффициент множественной корреляции имеет положительное смещение. Для одномерной регрессионной модели ($m = 1$) положительное смещение практически отсутствует.

– *Среднеквадратическое (стандартное) отклонение модели:*

$$\sigma_{y(x)} = \sqrt{\sum \frac{\left(y_i - \tilde{y}_i\right)^2}{\left(n-m-1\right)}}. \tag{7.40}$$

Можно показать, что данная величина функционально связана с линейным коэффициентом детерминации формулой:

$$\sigma_{y(x)} = \sigma_y \sqrt{1-R^2}. \tag{7.41}$$

– *Стандартные ошибки множественного коэффициента корреляции и коэффициентов регрессии:*

$$\sigma_R = \frac{1-R^2}{\sqrt{n-m-1}}, \tag{7.42}$$

$$\sigma_{bj} = \frac{\sigma_y}{\sigma_{xj}} \sqrt{\frac{\left(1-R^2\right)D_{yj}}{\left(n-m-1\right)D_{yy}}}. \tag{7.43}$$

где σ_{xj} – стандартное отклонение x_j предиктора, D_{yj} – минор главного определителя (детерминанта), у которого вычеркнута первая строка (y) и j-тый столбец, а D_{yy} – минор, у которого вычеркнута первая строка и первый столбец.

Строго говоря, использование формулы (7.42) правомерно только при условии, что выборочные значения R подчиняются нормальному закону, т. е. при сравнительно малых значениях R и

большой длине исходных рядов n. При больших значениях R и малых значениях n следует применять *z-преобразование Фишера*. Из формул (7.40)–(7.43) вытекает одно важное следствие. С увеличением длины рядов и уменьшением их числа точность модели МЛР повышается. Поэтому в практических расчетах необходимо соблюдать условие $n >> m$.

При проверке параметров R и b_j на *значимость*, т. е. насколько значимо (существенно) они отличаются от нуля, вначале формулируется нулевая гипотеза вида $H_0 : R = 0$ и $H_0 : |b_j| = 0$. Проверка этой гипотезы осуществляется также с помощью t-критерия:

$$R > t_\alpha \sigma_R,$$

$$|b_j| > t_\alpha \sigma_{bj}.$$

Если данные условия выполняются, то нулевая гипотеза отвергается как несостоятельная и выборочные оценки R и $|b_j|$ считаются значимыми, т. е. отклоняющимися от нуля неслучайным образом. В большинстве ППСП процедура проверки значений $|b_j|$ на значимость реализуется через *p-критерий* (*p-level*). Заметим, что проверка на значимость коэффициента множественной корреляции эквивалентна проверке на значимость всех коэффициентов регрессии, кроме свободного члена, т. е. $H_0 : |b_1| = |b_2| = \ldots = |b_m| = 0$. Если R значим, то хотя бы один из коэффициентов регрессии тоже является значимым.

– *Критерий Фишера,* используемый для оценки адекватности (значимости) всей модели МЛР. С этой целью проверяется нулевая гипотеза вида $H_0 : \sigma^2_{y(x)} = \sigma^2_\varepsilon$, т. е. дисперсия вычисленных по уравнению МЛР значений функции отклика равна дисперсии остатков. Нулевая гипотеза проверяется с помощью критерия Фишера, который по аналогии с моделью парной регрессии может быть представлен как:

$$F = \frac{Q_R\left(n-m-1\right)}{Q_\varepsilon\left(m\right)} = \frac{\sigma^2_{y(x)}}{\sigma^2_\varepsilon}. \qquad (7.44)$$

Вычисленное значение критерия Фишера сравнивается с его табличным (критическим) значением $\boldsymbol{F}_{кр}(\alpha, \nu_1, \nu_2)$ при заданном уровне значимости α и степенях свободы $\nu_1 = m$, $\nu_2 = n - m - 1$. Если выполняется неравенство $F > \boldsymbol{F}_{кр}$, то нулевая гипотеза о равенстве дисперсий вычисленных значений функции отклика и остатков отвергается и делается вывод, что дисперсия, описываемая моделью МЛР, неслучайным (существенным) образом отличается от

дисперсии ошибок. Это означает, что рассматриваемая модель является адекватной (значимой) и она хорошо соответствует исходным данным функции отклика. Если же $F < \boldsymbol{F}_{кр}$, то у нас есть основания полагать, что модель неадекватна, т. е. она плохо описывает исходные данные.

Следует иметь в виду, что вывод об адекватности модели полностью справедлив лишь в том случае, если все *коэффициенты регрессии значимы по критерию Стьюдента*. Такая модель может быть использована для принятия решений и осуществления прогнозов. Если же хотя бы один коэффициент модели не значим, то это свидетельствует о неполной (частичной) значимости модели. Использование такой модели в прогнозах нежелательно. Кроме того, оценка адекватности (значимости) модели МЛР при разработке методов прогноза гидрометеорологических характеристик по неравенству $F > \boldsymbol{F}_{кр}$ представляется «мягкой». В этом случае требуется гораздо более жесткое условие значимости модели МЛР, которое имеет вид $F > 4\,\boldsymbol{F}_{кр}$.

Заметим также, что критерий Фишера функционально связан с коэффициентом детерминации следующей формулой:

$$F = \frac{R^2\left(n-m-1\right)}{\left(1-R^2\right)m}. \qquad (7.45)$$

Отсюда следует, что критерий Фишера может использоваться для проверки нулевой гипотезы о значимости R^2 нулю ($H_0 : R^2 = 0$), которая полностью тождественна гипотезе оценки адекватности модели МЛР.

– *Частный коэффициент корреляции* ρ, представляющий собой аналог обычного парного коэффициента корреляции, показывает меру линейной связи функции отклика с какой-либо независимой переменной после исключения влияния на нее всех оставшихся $m - 1$ переменных. Очевидно, такое влияние можно интерпретировать как перенесение эффекта ложной (автоматической) корреляции на многомерный случай. Напомним, что если две случайных переменных X_1 и X_2 не содержат в себе информации о какой-либо третьей переменной, то такая корреляция называется истинной. В противном случае возникает эффект ложной корреляции, который тем выше, чем больше линейная связь с третьей переменной.

Оценка частных коэффициентов корреляции ρ производится через определители корреляционной матрицы. Чтобы понять суть

определения ρ, рассмотрим простейшую модель МЛР для двух переменных:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \varepsilon_i .$$

Вначале исключим из модели переменную X_1 и рассчитаем уравнение парной регрессии:

$$y'_i = b'_0 + b'_2 x_{i2} + \varepsilon'_i .$$

Затем рассчитаем уравнение парной регрессии между X_1 и X_2:

$$x_{i1} = b''_0 + b''_2 x_{i2} + e_i .$$

Остатки ε'_i и e_i по существу составляют ту часть функции отклика и переменной X_1, которые полностью независимы от влияния на них переменной X_2. Следовательно, парный коэффициент корреляции между остатками соответствует частному коэффициенту корреляции:

$$\rho_{yx1} = \frac{1}{n\sigma_\varepsilon \sigma_e} \sum \left(\varepsilon'_i - \overline{\varepsilon}\right)\left(e_i - \overline{e}\right). \qquad (7.46)$$

Аналогичным образом может быть рассчитана величина ρ между Y и X_2. Понятно, если мы имеем набор из m предикторов, то получаем набор из m частных коэффициентов корреляции.

Отсюда следует, что квадрат частного коэффициента корреляции, называемый *частным коэффициентом детерминации,* может быть интерпретирован как доля остаточной дисперсии функции отклика, которая объясняется включением дополнительной переменной в модель МЛР. Очевидно, что частные коэффициенты детерминации должны быть функционально связаны с полным коэффициентом детерминации R^2. Эта связь выражается следующей формулой:

$$R^2 = 1 - \left(1 - \rho^2_{yx1}\right)\left(1 - \rho^2_{yx2}\right)\left(1 - \rho^2_{yx3}\right)\ldots\left(1 - \rho^2_{yxm}\right). \qquad (7.47)$$

Заметим, что значения ρ во многих ППСП (Statistica, SPSS и др.) используются в пошаговых алгоритмах для ранжирования предикторов по их вкладу в описание изменчивости предиктанта.

– *Анализ остатков модели МЛР.* Полное их описание дано в разделе 7.5.

7.10. Проблема мультиколлинеарности и структурные противоречия модели множественной линейной регрессии

Рассмотрим вначале проблему мультиколлинеарности. В статистике различают строгую (полную) и реальную (частичную) мультиколлинеарность. *Строгая мультиколлинеарность* заключается в том, что ранг матрицы исходных переменных меньше $m + 1$, т. е. $\text{rank}\boldsymbol{X} < (m + 1)$. Как уже указывалось выше, это означает, что хотя бы одна переменная матрицы $\boldsymbol{X}$ должна быть выражена линейной функциональной связью через остальные переменные. Вследствие этого матрица $\boldsymbol{X'X}$ оказывается вырожденной, т. е. ее определитель равен нулю. Так как в данном случае не существует обратной матрицы $(\boldsymbol{X'X})^{-1}$, то определение коэффициентов регрессии становится невозможным.

Поскольку строгая мультиколлинеарность на практике встречается весьма редко и может быть довольно легко исключена, то целесообразно оценивать только *реальную мультиколлинеарность.* Суть ее сводится к тому, что если в исходной матрице между большинством исходных факторов отмечается высокая коррелированность, то система нормальных линейных уравнений становится плохо обусловленной, вырождающейся. В результате ее детерминант (главный определитель) стремится (но не равен!) к нулю. Вследствие этого коэффициенты регрессии становятся неустойчивыми, причем ошибки их определения могут уже существенно превышать сами значения коэффициентов.

К сожалению, точных количественных критериев оценки реальной мультиколлинеарности не существует. Элементарный прием проверки мультиколлинеарности – это визуальный анализ корреляционной матрицы исходных переменных. Если между некоторыми переменными отмечается высокая корреляция ($|r_{ij}| \geq 0.8 \div 0.9$), то один их дублирующих факторов (X_i или X_j) может быть исключен. В этом случае объем независимой информации, содержащейся в исходной матрице $\boldsymbol{X}$, уменьшится незначительно, но зато улучшится обусловленность системы нормальных уравнений и повысится точность параметров регрессионной модели. Исключение дублирующих аргументов может осуществляться на основе физических соображений или с помощью формальных критериев. Например, для этого можно использовать оценки доли вклада переменных

в описание функции отклика (Δ_{xj}). Если выполняется условие $\Delta_{xj} < \frac{2\sigma_R}{R}$, то переменную X_j следует исключить из модели.

Естественно, имеются более точные способы выявления и устранения эффекта мультиколлинеарности. В частности, известен целый ряд численных критериев (VIF-показатель, критерий толерантности, число обусловленности, критерий Феррара–Глоубера, методы пошаговой и гребневой регрессии и др.), однако ни один из них не является универсальным. Радикальное устранение эффекта мультиколлинеарности возможно при ортогонализации переменных, т. е. в результате приведения их к взаимной независимости. Это достигается, например, с помощью метода главных компонент. Причем чем сильнее мультиколлинеарность, тем наблюдается более быстрая сходимость собственных чисел. В результате этого появляется возможность путем отбрасывания последних компонент, дающих малый вклад в дисперсию исходного поля, построить регрессионную модель на главных компонентах существенно меньшей размерности по сравнению с набором из m предикторов.

Обсудим теперь структурные противоречия модели МЛР. Предположим, что мы имеем выборку размером $m \times n$, где m – число предикторов, n – их длина. С помощью пошаговой процедуры методом включения переменных можно построить m моделей, каждая их которых имеет на один предиктор больше по сравнению с предыдущей моделью. С одной стороны, с включением новой переменной k должно выполняться следующее соотношение: $R^2_{k+1} \geq R^2_k$, где k – число варьируемых предикторов в модели ($k \leq m$). С другой стороны, при включении новой переменной в модель происходит ухудшение точности всех ее параметров, связанных с тем, что в знаменателе соответствующих формул находится выражение $n - k - 1$.

Таким образом, имеем очевидное противоречие: при неизменном объеме выборки с включением в модель новой переменной повышается качество описания функции отклика (R^2), но при этом ухудшается точность всех параметров модели. Особенно значительными ошибки параметров модели становятся, когда разность $n - m$ является малой.

Данное противоречие отмечается даже для идеальной модели МЛР, которая предполагает отсутствие статистической взаимосвязи между всеми предикторами. Однако в действительности гидрометеорологические переменные зачастую скоррелированы

друг с другом. Поэтому при включении в набор нового предиктора может оказаться, что его дисперсия будет полностью описана уже имеющимся набором из k переменных. В результате частный коэффициент корреляции нового предиктора с предиктантом будет равен нулю и, как следствие, $R_{k+1} = R_k$. При этом повышается уровень мультиколлинеарности модели и, следовательно, ухудшается ее точность. Поэтому добавление новой переменной может даже усилить отмеченное выше противоречие. Итак, из сказанного вытекает очевидный вывод о необходимости построения оптимальных в смысле критериев точности регрессионных моделей и детального оценивания параметров моделей на всех ее этапах.

7.11. Пошаговые методы построения оптимальной модели МЛР

Прежде всего, отметим, что перед построением эффективной модели необходим предварительный этап, который сводится к тому, чтобы набор исходных предикторов отвечал определенным требованиям. К ним относятся:

– нормальность;

– стационарность;

– длина выборки должна существенно превосходить число предикторов;

– линейность связей между функцией отклика и предикторами;

– вариабельность факторов;

– погрешности функции отклика и факторов должны быть одного порядка;

– независимость (некоррелированность) факторов между собой.

Многие из перечисленных требований являются очевидными. В частности, при формулировании модели МЛР требование нормального распределения исходных данных в явном виде не постулируется, однако оно неизбежно вытекает из самой сущности регрессионного анализа. Действительно, как известно, коэффициент корреляции является параметрическим коэффициентом связи, параметром двухмерного нормального закона распределения. Кроме того, хотя при определении коэффициентов регрессии методом наименьших квадратов формально многомерное нормальное распределение данных не требуется, но только в этом случае МНК обеспечивает получение несмещенных, асимптотически состоятельных и обладающих минимальной дисперсией (эффективных) оценок, совпадающих

с методом максимального правдоподобия. Наконец, проверка статистических гипотез для параметров модели с помощью разных критериев (например, критерии Стьюдента, Фишера) осуществляется в предположении нормальности проверяемых параметров и, следовательно, тех данных, которые использованы для их вычисления.

Под вариабельностью факторов понимается их изменчивость. Если изменчивость какого-либо фактора существенно меньше изменчивости других факторов, в то время как физическая связь его с функцией отклика не вызывает сомнений, данный фактор может оказаться незначимым в модели МЛР. Это означает, что факторы должны иметь изменчивость, сравнимую с функцией отклика.

Отметим также, что не все из указанных выше требований к исходным данным являются одинаково важными, причем в зависимости от характера поставленной задачи их приоритет может быть существенно различным. Например, если нас интересует только модель какого-либо процесса, то в этом случае условие стационарности исходных данных не представляется принципиальным. Напротив, если модель МЛР используется для прогноза, то стационарность приобретает исключительно важное значение. Действительно, добиваясь высокой точности описания предиктанта на зависимой выборке, при переходе к независимым данным можно получить результаты, далекие от первоначальной точности, если исходные данные существенно нестационарны по среднему значению и дисперсии. Очень важным, особенно при большом числе предикторов, является требование их независимости, напрямую связанное с проблемой мультиколлинеарности.

Соответствие исходных данных указанным требованиям является условием построения оптимальных моделей МЛР. В общем случае построение такой модели можно рассматривать как задачу выбора некоторой системы эффективных предикторов, обеспечивающих максимальную точность модели МЛР с минимально возможными погрешностями ее параметров. Следует иметь в виду, что каким бы способом не проводился отбор эффективных (существенных) предикторов, обусловленность матрицы $\boldsymbol{X'X}$ при этом улучшается с уменьшением числа переменных, включаемых в модель. Впрочем, процедура отбора наиболее существенных переменных имеет самостоятельное значение и может рассматриваться как процесс выбора размерности линейной модели. В настоящее время наиболее эффективным методом решения данной задачи, особенно при большом числе предикторов, считаются пошаговые процедуры,

которые в широком смысле включают в себя несколько различных алгоритмов, причем во многих пакетах прикладных статистических программ реализованы методы:

– включения переменных;

– исключения переменных.

Суть *метода включения переменных* заключается в том, что на первом шаге выбирается наиболее коррелированный с функцией отклика предиктор, и рассчитываются все параметры модели парной регрессии, т. е. $Y = f(X_1)$. После этого вычисляются, например, частные коэффициенты корреляции для оставшихся $m - 1$ предикторов, которые показывают «чистый» вклад каждой переменной в дисперсию функции отклика. Таким образом, выбирается вторая переменная, имеющая максимальный частный коэффициент корреляции и строится новая модель $Y = f(X_1, X_2)$. Данная процедура может повторяться до тех пор, пока не будут построены все m моделей, т. е. $Y = f(X_1, X_2, \ldots, X_m)$.

