Министерство образования и науки Российской Федерации
Федеральное агентство по образованию
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГИДРОМЕТЕОРОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ

В.Н. Малинин

СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ГИДРОМЕТЕОРОЛОГИЧЕСКОЙ ИНФОРМАЦИИ

Рекомендовано Учебно-методическим объединением в области гидрометеорологии в качестве учебника для студентов высших учебных заведений, обучающихся по специальности «Океанология» и другим гидрометеорологическим специальностям

РГГМУ
Санкт-Петербург
2008

УДК 551.46:519.2

Малинин В.Н. Статистические методы анализа гидрометеорологической информации. Учебник.– СПб.: изд. РГГМУ, 2008. – 408 с.

ISBN 978-5-86813-213-1

Дается систематизированное изложение начальных основ математической статистики применительно к решению гидрометеорологических задач. Теоретические сведения иллюстрируются значительным числом конкретных примеров. Рассматриваются первичный анализ данных, методы построения эмпирических зависимостей, анализ временных рядов и пространственных полей.

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, methods of empirical constraint creation and analysis of temporal series and spatial fields.

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.

Рецензенты: Д.А. Субетто, д-р геогр. наук, проф., завкафедрой физической географии и природопользования РГГПУ им. А.И. Герцена,
Ю.А. Трапезников, д-р физ.-мат. наук, ведущий науч. сотр. Института озероведения РАН

ISBN 978-5-86813-213-1

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.

С философской точки зрения, детерминированность и случайность – это объективная реальность, две противоположные категории, характеризующие с разных сторон процессы и явления окружающего нас мира. При этом каждая из указанных категорий может рассматриваться как предельный случай реализации (проявления) другой категории. Действительно, если хаос (случайность) есть нечто иное, как предельное проявление детерминированности, то одновременно жесткий порядок – это предельное состояние (с вероятностью равной единице) случайности.

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 основой его служит 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 не-

которых ППСП содержание Help столь полно, что его можно рассматривать как оперативное руководство по статистике, адаптированное к конкретному программному продукту. К сожалению, для большинства ППСП, как правило, иностранных, практически 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 78 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 лишь для некоторых из них в двух 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 в книге 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 книге рассматриваются главным образом те аспекты математической статистики, которые, по мнению автора, нашли широкое применение в гидрометеорологии и соответственно входят в программу курса «Методы статистического анализа гидрометеорологической информации». Исходя из этого, содержание данной книги включает в себя четыре раздела:

– первичный анализ данных;

– построение эмпирических зависимостей;

– анализ временных рядов;

– анализ пространственных 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 положения, то оно, вообще говоря, вполне уместно. 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. 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 веков 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 в книге материал полностью соответствует программе дисциплины «Методы статистического анализа гидрометеорологической информации», которая на протяжении уже нескольких десятилетий читается автором на океанологическом факультете РГГМУ. Предполагается, что в последующем будет издано учебное пособие «Основы многомерного статистического анализа», курс которого читается для магистров и является естественным продолжением данной книги.

Большинство примеров подготовлено специально для книги, некоторые взяты из научно-исследовательских работ или практических работ студентов. Автор безмерно благодарен своему многолетнему помощнику, старшему преподавателю кафедры промысловой океанологии С.М. Гордеевой, подготовившей большинство примеров, аспиранту О.И. Шевчуку за помощь в вычислениях и компьютерной подготовке рисунков к печати, а также другим аспирантам и студентам, общение с которыми постоянно стимулирует автора к более доступному изложению даже довольно сложных статистических методов.

Кроме того, автор признателен рецензентам: д-ру геогр. наук, проф., зав кафедрой физической географии и природопользования РГГПУ им. А.И. Герцена Д.А. Субетто и д-ру физ.-мат. наук, ведущему научн. сотр. Института озероведения РАН Ю.А. Трапезникову за конструктивные советы и полезные замечания по улучшению рукописи учебника.

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 вариантами, обозначаются строчными буквами с индексом. Например, если случайная величина X имеет n возможных значений, то они будут обозначены как: $x_1, x_2, \ldots, x_n$.

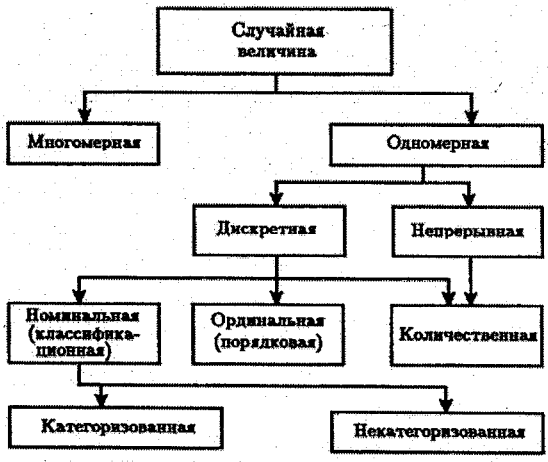


Рис. 1.1. Классификация случайной величины.

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 ре-

зультатом измерения (испытания) является регистрация целого набора характеристик, то случайную величину называют *многомерной*, которая всегда уже является 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, уже 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 и бесконечным. Примером могут служить многие 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, ледовитость 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 едини-

цами 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 и 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. 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 регла-

ментируется с помощью следующих трех свойств статистических оценок: несмещенности, состоятельности и эффективности.

Несмещенность. Оценка параметра θ называется несмещенной, если ее математическое ожидание, т.е. центр распределения генеральной совокупности случайной величины, равно истиной величине оцениваемого параметра. Сказанное можно записать как $M(\theta) = \boldsymbol{\theta}$. В противном случае оценка является смещенной. Если это равенство не выполняется, то оценка θ, полученная по разным выборкам, будет либо завышать значение параметра $\boldsymbol{\theta}$, либо занижать его. Следовательно, требование несмещенности гарантирует отсутствие систематических ошибок при оценивании параметров. По существу, требование несмещенности означает, что выборочная средняя должна совпадать с ее истинной оценкой, т.е. с математическим ожиданием. В результате имеем $\bar{\theta} = M(\theta) = m_\theta$.

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, т.е.

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

где 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 $\boldsymbol{\theta}$ 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, если оценка состоятельна, то с большой степенью достоверности можно считать, что при значительном объеме выборки $\theta \approx \boldsymbol{\theta}$.

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 же 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 не совпадает с эффективной оценкой параметра другого 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. Поэтому для выборок небольшого объема ($n < 25$–30) 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 формах.

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 *полиго-*

ном, 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. выборку, построенную в порядке возрастания ее отдельных значений.

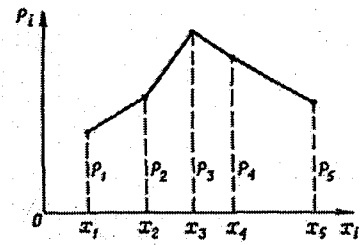


Рис. 1.2. Многоугольник (полигон) распределения.

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

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

$$\mathbf{F}(x) = p(X < x), \qquad (1.1)$$

где p – вероятность, понимаемая применительно к выборочным данным как *частота события*.

Геометрически это равенство можно истолковать так: $\mathbf{F}(x)$ 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 собой наиболее общую форму описания случайной величины и полностью характеризует ее с вероятностной точки зрения.

Функция распределения обладает следующими основными свойствами:

Свойство 1. Значения функции распределения заключены в диапазоне [0,1], т.е. $0 \le \mathbf{F}(x) \le 1$.

Это означает, что $\mathbf{F}(x) = 0$, если $x \to -\infty$ и $\mathbf{F}(x) = 1$, если $x \to \infty$.

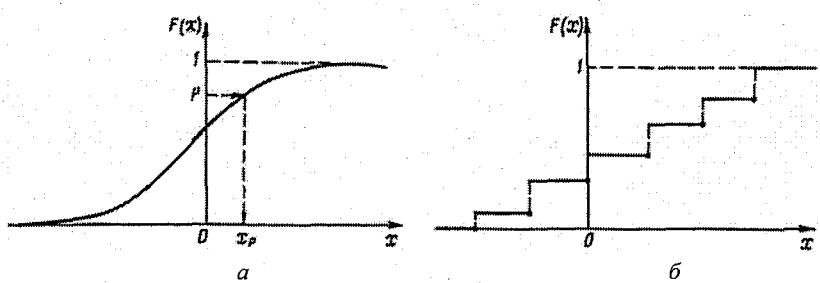
Свойство 2. Функция $\mathbf{F}(x)$ является неубывающей, т.е. если $x_1 < x_2$, то $\mathbf{F}(x_1) \le \mathbf{F}(x_2)$.

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 $F(x) = 0$ при $x \le a$ и $F(x) = 1$ при $x \ge b$.



а *б*

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

Для непрерывной случайной величины график этой функции представляет собой 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 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:

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

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

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 $\mathbf{F}(x)$:

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

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*.

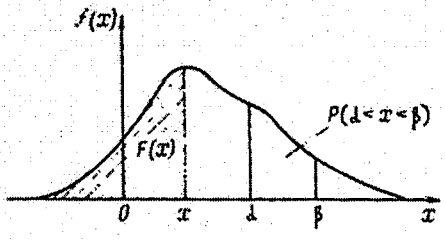


Рис. 1.4. График дифференциальной функции распределения.

К основным свойствам плотности распределения относятся:

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

Свойство 2. Интеграл от плотности распределения в бесконечных пределах равен 1, т.е.

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

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*. Однако во многих случаях истинные оценки вероятностей неизвестны, но 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, что

$$F(x) = n_x / n, \qquad (1.4)$$

где n_x – число вариант (значений), меньших x.

Таким образом, различие между $\mathbf{F}(x)$ и $F(x)$ состоит в том, что первая функция определяет вероятность события $X < x$, а вторая – относительную частоту этого же события. Заметим также, что свойства эмпирической 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 уже 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 ряде 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 тем или иным методом. Отметим, что основной в этом случае является вторичная информация по испарению.

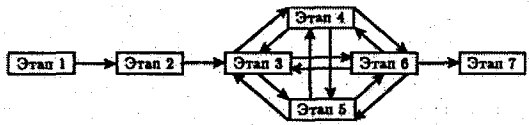


Рис. 1.5. Блок-схема структурных взаимосвязей основных этапов статистического анализа.

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

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 возни-

кали 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 данной книге. 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, этапы 3–6) 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 измерения их 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 и уровню. Очень слабо изучены 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. 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). 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 900 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 на 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 50-х годов. В Северной Атлантике работу осуществляли 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., 2° в.д.). Дело в том, что акватория Норвежского моря относится к числу важнейших энергоактивных зон 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 гидро-

логических, гидрохимических, а также метеорологических и даже аэрологических наблюдений. Общий период наблюдений, начатых в 1951 г., уже превышает 50 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 70-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 прошествии годового интервала времени публиковались атласы.

Кроме традиционных видов океанографических работ, в последние десятилетия все большее 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 температуры поверхности океана 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 °С. 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 табл. 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	Временная дискретность данных
NOAA NCEP/ NCAR(1) CDAS	1,875×1,875	Глобальная	с 1949	1 месяц
COADS	2×2	Глобальная	1854–1992	1 месяц
NOAA NCEP CMB GLOBAL (RESM)	1×1	Глобальная	с декабря 1981	5 дней 7 дней 1 месяц
UKMO	5×5	Глобальная	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.