Наиболее принципиальным моментом данной процедуры является выбор наилучшей или, другими словами, оптимальной в некотором смысле модели. В ППСП (Statistica, Statgraphics и др.) этот вопрос решается с помощью *частного F-критерия,* который представляет собой обычный F-критерий для каждой переменной при условии, что она оказывается последней переменной, включенной в модель регрессии. Частный F-критерий связан с коэффициентом частной корреляции следующим соотношением:

$$F_k = \frac{\rho^2_{yxj}\left(n-k-2\right)}{1-\rho^2_{yxj}}. \qquad (7.48)$$

Здесь k – число переменных, уже включенных в модель ($k \leq m$) с учетом последней переменной X_j, для которой и рассчитывается частный коэффициент корреляции ρ_{yxj}. На каждом шаге выполняется проверка адекватности (значимости) модели и сравнение с некоторым пороговым (критическим) значением $\boldsymbol{F}_{кр}$. Величина $\boldsymbol{F}_{кр}$ может быть задана самим исследователем. По умолчанию она обычно принимается в ППСП $\boldsymbol{F}_{кр} = 4{,}0$. Как только величина F_k становится меньше $\boldsymbol{F}_{кр}$, программа прекращает работу и последний шаг принимается за оптимальную модель регрессии. Заметим, однако, что при этом не все коэффициенты регрессии могут быть значимыми.

Метод исключения переменных реализует обратную процедуру. Вначале строится полная (из m переменных) модель МЛР. Затем

из нее исключается наименее значимый фактор, определяемый по минимальному коэффициенту частной корреляции. После этого из модели исключается следующий по значимости фактор. Так может продолжаться до тех пор, пока не останется самый значимый фактор. Выбор оптимальной модели также осуществляется по частному F-критерию, который на каждом шаге сравнивается с $\boldsymbol{F}_{кр}$. При выполнении условия $F_k > \boldsymbol{F}_{кр}$ полученное уравнение МЛР считается оптимальным. По сравнению с методом включения в данной процедуре в некоторых ППСП по умолчанию принимается чуть меньшее значение $\boldsymbol{F}_{кр}$ ($\boldsymbol{F}_{кр} = 3{,}9$).

Заметим, что если сравнивать результаты расчетов по обоим методам, то даже для одного и того же сравнительно большого набора переменных могут быть получены различные промежуточные регрессии. Это связано, прежде всего, с характером взаимных корреляционных связей между предикторами, а также частично при большой величине m с формальными (вычислительными) аспектами. Очевидно, при большом числе переменных в модели предпочтения заслуживает все же первый алгоритм. В этом случае нет необходимости строить полную модель МЛР, которая при большой величине m может быть очень сложной. Кроме того, этот подход значительно лучше соответствует общему принципу познания окружающего мира, а именно развитию от «простого к сложному».

Достоинство пошаговых процедур состоит в простоте алгоритмов, высокой скорости расчета на ЭВМ и возможности построения оптимального уравнения из очень большого числа потенциальных предикторов. Очевидный недостаток – раздельный анализ переменных. Возможны случаи, когда по отдельности переменные не являются значимыми, однако при совместном включении в модель они адекватно описывают функцию отклика. Этого недостатка можно избежать при использовании метода включения всех переменных, который реализует процедуру перебора всех возможных комбинаций между переменными. Однако при значительном числе m возникают трудности вычислительного характера даже для современных ЭВМ.

При использовании пошаговых процедур есть «тонкие» моменты. Прежде всего, это определенный волюнтаризм в выборе оптимальной модели. В современных пакетах, как правило, отсутствует математическое описание используемых алгоритмов. В результате приходится лишь предполагать, как работает та или иная процедура и какие критерии задействованы в расчетах. Например, можно лишь предположительно говорить о том, что при автоматическом отборе

оптимальной модели в ППСП используется частный критерий Фишера. Или, каким образом, осуществляется ранжирование факторов по их вкладу в дисперсию функции отклика. Весьма субъективным является и выбор пороговых (критических) значений различных статистик (прежде всего, Фишера и Стьюдента). Наконец, весьма опасно проводить отбрасывание незначимых регрессионных коэффициентов при построении оптимальной модели МЛР в случае, когда обратная матрица $(\boldsymbol{X'X})^{-1}$ не является ортогональной, т. е. при коррелированности факторов. Это приводит к смещенности значений функции отклика.

Однако главная проблема заключается все же в том, что нет единого объективного критерия для выбора наилучшей модели. Совершенно очевидно, что *нахождение оптимальной модели МЛР – задача неформальная. И чем более сложной является исходная модель, тем большее неформальное участие исследователя требуется для оценки ее оптимального вида.* Поэтому можно лишь предложить общую схему оценки оптимальности модели. Вначале целесообразно рассчитать полный комплекс (от 1 до m) моделей. Возможно даже с помощью разных пошаговых алгоритмов и, различным образом ранжируя предикторы. После этого необходим детальный анализ основных параметров моделей (коэффициент детерминации, стандартная ошибка модели, критерий Фишера, *p-level* коэффициентов регрессии). *Только комплексный анализ полученных моделей может позволить надежно определить оптимальный вид окончательной модели.* Но, к сожалению, следует отметить, что это удается не во всех случаях. Очень полезным представляется графический анализ параметров модели, когда ось абсцисс соответствует шагам модели, а по оси ординат откладываются указанные выше параметры модели.

При сравнении разных вариантов моделей нужно дополнительно принимать во внимание и неформальные критерии: стоимость информации, ее доступность, репрезентативность, оперативность получения и т.п. Если, например, при прогнозе среднегодовой температуры воды на основе ее среднемесячных значений на i-м шаге в модель будут включены значения температуры за декабрь или ноябрь, то какой бы сверхоптимальной эта модель не оказалась, она с физической точки зрения не имеет смысла, ибо практически отсутствует заблаговременность прогноза. Поэтому в качестве оптимальной может быть принята модель только на $i - 1$ шаге.

Возможно и неформальное участие самого исследователя в процедуре пошаговой регрессии. Исследователь может с помощью

специального входного параметра, называемого *уровнем принудительного включения,* задавать для каждой переменной либо инструкцию о способах ее включения, либо приоритет включения в зависимости от других переменных. Это дает возможность управлять отбором переменных и первыми включать в модель МЛР те предикторы, которые представляются исследователю наиболее важными.

Кроме того, важнейшим неформальным критерием является также и степень сложности модели. *Следует помнить, что чем проще модель, тем она надежнее.* Поэтому в тех случаях, когда приходится выбирать из нескольких моделей, нужно всегда предпочитать более простую модель. В частности, иногда за счет потери точности обучающей модели МЛР можно получить более работоспособную и точную модель при переходе к независимым данным, что особенно важно с точки зрения прогнозирования процессов.

Пример 7.4. Рассмотрим тестовый пример. Как известно, гидрометеорологические процессы и явления образуют единый комплекс, характеристики которого связаны между собой множеством связей, причем степень их тесноты обычно повышается с увеличением периода осреднения. Воспользуемся набором гидрометеорологических характеристик для области теплого Норвежского течения, распространяющегося от Фареро-Шетландского пролива вдоль побережья Скандинавии до мыса Нордкап. Этот набор включает средние годовые значения за период 1949−2001 гг. следующих характеристик: температура поверхности океана (T_w), температура приводного слоя атмосферы (T_a), зональная составляющая скорости ветра (U), меридиональная составляющая скорости ветра (V), атмосферное давление (P), осадки ($Prec$), облачность ($Cloud$), радиационный баланс (R) и тепловой баланс (TB).

Возьмем в качестве зависимой переменной величину T_w и попробуем подобрать к ней оптимальную модель МЛР методом включения переменных по матрице исходных факторов размером $m \times n$ ($m = 8$, $n = 53$). С этой целью рассчитаем последовательно 8 моделей и для каждой из них выделим основные параметры:

$$y_i^{(1)} = b_0 + b_1 x_{i1} \qquad \left(R^2_{(1)}, \sigma^{(1)}_{y(x)}, F^{(1)}, p - level^{(1)}_{\max}\right)$$

$$y_i^{(2)} = b_0 + b_1 x_{i1} + b_2 x_{i2} \qquad \left(R^2_{(2)}, \sigma^{(2)}_{y(x)}, F^{(2)}, p - level^{(2)}_{\max}\right)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots \qquad \ldots\ldots\ldots\ldots\ldots$$

$$y_i^{(8)} = b_0 + b_1 x_{i1} + \ldots + b_6 x_{i6} \qquad \left(R^2_{(8)}, \sigma^{(8)}_{y(x)}, F^{(8)}, p - level^{(8)}_{\max}\right).$$

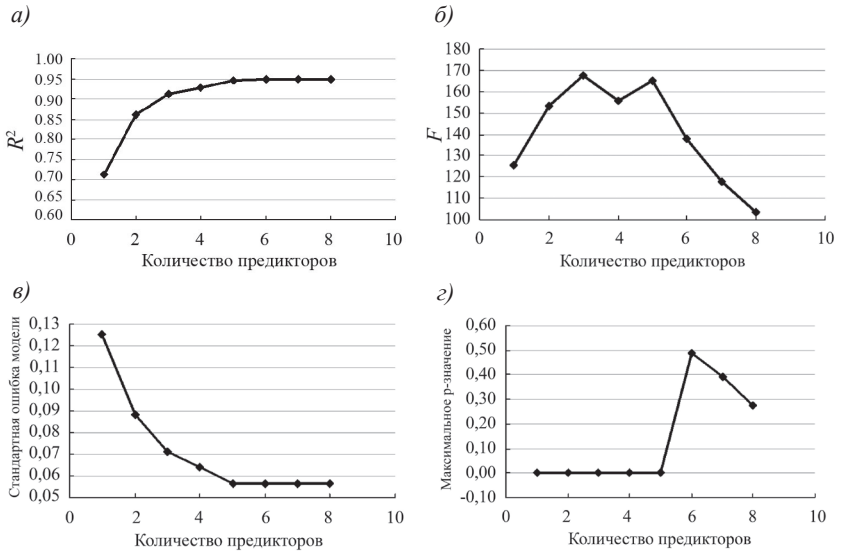

Рис. 7.15. Зависимость параметров регрессионной модели предвычисления температуры поверхности океана в области Норвежского течения от шага модели: *а)* $R^2 = f(m)$; *б)* $F = f(m)$; *в)* $\sigma_{y(x)} = f(m)$; *г)* $p\text{-}level_{(\max)} = f(m)$

Результаты оценивания этих моделей даны в таблице 7.7. Кроме того, графики параметров моделей в зависимости от шага (числа переменных) даны на рис. 7.15. Нетрудно видеть, что уже на первом шаге модель описывает более 70 % дисперсии исходного поля ТПО. До третьего шага коэффициент детерминации довольно заметно возрастает, после которого почти не меняется. Среднеквадратическая ошибка модели мала, начиная с первого сдвига, а критерий Фишера в несколько десятков раз превышает его критическую величину. Максимальное значение критерия Фишера отмечается на третьем шаге. По критерию *p-level* значимыми являются первые пять моделей. Выбор оптимальной модели в данном конкретном случае не представляет затруднений. Это модель на третьем шаге, которая имеет вид:

$$T_w = b_0 + b_1 T_a - b_2 TB - b_3 V + \varepsilon$$

или в стандартизированной форме:

$$z_{T_w} = 1{,}35 z_{T_a} - 0{,}54 z_{TB} - 0{,}28 z_V + \varepsilon.$$

Из последнего уравнения видно, что вклад температуры воздуха в описании изменчивости T_w более чем в два раза выше вклада *TB* и

Таблица 7.7

Статистические оценки параметров пошаговых моделей МЛР температуры поверхности океана в области Норвежского течения в зависимости от гидрометеорологических характеристик за 1949–2001 гг.

Шаг модели, параметр включения	$R_{(j)}^2$	$\sigma^{(j)}_{y(x)}$, °C	$F^{(j)}$,	$p\text{-}level^{(j)}_{\max}$
1-й, T_a	0,711	0,13	125,7	0,000
2-й, *TB*	0,860	0,09	153,3	0,000
3-й, *V*	0,911	0,07	167,3	0,000
4-й, *U*	0,928	0,06	155,8	0,001
5-й, *Prec*	0,946	0,06	165,1	0,000
6-й, *P*	0,947	0,06	137,8	0,487
7-й, *Cloud*	0,948	0,06	117,6	0,391
8-й, *R*	0,950	0,06	103,5	0,277

более чем в четыре раза выше вклада *V*. Очевидно, в приближенных расчетах вообще можно ограничиться только температурой воздуха, поскольку ошибка модели при этом составляет лишь $\sigma_{y(x)} = 0{,}13$ °C.

Пример 7.5. Рассмотрим задачу предвычисления средних годовых значений ТПО в районе судна погоды «М» по ее данным в отдельные месяцы за период 1951–2000 гг. Таким образом, матрица зависимых переменных имеет размер 12 × 50, где 12 – число месяцев, а 50 – количество лет наблюдений. Отметим, что данная задача будет носить прогностический характер, если мы по данным ТПО за несколько месяцев сможем с определенной заблаговременностью рассчитать среднюю годовую величину ТПО с достаточной для практических целей точностью. Следует также иметь в виду, что расчет полной модели ($m = 12$) не имеет смысла, так как в этом случае оценка средней годовой величины ТПО проще получить простым осреднением исходных данных. Итак, используя метод включения переменных, рассчитаем последовательно ряд моделей МЛР, основные параметры которых приведены в таблице 7.8.

Исходя из формальных соображений, за оптимальную модель МЛР мы должны принять 5-й шаг, т. е. модель с пятью предикторами. Действительно, начиная с первого шага по пятый включительно отмечается последовательное уменьшение среднеквадратической ошибки модели. На пятом шаге она равна 0,07 °C, т. е. становится сравнимой с ошибкой измерения. Все коэффициенты регрессии значимы по критерию Стьюдента. Критерий Фишера во много раз превышает его критическую величину ($\boldsymbol{F}_{\text{кр}} = 4{,}0$), а коэффициент детерминации близок к единице.

Таблица 7.8

Оценки основных параметров модели МЛР для средней годовой температуры поверхности океана в районе судна погоды «М»

Шаг модели	Месяц, включаемый в модель	$R_{(j)}^2$	$\sigma^{(j)}_{y(x)}$, °С	$F^{(j)}$	$p\text{-}level^{(j)}_{max}$
1-й	Август	0,616	0,24	77,0	0,0000
2-й	Март	0,823	0,16	109,1	0,0000
3-й	Июнь	0,904	0,12	143,7	0,0022
4-й	Октябрь	0,939	0,10	174,0	0,0060
5-й	Ноябрь	0,965	0,07	243,5	0,0043
6-й	Апрель	0,975	0,06	280,1	0,1101
7-й	Январь	0,984	0,05	370,2	0,2330

Однако модель МЛР с пятью предикторами не имеет практической значимости. Действительно, в этом случае в число предикторов входит ноябрь и, следовательно, заблаговременность предвычисления средней годовой ТПО уже практически отсутствует. Модель МЛР с четырьмя предикторами также имеет очень малую заблаговременность, поэтому принятие ее в качестве оптимальной тоже нецелесообразно. На наш взгляд, оптимальной является модель с тремя предикторами (август, март, июнь), которая обладает высокой точностью (R^2 = 0,90, а среднеквадратическая ошибка равна $\sigma_{y(x)}$ = 0,12 °С, т. е. всего на 0,05 °С превышает аналогичную ошибку на пятом шаге).

Глава 8. Анализ нелинейных зависимостей

8.1. Общая схема построения нелинейных зависимостей

Довольно часто связи между гидрометеорологическими процессами или явлениями могут носить нелинейный характер. В общем случае нелинейные модели делятся на два класса:

– нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам;

– нелинейные по оцениваемым параметрам.

К первому классу относятся уравнения, в которых функция отклика связана с параметрами линейно. Такими уравнениями, например, являются полиномиальные модели разных степеней и гиперболическая функция. Полиномиальная модель порядка m имеет вид:

$$y_i = a_0 + a_1 x_i + a_2 x_i^{\,2} + \ldots + a_m x_i^{\,m} + \varepsilon_i .$$

Отсюда видно, что функция отклика нелинейна по факторной переменной X и линейна по неизвестным коэффициентам модели. Поэтому для оценки коэффициентов данной модели может быть использован обычный метод наименьших квадратов.

Второй класс нелинейных моделей подразделяется на два типа:

– нелинейные модели внутренне линейные по параметрам;

– нелинейные модели внутренне нелинейные по параметрам.

Для первого типа моделей характерно то, что с помощью подходящих преобразований они могут быть приведены к линейному виду. Данная процедура называется линеаризацией. Коэффициенты *линеаризованных* моделей обычно определяются МНК. Для оценки параметров нелинейных моделей, которые не удается свести к линейному виду, используются, как правило, итеративные процедуры. К ним относятся квазиньютоновский метод, симплекс-метод, метод Хука-Дживса и др. Общая схема анализа в данном разделе нелинейных моделей представлена на рис. 8.1.

Одним из простейших способов построения нелинейной модели внутренне линейной по параметрам является подбор эмпирической формулы, аппроксимирующей связь между переменными по их известным значениям.

Отметим, что задача построения эмпирической формулы отлична от задачи интерполирования. При интерполировании, как правило, отыскивается такая функция, значения которой в заданных точках x_i совпадали бы с табличными значениями y_i. При нахождении эмпирической формулы этого обычно не требуется, вследствие чего осуществляется сглаживание исходных точек, приводящее к уменьшению дисперсии вычисленных по формуле значений y_i. Такой подход вполне правомерен, поскольку исходные эмпирические данные x_i и y_i, как правило, являются приближенными и содержат ошибки, величина которых обычно неизвестна. Поэтому построение эмпирической формулы, повторяющей эти ошибки, вряд ли целесообразно. Более того, правильный подбор сглаживающей эмпирической формулы может способствовать выявлению в исходных данных случайных погрешностей.

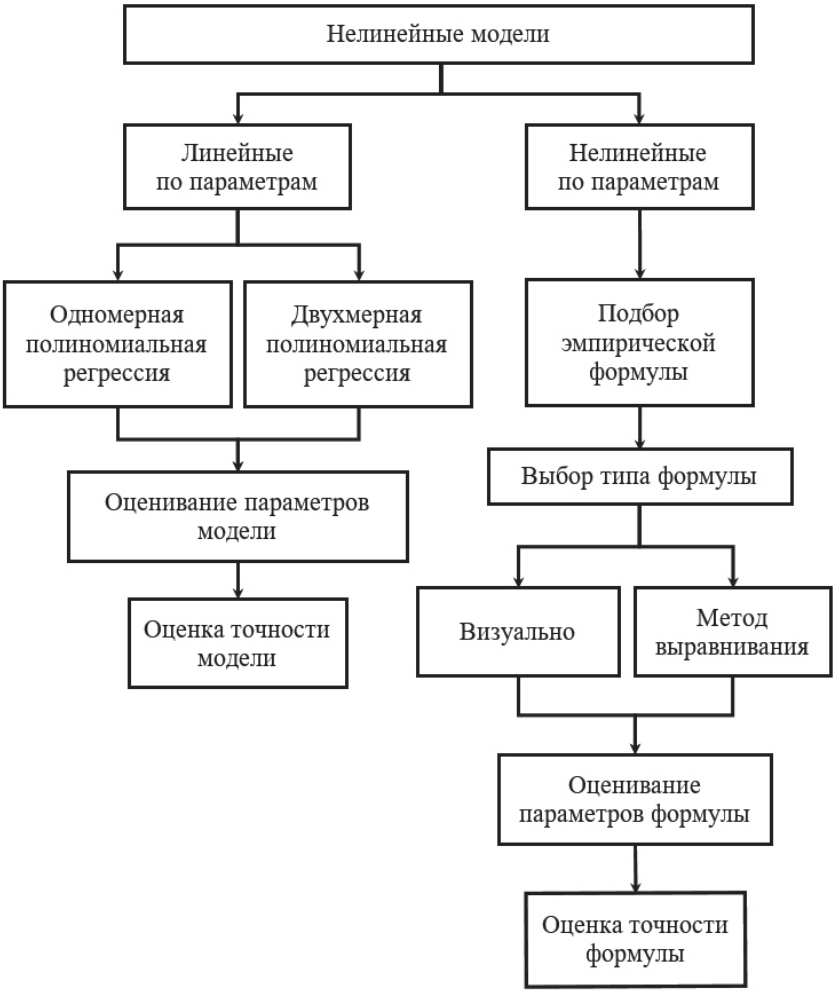

Рис. 8.1. Общая схема анализа нелинейных моделей

Необходимо отметить, что удачный подбор эмпирической формулы в значительной степени зависит от опыта и искусства исследователя.

«Эмпирические формулы не претендуют на роль законов природы, а являются лишь гипотезами, более или менее удовлетворительно согласующимися с наблюденными опытными данными. Однако значение их весьма велико; в истории науки известны

многочисленные примеры того, как получение удачной эмпирической формулы приводило к большим научным открытиям» (*Демидович Б.П., Марон И.А., Шувалова Э.З.* Численные методы анализа. М. 1967).

Здесь можно сослаться, например, на эмпирическую формулу Магнуса, связывающую насыщающее давление водяного пара с температурой:

$$E_0 = 6{,}1 \times 10^{\frac{7{,}45t}{235+t}}.$$

Эта формула, полученная много десятилетий назад, не потеряла своего значения в настоящее время и широко используется в численных расчетах.

Решение задачи построения эмпирической формулы можно разделить на 3 этапа: I этап – выбор типа формулы; II этап – определение параметров выбранной формулы; III этап – оценка достоверности полученной формулы.

Рассмотрим вкратце каждый из этапов.

Выбор типа формулы. Если нет каких-либо априорных теоретических соображений о выборе типа формулы, то строится график связи переменных в декартовой системе координат. Сравнение расположения точек на графике с различными кривыми, уравнения которых известны, может дать указание на тип формулы.

В простейшем случае, когда точки на графике близки к линейной зависимости, искомая формула обычно принимается в виде уравнения регрессии $y = a_0 + a_1 x$, в котором определению подлежат два параметра: a_0 и a_1.

При нелинейной связи переменных визуальное определение типа формулы не всегда оказывается возможным. Поэтому предварительно следует оценить достоверность выбранной формулы с помощью метода выравнивания.

Метод выравнивания состоит в том, что находятся некоторые величины $x' = \varphi(x)$, $y' = \psi(y)$, *которые должны быть связаны между собой линейной зависимостью.* Вычислив для заданных значений x_i и y_i соответствующие им новые значения x_i' и y_i' и нанеся их на график связи, нетрудно увидеть, насколько близка зависимость между x_i' и y_i' к линейному виду. Если все точки приблизительно ложатся на одну линию, то тип формулы выбран правильно.

Указания относительно выравнивания некоторых простейших формул с двумя параметрами приводятся в таблице 8.1. Эти формулы описывают довольно широкий класс нелинейных зависимостей.

Таблица 8.1

Математические функции и их линеаризация

Вид функции	Новые переменные		Уравнение в линейной форме
	y'	x'	
$y = a_0 + \frac{a_1}{x}$	y	$1/x$	$y = a_0 + a_1 x^{-1}$
$y = \frac{1}{(a_0 + a_1 x)}$	$1/y$	x	$y^{-1} = a_0 + a_1 x$
$y = \frac{x}{(a_0 + a_1 x)}$	x/y	x	$\frac{x}{y} = a_0 + a_1 x$
$y = a_0 a_1^x$	$\lg y$	x	$\lg y = \lg a_0 + x \lg a_1$
$y = a_0 e^{a_1 x}$	$\ln y$	x	$\ln y = \ln a_0 + a_1 x$
$y = \frac{1}{(a_0 + a_1 e^{-x})}$	$1/y$	e^{-x}	$y^{-1} = a_0 + a_1 e^{-x}$
$y = a_0 x^{a_1}$	$\lg y$	$\lg x$	$\lg y = \lg a_0 + a_1 \lg x$
$y = a_0 + a_1 \lg x$	y	$\lg x$	$y = a_0 + a_1 \lg x$
$y = \frac{a_0}{(a_1 + x)}$	$1/y$	x	$y^{-1} = \frac{a_1}{a_0} + \frac{1}{a_0} x$
$y = \frac{a_0 x}{(a_1 + x)}$	$1/y$	$1/x$	$y^{-1} = \frac{a_1}{a_0} + \frac{1}{a_0} x^{-1}$
$y = a_0 e^{\frac{a_1}{x}}$	$\ln y$	$1/x$	$\ln y = \ln a_0 + a_1 x^{-1}$
$y = a_0 + a_1 x^n$	y	x^n	$y = a_0 + a_1 x^n$

Значительно более сложным является вариант, когда необходимо подобрать три параметра эмпирической формулы, т. е. если точность описания исходных данных эмпирической формулой с двумя параметрами оказывается недостаточной.

В этом случае обычно вначале приближенно определяется один из неизвестных параметров выбранной формулы. Тогда для оставшихся двух параметров используется метод выравнивания, и таким

образом оценивается возможность применения этой формулы для описания исходных данных. Однако следует помнить, что при выборе типа формулы преимуществом при прочих равных условиях обладают те из них, которые имеют малое число неизвестных параметров. Большое число параметров, с одной стороны, затрудняет их определение, а с другой – затрудняет пользование формулой при выполнении расчетов. В общем случае выбор типа формулы облегчается знакомством с графиками элементарных функций, которые приводятся в различных справочниках по математике.

Определение параметров выбранной формулы. Если определению подлежат два неизвестных параметра эмпирической формулы, которая с помощью выравнивания может быть приведена к линейной зависимости, то используется метод наименьших квадратов. Этот метод, суть которого изложена в разделе 6, обеспечивает наиболее надежную оценку параметров.

Вначале составляется система нормальных уравнений и определяются параметры a_0' и a_1', а затем, используя их функциональную связь с параметрами выбранной формулы, находят a_0 и a_1.

Например, если выбранная формула имеет вид $y = a_0 x^{a_1}$, то после логарифмирования получим:

$$\lg y = \lg a_0 + a_1 \lg x.$$

Обозначим $y' = \lg y$, $a_0' = \lg a_0$, $a_1' = a_1$, $x' = \lg x$.

Далее составляется система нормальных уравнений:

$$a_0' n + a_1' \Sigma x_i' = \Sigma y_i',$$

$$a_0 \Sigma x_i' + a_1' \Sigma x_i'^2 = \Sigma x_i' y_i',$$

и вычисляются параметры a_0' и a_1':

$$a_1' = \frac{\sum \left(\overline{x}_i' y_i' - n \overline{x}' \overline{y}' \right)}{\sum x_i'^2 - n \overline{x}'^2}, \quad a_0' = \overline{y}' - a_1 \overline{x}'.$$

Зная a_1' и a_0', уже не представляет труда найти параметры исходной формулы a_0 и a_1.

В качестве иллюстрации сказанному приводится на рис. 8.2 типичный график функции $y = a_0 x^{a_1}$ и результат ее преобразования в логарифмической шкале. Нетрудно видеть, что произошла полная линеаризация данной зависимости. Поэтому мы можем для нахождения неизвестных коэффициентов использовать МНК. Вычислив оценку a_0' и перейдя к a_0, окончательно получим $y = 2x^{0,5}$.

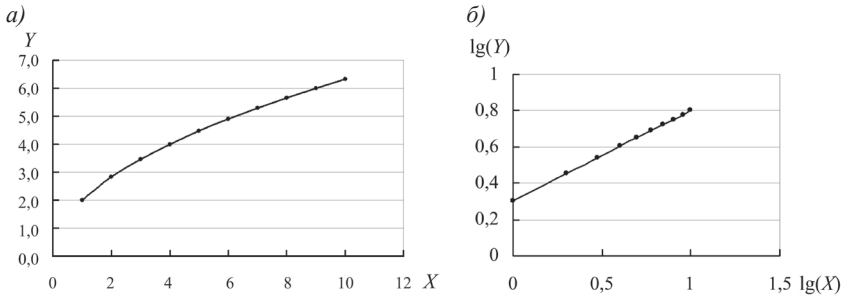

Рис. 8.2. График функции $y = a_0 x^{a_1}$
в исходной (*а*) и в логарифмической (*б*) системе координат

Однако следует помнить, что при определении неизвестных коэффициентов нелинейной формулы МНК используется не для минимизации суммы квадратов отклонений исходных переменных, а для *минимизации суммы квадратов отклонений преобразованных переменных,* что не одно и то же. В приведенном выше примере осуществляется минимизация суммы квадратов логарифмов переменных X и Y. Естественно, при этом происходит искажение структуры связи между переменными, которое, очевидно, тем заметнее, чем более нелинейной является связь. В результате полученные оценки параметров исходной формулы оказываются смещенными. Однако в большинстве случаев использование такого подхода на практике является вполне приемлемым, поскольку ошибка смещенности обычно находится в пределах точности исходных данных.

Оценка достоверности полученной формулы. Первичная проверка заключается в вычислении по найденной формуле значений $y_{(x)i}$ и их сравнении с наблюденными значениями y_i. При этом основное внимание уделяется анализу остатков, определяемых как $\varepsilon_i = y_{(i)x} - y_i$. Остатки – это то, что нельзя объяснить уравнением регрессии, если оно получено правильно. Поэтому остатки можно квалифицировать как шум, помехи или погрешности.

При проведении регрессионного анализа, как уже указывалось выше, принимается, что погрешности независимы, имеют нулевые средние, одинаковую (постоянную) дисперсию и подчиняются нормальному закону распределения. Подтверждение перечисленных свойств остатков служит доказательством того, что модель построена правильно. Следовательно, прежде всего, проверяется условие

$\overline{\varepsilon} = \sum_{i=1}^{n} \frac{\varepsilon_i}{n} = 0$. Затем проверке подлежит нормальность распределения остатков.

Кроме того, целесообразно построение графика остатков, который сразу же позволяет выявить степень случайности их хода. Если в ходе остатков наблюдаются какие-либо закономерные изменения, то следует пересмотреть регрессионную модель или же построить регрессионную модель для остатков.

Количественным критерием достоверности полученной формулы является *корреляционное отношение,* которое определяется по следующей формуле:

$$\eta = \sqrt{\frac{D_{y(x)}}{D_y}} = \sqrt{\frac{\sum\left(y_{(x)i} - \overline{y}\right)^2}{\sum\left(y_i - \overline{y}\right)^2}}, \qquad (8.1)$$

где $y_{(x)i}$ – вычисленные по нелинейной формуле значения переменной Y.

Перечислим **свойства корреляционного отношения.**

Свойство 1. Корреляционное отношение изменяется в пределах от 0 до 1, т. е. $0 \le \eta \le 1$.

Свойство 2. Если $\eta = 0$, то переменные Y и X являются взаимнонезависимыми.

Свойство 3. Если $\eta = 1$, то переменные Y и X связаны между собой функциональной зависимостью.

Свойство 4. Если $\eta = |r|$, то между переменными Y и X имеет место линейная корреляционная зависимость. В противном случае $\eta > |r|$, что означает наличие между этими переменными нелинейной зависимости.

Отсюда следует, что *корреляционное отношение служит безразмерным показателем тесноты связи между переменными любого вида.* В этом состоит его преимущество перед коэффициентом корреляции, который оценивает лишь тесноту связи линейной зависимости. Определенным недостатком корреляционного отношения является то, что оно не позволяет судить о характере связи между переменными X и Y (парабола, гипербола, экспонента и т.п.).

Кроме того, мерой точности нелинейной зависимости является среднеквадратическая ошибка, определяемая как:

$$\sigma_{y(x)} = \sqrt{\frac{\sum \varepsilon_i^2}{n-1}} = \sqrt{\frac{\sum\left(y_i - y_{(x)i}\right)^2}{n-1}}. \qquad (8.2)$$

Нетрудно показать, что данная формула может быть представлена в виде:

$$\sigma_{y(x)} = \sigma_y \sqrt{1 - \eta^2}. \qquad (8.3)$$

Однако следует помнить, что наиболее действенным способом оценки точности эмпирической формулы является расчет значений функции отклика *по независимым данным* (данным, не вошедшим в исходную выборку) и последующее их сравнение с наблюденными значениями y_i. Только в этом случае мы сможем достаточно надежно судить о точности полученной эмпирической формулы.

8.2. Особенности подбора эмпирической формулы

Рассмотрим более подробно процесс подбора эмпирической формулы. Как уже отмечалось выше, прежде всего, строится корреляционное поле – график связи переменных X и Y в декартовой системе координат и на нем наносится приближенная (эмпирическая) линия связи. Очень важным моментом является попытка установления физического характера связи между данными переменными. Если это не представляется возможным, то тогда следует обратиться к справочникам по математике, в которых приводятся графики элементарных функций и формулы, их аппроксимирующие. Например, весьма подробные сведения о них приведены в нескольких изданиях известной книги И.Н. Бронштейна, К.А. Семендяева «Справочник по математике для инженеров и учащихся втузов».

Сравнение эмпирической линии связи с теоретическими кривыми должно позволить определить тип формулы. В некоторых случаях сделать это оказывается весьма сложно, ибо эмпирическая линия связи может соответствовать сразу нескольким типам теоретических формул. Поэтому, чтобы не ошибиться, в любом случае, даже когда выбор формулы представляется очевидным, следует применить описанный выше метод выравнивания. Однако довольно часто на практике возникает ситуация, когда двухпараметрические зависимости $y_i = f(a_0, a_1; x_i)$ либо вообще не подходят к эмпирической кривой, либо, что встречается чаще, точность аппроксимации исходных данных двухпараметрическими зависимостями оказывается недостаточно высокой.