1.7. Общие сведения о временных рядах

При решении широкого круга гидрометеорологических задач наибольшее распространение получили временные ряды. Под *временным рядом обычно понимают некоторую конечную реализа-*

цию случайной величины, расположенную в хронологическом порядке. При этом его значения могут быть взяты как через равные, так и через неравные промежутки времени. В первом случае временной ряд называется *эквидистантным,* а во втором – *неэквидистантным.* Основной задачей нашего анализа являются эквидистантные (равноотстоящие) ряды. Следует иметь в виду, что существуют принципиальные отличия временного ряда от последовательности наблюдений $x_1, x_2, \ldots, x_n$, образующих случайную выборку. Эти отличия заключаются в следующем:

– члены временного ряда по сравнению с элементами случайной выборки не являются статистически независимыми;

– члены временного ряда не являются одинаково распределенными.

Это означает, что мы не можем распространять все свойства и правила статистического анализа случайной выборки на временные ряды. Действительно, в случайной выборке все ее значения могут быть представлены уже не в хронологическом, а в произвольном порядке. Вследствие этого происходит нарушение свойства внутренней корреляции (автокоррелированности), во многих случаях характерной для временных рядов. Отсюда вытекают и различия в формировании функции распределения различных сечений временной последовательности.

Промежутки времени, через которые берется временной ряд, называют *интервалом (шагом) дискретизации* Δt. Естественно, что возможен самый широкий спектр значений Δt: минута, час, сутки, месяц, год, сто лет, тысяча лет и т.д. Тогда каждому из интервалов дискретизации соответствует свой временной ряд, описывающий процессы того или иного масштаба (периода).

Колебания во времени гидрометеорологических величин создаются многообразными физическими процессами, протекающими в океане и атмосфере Земли, а также различными геофизическими силами (например, приливные силы, движение полюса Земли и т. п.). Исходя из понятия о спектрах временных колебаний, А.С. Мониным была предложена классификация океанологических процессов, состоящая из семи интервалов (классов).

1. *Мелкомасштабные явления* (периоды от долей секунды до десятков минут). К ним относятся: поверхностные и внутренние волны, турбулентность и процессы эволюции вертикальной микроструктуры океана.

2. *Мезомасштабные явления* (периоды от часов до суток) – приливные и инерционные колебания, возникающие под воздействием сил гравитационного притяжения Луны и Солнца и сил инерции при вращательном движении планеты.

3. *Синоптическая изменчивость* (периоды от нескольких суток до месяцев), заключающаяся прежде всего в непериодическом формировании в океане вихрей с масштабами порядка 100 км и процессами гидродинамической неустойчивости крупномасштабных океанских течений.

4. *Сезонные колебания* (годовой период и его гармоники), которые наиболее отчетливо проявляются в высоких широтах, а также в муссонной зоне Индийского океана.

5. *Межгодовая изменчивость* (периоды в несколько лет) – квазидвухлетний цикл, обнаруженный в колебаниях температуры воды, в интенсивности течений и других процессах взаимодействия океана с атмосферой; автоколебания системы океан – атмосфера, связанные с перемещением тепловых аномалий по гигантским океаническим круговоротам; явление Эль-Ниньо – Южное колебание (ЭНЮК), а также другие явления.

6. *Внутривековая изменчивость* (периоды в десятки лет), взаимосвязанная с внутривековыми колебаниями климата. Пример – потепление Арктики в 20–40-х годах прошлого столетия.

7. *Межвековая изменчивость* (периоды в сотни лет и более), взаимосвязанная с межвековыми колебаниями климата. Примером может служить так называемый «малый ледниковый период» (XVII–XIX вв.).

Разумеется, физические закономерности многих из отмеченных выше явлений и процессов настолько многообразны и сложны, что даже не всегда возможно их точное математическое описание. Однако описание внутренних закономерностей временных рядов не представляет каких-либо принципиальных затруднений и может быть осуществлено на основе статистических методов.

Следует также отметить, что в подавляющем большинстве архивы (базы) гидрометеорологических данных акцентированы на хранение информации в виде временных рядов, что, естественно, имеет принципиальное значение с точки зрения их статистической обработки и анализа.

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 случайной величины. 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. *Точечное оценивание* заключается в том, что с помощью статистических методов определяются конкретные (точечные) оценки выборочного параметра, около которого находятся его истинные значения. 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 среднюю квадратическую 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 неизвестен 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 $d\ln L/d\theta$;

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

3) найти вторую производную $d^2\ln L/d\theta^2$ 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 случайной величины X называется математическое ожидание величины x^k

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

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 случайной величины X называется математическое ожидание центрированной величины $(x - m_x)^k$:

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

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 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:

$$\mu_1 = 0$$
$$\mu_2 = \alpha_2 - \alpha_1^2$$
$$\mu_3 = \alpha_3 - 3\alpha_1\alpha_2 + 2\alpha_1^3$$
$$\mu_4 = \alpha_4 - 4\alpha_1\alpha_3 + 6\alpha_1^3\alpha_2 - 3\alpha_1^4$$

Итак, суть метода моментов точечного оценивания неизвестных параметров заключается в приравнивании теоретических моментов к 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: $\alpha_1 = \alpha_1$. Тогда с учетом формул (2.1) и (2.3) имеем:

$$M(x) = \bar{x}. \qquad (2.4)$$

Отсюда следует, что точечной оценкой математического ожидания является среднее арифметическое, полученное по ограниченной выборке. Аналогичным образом, приравняв центральный 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. Одна группа (математическое ожидание, медиана, мода и др.) определяет положение случайной величины на числовой оси и характеризует центр ее 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 два вида средних величин:

– степенные средние;

– структурные средние.

Степенные средние могут быть вычислены по следующей общей формуле:

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

где m – показатель степени; x_i – текущее значение (вариант) осредняемого признака.

В зависимости от величины m различают следующие виды степенных средних:

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

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

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

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

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

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.

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

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.

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

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 (варианта), деленной на общее число наблюдений, т.е.

$$\overline{x} = (x_1 + x_2 + \cdots + x_n)/n = n^{-1}\sum x_i. \qquad (2.8)$$

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:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 .$$

Свойство 4. Сумма квадратов отклонений членов ряда от центра их тяжести достигает минимума по сравнению с аналогичной суммой, вычисленной относительно любого числа $a \neq \bar{x}$, т. е.

$$S = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \min .$$

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:

$$\bar{x} = \frac{\sum_{k=1}^{m} n_k \bar{x}_k}{\sum_{k=1}^{m} n_k} .$$

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 значений признака. К структурным средним относятся мода и медиана.

Медианой называется такое значение, которое занимает среднее положение в ранжированном (вариационном) ряду. Если всем единицам ряда придать порядковые номера, то при нечетном числе членов ряда ($n = 2m + 1$) медиана определится как $\text{Me} = x_{m+1}$. При четном числе членов ряда ($n = 2m$) за медиану условно принимается среднее значение между центральными значениями величин ранжированного ряда, т. е.

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

Геометрически медиана – это абсцисса точки, в которой площадь, ограниченная кривой плотности вероятности, делится пополам (рис. 2.1). Сказанное означает, что справедливо следующее равенство $p(X < \text{Me}) = p(X > \text{Me}) = 0{,}5$.

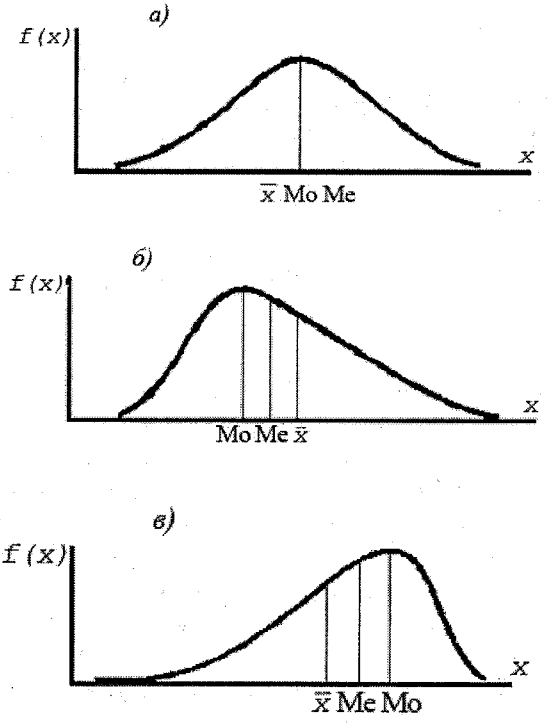


Рис. 2.1. Соотношение между средним арифметическим ($\bar{x}$), модой (Mo) и медианой (Me). *а* – симметричное распределение, *б* – положительно асимметричное распределение, *в* – отрицательно асимметричное распределение.

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

$$\Sigma| x_i - \text{Me}| = \min.$$

Модой называется *наиболее часто встречающаяся в данном статистическом ряду величина*. Геометрически мода представляет собой наибольшую ординату кривой плотности вероятности в случае одновершинного распределения (рис. 2.1). Поэтому 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:

$$\text{Mo} = \bar{x} + 3(\text{Me} - \bar{x}). \qquad (2.10)$$

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, среднее квадратическое (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:

$$D = \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \qquad (2.11)$$

и имеет размерность квадрата случайной величины. Выборочная оценка дисперсии иногда обозначается как s^2.

Среднее квадратическое отклонение случайной величины представляет собой корень квадратный из дисперсии и поэтому сохраняет размерность исходного ряда.

Коэффициент вариации $C = \sigma/m_x$ является безразмерной величиной, поэтому он очень удобен для анализа 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:

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

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

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

где r_{xy} – коэффициент корреляции между переменными x и y, определяемый 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 $(x_i - \bar{x})^2$.

Свойство 7. Если некоторая величина y_i связана с x_i линейным уравнением

$$y_i = ax_i + b, \text{ то } \sigma_y^2 = a^2\sigma_x^2 .$$

Перечисленные выше свойства дисперсии, так же 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 $d = \sum|x_i - \bar{X}|/n$, а для сгруппиро-

ванных – по формуле $d = \sum|x_i - \bar{X}|f_i / \sum f_i$. 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 характеристик, позволяющих оценить степень асимметрии и крутости распределения.

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

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

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 $\bar{x} = Mo = Me$ (рис. 2.1, *а*).

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 $\bar{x} > Mo$ и $\bar{x} > Me$. (рис. 2.1, *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. 2.1, *в*).

Характеристикой крутости (островершинности или плосковершинности) кривой распределения является безразмерный коэффициент эксцесса:

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

Величина коэффициента эксцесса характеризует отклонение крутости эмпирической кривой от нормальной кривой распределения, так как в последнем случае принимается Ex = 0.

Если эмпирическая кривая распределения является более островершинной по сравнению с нормальной кривой, то Ex > 0 (рис. 2.2). В результате мода эмпирического распределения должна быть больше моды нормального распределения ($Mo > Mo_N$).

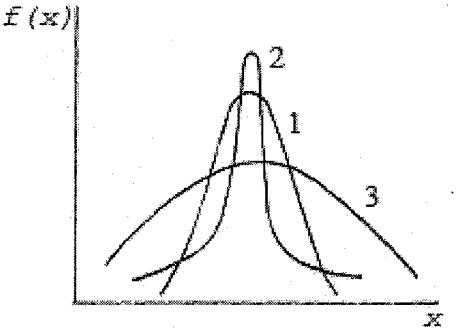


Рис. 2.2. Виды распределений с различной крутостью.
1 – нормальное распределение, *2* – распределение с положительным эксцессом, *3* – распределение с отрицательным эксцессом.