Очевидно, в этом случае необходимо уже переходить к более сложным многопараметрическим зависимостям, содержащим три и более неизвестных параметров. При этом непосредственное

использование МНК оказывается не всегда возможным, ибо далеко не во всех случаях удается привести эмпирическую формулу к виду линейному по параметрам. Отсюда вытекает необходимость построения нелинейных моделей, представляющих собой функциональную зависимость, в которую нелинейно входит один или несколько неизвестных параметров. К сожалению, до сих пор не существует эффективных теоретических подходов к оцениванию параметров подобного вида нелинейных моделей. Поэтому в практических расчетах приходится ограничиваться применением численных итерационных процедур.

Пример 8.1. Рассмотрим конкретный пример построения эмпирической формулы двухпараметрической зависимостью. Если принять, что взаимное приспособление полей температуры и влажности в приводном слое над океаном значительно меньше месяца, то можно предположить наличие зависимости между безразмерными вертикальными профилями влажности и температуры:

$$\alpha_e = f(\alpha_T),$$

где $\alpha_e = \Delta e/e_0$, $\alpha_T = \Delta T/T_0$, Δe и ΔT – вертикальные градиенты влажности и температуры воздуха в приводном слое, e_0 – насыщающая упругость водяного пара при температуре поверхности океана T_0.

На основе данных по влажности, температуре воздуха и воды для 9 судов погоды за 20-летний период путем пространственного осреднения была получена зависимость перепада влажности от перепада температуры (кривая *1* на рис. 8.3 и таблица 8.2). Проверка всех эмпирических формул, приведенных в таблице 8.1, показала, что полного выравнивания нет ни в одном случае.

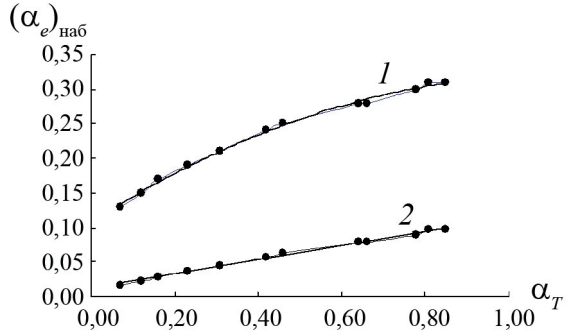

Рис. 8.3. Зависимости безразмерного вертикального профиля влажности (α_e) от вертикального профиля температуры (α_T): *1)* $\alpha_e = f(\alpha_T)$, *2)* $\alpha_e^2 = f(\alpha_T)$

Таблица 8.2

Сезонные изменения осредненных за многолетний период и для девяти судов погоды безразмерных вертикальных профилей влажности и температуры (α_e и α_T)

Параметр	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
$\alpha_T \cdot 10^2$	0,85	0,81	0,66	0,42	0,23	0,12	0,07	0,16	0,31	0,46	0,64	0,78
$(\alpha_e)_{наб}$	0,31	0,31	0,28	0,24	0,19	0,15	0,13	0,17	0,21	0,25	0,28	0,30
$(\alpha_e)_{выч}$	0,31	0,30	0,28	0,24	0,19	0,16	0,14	0,17	0,21	0,24	0,28	0,30

В результате анализа графиков элементарных функций было установлено, что зависимость $\alpha_e = f(\alpha_T)$ целесообразно аппроксимировать следующим выражением:

$$\alpha_e = (a_0 + a_1\alpha_T)^{1/2}.$$

Для выравнивания данного выражения достаточно α_e возвести в квадрат, т. е. $y' = \alpha_e^2$.

Как следует из графика связи величин α_e^2 и α_T (рис. 8.3, зависимость *2*), все точки находятся практически на прямой линии. Это свидетельствует о том, что данная формула может быть использована для аппроксимации зависимости $\alpha_e = f(\alpha_T)$. Численные значения коэффициентов a_0 и a_1, найденные методом наименьших квадратов, составляют: $a_0 = 0{,}012$, $a_1 = 10$, т. е.

$$\alpha_e = \left(0{,}012 + 10\alpha_T\right)^{1/2}. \qquad (8.4)$$

Вычисленные по формуле (8.4) значения α_e также представлены в таблице 8.2. Нетрудно видеть, что они почти полностью совпадают с заданными значениями α_e, поэтому анализ остатков не имеет смысла. Корреляционное отношение оказалось равным $\eta = 0{,}99$.

Пример 8.2. Рассмотрим теперь один из возможных способов определения неизвестных коэффициентов трехпараметрической зависимости на примере формулы Магнуса, которую запишем в следующем виде:

$$y = a_0 10^{\frac{a_1 x}{a_2 + x}}. \qquad (8.5)$$

Логарифмируя формулу (8.5), получим:

$$\lg y = \lg a_0 + \frac{a_1 x}{\left(a_2 + x\right)}.$$

Или после некоторых преобразований:

$$\frac{1}{\lg\frac{y}{a_0}} = \frac{a_2 + x}{a_1 x} = \frac{a_2}{a_1 x} + \frac{1}{a_1}.$$

Выполним теперь замену переменных:

$$y' = \frac{1}{\lg\frac{y}{a_0}},\ x' = x^{-1},\ a_1' = \frac{a_2}{a_1},\ a_0' = a_1^{-1}.$$

Итак, получаем линейное относительно неизвестных параметров уравнение, в котором определению подлежит три коэффициента: a_0, a_1 и a_2. Однако с помощью МНК мы можем определить только два из них. Поэтому поступим следующим образом. На графике связи между переменными x_i и y_i проводим приближенную нелинейную зависимость, аппроксимирующую эту связь (рис. 8.4). Нетрудно видеть, что из неизвестных коэффициентов легче всего поддается определению a_0. Действительно, при $x = 0$ $y = a_0$, т. е. величина a_0 представляет точку пересечения искомой зависимости с осью ординат. Выберем из графика связи приближенное значение a_0. Пусть $a_0 = 6{,}0$. С помощью МНК составляем систему из двух нормальных линейных уравнений, определяем из них a'_0 и a'_1, а затем $a_1 = (a'_0)^{-1}$ и $a_2' = \frac{a_1'}{a_0'}$. После этого рассчитываем среднюю квадратическую ошибку модели как:

$$\sigma_{y(x)} = \sqrt{\frac{\sum \varepsilon_i^{\,2}}{n-1}}.$$

Однако полученное уравнение еще нельзя назвать оптимальным, ибо один параметр нами был задан приближенно. Поэтому далее задаем с некоторым шагом Δ новые значения a_0 и повторяем все расчеты. Оптимальным будем считать такое уравнение, когда среднеквадратическая ошибка модели $\sigma_{y(x)}$ окажется минимальной, т. е. $\sigma_{y(x)} = \min$.

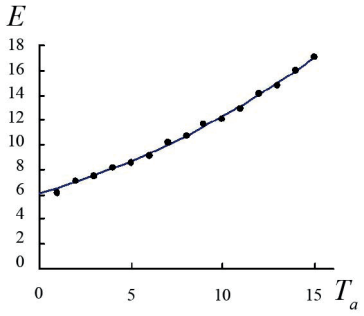

Рис. 8.4. Зависимость насыщающего давления водяного пара от температуры воздуха (формула Магнуса)

Пример 8.3. Рассмотрим зависимость средних годовых значений влагосодержания атмосферы над океаном от средних годовых значений температуры воздуха в приводном слое, осредненных по 10-градусным широтным зонам Мирового океана, т. е. $[W] = f[T_a]$.

Если построить график связи между указанными характеристиками, то из него следует, что эмпирическая зависимость очень близка по своему виду к формуле Магнуса (8.5). Определяем по графику начальное значение a_0 = 8,5 мм/год. После этого методом наименьших квадратов рассчитываем коэффициенты a_1 = 3,83 и a_2 = −168 °С. Уточнение полученной таким образом зависимости осуществляем путем подгонки параметра a_0. Зададим шаг Δ = 0,1. Пересчитываем после этого уравнение (8.5). Минимального значения среднеквадратическая ошибка достигает при a_0 = 8,7 мм/год. Итак, окончательно имеем:

$$[W] = 8,7^{\frac{3,83[T_a]}{[T_a]-168}}. \qquad (8.6)$$

Коэффициент детерминации данной зависимости равен $\eta^2 = 0,96$, а ее среднеквадратическая ошибка равна $\sigma_{y(x)} = 0,17$ мм/год или 0,8 %.

8.3. Одномерная полиномиальная регрессия

В тех случаях, когда при построении эмпирической зависимости требуется высокая точность аппроксимации и в то же время не играет большой роли физическая суть связи между переменными, целесообразно использовать метод полиномиальной регрессии, т. е.

$$y = \sum_{i=0}^{m} a_i x^i = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m + \varepsilon, \qquad (8.7)$$

где m – максимальная степень полинома, называемая степенью уравнения. Если в (8.7) положить m = 1, то имеем уравнение первой степени (линейная регрессия), при m = 2 имеем уравнение второй степени (парабола) и т.д.

Полиному первой степени соответствует прямая линия, второй степени – квадратичная кривая (парабола) с одной точкой экстремума, третьей степени – кубическая кривая с двумя точками экстремума, четвертой степени – кривая с тремя точками экстремума (рис. 8.5). Следовательно, $q = m - 1$, где q – число экстремумов функции (8.7).

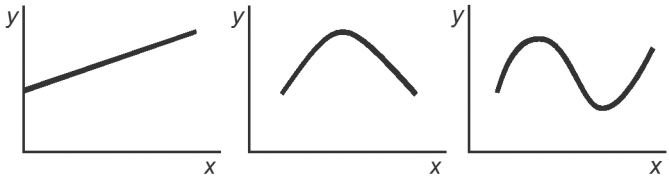

Рис. 8.5. Вид полиномиальной кривой в зависимости от степени полинома

При малом числе исходных данных с возрастанием степени полинома кривая все ближе подходит к исходным точкам и при $m = n - 1$ (n – число исходных точек) кривая точно пройдет через каждую данную точку. Однако в построении такого полинома мало смысла, так как он не является более эффективным, чем сами исходные данные. Кроме того, он содержит погрешности исходных данных и может давать в промежутках между точками заведомо абсурдные результаты. В то же время задавая полином более низкого порядка, можно тем самым сгладить случайные погрешности и в результате получить более точную аппроксимацию исходных точек. Поэтому на практике обычно редко используют максимальную степень полинома $m > 3$–4.

Для нахождения коэффициентов a_0, a_1, ... , a_m применяется метод наименьших квадратов:

$$S = \sum_{i=1}^{n} \left[y_i - \left(a_0 + a_1 x + ... + a_m x^m \right)_i \right]^2 = \min. \qquad (8.8)$$

Приравняв частные производные от S по всем неизвестным параметрам a_0, a_1, ... , a_m к нулю, т. е.

$$\frac{\partial S}{\partial a_0} = 0, \ \frac{\partial S}{\partial a_1} = 0, \ \ldots, \ \frac{\partial S}{\partial a_m} = 0.$$

получим систему нормальных уравнений, решая которую, можно определить параметры a_0, a_1, ... , a_m.

Оценивание параметров полиномиальной регрессии осуществляется точно таким же образом, как для линейной регрессии. В частности, для оценки качества полиномиальной регрессии находятся:

– нелинейный коэффициент детерминации η^2;
– корреляционное отношение η;
– среднеквадратическая ошибка модели $\sigma_{y(x)}$.

– среднеквадратические ошибки коэффициентов регрессии;
– критерий Фишера;
– *p*-критерий;
– критерий Дарбина–Уотсона.

Значимость корреляционного отношения проверяется по критерию Стьюдента аналогично проверке на значимость коэффициента корреляции с той лишь разницей, что критическое значение статистики Стьюдента определяется как $\boldsymbol{t}_{кр}(\alpha, \nu = n - m - 1)$, где m – степень полинома. Аналогичным образом определяется и критическое значение статистики Фишера $\boldsymbol{F}_{кр}(\alpha, \nu_1 = m, \nu_2 = n - 1)$.

Отметим весьма важное следствие, которое вытекает из анализа уравнения (8.7). При $m = 1$ имеем линейную регрессию. В этом случае $\eta = |r|$. Для нелинейной регрессии, когда $m \geq 2$, $\eta > |r|$. Таким образом получаем очевидное соотношение:

$$\eta \geq |r|. \tag{8.9}$$

Кроме того, можно ввести величину $\Delta = \eta - |r|$, представляющую собой меру нелинейности (кривизны) регрессионной зависимости (8.7). Естественно, чем больше величина Δ, тем более «кривой» является зависимость (8.7).

Обычно процедура расчетов при полиномиальной аппроксимации состоит в следующем. Вначале принимается $m = 1$, и для этого случая определяются параметры регрессии и корреляционное отношение η_1. Затем принимается $m = 2$, и заново рассчитываются параметры регрессии и корреляционное отношение η_2. Если выполняется условие $(\eta_2 - \eta_1) > \varepsilon$, где ε – некоторое заданное положительное число, то делается вывод о продолжении расчетов. Принимается $m = 3$, и вся процедура повторяется. Когда добавление члена полинома более высокого порядка уже не будет давать существенного увеличения точности аппроксимации, делается вывод о прекращении расчетов. Естественно, что это зависит и от выбора числа ε, которое в зависимости от поставленной задачи может быть различным. Обычно ε принимается в пределах от 0,01 до 0,1.

В принципе возможен другой вариант оценки перехода к уравнению регрессии более высокой степени на основе использования критериев проверки гипотез. Например, требуется проверить, является ли зависимость между изучаемыми переменными линейной ($m = 1$) или она носит нелинейный характер ($m = 2$). В этом случае нулевая гипотеза примет вид $H_0 : \eta^2 = r^2$ при альтернативной гипотезе $H_1 : \eta^2 \neq r^2$. Установлено, что проверка нулевой гипотезы может

быть выполнена с помощью критерия Стьюдента, который в данном случае рассчитывается как:

$$t = \frac{\eta^2 - r^2}{\delta_{(\eta^2 - r^2)}}, \qquad (8.10)$$

где $\delta_{(\eta 2 - r2)}$ – величина ошибки разности $\eta^2 - r^2$.

После этого проверяется неравенство $t > \boldsymbol{t}_{кр}(\alpha, \nu = n - m - 1)$. Если данное неравенство выполняется, то нулевая гипотеза отвергается и делается вывод, что между изучаемыми переменными существует нелинейная связь. В противном случае мы можем предположить наличие линейной связи. Аналогичным образом выполняется проверка целесообразности перехода от уравнения регрессии второй степени к уравнению третьей степени.

Однако отметим, что процесс выбора размерности нелинейной модели является наиболее «тонким» моментом при построении полиномиальной регрессии. Далеко не во всех случаях анализ одного корреляционного отношения или связанного с ним коэффициента нелинейной детерминации дает возможность определить порядок «лучшей» модели. Очень важно, чтобы критерий Фишера всегда был значимым, стандартная ошибка модели как можно меньше, а все коэффициенты регрессии значимы по критерию Стьюдента, что, вообще говоря, не всегда оказывается реальным. Поэтому возникает задача построения оптимальной в некотором смысле модели, т. е. такой модели, которая бы минимизировала отмеченные противоречия между указанными параметрами. Кроме того, следует принимать во внимание и неформальные аспекты. Например, чем меньше степень модели, тем она надежнее. Более подробно задача определения оптимальной степени модели рассматривается далее, в примере 8.4.

Пример 8.4. Один из возможных способов применения полиномиальной регрессии – это аппроксимация вертикальных профилей гидрометеорологических элементов. Воспользуемся данными о вертикальном профиле солености, измеренного с помощью гидрозонда через 1 м до глубины 15 м на гидрологической станции в Кандалакшской губе Белого моря (рис. 8.6).

Вертикальный профиль солености является весьма типичным и имеет четыре характерные особенности:

– верхний квазиоднородный слой, который отмечается примерно до горизонта 4 м;

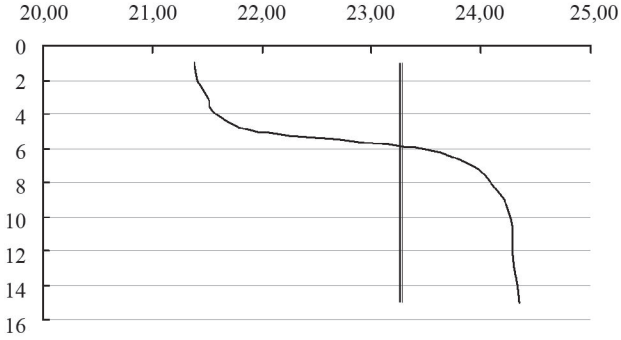

Рис. 8.6. Вертикальный профиль солености
на гидрологической станции в Кандалакшской губе Белого моря.
Вертикальная линия – среднее значение солености

– сезонный халоклин, характеризующийся резким градиентом солености, находится в слое 4−6 м;

– главный (постоянный) халоклин прослеживается примерно в слое 6−10 м;

– придонный квазиоднородный слой, который располагается ниже 10 м.

Таким образом, функцией отклика являются значения солености, а регрессором – глубины через 1 м (n = 15). Основные статистические параметры модели (8.7) от m = 1 до m = 6 представлены в таблице 8.3.