Если эмпирическая кривая распределения является более плосковершинной по сравнению с нормальной кривой, то Ex < 0 (рис. 2.2). В этом случае $Mo < Mo_N$.

Пример 2.1. Оценим числовые характеристики среднемесячных значений температуры поверхности океана (ТПО) в районе судна погоды «М» (66° с.ш., 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 табл. 2.1, а ее межгодовой ход дан на рис. 2.3.

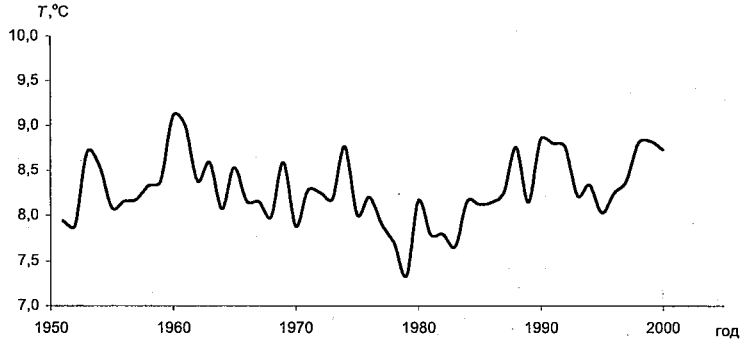


Рис. 2.3. Межгодовой ход температуры поверхности океана в районе судна погоды «М».

Представленные в табл. 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.

Таблица 2.1

Первичные статистические оценки среднемесячных значений ТПО 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

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 среднего квадратического отклонения ТПО почти повторяет годовой ход величины R. Аналогичный характер годового хода в $\bar{X}$ 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 коэффициент ассиметрии As мал, то коэффициент эксцесса Ex весьма велик (например, март), и наоборот (например, июнь).

В распределении As преобладают положительные значения. 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 отрицательные отклонения. Отсюда следует, что должно выполняться неравенство $\bar{x} >$ Me. Из табл. 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 находится его истинное неизвестное значение. Аналитически интервальная оценка произвольного выборочного параметра θ, имеющего некоторое теоретическое распределение, может быть записана в виде:

$$p(\theta_н < \theta < \theta_в) = 1 - \alpha, \qquad (2.15)$$

где $\theta_н$ 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) $\gamma = 1 - \alpha$, где α – уровень значимости, представляющий собой вероятность события, которым решено пренебречь.

При симметричности доверительного интервала относительно оценки θ его нижняя и верхняя границы определяются как

$$\theta_н = \theta - \varepsilon, \quad \theta_в = \theta + \varepsilon,$$

где ε – половина длины доверительного интервала при заданном уровне значимости, означающим вероятность принятия ошибочного решения.

Величину | θ – θ | 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 по выборочной средней $\bar{x}$ при известном среднем квадратическом отклонении σ генеральной совокупности служит доверительный интервал вида:

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

где z – значение аргумента функции Лапласа $\Phi(z)$ (см. *Приложение 1*), при котором $\Phi(z) = (1 - \alpha)/2$.

Данный критерий на практике используется весьма редко, так как истинная оценка величины σ обычно неизвестна.

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

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

где s – выборочная оценка стандартного отклонения σ, рассчитываемая как $s = [\sum(x_i - \bar{x})^2/(n-1)]^{1/2}$; 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 п 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,..., X_n при увеличении n стремится к нормальному закону распределения. Однако при малых значениях n расхождения в доверительных границах уже существенны. Исходя из сказанного, на практике для построения интервальной оценки для математического ожидания при любых n можно ограничиться формулой (2.17).

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

$$p(ns^2/\chi^2_2 < D_x < ns^2/\chi^2_1) = 1 - \alpha, \qquad (2.18)$$

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

$$p[(n-1)s^2/\chi^2_{2*} < D_x < (n-1)s^2/\chi^2_{1*}] = 1 - \alpha, \qquad (2.19)$$

где χ^2_{2*} и χ^2_{1*} – табличные значения статистики χ^2 при числе степеней свободы $\nu = n - 1$.

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

$$(n-1)^{1/2}s/\chi_{2*} < \sigma_x < (n-1)^{1/2}s/\chi_{1*}, \qquad (2.20)$$

Заметим, что, кроме данного выражения, интервальная оценка среднего квадратического отклонения нормально распределенной случайной величины x может быть получена исходя из следующих формул:

$$s(1-q) < \sigma_x < s(1+q) \text{ (при } q < 1),$$
$$0 < \sigma_x < s(1+q) \text{ (при } q > 1)$$

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

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 $\bar{S}$ = 34,2 ‰, среднее квадратическое отклонение s = 21,8 ‰, период наблюдений n = 28 лет. В предположении нормального распределения исходных данных при построении доверительного интервала для математического ожидания солености воспользуемся формулой (2.17). Из таблицы распределения Стьюдента принимая $\alpha = 0{,}05$ и $\nu = n - 1 = 27$ находим $t_\alpha = 2{,}05$. После этого определяем нижнюю и верхнюю доверительные границы:

$$\bar{x} - t_\alpha[s/(n-1)^{1/2}] = 34{,}2 - 2.05\times 21{,}8/(27)^{1/2} = 25{,}6 \text{ ‰},$$
$$\bar{x} + t_\alpha[s/(n-1)^{1/2}] = 34{,}2 + 2{,}05\times 21{,}8/(27)^{1/2} = 42{,}8 \text{ ‰}.$$

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 будем использовать формулу (2.20). Неизвестными параметрами в ней являются χ_{1*} и χ_{2*}. Из распределения χ^2 по значениям $\alpha = 0{,}05$ и $\nu = n - 1 = 27$ находим $\chi^2_{1*} = 16{,}8$ и $\chi^2_{2*} = 47$. Теперь определяем нижнюю и верхнюю доверительные границы:

$$(n-1)^{1/2}s/\chi_{2*} = 21{,}8(27)^{1/2}/(47)^{1/2} \approx 16{,}8\ ‰,$$

$$(n-1)^{1/2}s/\chi_{1*} = 21{,}8(27)^{1/2}/(16{,}8)^{1/2} \approx 27{,}7\ ‰.$$

Таким образом, на уровне значимости $\alpha = 0{,}05$ можно утверждать, что генеральное значение среднего квадратического отклонения годовых значений солености лежит в интервале $16{,}8 < \sigma_S < 27{,}7$ ‰.

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 $\overline{x}$ и σ. Нетрудно задать такие пределы

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

которые с вероятностью p могут 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:

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

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, задав 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. Более эффективной оценкой является медиана, которая очень 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 выборке) 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.

К сожалению, в статистике нет строгого определения малой выборки. Интуитивно понятно, что выборка длиной 10 значений является малой, а объемом 100 значений – большой. Возникает вопрос, где провести верхнюю 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, несвязанных с оценкой функции распределения исходных данных. Например, как будет показано в главе 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.* Однако и в данном случае присутствует некоторая доля субъективизма, ибо понятия заданной точности и достоверно-

сти являются неоднозначными и могут быть различными в зависимости от поставленной задачи даже для одной и той же выборки. Так, для выборки длиной $n = 40$ практически невозможно 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 не используются методы группирования данных, не вычисляются статистические моменты выше второго порядка и применяются 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. Наиболее интенсивное развитие она получила во второй половине 20-го столетия. Что касается квантильного анализа, 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.

$$F(x_p) = p. \qquad (2.22)$$

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

персентили: $P_1, P_2, \ldots, P_{99}$ – квантили порядков 0,01; 0,02; …; 0,99;
децили: $D_1, D_2, \ldots, D_9$ – 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 $x_{0,25}$ – это значение, ниже

которого лежит 25 % наблюдений; Q_2 или $x_{0,5}$ – 50 % наблюдений, Q_3 – 75 % наблюдений. Указанные квартили имеют особые названия. Так, медианой называется квартиль, отвечающий доверительной вероятности $p = 0,5$, т.е. $x_{0,5}$. 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: $x_{\min}$, $x_{\max}$, $x_{0,25}$, $x_{0,75}$, $x_{0,5}$ и др.

Наглядной формой представления результатов квантильного анализа является предложенный Тьюки так называемый «ящик с усами» (рис. 2.4). 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.

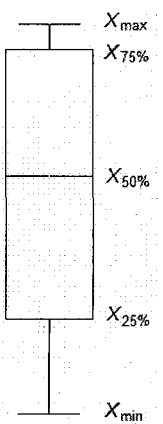


Рис. 2.4. Графическое представление «ящика с усами».

Несмотря на видимую простоту построения ящика с усами, в нем содержится большое количество 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

$$X_3 = (x_{0,25} + 2\text{Me} + x_{0,75})/ 4. \qquad (2.23)$$

Основной характеристикой рассеяния служит интерквартильное расстояние, представляющее аналог среднего квадратического 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 расстояние $1,5Q$, внешние барьеры – на расстояние $3Q$. 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 средним квадратическим отклонением существует следующее соотношение:

$$Q = 1{,}34\sigma. \qquad (2.24)$$

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

$$\mathrm{As} = (x_{0,75} + x_{0,25} - 2x_{0,5}) / (x_{0,75} - x_{0,25}). \qquad (2.25)$$

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. 2.5), 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.

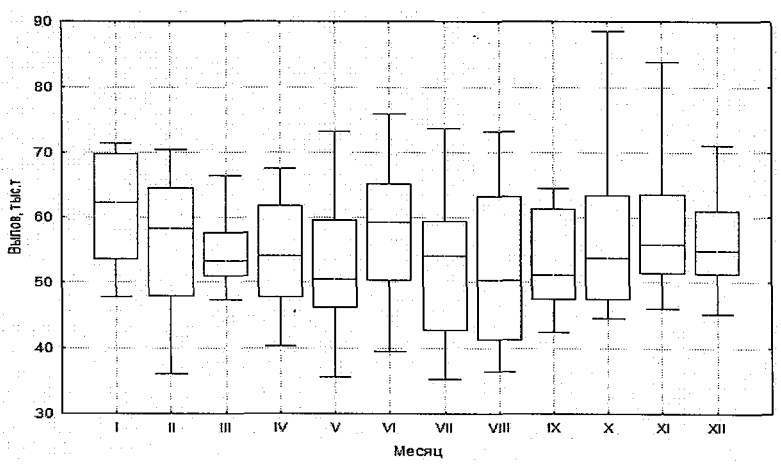


Рис. 2.5. Квантильный анализ вылова ставриды в юго-восточной части Тихого океана для отдельных месяцев года за период 1979–1990 гг.

Из рис. 2.5 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.

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 две 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 определяется следующей формулой:

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

Как видно из данной формулы, достаточно знать всего два параметра, а именно – математическое ожидание и генеральное стандартное отклонение (m_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 $f(x)_{max} = 1/\sigma_x(2\pi)^{1/2}$ и проходящей через m_x (рис. 3.1). При $m_x = 0$ нормальная кривая будет симметрична началу координат.

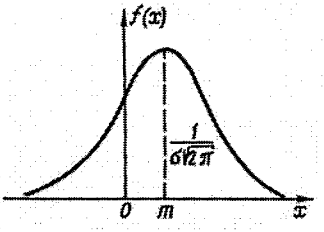


Рис. 3.1. Функция плотности вероятности нормального закона распределения.

Если положить σ_x = const, но изменять параметр m_x, то кривая нормального распределения будет смещаться параллельно оси абсцисс, не меняя своей формы. При изменении параметра σ_x (m_x = const) происходит изменение формы кривой нормального закона. С возрастанием σ_x она становится все более плоской, растягиваясь вдоль оси абсцисс. При уменьшении σ_x, наоборот, кривая распределения сжимается с боков и вытягивается вдоль оси ординат.

Преобразуем выражение (3.1) к интегральному виду:

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

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:

$$t = (x - m_x)/\sigma_x,$$

где t – *стандартизованная случайная величина*, обладающая тем важным свойством, что при любом распределении случайной величины $M[t] = 0$, $D[t] = 1$. 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 величины 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 функция нормального распределения может быть представлена через функцию Лапласа по следующей формуле:

$$F(x) = 0{,}5 + 0{,}5\Phi[(x - m_x)/\sigma_x]. \qquad (3.4)$$

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, использовав 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

$$p(|X - m_x| < \varepsilon) = 2\Phi(\frac{\varepsilon}{\sigma_x}),$$

где ε – некоторая заранее заданная величина.

Рассчитаем теперь по этой формуле вероятность попадания X в интервалы $\pm\sigma_x$, $\pm 2\sigma_x$, $\pm 3\sigma_x$:

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

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

Естественно, что вероятность подобного события чрезвычайно мала. Это позволяет сформулировать следующее правило «трех сигм». Если распределение случайной величины неизвестно, но в интервале $\pm 3\sigma$ 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 и область ее существования $-\infty < x < \infty$.

Свойство 2. Функция $f(x)$ является четной, т.е. $f(x) = f(-x)$.

Свойство 3. Нормальная кривая не пересекается с осью x.

Свойство 4. Математическое ожидание, мода и медиана совпадают, а 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 m_x, σ_x и m_y, σ_y, то их сумма $Z = X + Y$ будет также иметь нормальное распределение с параметрами $m_z = m_x + m_y$ и

$$\sigma_z = \sqrt{\sigma_x^2 + \sigma_y^2}\,.$$

Отметим, что при переходе от генеральной совокупности к выборочным данным и соответственно от математического ожидания и генерального стандартного отклонения к их выборочным аналогам все свойства и закономерности нормального закона распределения сохраняются.

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 пре-*

дельной 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, ..., X_n *имеют конечные абсолютные центральные моменты третьего порядка и если при* $n \to \infty$ *выполняется условие*

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

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

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,..., X_n имеют одинаковые законы распределения с параметрами m и σ, 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 $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$ будет иметь асимптотически нормальное распределение с параметрами $m_{\bar{x}} = m$ и $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

Итак, когда случайная величина представляет собой результат взаимодействия большого числа сравнительно слабых и 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 фактора, то 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 показывают результаты практических расчетов, для многих природных процессов достаточно четырех-пяти равноценных факторов, чтобы распределение случайной величины стало близким к нормальному закону.

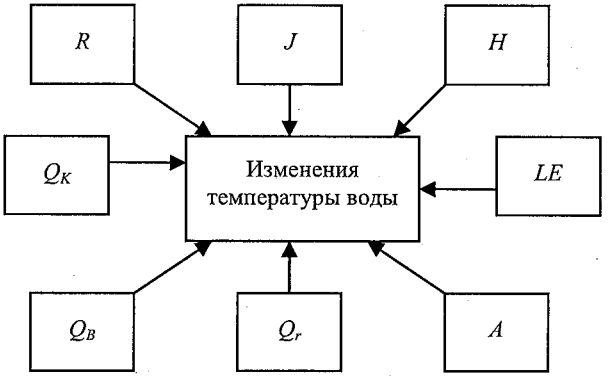


Рис. 3.2. Основные факторы изменения температуры поверхностного слоя воды в океане.

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 ($Q_к$);

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

– 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);
– длинноволновым потоком радиации из атмосферы к океану (Q_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.

Естественно полагать, что годовой ход температуры обусловлен главным образом годовым ходом солнечной радиации. Для исключения влияния радиации можно рассчитать аномалии температуры воды

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

где n – количество лет; m – количество месяцев (m = 12); $\overline{t}_j$ – среднемноголетняя норма температуры для j-го месяца.

В результате такой процедуры обычно принимается, что в аномалиях температуры воды уже отсутствует годовой цикл солнечной радиации. 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 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 видеть, что распределение среднемесячных значений ТПО является двухмодальным и, естественно, абсолютно не соответствует нормальному закону распределения. 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.

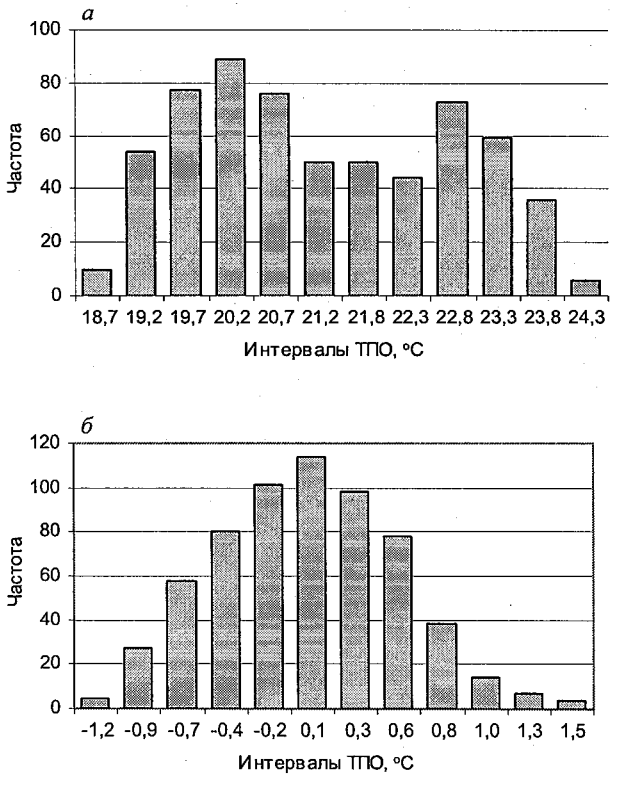


Рис. 3.3. Гистограммы среднемесячных (*а*) значений температуры поверхности океана и их аномалий (*б*) для района Канарского апвеллинга.

3.2. Законы распределения, используемые в гидрометеорологии

Логарифмически нормальное распределение. Вообще говоря, на практике довольно часто встречается такая ситуация, когда случайная величина X сама не является нормально распределенной, однако путем несложного ее 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 ее логарифм $Y = \log_a X$. В соответствии с этим плотность вероятности выражается формулой:

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

где $m_y = M[\log_a X]$ – математическое ожидание; $\sigma_y^2 = D[\log_a X]$ – дисперсия; a – основание логарифма, причем наиболее часто принимается $a = e$, 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:

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

где $m_y = M[\ln X]$, $\sigma_y^2 = D[\ln X]$.

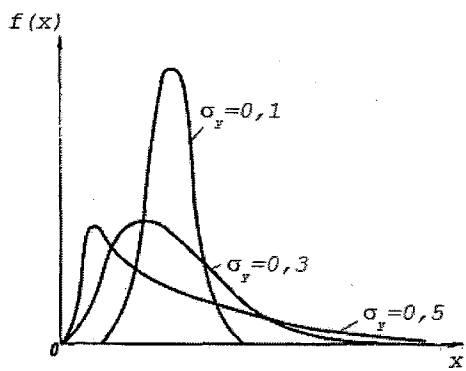


Рис. 3.4. Функция плотности вероятности логарифмически нормального распределения при различных значениях параметра σ_y и $m_y = 1$.

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 σ_y. Естественно, чем меньше σ_y, тем ближе друг к другу значения моды, медианы и

математического ожидания и тем ближе кривая распределения к нормальному закону. Оно свойственно таким случайным величинам, формирование которых происходит в результате умножения большого числа влияющих на них независимых равнозначных факторов.

Между параметрами нормального и логарифмически нормального распределений существуют следующие соотношения:

$$m_x = \exp(m_y + 0{,}5\sigma_y^2),$$
$$\sigma_x = m_x[\exp(\sigma_y^2 - 1)]^{1/2},$$

или

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

где C_x – коэффициент вариации величины X.

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

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

С увеличением коэффициента вариации различия между Mo и m_x возрастают.

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

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

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:

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

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

$$f(x) = \begin{cases} 0, & \text{при} \quad x < 0, \\ c\exp(-cx), & \text{при} \quad x \geq 0, \end{cases} \tag{3.10}$$

а функция распределения показательного (экспоненциального) закона имеет вид:

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

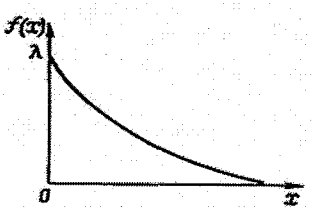


Рис. 3.5. Функция плотности вероятности показательного распределения.

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

$$f(x) = \begin{cases} 0, & \text{при} \quad x < 0, \\ (x/\sigma_x^2)\exp(-x^2/2\sigma_x^2), & \text{при} \quad x \geq 0, \end{cases} \tag{3.12}$$

где σ_x – единственный определяемый параметр.

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

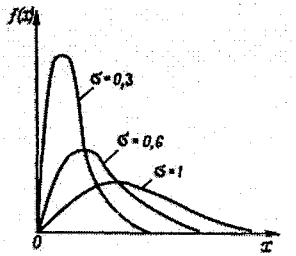


Рис. 3.6. Функция плотности вероятности распределения Релея при различных значениях параметра σ.

Интегральная функция распределения закона Релея выражается формулой:

$$F(x)=\begin{cases}0, & \text{при} \quad x<0,\\ 1-\exp(-x^2/2\sigma_x^2), & \text{при} \quad x\geq 0.\end{cases} \tag{3.13}$$

При этом основные числовые характеристики имеют вид

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

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 ее возможных значений постоянна, а за его пределами равна нулю.* В соответствии с этим плотность вероятности в интервале $[a, b]$ может быть представлена в виде:

$$f(x)=\begin{cases}0, & \text{при} \quad x\leq a,\\ \dfrac{1}{b-a}, & \text{при} \quad a<x<b,\\ 0, & \text{при} \quad x\geq b,\end{cases} \tag{3.14}$$

а функция распределения записана как

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

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

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

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:

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

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

$$m_x = np,\ D_x = npq,\ \text{As} = \frac{q - p}{\sqrt{npq}},\ \text{Ex} = (1 - 6pq)/npq.$$

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 распределения Пирсона χ^2, 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 распределением суммы квадратов независимых случайных величин $\chi^2 = X^2_1 + X^2_2 + \ldots + X^2_n$, 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 всегда равно $\nu = n - 1$.

Плотность вероятности распределения χ^2 имеет вид:

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

где $\Gamma(\nu/2)$ – гамма-функция Эйлера, определяемая как $\Gamma(\nu/2) = \int_0^\infty t^{(\nu/2-1)} \exp(-t)dt$.

Таким образом, распределение χ^2 зависит лишь от одного параметра – числа степеней свободы, который 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.

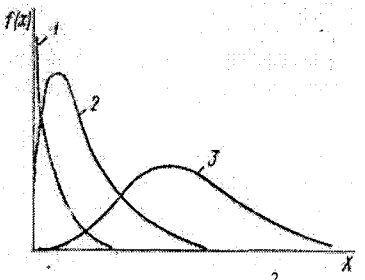


Рис. 3.7. Функция плотности вероятности χ^2-распределения Пирсона для различных степеней свободы (ν): *1* – $\nu = 1$, *2* – $\nu = 4$, *3* – $\nu = 20$.