Таблица 8.3

Оценки статистических параметров модели (8.7)
от m = 1 до m = 6 по расчету вертикального профиля солености

Степень полинома модели (8.7)	Коэффициент детерминации	Критерий Фишера	Стандартная ошибка модели, ‰	Максимальный p-критерий
1	0,798	51,5	0,60	$7{,}2\times10^{-6}$
2	0,901	54,4	0,44	0,004
3	0,921	42,9	0,41	0,54
4	0,967	73,2	0,28	0,01
5	0,968	54,1	0,29	0,64
6	0,979	61,8	0,25	0,44

Как видно из таблицы 8.3, уже при линейной аппроксимации (m = 1) модель описывает 80 % дисперсии вертикального профиля

солености. До $m = 4$ наблюдается рост нелинейного коэффициента детерминации, затем его рост почти прекращается. Критерий Фишера при всех значениях m на порядок превышает его критическую величину $\boldsymbol{F}_{кр}$. Стандартная ошибка модели с увеличением степени полинома уменьшается, достигая минимального значения при $m = 6$. Оценка *p-критерия,* характеризующего значимость коэффициентов регрессии по критерию Стьюдента, выбирается как максимальное значение для каждого из 6 уравнений модели. Из таблицы 8.3 следует, что значимыми являются всего 3 уравнения модели: первой, второй и четвертой степени.

Итак, принимая во внимание противоречивый характер распределения параметров модели (8.7), необходимо выбрать такую степень уравнения, чтобы оно наилучшим (оптимальным) образом аппроксимировало вертикальный профиль солености. Очевидно, если исходить только из чисто формальных условий, то для оптимальной модели ход ее основных параметров должен быть следующим:

– коэффициент детерминации резко возрастает до определенного шага, затем почти не меняется;

– стандартная ошибка на определенном шаге становится минимальной;

– оценка критерия Фишера больше его критического значения;

– *p-level* меньше заданного уровня значимости (например, $\alpha = 0{,}05$).

В распределении коэффициента детерминации проявляется 2 скачка: резкое уменьшение от $m = 2$ к $m = 3$ и от $m = 4$ к $m = 5$. Начиная с $m = 4$, величина η^2 почти не меняется. Использование *t*-критерия в виде (8.10) показало, что следует ограничиться уравнением модели четвертой степени. Критерий Фишера, как уже указывалось выше, является значимым для уравнения любой степени модели. Минимальная стандартная ошибка модели наблюдается при $m = 6$. Максимальный *p-критерий,* меньший $\alpha = 0{,}05$, отмечается для уравнений первой, второй и четвертой степени. Таким образом, нет ни одного уравнения, для которого бы все требования к параметрам соответствовали условиям оптимальности. Наиболее близким к оптимальному является уравнение четвертой степени, для которого только стандартная ошибка незначительно превышает ее минимальное значение. Поэтому, очевидно, именно уравнение четвертой степени следует признать наилучшим. Сопоставление фактических и вычисленных по модели значений солености приводится на рис. 8.7.

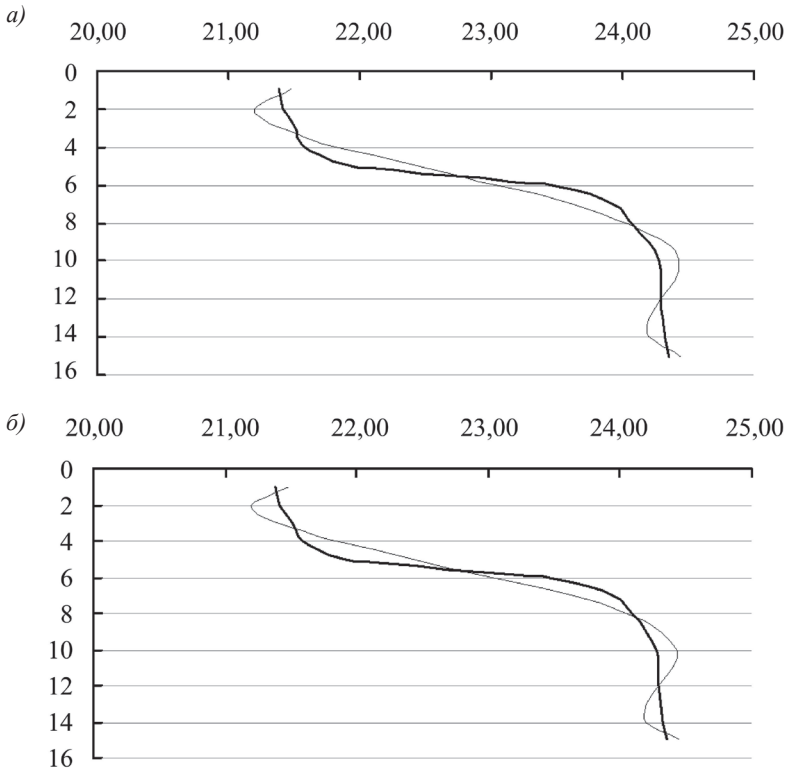

Рис. 8.7. Аппроксимация вертикального профиля солености на гидрологической станции в Кандалакшской губе Белого моря полиномиальной регрессией (*а*) и полиномами Чебышева и 4-го порядка (*б*)

8.4. Ортогональная регрессия

Существенным неудобством классического уравнения одномерной полиномиальной регрессии (8.7) является необходимость пересчета на каждом шаге всех коэффициентов регрессии при увеличении степени полинома. Этого недостатка можно избежать, если воспользоваться ортогональными многочленами Чебышева, благодаря которым удается выполнять добавление новых слагаемых более высокого порядка, не изменяя при этом вычисленные ранее коэффициенты.

Суть способа Чебышева заключается в том, что аппроксимирующий многочлен отыскивают не непосредственно в виде суммы

степеней переменной X, а как некоторую комбинацию многочленов. В результате уравнение (8.7) можно переписать в виде уравнения ортогональной регрессии:

$$y = a_0\varphi_0(x) + a_1\varphi_1(x) + ... + a_m\varphi_m(x) + \varepsilon. \qquad (8.11)$$

Многочлены Чебышева $\varphi_0(x)$, $\varphi_1(x)$, ... , $\varphi_m(x)$ зависят только от объема выборки n. Первые два из них имеют вид:

$$\varphi_0(x) = 1; \quad \varphi_1(x) = x - \frac{n+1}{2}.$$

Остальные многочлены определяются по формуле:

$$\varphi_{m+1}(x) = \varphi_1(x)\varphi_m(x) - \frac{m^2\left(n^2 - m^2\right)}{4\left(4m^2 - 1\right)}\varphi_{m-1}(x). \qquad (8.12)$$

Например, многочлен $\varphi_2(x)$ будет иметь вид:

$$\varphi_2(x) = x^2 - \left(n+1\right)x + \frac{\left(n+1\right)\left(n+2\right)}{6}.$$

Следовательно, многочлен $\varphi_{m+1}(x)$, зависящий лишь от объема выборки, может быть вычислен заранее, и при каждом увеличении степени регрессии необходимо рассчитывать только один коэффициент a_{m+1}.

Неизвестные коэффициенты a_0, a_1, ... , a_m определяются непосредственно на основе многочленов Чебышева. Опуская достаточно громоздкие промежуточные выкладки, приведем сразу окончательные формулы:

$$a_0 = \frac{\sum y_i}{n}, \; a_1 = \frac{\sum y_i\varphi_1(x_i)}{\sum \varphi_1^2(x_i)},$$

$$a_2 = \frac{\sum y_i\varphi_2(x_i)}{\sum \varphi_2^2(x_i)}, \; ..., \; a_m = \frac{\sum y_i\varphi_m(x_i)}{\sum \varphi_m^2(x_i)}. \qquad (8.13)$$

Итак, если мы задаем многочлен при a_0, то решение уравнения (8.11) соответствует средней арифметической исходных данных, добавляя многочлен при a_1, получаем уравнение прямой линии, при a_2 – параболу и т.д.

Как видно из формул (8.13), вычисленные коэффициенты не зависят от того, каков будет порядок разыскиваемого уравнения регрессии. Уравнение регрессии составляется методом последовательных

приближений, при этом повышение на один порядок регрессии связано с нахождением только одного коэффициента a_j.

В принципе, оценка числа используемых многочленов в уравнении (8.11) может быть осуществлена аналогично оценке степени полинома в уравнении (8.7), т. е. путем оптимизации разности корреляционных отношений $\eta_m - \eta_{m-1}$. Когда эта разность достигает некоторого заданного положительного числа Δ, то делается вывод о прекращении расчетов. Необходимо помнить, что поскольку неизвестные коэффициенты a_j вычисляются в уравнениях (8.7) и (8.11) различными способами, то коэффициенты детерминации полиномиальной и ортогональной регрессий не совпадают. Так как на ортогональную регрессию не распространяются свойства корреляционного отношения, полученного по МНК, то оно в отдельных случаях может даже превысить $\eta > 1$. Однако это является лишь свидетельством ошибок расчета и ничего более.

Несомненное достоинство ортогональной регрессии заключается в том, что она является линейной относительно зависимой переменной, в качестве которой выступают многочлены Чебышева, и тем самым позволяет избежать использования высоких степеней полинома. Естественно, это уменьшает вычислительные ошибки, возникающие при непосредственном использовании МНК к уравнению (8.7). Однако в статистическом плане МНК является несравненно более мощным аппаратом по сравнению с многочленами Чебышева, что необходимо учитывать в практических расчетах.

Пример 8.5. Рассмотрим применение полиномов Чебышева для аппроксимации вертикального профиля солености (рис. 8.5). Вначале определялись коэффициенты $a_0, a_1, \ldots, a_m$ до $m = 6$, затем последовательно рассчитывались значения солености для всех горизонтов. Это позволило найти разности между фактическими и вычисленными значениями солености, рассчитать стандартную ошибку модели и дисперсию ошибки для ее каждого вертикального профиля. После этого уже нетрудно оценить нелинейный коэффициент детерминации как $\eta^2 = 1 - D_\varepsilon / D_y$. В таблице 8.4 приведены оценки нелинейных коэффициентов детерминации и стандартной ошибки модели для различных номеров полиномов Чебышева от $m = 1$ до $m = 6$. Таким образом, мы имеем возможность сравнивать результаты расчета вертикального профиля солености по моделям (8.7) и (8.11).

Итак, до $m = 4$ результаты расчета вертикального профиля солености по обеим моделям полностью совпадают. Однако, начиная с $m = 5$, результаты начинают расходиться. Из таблицы 8.4 видно,

Таблица 8.4

Оценки параметров модели (8.11)
от $m = 1$ до $m = 6$ по расчету вертикального профиля солености

Номер полинома Чебышева	Коэффициент детерминации	Стандартная ошибка модели, ‰
1	0,798	0,60
2	0,900	0,44
3	0,921	0,41
4	0,967	0,28
5	0,988	0,32
6	0,999	0,29

что при $m = 6$ коэффициент детерминации практически равен единице. Из этого следует, что стандартная ошибка должна стремиться к нулю. Однако она, напротив, возросла. Отметим, что для более высоких номеров полиномов Чебышева коэффициент детерминации может даже превышать единицу. Это означает, что с увеличением числа полиномов Чебышева происходит возрастание ошибок аппроксимации и, следовательно, использование в таком случае ортогональной регрессии теряет смысл.

Если исходить из результатов, представленных в таблице 8.4, то наилучшей следует признать модель при $m = 4$. Это совпадает с результатами, полученными при аппроксимации профиля солености с помощью полиномиальной регрессии (рис. 8.6). Однако из сравнения примеров 8.4 и 8.5 достаточно очевидным становится, что в статистическом плане МНК является несравненно более мощным аппаратом, чем разложение по полиномам Чебышева.

8.5. Двухмерная полиномиальная регрессия

Во многих случаях случайная величина может зависеть не от одной переменной, а от двух. Характерный пример – анализ пространственных полей. Действительно, любую карту можно представить в следующем виде:

$$G(x,y) = f(a_1, a_2, \ldots, a_m, x, y), \qquad (8.14)$$

где x и y – пространственные координаты. Отсюда нетрудно видеть, что определение неизвестных коэффициентов $a_1, a_2, \ldots, a_m$ может быть осуществлено с помощью МНК. Аналитическая аппроксимация заданного случайного поля представляет интерес не только с точки зрения построения модели, но и с точки зрения объективного анализа, являющегося важным этапом численного моделирования

гидрометеорологических полей. Под объективным анализом, как известно, понимают процедуру перевода данных из нерегулярной сети точек в регулярную.

Другой задачей использования зависимости (8.14) является аппроксимация некоторой таблично заданной функции, когда она зависит от двух переменных. Таким образом, приходим к двухмерной полиномиальной регрессии. Отметим, что в геологии она получила название *тренд-анализа – математического метода разделения эмпирических данных на две части: систематическую и случайную*. Систематическая часть трактуется как поверхность тренда, случайная часть – отклонения поверхности тренда от системы исходных эмпирических данных. При этом тренд представляет собой некоторую функцию пространственных координат, построенную по эмпирическим данным таким образом, чтобы сумма квадратов отклонений их от поверхности тренда была бы минимальна.

Основная формула двухмерной полиномиальной регрессии может быть записана следующим образом:

$$G(x, y) = \sum_{i=0}^{m} \sum_{j=0}^{m-1} a_{ij} x^i y^j + \varepsilon, \qquad (8.15)$$

где m – показатель степени. Очевидно, что с увеличением m точность аппроксимации пространственного поля $G(x, y)$ возрастает. Однако при $m > 4$ возникают трудности вычислительного характера, связанные с процессом обращения матриц высокого порядка.

Для многих гидрометеорологических полей даже при малых значениях m могут быть получены результаты с достаточной для практических целей точностью. Если, например, принять $m = 3$, то основное уравнение (8.15) приобретет следующий вид:

$$G(x, y) = a_0 + a_{10}x + a_{01}y + a_{20}x^2 + a_{02}y^2 + a_{11}xy + a_{30}x^3 + \\ + a_{03}y^3 + a_{21}x^2y + a_{12}xy^2 + \varepsilon. \qquad (8.16)$$

В данном выражении первые три слагаемых дают линейное уравнение двумерной полиномиальной регрессии, первые шесть слагаемых – уравнение двумерной полиномиальной регрессии второго порядка, а все выражение (8.16) представляет уравнение двумерной полиномиальной регрессии третьего порядка.

Коэффициенты a_{ij} в (8.15) находятся методом наименьших квадратов. Так, например система линейных нормальных уравнений для определения линейного уравнения тренда имеет вид:

$$\Sigma G = a_0 n + a_1 \Sigma x + a_2 \Sigma y,$$
$$\Sigma xG = a_0 \Sigma x + a_1 \Sigma x^2 + a_2 \Sigma Ey,$$
$$\Sigma yG = a_0 \Sigma y + a_1 \Sigma xy + a_2 \Sigma y^2.$$

Аналогичным образом составляются системы нормальных уравнений для поверхностей тренда более высокого порядка. Решением линейного уравнения двумерной полиномиальной регрессии является семейство прямых линий в декартовой системе координат, а уравнения двумерной полиномиальной регрессии второго порядка – семейство парабол (рис. 8.8). Более сложный характер имеет решение уравнения третьего порядка. Во многом это определяется знаками переменных X и Y в последних четырех слагаемых в формуле (8.16).

Для оценки точности аппроксимации рассчитывается дисперсия фактических значений поля:

$$D_T = \frac{1}{n-1} \sum_{i=1}^{n} \left(G_{Ti} - \bar{G} \right)^2, \qquad (8.17)$$

где n – число исходных точек, а также дисперсия вычисленных по уравнению (8.16) значений G_R:

$$D_R = \frac{1}{n-1} \sum_{i=1}^{n} \left(G_{Ri} - \bar{G} \right)^2. \qquad (8.18)$$

Отношение D_R/D_T, представляющее собой коэффициент детерминации, дает качество приближения вычисленных значений аппроксимированного поля к его фактическим значениям.

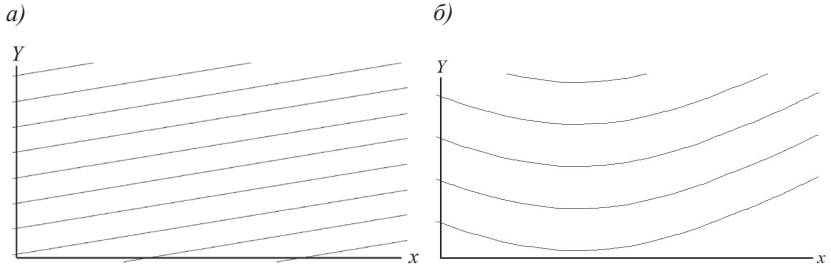

Рис. 8.8. Графический вид решения уравнения двумерной полиномиальной регрессии: *а)* линейное уравнение; *б*) уравнение второго порядка

Другим количественным критерием достоверности полученного уравнения является корреляционное отношение, вычисляемое по формуле (8.1). Кроме того, необходимо рассчитывать среднюю квадратическую ошибку модели. Однако, как уже указывалось выше, наиболее точным способом оценки точности полиномиальной регрессии является расчет значений $y_{(x)i}$ *по независимым данным* (данным, не вошедшим в исходную выборку) и последующее сравнение с наблюденными значениями y_i.