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:

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

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.

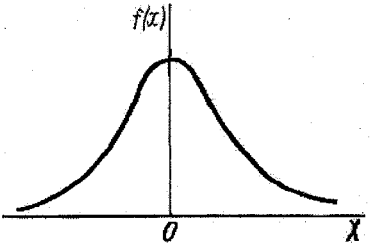


Рис. 3.8. Функция плотности вероятности t-распределения Стьюдента для степеней свободы $\nu = 9$.

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.

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:

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

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.

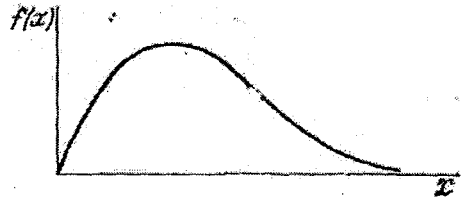


Рис. 3.9. Функция плотности вероятности F-распределения Фишера для степеней свободы: $\nu_1 = 10$, $\nu_2 = 25$.

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

Заметим, что эта таблица дана 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 в данном интервале (классе) k, т. е. эмпирическая повторяемость; 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).

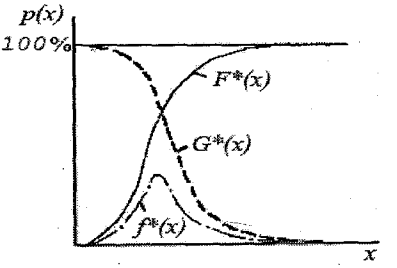


Рис. 3.10. Распределение эмпирических функций распределения $F^*(x)$, функции плотности распределения $f^*(x)$ и функции обеспеченности $G^*(x)$.

Если объем выборки n весьма велик, то построение надежных в статистическом смысле эмпирических функций распределения и обеспеченности не представляет затруднений. Однако, если в каждом интервале $m < 7$–8, 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 определения числа интервалов (классов). 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 при этом 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:

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

$$k \approx 5 \lg n.$$

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 классов. Границы градаций (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 $c_1 = x_{min} - \Delta c/2$.

В процессе группирования выборки могут быть случаи точного совпадения отдельных наблюдений с границами градаций. Если число точно совпадающих членов выборки четное, то тогда 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.

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,

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

Все эти формулы в какой-то степени учитывают выборочность имеющихся наблюдений, что выражается в асимптотическом приближении F^* к F 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.

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 $\Delta c = (14{,}1-$

–9,7)/8 = 0,57 °С. Так как рассчитанное значение Δc не очень удачно характеризует ширину градации, то округляем его в большую сторону до Δc = 0,6 °С. За начальное значение первого интервала примем величину $c_1 = x_{min} - \Delta c/2 = 9,7 - 0,6/2 = 9,4$ °С.

Распределение значений температуры воды по градациям, т.е. оценки эмпирической частоты, приведено в табл. 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 табл. 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, °С	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 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:

$$x' = \lg x \cdot 10^{a}.$$

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:

$$x' = 1/x, \quad x' = 1/(x)^{1/2}.$$

Отметим, что обратная величина $1/x$ является наиболее «сильным» преобразованием, нормализующим выборки с весьма 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 вида:

$$x' = x^{\xi},$$

где показатель степени может принимать различные положительные значения больше 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 естественно сразу же 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. п. 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. 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, критерии Уилкоксона, серий, знаков и др.

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

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

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:

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

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

$$H_0 : \bar{x}_1 = a_1 < x < a_2$$

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 гипотезы 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.*

В результате проверки нулевая гипотеза или принимается как правдоподобная, или отвергается как несостоятельная, причем третьего не дано. Однако сформулированная гипотеза может быть

истинной или 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 табл. 4.1.

Таблица 4.1

Возможные комбинации принятия (отвержения) нулевой гипотезы

Гипотеза H_0	Гипотеза верна	Гипотеза неверна
Гипотеза принимается	Правильное решение	Ошибка второго рода
Гипотеза отвергается	Ошибка первого рода	Правильное решение

Только принятие правильной или отклонение неправильной гипотезы можно считать верным решением. При этом правило, по которому гипотеза H_0 отвергается или принимается, называется статистическим критерием. 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, вообще говоря, 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 $\alpha = 5$ %. 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 опровергается.

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 обычно 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 будет вероятность ошибки второго рода, тем эффективнее будет критерий. Другими словами, вероятность сделать правильный выбор в этом случае будет 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. отвергнуть 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. 4.1). 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.

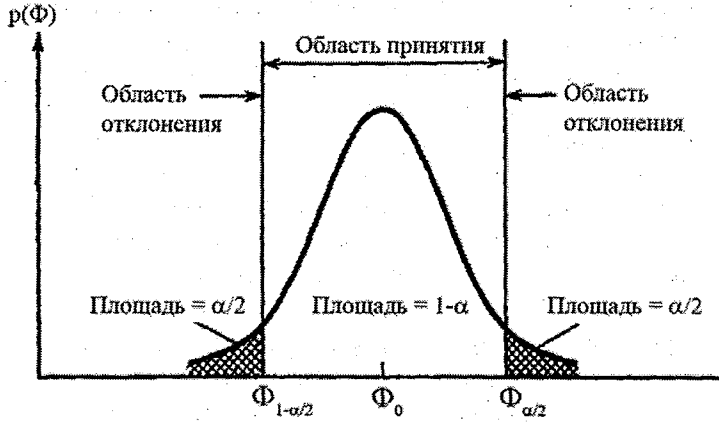


Рис. 4.1. Формирование области допустимых значений и критической области гипотезы.

Кроме того, как следует из рис. 4.1, 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 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, т.е. 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 опровергнута, но никогда 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.

Итак, общая схема проверки нулевой гипотезы состоит в следующем:

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 $H_1 : \bar{\theta}_1 \neq \bar{\theta}_2$, то в этом случае строится двусторонняя критическая область. Если же альтернативная гипотеза принимается в виде неравенств $H_1 : \bar{\theta}_1 > \bar{\theta}_2$ или $H_1 : \bar{\theta}_1 < \bar{\theta}_2$, то соответственно строится правосторонняя (левосторонняя) критическая область.

3. Выбирается какой-либо статистический критерий θ, наилучшим образом отвечающий, по мнению исследователя, проверке нулевой гипотезы.

4. Рассчитывается по экспериментальным данным выборочное значение параметра θ.

5. Осуществляется проверка неравенства $\theta > \theta_{кр}(\alpha, \nu)$, где $\theta_{кр}(\alpha, \nu)$ – критическое (пороговое) значение статистики θ, выбираемое из соответствующей таблицы 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 $\theta = \theta_{кр}(\alpha, \nu)$, 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 главе 9, приближенно можно понимать постоянство во времени выборочных средних и дисперсии. Понятие стационарности является одним из ключевых при анализе случайных процессов. Одним из простейших способов проверки стационарности является использование статистических гипотез о равенстве выборочных средних и дисперсий, причем проверку нужно начинать с равенства 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 их проверки служат параметрические критерии Стьюдента и Фишера.

Гипотеза о равенстве средних при неизвестных генеральных дисперсиях.

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

Поскольку указанные выборочные средние имеют нормальное распределение, то естественно считать, что их разность также должна быть распределена по нормальному закону. В этом случае

для проверки нулевой гипотезы может быть использована статистика Стьюдента, рассчитываемая по следующей формуле:

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

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 $t > \mathbf{t}_{кр}(\alpha, \nu)$, где $\mathbf{t}_{кр}(\alpha, \nu)$ – критическое значение статистики Стьюдента, соответствующее уровню значимости α и числу степеней свободы $\nu = n + m - 2$. *Если данное соотношение выполняется, то нулевая гипотеза о равенстве средних значений отвергается и можно сделать 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. $t < \mathbf{t}_{кр}$, у нас есть основания считать, что расхождения между выборочными средними не являются значимыми.

Гипотеза о равенстве 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

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

где D_x и D_y – генеральные дисперсии двух выборок.

Затем по таблице функции Лапласа находится критическая точка $\mathbf{Z}_{кр}$ из равенства:

$$\Phi(\mathbf{Z}_{кр}) = (1 - \alpha)/2.$$

Если выполняется неравенство $Z > \mathbf{Z}_{кр}$, то нулевая гипотеза о равенстве средних отвергается, если $Z < \mathbf{Z}_{кр}$, 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 они могут быть использованы 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 $\bar{x}$ и $\bar{y}$ распределены асимптотически нормально. Отметим, что генеральные дисперсии известны редко, поэтому данный критерий не нашел широкого применения.

Гипотеза о равенстве дисперсий при неизвестных средних.

Рассмотрим опять две независимые выборки 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 $H_0 : s^2_x = s^2_y$ при альтернативе $H_1 : s^2_x \neq s^2_y$. Наиболее точным критерием ее проверки, как известно, является статистика Фишера (дисперсионное отношение), определяемое по формуле:

$$F = s^2_x / s^2_y, \tag{4.3}$$

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

$$s^2_x = (m-1)^{-1}\sum_{i=1}^{m}(x_i - \bar{x})^2, \qquad s^2_y = (n-1)^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Далее осуществляется проверка неравенства $F > \mathbf{F}_{\text{кр}}(\alpha; \nu_1, \nu_2)$, где $\nu_1 = n - 1$ $\nu_2 = m - 1$ (*Приложение* 4). Если оно выполняется, то *нулевая гипотеза о равенстве выборочных дисперсий отвергается и можно сделать вывод, что выборочные дисперсии, извлеченные из нормальных генеральных совокупностей, имеют значимые расхождения (не равны друг другу) при заданном уровне значимости.* Если оно не выполняется, то у нас нет оснований для отвержения нулевой гипотезы.

Гипотеза о равенстве дисперсий при известных средних.

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.

$$s^2_x = m^{-1}\sum_{i=1}^{m}(x_i - m_x)^2, \qquad s^2_y = n^{-1}\sum_{i=1}^{n}(y_i - m_y)^2.$$

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

Гипотеза о равенстве нескольких дисперсий.

Для сравнения нескольких дисперсий нормальных генеральных совокупностей по выборкам одинакового объема может быть использован критерий Кочрена, а различного объема – критерий Бартлетта. Однако оба критерия обладают недостатками. Так, критерий Бартлетта является весьма приближенным, а распределение критерия Кочрена хотя известно точно, но имеет существенно меньшую мощность, чем, например, критерий Фишера. Поэтому, на наш взгляд, для сравнения нескольких дисперсий все же целесообразно пользоваться критерием 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 будет тем более незначимо. В противном случае выбирается следующая пара дисперсий, имеющая максимальную разность и процедура сравнения их 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 $\alpha = 0{,}05$ выяснить, насколько существенно влияние о. Котлин на распределение средней температуры воды и дисперсии в пределах проведения гидрологической съемки.

Прежде всего рассмотрим равенство выборочных дисперсий. Нулевая гипотеза имеет вид $H_0 : s^2_x = s^2_y$, а альтернативную гипотезу примем $H_1 : s^2_x \neq s^2_y$. 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

$\alpha = 0{,}05$. Из *Приложения 4* находим, что $\mathbf{F}_{кр}(\alpha;\ \nu_1,\ \nu_2) = 3{,}73$. Так как $F < \mathbf{F}_{кр}$, 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. Наилучшим образом проверке гипотезы отвечает критерий Стьюдента. Поэтому рассчитываем его фактическое значение по формуле (4.1). Получаем $t = 1{,}62$. Теперь определяем критическое значение статистики Стьюдента при числе степеней свободы $\nu = 9 + 8 - 2 = 15$ для односторонней области, соответствующей удвоенному уровню значимости, т.е. 2α. Из *Приложения 3* находим $\mathbf{t}_{кр}(2\alpha = 0{,}10,\ \nu = 15) = 1{,}75$. Поскольку $t < \mathbf{t}_{кр}$, 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.

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

В табл. 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 значительное летом. Возникает вопрос – насколько существенны указанные расхождения в оценках средних и дисперсий. Отметим, что критическое значение критерия Стьюдента при $\alpha = 0,05$ и $\nu = 101$ равно $\mathbf{t}_{кр} = 1,98$, а критерия Фишера при $\alpha = 0,05$ и $\nu_1 = 50$, $\nu_2 = 50$ равно $\mathbf{F}_{кр} = 1,60$.

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

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

Достаточно 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 табл. 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 с помощью относительно простых тестов (критериев). 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, которые эта модель должна описать. Другими словами, они выясняют, 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, критерии согласия прежде 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 $n \geq 40$, причем число градаций должно быть не меньше 5–7, а в каждой из градаций должно быть минимум 5–7 наблюдений. Последнее требование на практике обычно очень сложно выполнить, так как «хвосты» (края) эмпирического распределения имеют значительно более низкую повторяемость по сравнению с его центральной частью.

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

$$H_0 : f(\bar{x}, s^2) = f(m_x, D_x).$$

Альтернативную гипотезу зададим в обычном виде

$$H_1 : f(\bar{x}, s^2) \neq f(m_x, D_x).$$

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

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

где p_i – эмпирическая вероятность в i-й градации; $\mathbf{p}_i$ – 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 $\chi^2 > \chi^2_{кр}(\alpha, \nu)$ (*Приложение 2*). При этом число степеней свободы определяется как $\nu = k - \xi - 1$, где ξ – число параметров тео-

ретического распределения. Поскольку для нормального закона $\xi = 2$, то имеем $\nu = k - 3$. Если данное неравенство выполняется, то нулевая гипотеза о соответствии эмпирического распределения нормальному закону отвергается. Если $\chi^2 < \chi^2_{кр}(\alpha, \nu)$, то у нас уже нет оснований отвергать нулевую гипотезу о нормальном распределении генеральной совокупности. 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 может быть использован для проверки любой теоретической функции распределения. В отличие от критерия Пирсона его удобнее использовать для интегральных функций распределения. Нулевая гипотеза записывается аналогично предшествующему случаю. Проверка ее осуществляется с помощью статистики D, представляющей собой модуль максимального уклонения между эмпирической $F(x)$ и теоретической $\mathbf{F}(x)$ функциями распределения, т.е.

$$D = \max_{x \in (-\infty, \infty)} | F(x) - \mathbf{F}(x)|. \qquad (4.5)$$

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:

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

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

$$\lambda = D \times n^{1/2}. \qquad (4.6)$$

Оценка величины D как максимального уклонения между $F(x)$ и $\mathbf{F}(x)$ демонстрируется на рис. 4.2. Значения статистики $\lambda_{кр}$, зависящие лишь от уровня значимости, затабулированы и приводятся в табл. 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

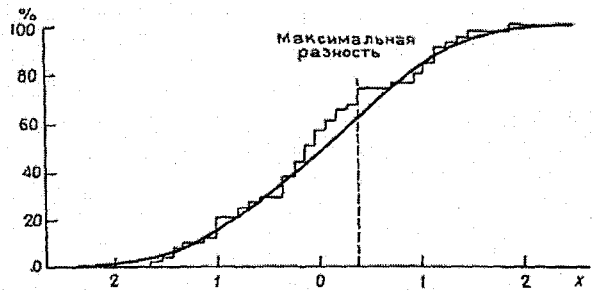


Рис. 4.2. Оценка максимального уклонения между эмпирической $F(x)$ и теоретической $F(x)$ функциями распределения.

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 заключается в следующем:

1. Строятся эмпирическая функция распределения $F(x)$ и предполагаемая теоретическая функция $\mathbf{F}(x)$.
2. Определяется статистика D и вычисляется величина λ.
3. Если выполняется неравенство $\lambda > \lambda_{кр}(\alpha)$, то нулевая гипотеза о том, что случайная величина X соответствует заданному теоретическому закону распределения отвергается. В противном случае у нас нет оснований отвергать нулевую гипотезу и, следовательно, она не противоречит тому, что опытные 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 далеко не всю информацию, заключающуюся в исходной выборке. Действительно, нетрудно представить себе, что эмпирические данные систематически уклоняются от принятой теоретической кривой в разные стороны, но не настолько, чтобы повысить максимальное уклонение, т. е. величину D. В этих случаях критерий Колмогорова будет показывать на хорошее согласие теоретической и эмпирической функций распределения.

Если к этой же выборке применить критерий Пирсона, то в соответствии с ним 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 получаем 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 представлен в табл. 3.1, поэтому воспользуемся оценками эмпирической частоты, которые перенесем в табл. 4.4.

Далее по формуле $f(x) = \frac{1}{\sigma_x\sqrt{2\pi}}\exp\left[-\frac{(x-m_x)}{2\sigma_x^2}\right]$ рассчитываем теоретические оценки вероятности нормальной функции распределения для середин интервалов (табл. 4.4). Отметим, что в каче-

стве m_x и σ_x берутся выборочные оценки среднего арифметического и стандартного отклонения ($\bar{x} = 11{,}9$ °С, $s = 0{,}9$ °С).

Таблица 4.4

Проверка соответствия эмпирических данных нормальному закону распределения по критерию Пирсона

Номер градации	Градация, °С	Эмпирическая частота, m_i	Вероятность, p_i	Теоретическая частота, np_i	$(m_i - np_i)^2$	$(m_i - np_i)^2/np_i$
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

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 табл. 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 $\chi^2 > \chi^2_{кр}(\alpha,\nu)$, причем число степеней свободы $\nu = k - 3 = 6 - 3 = 3$. Принимая уровень значимости $\alpha = 0{,}05$, находим по распределению Пирсона $\chi^2_{кр} = 7{,}82$. Нетрудно видеть, что $\chi^2 < \chi^2_{кр}$. 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 дифференциальную эмпирическую функцию $f(x)$ в интегральную функцию $F(x)$. Эмпирические оценки функции $F(x)$, которые соответствуют накопленной частости (см. табл. 3.1), приведены в табл. 4.5.

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
$\mathbf{F}(x)$	0,004	0,021	0,080	0,221	0.449	0,695	0,878	0,964	0,993

Далее следует рассчитать теоретические оценки $\mathbf{F}(x)$. Для этого можно воспользоваться, например, формулой (3.4):

$$\mathbf{F}(x) = 0{,}5 + 0{,}5\Phi[(x - m_x)/\sigma_x].$$

Исходя из этой формулы, для первого значения ПТВ получим:

$$\boldsymbol{F}(9{,}4) = 0{,}5 + 0{,}5\Phi[9{,}4 - 11{,}9)/0{,}93] = 0{,}5 + 0{,}5\Phi(-2{,}69) =$$
$$= 0{,}5 - 0{,}5 \times 0{,}9928 \approx 0{,}004.$$

Аналогичным образом рассчитываются все остальные оценки функции $\mathbf{F}(x)$. Сравнение значений эмпирической и теоретической функций распределения, указанных в табл. 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 $\lambda < \lambda_{кр}$ при любом числе степеней свободы, то можно полагать, что нулевая гипотеза о выбранном нормальном распределении согласуется с опытными данными.

4.4. 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 эмпирических распределений $F_1(x)$ и $F_2(x)$, принадлежащих общему теоретическому распределению $\mathbf{F}(x)$, 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 различия между этими законами распределения незначимы, 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 на уже рассмотренной выше статистике D, которая в отличие от критерия согласия сравнивает две эмпирические функции распределения, т.е.

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

Затем вычисляется величина

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

где n_1 и n_2 – объемы выборок, причем необязательно $n_1 = n_2$. Далее проверяется неравенство $\lambda' > \lambda'_{кр}(\alpha)$.

Показано, что для довольно длинных выборок ($n_1 \geq 50$, $n_2 \geq 50$) распределение статистики λ' сходится к распределению статистики λ. Поэтому в данном случае можно воспользоваться распределением $\lambda_{кр}$ (см. табл. 4.2). Для более коротких выборок используются специальные таблицы.

Пример 4.4. Как известно, при измерении осадков на метеостанциях неоднократно происходила смена приборов. В частности, в России в течение довольно длительного периода систематические измерения осадков осуществлялись дождемером с защитой Нифера. Затем была произведена замена этого дождемера на осадкомер Третьякова, обладающего значительно более лучшими 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 табл. 4.6.

Прежде всего рассчитываем накопленные частоты для обеих частей выборок Σm_i, используемых для оценок эмпирических

функций распределения: $F_1(x) = \Sigma m_i/n_1$ и $F_2(x) = \Sigma m_i/n_2$, распределение которых дается в табл. 4.7. Теперь определяем максимальное уклонение между ними, которое отмечается для шестой градации и составляет $D = 0{,}089$.

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

По формуле (4.7) рассчитываем $\lambda' = 0{,}644$. По табл. 4.3 находим, что при уровне значимости $\alpha = 0{,}05$ $\lambda_{кр} = 1{,}36$. Поскольку $\lambda' < \lambda_{кр}$, то у нас есть основание считать, что различия между этими законами распределения незначимы, т.е. выборки принадлежат одной и той же генеральной совокупности и, следовательно, общая выборка является однородной.

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	3	5	0,027	0,050	0,023
35	13	17	0,118	0,170	0,052
40	28	25	0,254	0,250	0,004
45	48	50	0,436	0,500	0,064
50	60	60	0,545	0,600	0,550
55	65	68	0,591	0,680	0,089
60	90	88	0,818	0,880	0,072
65	105	95	0.955	0,950	0,005
70	110	100	1,000	1,000	0,000

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 можно перенумеровать. Затем каждому значению объединенного ряда присвоим свой ранг (порядковый номер). Пусть, например, общий вариационный ряд имеет вид:

$$\begin{array}{cccccccccccc} x_1 & y_1 & x_2 & x_3 & y_2 & x_4 & y_3 & y_4 & x_5 & x_6 & y_5 & y_6. \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \end{array}$$

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:

$$w_x + w_y = (m + n)(m + n + 1)/2.$$

В рассматриваемом случае имеем $33 + 45 = 78$, $12(12 + 1)/2 = 78$.