Пример 8.6. Рассмотрим использование двухмерной полиномиальной регрессии для аппроксимации некоторой таблично заданной функции в зависимости от двух переменных. Так, при расчете испарения с поверхности океана за длительные интервалы времени обычно применяется аэродинамический метод $E = c_E \rho \Delta q U$, где ρ – плотность воздуха; c_E – коэффициент влагообмена; Δq – перепад важности в приводном слое атмосферы; U – скорость ветра. Наибольшие трудности при расчете испарения по этой формуле связаны с тем, что до сих пор еще не найдена универсальная зависимость коэффициента влагообмена от внешних факторов. Поэтому в расчетах используются самые различные варианты, начиная от принятия c_E постоянной величиной, до сложных многопараметрических зависимостей c_E от характеристик приводного слоя. Довольно широко распространено мнение, что коэффициент влагообмена должен зависеть от скорости ветра и разности температур между водой и воздухом, т. е. $c_E = f(U, \Delta T)$.

Вид функции f может быть найден с помощью двухмерной полиномиальной регрессии. В результате расчетов было установлено, что коэффициент влагообмена может быть аппроксимирован поверхностью тренда второй степени:

$$c_E = 0{,}85 \times 10^{-3} + 0{,}762 \times 10^{-4} U_{10} + 0{,}882 \times 10^{-4} \Delta T_{10} -$$
$$-0{,}591 \times 10^{-6} U_{10}^2 - 0{,}11 \times 10^{-5} \Delta T_{10}^2 - 0{,}191 \times 10^{-5} U_{10} \Delta T_{10}, \quad (8.19)$$

где U_{10} – скорость ветра на высоте 10 м, м/с, ΔT_{10} – разность между температурой поверхности океана и температурой на высоте 10 м. Корреляционное отношение оказалось равным $\eta = 0{,}96$.

Корреляционное отношение, как известно, показывает лишь качество аппроксимации. Более важным для нас является оценка достоверности величин испарения, рассчитанных на основе полученной выше аппроксимации значений c_E. В этом случае можно воспользоваться, например, сравнением с другим методом, погрешности которого известны (см. разд. 5.1).

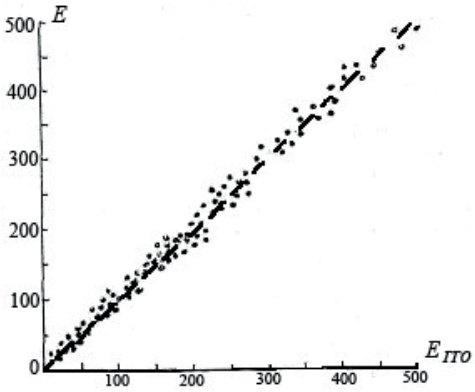

Рис. 8.9. Сравнение среднемесячных величин испарения, рассчитанных по методике ГГО и с использованием формулы (8.16) в мм

Сравнение среднемесячных величин испарения для пяти судов погоды, вычисленных с использованием зависимости (8.19), с аналогичными значениями испарения, рассчитанными по методике ГГО, показало (рис. 8.9), что систематической погрешностью между ними можно практически пренебречь. Случайная погрешность при этом составила около 5 %, что значительно ниже погрешностей определения среднемесячного испарения с поверхности океана.

8.6. Понятие о кубических сплайнах

Сплайн – это кусочно-сопряженная функция, кривая которой состоит из отрезков полиномиальных кривых, состыкованных таким образом, чтобы производные полученной функции были бы непрерывны на всем рассматриваемом промежутке. При этом непрерывность производных осуществляется до максимально высокого возможного порядка при выполнении условия, что степень многочленов, используемых для сглаживания исходных данных, ниже степени единственного многочлена, кривая которого проходит через все заданные точки. В связи с этим данная процедура обеспечивает гораздо большую гладкость, чем традиционная кусочно-линейная интерполяция, при которой интерполяционная функция терпит разрывы даже в первой производной.

Отметим, что сплайны не являются ни аналитическими функциями, ни статистическими моделями, такими, например, как рассмотренная выше полиномиальная регрессия. Однако в силу своих

свойств они обеспечивают высокую точность интерполяции или аппроксимации исходных данных.

Наиболее широкое распространение, в силу их простоты, получили *кубические сплайны.* Основные идеи теории кубических сплайнов сформировались в результате попыток описать математически гибкие рейки из упругого материала, которыми издавна пользовались чертежники в тех случаях, когда возникала необходимость проведения через заданные точки достаточно гладкой кривой. Известно, что рейка из упругого материала, закрепленная в некоторых точках и находящаяся в состоянии равновесия, принимает форму, при которой ее энергия является минимальной. Другими словами, рейка (сплайн) ограничена определенными точками, но между ними она изгибается так, чтобы в результате получилась гладко изменяющаяся линия. Это фундаментальное свойство позволяет эффективно использовать сплайны при решении задач обработки экспериментальных данных.

Для понимания сути построения сплайнов обратимся к рис. 8.10, на котором представлено множество из четырех наблюдений, связанных между собой сплайн-функцией. Наблюдения представлены точками P_i в декартовой системе координат, т. е. $P_i = [X_i, Y_i]$. Интервалы между точками можно измерить хордой (прямолинейным отрезком, соединяющим две точки), которую обозначим как t_i, где i – номер второй точки, а касательная к сплайну во внутренней точке

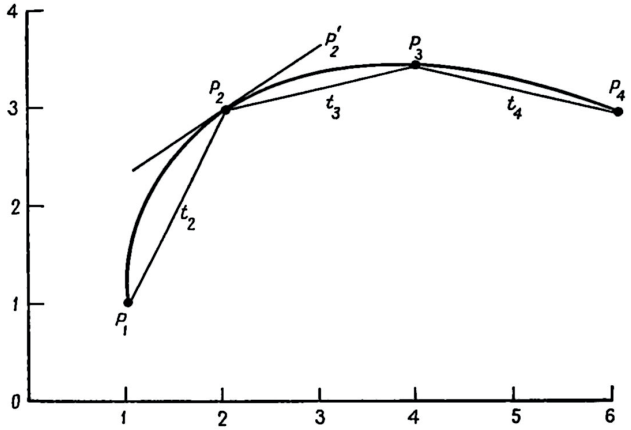

Рис. 8.10. Построение кубической сплайн-функции для четырех точек

P_i обозначена через P_i'. Кубическая сплайн-функция строится по каждой паре точек.

В общем виде уравнение кубического сплайна можно записать как:

$$P_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3,$$

то есть оно внешне полностью совпадает с уравнением полиномиальной регрессии. Однако определение коэффициентов β_j принципиально отличается от МНК. Для их нахождения требуется знать координаты точек, определяющих концы сплайна и наклоны касательных прямых в этих точках. Кроме того, мы должны дополнительно указать граничные условия, определяющие поведение аппроксимирующей линии на первом и последнем участках. При этом координаты точек считаются заданными. По этим данным требуется определить наклоны касательных векторов. Основная проблема заключается в выборе граничных условий.

Основоположником теории сплайнов можно считать Л. Эйлера, который еще в XVIII в. разработал «метод ломаных» для интегрирования дифференциальных уравнений. Этот метод представляет собой решение дифференциального уравнения с помощью ломаной линии, которая, по существу, является простейшим сплайном первой степени.

Предположим, что мы имеем неизвестную функцию $y = f(x)$, заданную значениями $y_1, \ldots, y_n$ на отрезке $[a,b]$ в точках $x_1, \ldots, x_n$, называемых узлами интерполяции. Для функции $y = f(x)$ требуется найти приближение $y = \varphi(x)$ таким образом, чтобы $f(x_i) = \varphi(x_i)$ в узлах интерполяции, а в остальных точках отрезка $[a,b]$ значения этих функций были бы близки друг другу. Данная задача реализуется с помощью *интерполяционного сплайна.* Кубическим сплайном на отрезке $[a,b]$ называется дважды непрерывно дифференцируемая функция $y = \varphi(x)$, на каждом из отрезков $\Delta_j = [x_{j-1}, x]$ совпадающая с кубическим полиномом и удовлетворяющая условиям интерполяции $\varphi(x_j) = y_j; j = 1, \ldots, N$.

Для построения интерполяционного сплайна положим $K_j = \varphi''(x_j)$, где $\varphi''(x_j)$ – вторая производная сплайна. Поскольку кубический сплайн на каждом из отрезков Δ_j совпадает с кубическим полиномом, то для этих отрезков $\varphi''(x_j)$ должна быть линейной функцией. Если ее график в декартовой системе координат проходит через точки (x_{j-1}, K_{j-1}), (x_j, K_j), то эту функцию можно представить как:

$$\frac{\varphi''(x)-K_{j-1}}{x-x_{j-1}}=\frac{K_j-K_{j-1}}{x_j-x_{j-1}}.$$

Отсюда следует, что

$$\varphi''(x)=\frac{K_{j-1}\left(x_j-x\right)}{h_j}+\frac{K_j\left(x-x_{j-1}\right)}{h_j},$$

где $h_j=x_j-x_{j-1}, j=2, \ldots, N$.

Интегрируя данное равенство дважды на отрезке x–x_{j-1} и определив константы интегрирования, получим аналитическое выражение кубического сплайна:

$$\varphi(x)=\frac{K_{j-1}\left(x_j-x\right)^3}{6h_j}+\frac{K_j\left(x-x_{j-1}\right)^3}{6h_j}+$$

$$+\frac{x_j-x}{h_j}\left(y_{j-1}-\frac{1}{6}K_{j-1}h_j^2\right)+\frac{x-x_{j-1}}{h_j}\left(y_j-\frac{1}{6}K_jh_j^2\right). \qquad (8.20)$$

На практике вместо (8.20) предпочитают пользоваться обычным кубическим полиномом. Если $E\in\Delta_j=\left[x_{j-1},x_j\right]$, то аналитическое выражение кубического сплайна (8.17) можно переписать в виде:

$$y(x)=y_{j-1}+a_{1.j-1}\left(x-x_{j-1}\right)+a_{2.j-1}\left(x-x_{j-1}\right)^2+a_{3.j-1}\left(x-x_{j-1}\right)^3, \quad (8.21)$$

где коэффициенты полинома (8.21) связаны с коэффициентами сплайна (8.20) следующими формулами:

$$a_{1.j-1}=h_j^{-1}\left(y_j-y_{j-1}\right)-h_j^{-1}\left(\frac{K_j}{6}+\frac{K_{j-1}}{3}\right),$$

$$a_{2.j-1}=\frac{K_{j-1}}{2},$$

$$a_{3.j-1}=\frac{\left(K_j-K_{j-1}\right)}{6h_j}.$$

Коэффициенты сплайна находятся различными численными методами.

Ясно, что интерполяционный сплайн для решения задачи аппроксимации эмпирических данных не годится, поскольку его главная

цель состоит в максимально точном восстановлении значений искомой функции в узлах интерполяции. Поэтому в промежутках между узлами он может довольно сильно искажать «поведение» заданной функции. Очевидно, для аппроксимации эмпирических данных целесообразно использовать *сглаживающий сплайн,* который осуществляет построение более гладких кривых, не обязательно проходящих через заданные узлы.

Предположим, что экспериментальные значения y_j функции $y = f(x)$ известны с некоторыми погрешностями:

$$\left| y_j - f(x_j) \right| \le \delta, \quad j = 1, \ \dots \ , N.$$

Так как кубический сплайн имеет минимальную кривизну, то при построении сглаживающего сплайна естественно потребовать, чтобы он минимизировал характеризующий кривизну интеграл:

$$\Phi(\varphi) = \int_a^b \left[\varphi''(x) \right]^2 dx \to \min, \tag{8.22}$$

и при этом удовлетворялись условия:

$$\left| f(x_j) - y_j \right| \le \delta, \quad j = 1, \ \dots \ , N. \tag{8.23}$$

Отсюда следует, что задача построения сглаживающего сплайна является задачей нелинейного программирования. При решении данной задачи возникает ряд трудностей. Одна из основных состоит в том, что численные методы оптимизации, с помощью которых осуществляется поиск минимума функции (8.22), малоэффективны в областях типа (8.21). Чтобы избежать этого, может быть использован следующий подход. Эффективный и в то же время легко реализуемый на ЭВМ сплайн возникает при минимизации функционала:

$$\Phi(\varphi) = \int_a^b \left[\varphi''(x) \right]^2 dx + \sum_{j=1}^{N} \frac{\left[f(x_j) - y_j \right]^2}{\rho_j}, \tag{8.24}$$

где $f(x_j)$ – значения в узлах x_j, $\rho_j > 0$ – заданные весовые коэффициенты. Можно показать, что чем меньшее значение имеет коэффициент ρ_j, тем ближе проходит функция $f(x)$ к экспериментальному значению y_j. Если для некоторого номера j коэффициент $\rho_j = 0$, то $f(x_j) = y_j$, т. е. в точке x_j значение сглаживающего сплайна совпадает со значением функции в этой точке. Это означает, что сглаживающий сплайн становится интерполяционным.

Пусть $K_i = \varphi''(x)$. Тогда коэффициенты сглаживающего сплайна можно найти путем решения системы линейных алгебраических уравнений:

$$
\begin{gathered}
c_1K_1 + b_1K_2 + a_1K_3 = d_1,\\
b_1K_1 + c_2K_2 + b_2K_3 + a_2K_4 = d_2,\\
a_{i-2}K_{i-2} + b_{i-1}K_{i-1} + c_iK_i + b_iK_{i+1} + a_iK_{i+2} = d_i,\\
a_{N-3}K_{N-3} + b_{N-2}K_{N-2} + c_{N-1}K_{N-1} + b_{N-1}K_N = d_{N-1},\\
a_{N-2}K_{N-2} + b_{N-1}K_{N-1} + c_NK_N + d_N. \qquad (8.25)
\end{gathered}
$$

Матрица этой системы является положительно определенной, поэтому данная система уравнений имеет единственное решение. Отсюда следует, что сглаживающий сплайн является единственным. Решение данной системы может быть осуществлено, например, с помощью метода факторизации.

Пример 8.8. На рис. 8.11 приводятся графики аппроксимации с помощью сглаживающего сплайна функции $y = 0{,}3x + 7\sin 5x - 5\cos 8x$ при $\rho = 0{,}1$ (рис. 8.11-*а*) и $\rho = 0{,}005$ (рис. 8.11-*б*).

Нетрудно видеть, что уменьшение параметра ρ существенно повышает точность аппроксимации данной функции. Однако вопрос заключается в том, насколько в действительности необходима такая точность. Дело в том, что с повышением точности одновременно аппроксимируются и случайные ошибки экспериментальных данных. Ясно, что для их обнаружения необходим физический анализ. Если на основе физических соображений будет установлено, что отклонение точек от линии связи действительно связано с наличием в них существенных ошибок, то тогда можно ограничиться аппроксимацией при $\rho = 0{,}1$. В противном случае – аппроксимацией при $\rho = 0{,}005$.

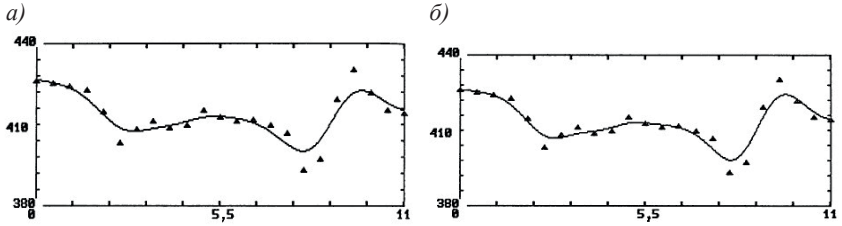

Рис. 8.11. Аппроксимация функции $y = 0{,}3x + 7\sin 5x - 5\cos 8x$ с помощью кубического сглаживающего сплайна при $\rho = 0{,}1$ (*а*) и $\rho = 0{,}005$ (*б*)

Список литературы

1. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. – М.: ЮНИТИ, 1998. – 1022 с.
2. *Вайновский П.А., Малинин В.Н.* Методы обработки и анализа океанологической информации. – Ч. 1. Одномерный анализ. – Л.: РГГМИ, 1991. – 136 с.
3. *Вайновский П.А., Малинин В.Н.* Методы обработки и анализа океанологической информации. – Ч. 2. Многомерный анализ. – СПб.: РГГМИ, 1992. – 96 с.
4. *Григоркина Р.Г., Губер П.К., Фукс В.Р.* Прикладные методы корреляционного и спектрального анализа крупномасштабных океанологических процессов. – Л.: ЛГУ, 1973. – 172 с.
5. *Вуколов Э.И.* Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL. – М.: ФОРУМ, 2008. – 464 с.
6. *Гмурман В.Е.* Теория вероятностей и математическая статистика. – М.: Высшая школа, 2002. – 479 с.
7. *Девис Дж.С.* Статистический анализ данных в геологии. – М.: Недра. 1990. – Кн. 1. – 319 с.; Кн. 2 – 427 с.
8. *Дегтярев А.С., Драбенко В.А., Драбенко В.А.* Статистические методы обработки метеорологической информации: учебник. – СПб: ООО «Андреевский издательский дом», 2015. – 225 с.
9. *Демьянов В.В., Савельева Е.А.* Геостатистика: теория и практика. – М.: Наука, 2010. – 329 с.
10. *Дженкинс Г., Ваттс. Д.* Спектральный анализ и его приложения. – Вып. 1. – М.: Мир, 1971. – 316 с.; Вып. 2. – М.: Мир, 1972. – 287 с.
11. *Добровольский С.Г.* Климатические изменения в системе «гидросфера–атмосфера». – М.: Геос, 2002. – 231 с.
12. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. Кн. 1. – М.: Финансы и статистика, 1986. – 366 с.; Кн. 2 – 1987. – 351 с.
13. *Казакевич Д.Л.* Основы теории случайных функций в задачах гидрометеорологии. – Л.: Гидрометеоиздат, 1989. – 230 с.
14. *Кремер Н.Ш.* Теория вероятностей и математическая статистика. – М.: ЮНИТИ, 2003. – 543 с.
15. *Львовский Е.Н.* Статистические методы построения эмпирических формул. – М.: Высшая школа, 1982. – 224 с.
16. *Макарова Н.В., Трофимец В.Я.* Статистика в Excel. – М.: Финансы и статистика, 2002. – 365 с.
17. *Малинин В.Н., Гордеева С.М.* Физико-статистический метод прогноза океанологических характеристик. – Мурманск, Изд-во ПИНРО, 2003. – 164 с.
18. *Матерон Ж.* Основы прикладной геостатистики. М., 1968. − 407 с.
19. *Мюллер П., Нойман П., Шторм Р.* Таблицы по математической статистике. – М.: Финансы и статистика, 1982. – 272 с.
20. *Носач В.В.* Решение задач аппроксимации с помощью персональных компьютеров. – М.: МИКАП, 1994. – 382 с.
21. *Пановский Г.А., Брайер Г.В.* Статистические методы в метеорологии. – Л.: Гидрометеоиздат, 1967. – 242 с.

22. *Поляк И.И.* Методы анализа случайных процессов и полей в климатологии. – Л.: Гидрометеоиздат, 1979. – 255 с.
23. *Привальский В.Е., Панченко В.А., Асарина Е.Ю.* Модели временных рядов с приложениями в гидрометеорологии. – СПб.: Гидрометеоиздат, 1992. – 226 с.
24. *Пузаченко Ю.Г.* Математические методы в экологических и географических исследованиях. – М.: Академия, 2004. – 407 с.
25. *Рожков В.А.* Теория и методы статистического оценивания вероятностных характеристик случайных величин и функций с гидрометеорологическими приложениями. Кн.1. – СПб.: Гидрометеоиздат, 2001. – 340 с.; Кн. 2 – 2002. – 440 с.
26. *Сикан А.В.* Методы статистической обработки гидрометеорологической информации. Учебник. – СПб.: РГГМУ, 2007. – 279 с.
27. *Смирнов Н.П., Вайновский П.А., Титов Ю.Э.* Статистический диагноз и прогноз океанологических процессов. – СПб.: Гидрометеоиздат, 1992. – 199 с.
28. *Смоленцев Н.К.* Основы теории вейвлетов. Вейвлеты в MATLAB. – Изд. 2-е, доп. и перераб. – М.: ДМК Пресс, 2005. – 304 с.
29. *Тьюки Дж.* Анализ результатов наблюдений. Разведочный анализ. – М.: Мир, 1981. – 693 с.
30. *Тюрин Ю.Н., Макаров А.А.* Статистический анализ данных на компьютере. – М.: ИНФРА-М, 1998. – 528 с.
31. *Хьюбер П.* Робастность в статистике. – М.: Мир, 1984. – 303 с.

Приложения

Приложение 1. Функция Лапласа

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,0000	0,33	0,1293	0,66	0,2454
0,01	0,0040	0,34	0,1331	0,67	0,2486
0,02	0,0080	0,35	0,1368	0,68	0,2517
0,03	0,0120	0,36	0,1406	0,69	0,2549
0,04	0,0160	0,37	0,1443	0,70	0,2580
0,05	0,0199	0,38	0,1480	0,71	0,2611
0,06	0,0239	0,39	0,1517	0,72	0,2642
0,07	0,0279	0,40	0,1554	0,73	0,2673
0,08	0,0319	0,41	0,1591	0,74	0,2703
0,09	0,0359	0,42	0,1628	0,75	0,2734
0,10	0,0398	0,43	0,1664	0,76	0,2764
0,11	0,0438	0,44	0,1700	0,77	0,2794
0,12	0,0478	0,45	0,1736	0,78	0,2823
0,13	0,0517	0,46	0,1772	0,79	0,2852
0,14	0,0557	0,47	0,1808	0,80	0,2881
0,15	0,0596	0,48	0,1844	0,81	0,2910
0,16	0,0636	0,49	0,1879	0,82	0,2939
0,17	0,0675	0,50	0,1915	0,83	0,2967
0,18	0,0714	0,51	0,1950	0,84	0,2995
0,19	0,0753	0,52	0,1985	0,85	0,3023
0,20	0,0793	0,53	0,2019	0,86	0,3051
0,21	0,0832	0,54	0,2054	0,87	0,3078
0,22	0,0871	0,55	0,2088	0,88	0,3106
0,23	0,0910	0,56	0,2123	0,89	0,3133
0,24	0,0948	0,57	0,2157	0,90	0,3159
0,25	0,0987	0,58	0,2190	0,91	0,3186
0,26	0,1026	0,59	0,2224	0,92	0,3212
0,27	0,1064	0,60	0,2257	0,93	0,3238
0,28	0,1103	0,61	0,2291	0,94	0,3264
0,29	0,1141	0,62	0,2324	0,95	0,3289
0,30	0,1179	0,63	0,2357	0,96	0,3315
0,31	0,1217	0,64	0,2389	0,97	0,3340
0,32	0,1255	0,65	0,2422	0,98	0,3365

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,99	0,3389	1,36	0,4131	1,73	0,4582
1,00	0,3413	1,37	0,4147	1,74	0,4591
1,01	0,3438	1,38	0,4162	1,75	0,4599
1,02	0,3461	1,39	0,4177	1,76	0,4608
1,03	0,3485	1,40	0,4192	1,77	0,4616
1,04	0,3508	1,41	0,4207	1,78	0,4625
1,05	0,3531	1,42	0,4222	1,79	0,4633
1,06	0,3554	1,43	0,4236	1,80	0,4641
1,07	0,3577	1,44	0,43	1,81	0,4649
1,08	0,3599	1,45	0,4265	1,82	0,4656
1,09	0,3621	1,46	0,4279	1,83	0,4664
1,10	0,3643	1,47	0,4292	1,84	0,4671
1,11	0,3665	1,48	0,4306	1,85	0,4678
1,12	0,3686	1,49	0,4319	1,86	0,4686
1,13	0,3708	1,50	0,4332	1,87	0,4693
1,14	0,3729	1,51	0,4345	1,88	0,4699
1,15	0,3749	1,52	0,4357	1,89	0,4706
1,16	0,3770	1,53	0,437	1,90	0,4713
1,17	0,3790	1,54	0,4382	1,91	0,4719
1,18	0,3810	1,55	0,4394	1,92	0,4726
1,19	0,3830	1,56	0,4406	1,93	0,4732
1,20	0,3849	1,57	0,4418	1,94	0,4738
1,21	0,3869	1,58	0,4429	1,95	0,4744
1,22	0,3883	1,59	0,4441	1,96	0,475
1,23	0,3907	1,60	0,4452	1,97	0,4756
1,24	0,3925	1,61	0,4463	1,98	0,4761
1,25	0,3944	1,62	0,4474	1,99	0,4767
1,26	0,3962	1,63	0,4484	2,00	0,4772
1,27	0,398	1,64	0,4495	2,02	0,4783
1,28	0,3997	1,65	0,4505	2,04	0,4793
1,29	0,4015	1,66	0,4515	2,06	0,4803
1,30	0,4032	1,67	0,4525	2,08	0,4812
1,31	0,4049	1,68	0,4535	2,10	0,4821
1,32	0,4066	1,69	0,4545	2,12	0,483
1,33	0,4082	1,70	0,4554	2,14	0,4838
1,34	0,4099	1,71	0,4564	2,16	0,4846
1,35	0,4115	1,72	0,4573	2,18	0,4854

x	$\Phi(x)$
2,20	0,4861
2,22	0,4868
2,24	0,4875
2,26	0,4881
2,28	0,4887
2,30	0,4893
2,32	0,4898
2,34	0,4904
2,36	0,4909
2,38	0,4913
2,40	0,4918
2,42	0,4922
2,44	0,4927
2,46	0,4931
2,48	0,4934
2,50	0,4938

x	$\Phi(x)$
2,52	0,4941
2,54	0,4945
2,56	0,4948
2,58	0,4951
2,60	0,4953
2,62	0,4956
2,64	0,4959
2,66	0,4961
2,68	0,4963
2,70	0,4965
2,72	0,4967
2,74	0,4969
2,76	0,4971
2,78	0,4973
2,80	0,4974
2,82	0,4976

x	$\Phi(x)$
2,84	0,4977
2,86	0,4979
2,88	0,498
2,90	0,4981
2,92	0,4982
2,94	0,4984
2,96	0,4985
2,98	0,4986
3,00	0,49865
3,20	0,49931
3,40	0,49966
3,60	0,499841
3,80	0,499928
4,00	0,499968
4,50	0,499997
5,00	0,499997

Приложение 2. Распределение Пирсона χ^2

Число степеней свободы ν	Уровень значимости α (двусторонняя критическая область)					
	0,01	0,025	0,05	0,95	0,975	0,99
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,3	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

Приложение 3. Распределение Стьюдента

Число степеней свободы ν	Уровень значимости α (двусторонняя критическая область)					
	0,1	0,05	0,02	0,01	0,002	0,001
1	6,31	12,71	31,82	63,66	318,31	636,62
2	2,92	4,30	6,96	9,92	22,33	31,60
3	2,35	3,18	4,54	5,84	10,21	12,92
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,02	2,57	3,36	4,03	5,89	6,87
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3,50	4,79	5,41
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,02	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,74	2,11	2,57	2,90	3,65	3,97
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,72	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,50	3,79
23	1,71	2,07	2,50	2,81	3,48	3,77
24	1,71	2,06	2,49	2,80	3,47	3,75
25	1,71	2,06	2,49	2,79	3,45	3,73
26	1,71	2,06	2,48	2,78	3,43	3,71
27	1,70	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,47	2,76	3,41	3,67
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,16	3,37
∞	1,64	1,96	2,33	2,58	3,09	3,29
	0,05	0,025	0,01	0,005	0,001	0,0005
	Уровень значимости α (односторонняя критическая область)					

Приложение 4. Распределение Фишера

ν_2 \ ν_1	Уровень значимости $\alpha = 0{,}05$															
	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60
1	161	199	216	225	230	234	237	239	241	242	244	246	248	250	251	252
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,4	19,4	19,4	19,5	19,5	19,5
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,62	8,59	8,57
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,75	5,72	5,69
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,50	4,46	4,43
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,81	3,77	3,74
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,38	3,34	3,30
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,08	3,04	3,01
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,86	2,83	2,79
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,70	2,66	2,62
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,57	2,53	2,49
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,47	2,43	2,38
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,38	2,34	2,30
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,31	2,27	2,22
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,25	2,20	2,16
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,19	2,15	2,11
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,15	2,10	2,06
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,11	2,06	2,02
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,07	2,03	1,98
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,04	1,99	1,95
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	1,98	1,94	1,89
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,94	1,89	1,84
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,90	1,85	1,80
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,87	1,82	1,77
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,84	1,79	1,74
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,74	1,69	1,64
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,65	1,59	1,53
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,55	1,50	1,43
∞	3,84	3,00	3,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,46	1,39	1,32

Приложение 5. Значения величины *z* для значений коэффициента корреляции *r* от 0,00 до 0,99

r	0	1	2	3	4	5	6	7	8	9
0,0	0,000	0,010	0,020	0,030	0,040	0,050	0,060	0,070	0,080	0,090
1	0,100	0,110	0,121	0,131	0,141	0,151	0,161	0,172	0,182	0,192
2	0,203	0,213	0,224	0,234	0,245	0,255	0,266	0,277	0,288	0,299
3	0,309	0,321	0,332	0,343	0,354	0,365	0,377	0,388	0,400	0,412
4	0,424	0,436	0,448	0,460	0,472	0,485	0,497	0,510	0,523	0,536
5	0,549	0,563	0,576	0,590	0,604	0,618	0,633	0,647	0,662	0,678
6	0,693	0,709	0,725	0,741	0,758	0,775	0,793	0,811	0,829	0,848
7	0,867	0,887	0,908	0,929	0,950	1,973	0,996	1,020	1,045	1,071
8	1,099	1,127	1,157	1,188	1,221	1,256	1,293	1,333	1,376	1,422
9	1,472	1,527	1,589	1,658	1,738	1,832	1,946	2,092	2,298	2,647

Приложение 6. Словарь статистических терминов

Амплитуда гармоники – разность между максимальным и минимальным значениями гармоники.

Анализ объективный – процедура перевода данных из нерегулярной сети точек в регулярную сетку.

Вариограмма – статистический момент второго порядка, использующийся в геостатистике для анализа и моделирования пространственной корреляции.

Величина случайная – переменная величина, которая в результате испытания (измерения) в одинаковых условиях может принимать то или иное заранее неизвестное значение.

Величина случайная количественная – случайная величина, выражаемая в метрической шкале.

Величина случайная ординальная – случайная величина, соответствующая порядковой (ординальной) шкале.

Величина случайная номинальная – случайная величина, соответствующая номинальной шкале

Величина случайная стандартизированная – величина, полученная преобразованием $t = (x - m_x)/s_x$, имеющая математическое ожидание равное нулю и дисперсию равную 1.

Величина случайная центрированная – отклонение от математического ожидания (среднего арифметического).

Вероятность доверительная – степень надежности определения истинной оценки по выборочной оценке.

Вероятность теоретическая – частота события, свойственная генеральной совокупности.

Вероятность эмпирическая – частота события, свойственная выборочной совокупности

Выборка (выборочная совокупность) – любая последовательность конечного объема, извлеченная из генеральной совокупности.

Выборка представительная – выборка, достаточно точно отражающая основные закономерности генеральной совокупности.

Выборочное среднее – сумма значений выборки, деленная на ее длину.

Выброс – резко отличающееся от других значение временного ряда.

Гармоника – слагаемое в разложении Фурье.

Генеральная совокупность – весь мыслимо возможный набор случайной величины.

Гетероскедастичность – непостоянство дисперсии остатков в регрессионной модели.

Гипотеза нулевая – предположение об отсутствии различий в тех или иных свойствах случайного процесса.

Гипотеза альтернативная – логическое отрицание нулевой гипотезы.

Гистограмма распределения – график, представляющий распределение частот по интервалам вариационного ряда.

Гомоскедастичность – постоянство дисперсии остатков в регрессионной модели.

Дециль – квантиль, соответствующий одной из вероятностей 0,10; 0,20; …; 0,90.

Дисперсия генеральная (выборочная) – мера изменчивости случайной величины в генеральной (выборочной) совокупности, имеющая размерность ее квадрата.

Доверительный интервал – область значений случайной величины внутри доверительных границ.

Закон распределения – любое соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями.

Закон распределения теоретический – распределение истинных значений вероятностей случайной величины.

Закон распределения эмпирический – распределение вероятностей, полученных из опытных (эмпирических) данных достаточно большого объема.

Изокорреляты – линии равной корреляции.

Интервал дискретизации – промежуток времени, через который берется временной ряд.

Интерквартильное расстояние – разность между верхним $x_{0,75}$ и нижним $x_{0,25}$ квартилями.

Информация – любые сведения (в количественной и качественной форме) об исследуемом объекте.

Информация первичная – результаты непосредственного измерения характеристик природной среды.

Информация вторичная – результаты расчетов, выполненных на основе первичной информации.

Квантиль – элемент ряда, при котором функция распределения принимает значение, равное вероятности p.

Квантильный анализ – непараметрический метод анализа малых выборок, основанный на использовании квантилей.

Квартиль – квантиль, соответствующий одной из вероятностей: 0,25, 0,50, 0,75.

Когерентность – характеристика линейной статистической связи спектральных компонент одинаковой частоты.

Колебание циклическое – колебание, параметры которого (период, амплитуда, фаза) испытывают нерегулярные изменения во времени в пределах некоторого диапазона.

Коррелограмма – график автокорреляционной функции.

Корреляционное отношение – безразмерная мера нелинейной связи двух случайных величин.

Корреляционное поле – график двух случайных величин в декартовой системе координат.