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:

$$\bar{w} = (m + n)(m + n + 1)/4.$$

В общем случае проверка нулевой гипотезы $H_0 : F(x) = F(y)$ при альтернативе $H_1 : F(x) \neq F(y)$ осуществляется путем построения доверительных интервалов $w_{\text{ниж}} < w < w_{\text{вер}}$. Если окажется, что сумма рангов $w < w_{\text{ниж}}$ или $w > w_{\text{вер}}$, т.е. оно 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 $w_{\text{ниж}}(q = \alpha/2, m, n)$ определяется по формуле:

$$w_{\text{ниж}} = [(m + n + 1)m - 1]/2 - z_{\text{кр}}\psi, \qquad (4.8)$$

где $z_{\text{кр}}$ – квантиль функции Лапласа, определяемый по *Приложению 1* в соответствии с равенством $\Phi(z_{\text{кр}}) = (1 - \alpha)/2$, а величина ψ, имеющая смысл среднего квадратического отклонения суммы рангов, равна

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

После этого находится верхняя критическая точка $w_{\text{вер}}$ как

$$w_{\text{вер}} = [(m + n + 1)m - 1] - w_{\text{ниж}}. \qquad (4.9)$$

В том случае, если объем обеих выборок не превышает 25, то для нахождения нижней критической точки $w_{\text{ниж}}$ используются 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 $w_{\text{вер}}$. 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 $w_{\text{ниж}}$ при $q = \alpha/2 = 0{,}025$, которая равна $w_{\text{ниж}} = 26$. Верхняя критическая точка равна $w_{\text{вер}} = 52$. Нетрудно ви-

деть, что $26 < 33 < 52$. Итак, у нас нет оснований отвергнуть нулевую гипотезу.

Следует иметь в виду, что данный критерий наиболее чувствителен к различию выборок по характеристикам положения и весьма 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. Вначале 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 $w_x + w_y = (m + n)(m + n + 1)/2 = (210 \cdot 211)/2 = 22\,155$. Нетрудно видеть, что сумма рангов подсчитана правильно. Теперь определяем $z_{кр}$ по равенству $\Phi(z_{кр}) = (1 - \alpha)/2 = (1 - 0{,}05)/2 = 0{,}4975$. По таблице функции Лапласа находим $z_{кр} = 2{,}81$. После этого вычисляем нижнюю критическую точку $w_{ниж}$ при $q = \alpha/2 = 0{,}025$. Величина ψ равна $\psi = [mn(m + n + 1)/12]^{1/2} = 60{,}8$. В результате имеем:

$$w_{ниж} = [(m + n + 1)m - 1]/2 - z_{кр}\psi = (211 \times 100 - 1)/2 - 2{,}81 \times 60{,}8 = 10\,379.$$

Осталось найти верхнюю критическую точку $w_{вер}$:

$$w_{вер} = [(m + n + 1)m - 1] - w_{ниж} = 21\,099 - 10\,379 = 10\,720.$$

Итак, $10\,379 < 10\,504 < 10\,720$. Следовательно, у нас нет оснований отвергнуть 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 последователь-*

ности, 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 совокупностей.

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

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

где $T_1 = (2n_1n_2) / (n_1 + n_2)$,

$T_2 = 2n_1n_2(2n_1n_2 - n_1 - n_2)$,

$T_3 = (n_1 + n_2)^2(n_1 + n_2 - 1)$.

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 $N \leq N_1$ и $N \leq N_2$, 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 получены из одной генеральной совокупности. Альтернативная гипотеза H_1 : 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 не 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» величина 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).

В самом общем случае невязка 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 неучитываемых членов.

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 неучет 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:

$$Q + P - E - \Delta V = \eta, \qquad (5.4)$$

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 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.

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

Оценить подземный приток и особенно его межгодовую изменчивость весьма сложно. Хотя в оценках величины U, полученных разными авторами, отмечаются заметные расхождения, все же достаточно уверенно можно принять ее норму близкой к 3–4 км^3/год. Учитывая, что приток речных вод Q за разные многолетние периоды времени равен 275–305 км^3/год, 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, прибавив 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 в изменения объема дает ΔV_ρ = 1,5 км^3/год. Но поскольку во временном ходе температуры воды отсутствует 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 измене-

ния составляющих водного баланса. В этом случае, вследствие отчетливо выраженного сезонного хода температуры воды, амплитуда $\Delta V_р$ уже будет составлять десятки кубических километров в год.

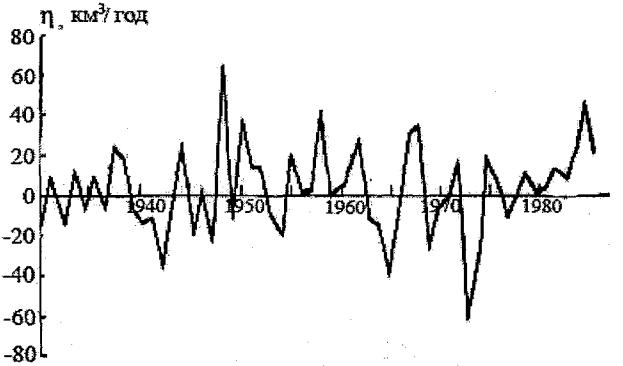


Рис. 5.1. Межгодовой ход значений невязки η уравнения водного баланса Каспийского моря за 1930–1989 гг.

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. 5.1. Ее положительным значениям соответствует повышение притока речных вод и осадков 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. Кроме того, можно отметить, 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 (табл. 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 прежде всего 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
X	290	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
X_{min}	213,0	49,1	303	–114,5	–98

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

Таблица 5.2

Корреляционная матрица межгодовых колебаний составляющих водного баланса Каспийского моря

Составляющие	η	P	Q	E	ΔV
P	–0,11	1,00			
Q	0,18	–0,03	1,00		
E	–0,26	0,09	–0,13	1,00	
ΔV	–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 величину x_i 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 среднего квадратического отклонения случайной величины x_i, определяемая по формуле:

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

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

$$A_\delta = |x - x_{ист}|.$$

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

$$\varepsilon = A_\delta / x_{ист}.$$

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:

$$\sigma_в \approx (2/3)\sigma_x.$$

При условии, что исходные данные распределены по нормальному закону, а внутрирядная связь отсутствует, что соответствует модели временного ряда «белый шум», случайные погрешности разных характеристик первых четырех статистических моментов определяются следующим образом:

– ошибка выборочного среднего

$$\sigma_{\bar{x}} = \sigma_x/(n-1)^{1/2}; \qquad (5.5)$$

– ошибка среднего квадратического отклонения

$$\sigma_\sigma \approx \sigma_x/(2n-1)^{1/2}; \qquad (5.6)$$

– ошибка коэффициента асимметрии

$$\sigma_A = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \approx \sqrt{\frac{6}{n}}; \qquad (5.7)$$

– ошибка коэффициента эксцесса

$$\sigma_Э = \sqrt{\frac{24\,n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}} \approx \sqrt{\frac{24}{n}}. \qquad (5.8)$$

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:

$$\sigma_{\bar{x}} = \sigma_x/(n^*)^{1/2}, \qquad (5.9)$$

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:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n^*}} \approx \frac{\sigma_x}{\sqrt{n}}\sqrt{\frac{1+r_1}{1-r_1}}. \qquad (5.10)$$

Аналогичным образом осуществляется учет связности при оценке ошибки стандартного отклонения.

Для временного ряда, имеющего автокорреляцию между смежными значениями, случайная ошибка среднего стандартного отклонения 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 на первый взгляд такой легко измеряемой 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 целого ряда факторов. Так, суммарная систематическая погрешность определения осадков может быть выражена как

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

где $\Delta P_{\text{в}}$ – ветровая поправка, обусловленная искажением попадания осадков в ведро при ветре; $\Delta P_{\text{и}}$ – поправка на испарение части выпавших осадков; $\Delta P_{\text{с}}$ – поправка на смачивание; $\Delta P_{\text{м}}$ – 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 как раз 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 за счет 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 обычно в уравнении (5.2) присутствуют все виды систематических ошибок, поэтому 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 «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 экспе-

риментальных 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 ($\sigma_1 << \bar{x}_1$, ..., $\sigma_m << \bar{x}_m$) и некоррелированы друг с другом. В этом случае для оценки косвенных погрешностей может быть использован так называемый метод *частных погрешностей*.

Предположим, что погрешности подчиняются нормальному закону распределения, по абсолютной величине значительно уступают $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m$ ($\sigma_1 << \bar{x}_1$ и т. д.) и некоррелированы друг с другом. В этом случае для оценки случайных погрешностей может быть использован так называемый метод частных погрешностей:

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

Придав функции f конкретный вид, можно получить аналитические формулы для оценки случайных погрешностей. Формулы для некоторых простейших зависимостей приводятся в табл. 5.3. Заметим, что особенностью метода частных производных является то, что он оказывается корректным только для абсолютных погрешностей. Относительные их значения должны находиться соответствующим пересчетом. Конкретные примеры такого пересчета также указаны в табл. 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 $r_{1,2} \le 0$ 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.

Таблица 5.3

Формула для оценки случайных погрешностей различных зависимостей

Вид зависимости	Абсолютная погрешность	Относительная погрешность
$y = x_1 + x_2$	$\sigma_y = \sqrt{\sigma_1^2 + \sigma_2^2}$	$\varepsilon_y = \dfrac{\sqrt{\sigma_1^2 + \sigma_2^2}}{x_1 + x_2}$
$y = x_1 - x_2$	$\sigma_y = \sqrt{\sigma_1^2 + \sigma_2^2}$	$\varepsilon_y = \dfrac{\sqrt{\sigma_1^2 + \sigma_2^2}}{x_1 - x_2}$
$y = x_1 x_2$	$\sigma_y = \sqrt{x_2^2\sigma_1^2 + x_1^2\sigma_2^2}$	$\varepsilon_y = \sqrt{\dfrac{\sigma_1^2}{x_1^2} + \dfrac{\sigma_2^2}{x_2^2}}$
$y = x_1 / x_2$	$\sigma_y = \sqrt{\dfrac{\sigma_1^2}{x_2^2} + \dfrac{x_1^2}{x_2^4}\sigma_2^2}$	$\varepsilon_y = \sqrt{\dfrac{\sigma_1^2}{x_1^2} + \dfrac{\sigma_2^2}{x_2^2}}$
$y = x_1 x_2 x_3$	$\sigma_y = \sqrt{x_2^2 x_3^2\sigma_1^2 + x_3^2 x_1^2\sigma_2^2 + x_1^2 x_2^2\sigma_3^2}$	$\varepsilon_y = \sqrt{\dfrac{\sigma_1^2}{x_1^2} + \dfrac{\sigma_2^2}{x_2^2} + \dfrac{\sigma_3^2}{x_3^2}}$
$y = ax_1 + bx_2 + cx_3$	$\sigma_y = \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2 + c^2\sigma_3^2}$	$\varepsilon_y = \sqrt{\dfrac{a^2\sigma_1^2 + b^2\sigma_2^2 + c^2\sigma_3^2}{ax_1 + bx_2 + cx_3}}$

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:

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

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

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

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

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

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

$$S = f(t, \chi, c), \quad \sigma_S = 0{,}0096\ ‰,$$

$$S = f(t, c, P), \quad \sigma_S = 0{,}075\ ‰,$$

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

Нетрудно видеть, что наиболее точные результаты при расчете солености могут быть получены, если в качестве первичных параметров служат 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 экстремальные свойства изучаемого явления (переменной) или 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 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. В связи с этим 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 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 гипотеза вида $H_0 : \bar{x} = x_n$, проверка которой при условии нормальности исходных данных осуществляется с помощью критерия Стьюдента:

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

Здесь выборочные характеристики $\bar{x}$ и σ вычисляются без учета величины x_n. Затем проверяется выполнение неравенства

$$t > \mathbf{t}_{кр.}(\alpha, \nu = n - 1).$$

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 исключать нецелесообразно. После исключения крайнего значения, данную процедуру можно повторить и для следующего по абсолютной величине максимального отклонения, но предварительно необходимо пересчитать $\bar{x}$ и σ для выборки нового объема $n - 1$.