Корреляция ложная – линейная стохастическая связь между двумя переменными, если они связаны с третьей переменной.

Корреляция ранговая – линейная стохастическая связь между порядковыми переменными.

Корреляция сериальная – линейная стохастическая связь между остатками в регрессионной модели.

Корреляция частная – корреляция зависимой переменной с какой-либо независимой переменной после исключения влияния на нее присутствующих в модели переменных.

Косинус-спектр – вещественная часть взаимной спектральной функции.

Коэффициент автокорреляции – коэффициент корреляции между значениями данного ряда и его же значениями, относящимися к некоторому сдвигу τ.

Коэффициент автокорреляции частный – коэффициент корреляции между значениями данного ряда и его же значениями, относящимися к некоторому сдвигу τ после исключения влияния на него присутствующих в модели переменных.

Коэффициент асимметрии – безразмерная характеристика скошенности кривой плотности распределения случайной величины.

Коэффициент вариации – безразмерная мера изменчивости случайной величины в генеральной (выборочной) совокупности.

Коэффициент взаимной корреляции – коэффициент корреляции двух переменных при некотором сдвиге одной из них относительной другой.

Коэффициент детерминации линейный – доля объясненной дисперсии функции отклика в уравнении линейной парной (множественной) регрессии.

Коэффициент детерминации нелинейный – доля объясненной дисперсии функции отклика в уравнении нелинейной (полиномиальной) регрессии.

Коэффициент детерминации частный – доля остаточной дисперсии функции отклика в уравнении множественной регрессии, объясненная включением дополнительной переменной в модель.

Коэффициент корреляции бисериальный – непараметрическая безразмерная характеристика линейной взаимосвязи качественной альтернативной (да, нет) и количественной переменных.

Коэффициент корреляции Кендалла – непараметрическая безразмерная характеристика линейной взаимосвязи двух случайных величин.

Коэффициент корреляции множественный – безразмерная параметрическая характеристика линейной взаимосвязи фактических и вычисленных по модели множественной регрессии функции отклика

Коэффициент корреляции парный – безразмерная параметрическая характеристика линейной взаимосвязи двух случайных величин.

Коэффициент корреляции Спирмена – непараметрическая безразмерная характеристика линейной взаимосвязи двух случайных величин.

Коэффициент корреляции частный – мера линейной связи функции отклика с независимой переменной в модели множественной регрессии после исключения влияния на нее остальных переменных.

Коэффициент регрессии – коэффициент пропорциональности между зависимой и независимой переменными.

Коэффициент эксцесса – характеристика крутости кривой плотности распределения случайной величины.

Критерий Дарбина-Уотсона – безразмерная характеристика взаимосвязи между смежными значениями остатков функции отклика.

Критерий статистический – свод правил, указывающих, при каких результатах наблюдений нулевая гипотеза отклоняется.

Критерий Стьюдента регрессионный – критерий для оценки значимости параметров модели регрессии.

Критерий Фишера регрессионный – критерий для оценки адекватности (значимости) модели регрессии.

Критерий Фишера частный – обычный F-критерий для каждой переменной при условии, что она оказывается последней переменной, включенной в модель регрессии.

Критическая область статистики – область значений статистики, вероятность которых меньше заданного уровня значимости.

Кумулятивная кривая – сумма накопленных частот, показывающая степень приближения к 1 или 100 % ряда распределения.

Линеаризация – процедура перевода нелинейной зависимости к линейному виду.

Медиана – величина, занимающая среднее положение вариационного ряда.

Математическое ожидание – центр распределения генеральной совокупности случайной величины.

Метод наименьших квадратов – метод отыскания неизвестных коэффициентов эмпирической зависимости.

Мода – наиболее часто встречающаяся в данном статистическом ряду величина.

Модель авторегрессионная – модель временного ряда, в которой его значения в данный момент времени линейно зависят от предыдущих значений этого же ряда.

Момент начальный порядка *k* – математическое ожидание случайной величины x^k.

Момент центральный порядка *k* – математическое ожидание центрированной величины $(x - m_x)^k$.

Мощность критерия – вероятность попадания заданной статистики в критическую область, когда верна альтернативная гипотеза.

Мультиколлинеарность реальная – корреляционная матрица, в которой между большинством исходных переменных отмечается высокая коррелированность.

Мультиколлинеарность строгая – корреляционная матрица, в которой имеет место линейная функциональная связь хотя бы между двумя переменными.

Невязка (дисбаланс) – суммарная погрешность определения всех компонент какого-либо уравнения.

Область допустимых значений статистики – область значений статистики, вероятность которых больше заданного уровня значимости.

Огива – кривая суммы накопленных частот, обратная кумулятивной кривой.

Оценка – любое числовое значение случайной величины или случайной функции.

Оценка адекватности регрессионной модели – проверка значимости регрессионной модели по критерию Фишера.

Оценка значимости коэффициента корреляции – проверка нулевой гипотезы на равенство коэффициента корреляции нулю.

Оценка несмещенная – оценка, для которой ее математическое ожидание равно оцениваемому параметру.

Оценка робастная – оценка, которая является устойчивой к существенным отклонениям в значениях данных.

Оценка состоятельная – оценка, которая при неограниченном возрастании объема выборки сходится по вероятности к оцениваемому параметру.

Оценка эффективная – оценка, которая при заданном объеме выборки имеет наименьшую дисперсию среди всех возможных несмещенных оценок.

Оценивание – определение числовых характеристик или свойств случайной величины или случайной функции.

Оценивание точечное – определение конкретных оценок выборочного параметра, около которого находятся его истинные значения.

Оценивание интервальное – определение диапазона оценок выборочного параметра, внутри которого с большой заданной вероятностью находится его истинное неизвестное значение.

Оценивание гипотез параметрическое – проверка гипотез, когда предполагается известным вид функции распределения и отдельные параметры, а проверка относится к неизвестному параметру.

Оценивание гипотез непараметрическое – проверка гипотез, когда знание законов распределения случайной величины не требуется.

Персентиль – квантиль, соответствующий одной из вероятностей 0,01; 0,02; …; 0,99.

Погрешность – ошибка измерений или расчетов.

Погрешность грубая – погрешность, резко выделяющаяся от всех других.

Погрешность косвенная – погрешность, которая может быть вычислена через измеряемые параметры.

Погрешность модели среднеквадратическая – случайная ошибка описания функции отклика в регрессионной модели.

Погрешность систематическая – погрешность, изменяющаяся по определенному закону.

Погрешность случайная – погрешность, которая при испытаниях в одинаковых условиях меняется произвольным образом.

Поле случайное – случайная функция, изменяющаяся в пространстве.

Поле случайное однородное (в широком смысле) – поле, для которого математическое ожидание является постоянной величиной, а корреляционная функция зависит только от одного аргумента – разности векторов $l = \rho_2 - \rho_1$.

Поле случайное однородное (в узком смысле) – поле, для которого все его n-мерные законы распределения не изменяются при переносе системы точек $\rho_1, \rho_2, \ldots, \rho_n$ на один и тот же вектор.

Поле случайное однородное изотропное – поле, для которого все его n-мерные законы распределения не изменяются при всевозможных вращениях системы точек $N_1(\rho_1), N_2(\rho_2), \ldots, N_n(\rho_n)$ вокруг любой оси, проходящей через начало координат, и при зеркальном их отражении относительно любой плоскости, проходящей через начало координат.

Поле случайное однородное изотропное, обладающее эргодическим свойством – поле, для которого пространственное среднее и корреляционная функция, полученная осреднением по одной реализации поля при безграничном увеличении диаметра области, может быть с вероятностью, сколь угодно близкой к единице, приближены к соответствующим характеристикам, полученным осреднением по всему множеству реализаций.

Полигон распределения – ломаная линия, соединяющая частоты вариационного ряда.

Преобразование Фишера – функциональное преобразование вида $z = 0{,}5 \times \ln(1 + r) / (1 - r)$, позволяющее нормализовать коэффициенты корреляции при их большой величине и малой длине выборки.

Процесс гармонический – колебание, все основные параметры которого (амплитуда, период, фаза) остаются строго постоянными во времени.

Процесс детерминированный – временной ряд, значения которого изменяются по строго определенному, как правило, физическому закону.

Процесс случайный – случайная функция, изменяющаяся во времени.

Процесс случайный стационарный (в широком смысле) – процесс, у которого выборочные оценки среднего и дисперсии случайного процесса постоянны во времени и соответствуют математическому ожиданию и генеральной дисперсии, а его автокорреляционная функция является только функцией интервала времени $\tau = t_2 - t_1$ и не зависит от значения каждого аргумента t_1 и t_2 в отдельности.

Процесс случайный стационарный (в узком смысле) – процесс, у которого многомерные распределения при одновременном прибавлении ко всем аргументам t_1, t_2, … , t_n одного и того же сдвига τ остаются неизменными.

Процесс случайный стационарный эргодический – процесс, одна реализация которого достаточно большой длины содержит в себе фактически всю информацию об основных свойствах случайного процесса, т. е. может заменить при обработке множество реализаций той же продолжительности.

Разложение Фурье – представление в виде тригонометрического ряда по синусам и косинусам, обеспечивающее при его фиксированной длине наименьшую среднеквадратическую ошибку.

Разложение Чебышева – представление в виде ортогональных многочленов, позволяющее производить добавление новых слагаемых более высокого порядка, не изменяя при этом вычисленные ранее коэффициенты.

Размах – разность между максимальным и минимальным значениями выборки (ряда).

Ранг матрицы – наибольший порядок ее отличного от нуля минора, совпадающий с максимальным числом линейно независимых столбцов матрицы.

Ранг случайной величины – порядковый номер значения признака ранжированного ряда.

Реализация случайной функции – конкретный вид, который случайная функция принимает в результате испытаний или наблюдений.

Регрессионный анализ – совокупность статистических методов исследования влияния одной или нескольких независимых переменных на зависимую переменную.

Регрессия множественная линейная – уравнение, описывающее линейную зависимость функции отклика от множества факторов.

Регрессия парная линейная – уравнение линейной зависимости между двумя случайными переменными.

Регрессия полиномиальная одномерная – уравнение, описывающее нелинейную зависимость функции отклика от одного фактора.

Регрессия полиномиальная двухмерная – уравнение, описывающее нелинейную зависимость функции отклика от двух факторов.

Регрессия пошаговая – процедура отбора наиболее существенных переменных в многофакторной регрессионной модели.

Регрессия робастная – регрессия, устойчивая к выбросам в исходных данных.

Ряд временной – конечная реализация случайной величины, расположенная в хронологическом порядке.

Ряд распределения атрибутивный – ряд распределения, построенный по качественному признаку.

Ряд распределения вариационный – ряд распределения, построенный в порядке возрастания по количественному признаку.

Ряд распределения статистический – упорядоченное распределение единиц совокупности на группы по определенному варьирующему признаку.

Серия – участок двух совмещенных вариационных рядов, состоящий из идущих подряд одинаковых кодов и ограниченный с обеих сторон противоположными кодами

Связь случайная – если каждому значению одной переменной может соответствовать практически любое значение другой переменной.

Связь стохастическая – если каждому значению одной переменной с определенной вероятностью соответствует значение другой переменной.

Связь функциональная – если каждому значению одной переменной соответствует единственное значение другой переменной.

Скользящее осреднение – вид фильтрации временного ряда, основанный на последовательном осреднении членов ряда за интервал сглаживания.

Спектр амплитудный – модуль взаимной спектральной плотности.

Спектр квадратурный – мнимая часть взаимной спектральной функции.

Спектр фазовый – характеристика отставания по фазе одного случайного процесса от другого.

Спектральная плотность – прямое преобразование Фурье автокорреляционной функции.

Спектральная плотность взаимная – прямое преобразование Фурье взаимной корреляционной функцией двух стационарных случайных процессов.

Спектральная плотность нормированная – прямое преобразование Фурье нормированной автокорреляционной функции.

Спектральная плотность взаимная нормированная – прямое преобразование Фурье нормированной взаимной корреляционной функцией двух стационарных случайных процессов.

Спектрограмма – график кривой спектральной плотности.

Среднеквадратическое (стандартное) отклонение генеральное (выборочное) – мера изменчивости случайной величины в генеральной (выборочной) совокупности, имеющая размерность случайной величины.

Статистика параметрическая – статистика, требующая предварительного знания теоретического закона распределения при проверке свойств случайной величины.

Статистика непараметрическая – статистика, не требующая предварительного знания теоретического закона распределения при проверке свойств случайной величины.

Статистика порядковая – каждый член вариационного ряда.

Схема расположения точек равномерная – плотность точек в любой подобласти пространства равна плотности точек во всех других подобластях.

Схема расположения точек регулярная – точки образуют в пространстве какой-либо вид сети, расстояния в которой между любыми точками подчиняются определенному закону.

Схема расположения точек случайная – точки в пространстве размещены совершенно произвольным образом и появление (исключение) одной или нескольких точек никак не сказывается на характере распределения всей совокупности точек в целом.

Точка отсечения автокорреляционной функции – максимальный сдвиг, до которого осуществляется расчет автокорреляционной функции.

Тренд – медленное изменение случайного процесса без образования циклов.

Тренд линейный – линейное уравнение, описывающее зависимость искомого параметра от времени.

Тренд нелинейный – нелинейное уравнение, описывающее зависимость искомого параметра от времени.

Уровень значимости – вероятность события, которым решено пренебречь в данном исследовании.

Фаза гармоники – временной интервал наступления первого максимума от начала отсчета.

Фильтрация ряда высокочастотная – подавление долгопериодных колебаний и выделение высокочастотных колебаний временного ряда до определенного предела частоты.

Фильтрация ряда низкочастотная – выделение долгопериодных колебаний и подавление высокочастотных колебаний временного ряда.

Фильтрация ряда полосовая – выделение колебаний в определенном диапазоне частот временного ряда.

Функция автокорреляционная – неслучайная функция двух независимых фиксированных аргументов, равная корреляционному моменту сечений этих аргументов.

Функция автокорреляционная нормированная – безразмерная функция линейной связи между сечениями случайной функции.

Функция корреляционная взаимная – неслучайная функция двух случайных процессов, соответствующих одному и тому же значению аргумента.

Функция корреляционная взаимная нормированная – безразмерная характеристика линейной связи между сечениями двух случайных функций.

Функция кросскорреляционная – безразмерная характеристика линейной связи между значениями случайного поля в различных точках пространства и в различные моменты времени.

Функция обеспеченности эмпирическая – закон изменения частоты события $X \geq x$ в статистической выборке.

Функция распределения дифференциальная – предел отношения вероятности попадания случайной величины X в интервал $[x, x + \Delta x]$ к величине Δx при $\Delta x \to 0$.

Функция распределения интегральная – вероятность того, что случайная величина примет значение, которое изображается на числовой оси точкой, лежащей левее точки x.

Функция распределения эмпирическая – закон изменения частоты события $X < x$ в статистической выборке.

Функция случайная – функция, значения которой при каждом значении аргумента представляют случайную величину.

Функция спектральная – функция, характеризующую интегральную долю дисперсии, приходящейся на некоторый интервал частот.

Функция частотная весовая – функция, которая уравнивает выборочные и истинные оценки автокорреляционной функции.

Характеристика фильтра частотная – функция, определяющая характер изменения амплитуд случайного процесса при прохождении ряда через фильтр.

Цепь Маркова – последовательность событий $A_i^{(t)}$ называется цепью Маркова k-того порядка, если для каждого момента времени t условная вероятность события $A_i^{(t+1)}$ зависит только от

того, какие события произошли в k предыдущих моментах времени и не зависит от поведения последовательности до момента $t - k + 1$.

Цепь Маркова простая – последовательность событий, при котором вероятность любого состояния системы в будущем зависит только от состояния системы в настоящий момент и не зависит от того, каким образом эта система пришла в это состояние.

Цепь Маркова сложная – последовательность событий, при котором вероятность системы в настоящий момент зависит от некоторого множества состояний системы в предшествующие моменты времени.

Частота – эмпирическая повторяемость, выражаемая суммой значений случайной величины в каждой группе вариационного ряда.

Частость – относительная частота (в долях единицы).

Число степеней свободы – количество значений выборки, функционально не связанных между собой.

Шум белый – случайный процесс, представляющий набор случайных чисел, некоррелированных друг с другом.

Шум красный – случайный процесс, которому свойственна корреляция только между смежными (соседними) значениями временного ряда.

Содержание

Часть 2. Построение эмпирических зависимостей

Приложения

Учебное издание

Малинин Валерий Николаевич, д.г.н., проф.

Статистические методы анализа гидрометеорологической информации

Том 1. Первичный анализ и построение эмпирических зависимостей

Начальник РИО А.В. Ляхтейнен
Редактор Л.Ю. Кладова
Верстка М.В. Ивановой

Подписано в печать 14.10.2020. Формат 60×90 $^{1}/_{16}$. Гарнитура Times New Roman.
Печать цифровая. Усл. печ. л. 16. Тираж 50 экз. Заказ № 961/1.
РГГМУ, 192007, Санкт-Петербург, Воронежская ул., 79.