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

$$t = | \bar{x} - x_n | / \sigma \sqrt{(n-1)/n}\,. \qquad (5.14)$$

Далее процедура обнаружения грубых погрешностей аналогична изложенной выше. Введение множителя $\sqrt{(n-1)/n}$ 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 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 собой 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:

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

где x_i – выборка i наблюдений, распределенных 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:

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

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

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

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

где $\bar{x}_{k'}$ – 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:

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

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 $|x_m - \bar{x}|$ значения 3σ ничтожно мала и составляет менее 0,3 %.

Поэтому если выполняется условие

$$|\bar{x} - x_n| > 3\sigma, \tag{5.19}$$

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:

$$|\bar{x} - x_n| > 2\sigma, \tag{5.20}$$

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 п. 2.7. В квантильном анализе точка, резко выделяющаяся от остальной совокупности, называется выбросом, если для нее выполняются следующие условия:

– для положительных аномалий: $x_{выб} > x_{0,75} + kQ$,

– для отрицательных аномалий: $x_{выб} < x_{0,25} - kQ$,

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.

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 трудной 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.

Пример 5.3. В январе 1989 г. температура воздуха на одной из метеостанций Ленинградской области составила 0 °С. При этом средняя многолетняя величина и стандартное отклонение температуры воздуха за 50 лет соответственно равны –8 °С и 2,5 °С. По формуле (5.13) получаем $t = 3{,}5$, а критерий Стьюдента при уровне значимости $\alpha = 10$ % составляет $t_{кр} = 1{,}96$. 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 табл. 5.4 представлены первичные статистические характеристики температуры воды для станции Болванский Нос. Нетрудно видеть, что длина наблюдений в разные месяцы существенно различна, причем она минимальна в зимний период. Однако, учитывая, что именно зимой температура практически постоянна, это обстоятельство не должно нас беспокоить. Здесь же даны критиче-

ские 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 $\mathbf{t}_{кр} = 2$, т.е. критерий Стьюдента совпадает с критерием «двух сигм». Увеличение критерия значимости существенно занижает величину $\mathbf{t}_{кр}$ и тем самым будет завышать число выбросов. Уменьшение его до $\alpha \geq 0{,}005$ или 0,001, наоборот, приближает $\mathbf{t}_{кр}$ к критерию «трех сигм», что является малоинформативным. Поэтому мы приняли промежуточный вариант: α = 0,01.

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

При анализе значений температуры воды прежде всего обращает на себя внимание наличие ее больших отрицательных значений в течение периода с января по март (–3,5, –3,7 °С). Всего таких значений зарегистрировано 9. Оценить реальность существования подобных выбросов нетрудно, поскольку температура замерзания морской воды в зависимости от солености приближенно описывается следующей формулой:

$$T_з = -0{,}003 - 0{,}0527S - 0{,}00004S^2 - 0{,}0000004S^3.$$

Подставив в эту формулу среднее значение солености, получим, что температура замерзания составляет $T_з = -1{,}9$ °С. Таким образом, совершенно очевидно, что значения температуры воды ниже –1,9 °С 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 табл. 5.5 приводится число выбросов для температуры воды для всех месяцев этой ГМС, полученных с помощью этих критериев.

Таблица 5.5

Число выбросов в межгодовом ходе температуры воды на ГМС Болванский Нос

Месяц	Критерий			
	Квантильный анализ	Стьюдента	«трех сигм»	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 жестким, то есть выбросов он не выделяет. Как следует из табл. 5.5, в общем 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. 5.2, на котором приведены оценки выбросов, определенные квантильным анализом. При этом использовались два варианта коэффициента k ($k = 1{,}5$ и $k = 3{,}0$). Отметим, что все выбросы в августе регистрируются по величине $k = 1{,}5$.

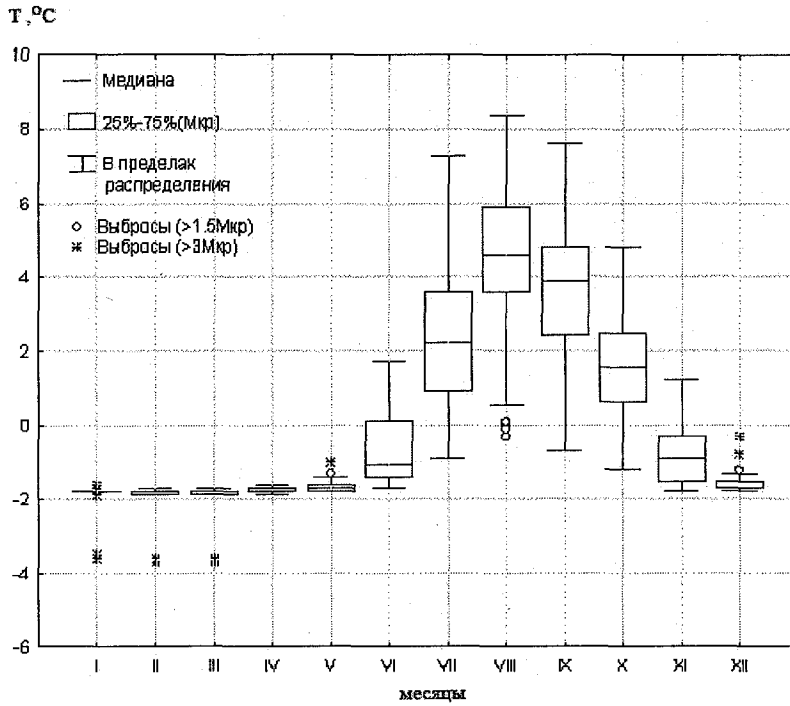


Рис. 5.2. Квантильный анализ наблюдений за температурой воды на ГМС Болванский Нос.

Дополнительный анализ градиентов $\Delta x = x_{i+1} - x_i$, где i – номер месяца, показал, что градиент для выбросов, приведенных в табл. 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

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 Кронштадтскому 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. 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 за уровень C на интервале T может быть вычислено по следующей приближенной формуле

$$\bar{N}_c = \bar{N}_a \exp[-(C-a)(C+a-2m_x)/2\sigma^2_x], \qquad (5.21)$$

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:

$$\bar{T}_c = T[0{,}5 - \Phi(t)], \qquad (5.22)$$

где $\Phi(t)$ – функция Лапласа, определяемая по формуле (3.3); T – продолжительность интервала, на котором оцениваются выбросы.

Зная значения $\bar{N}_c$ и $\bar{T}_c$, можно оценить среднюю длительность единичного выброса:

$$\bar{\theta} = N_c / \bar{T}_c. \qquad (5.23)$$

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:

$$g(S_c) = (1/3)\lambda^{2/3} S_c^{-1/3} \exp[-0{,}5(\lambda S_c)^{2/3}], \qquad (5.24)$$

где $\lambda = 3(C - m_x)^2[-r''(\tau)]^{1/2} / 2\sigma^3{}_x$; $r''(\tau)$ – 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

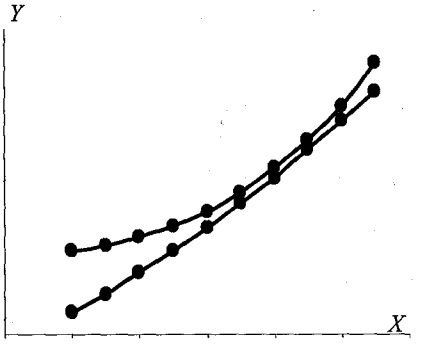
6.1. Виды связей между двумя переменными

При анализе гидрометеорологических явлений или процессов очень часто возникает необходимость установить между ними связь. В общем случае эта связь может быть трех типов: функциональной (детерминированной), стохастической (вероятностной) и случайной, характеризующей полное отсутствие связи.

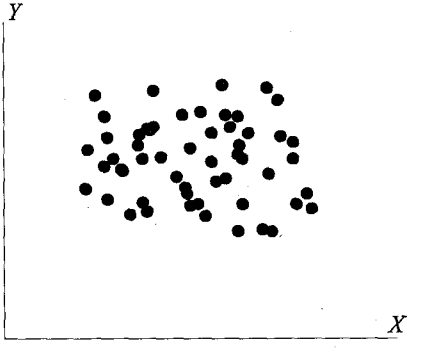
Функциональная связь. Если каждому значению одной переменной соответствует единственное значение другой переменной, то такая зависимость носит название функциональной (рис. 6.1, *а*). Функциональные зависимости обычно могут быть доказаны теоретически (с помощью логических рассуждений) и не нуждаются в опытной проверке. Естественно, что доверительная вероятность такой связи равна единице ($p = 1$).

Отсутствие связи. Если любому значению одной переменной соответствует 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 связи при этом равна нулю ($p = 0$).

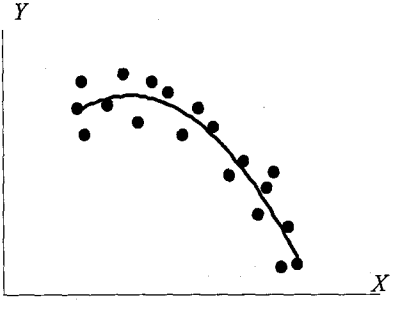
Стохастическая (вероятностная) связь. Если каждому значению одной переменной с определенной вероятностью ($0 < p < 1$) соответствует значение другой переменной, то такая зависимость называется стохастической. Естественно, что она носит промежуточный характер между функциональным и случайным типами связи (рис. 6.1, *в*). Стохастическая зависимость характеризуется тесностью 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.



а



б



в

Рис. 6.1. Виды связи 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 изучение 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 посвящена глава 8.

6.2. Коэффициент корреляции и его свойства

В настоящее время известно много различных показателей тесноты статистических 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, ранговый коэффициент Спирмэна, 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 вытекает из того, что он является характеристикой двухмерного нормального распределения. Следовательно, прежде чем рассчитывать коэффициенты корреляции, надо убедиться, что система двух случайных величин имеет нормальное распределение.

Наибольшее распространение в практических расчетах получил *коэффициент корреляции, который является безразмерной параметрической характеристикой линейной взаимосвязи двух случайных величин X и Y*, т. е.

$$r = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}. \qquad (6.1)$$

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

$$r = \frac{\sum xy - (n^{-1})(\sum x)(\sum y)}{\sqrt{[\sum x^2 - (n^{-1})(\sum x)^2][\sum y^2 - (n^{-1})(\sum y)^2]}}. \qquad (6.2)$$

Отметим некоторые важные свойства коэффициента корреляции.

Свойство 1. Коэффициент корреляции не изменится, если: а) прибавить (вычесть) к величинам X и Y какие-либо постоянные слагаемые; б) умножить (разделить) величины 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 величины 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 дру-

гом ломаной линией, которая называется *эмпирической линией* (функцией) *связи*. По ее виду можно судить не только о наличии, но и о форме зависимости между рассматриваемыми переменными.

На рис. 6.2 приводится несколько корреляционных графиков. В том случае, когда между переменными X_1 и X_2 отмечается ярко выраженная прямо пропорциональная зависимость (рис. 6.2, *а*), коэффициент корреляции близок к единице. Менее отчетливая положительная корреляция (r = 0,54) наблюдается между переменными на рис. 6.2, *б*. Практически случайная связь на рис. 6.2, *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 рис. 6.2, *г*. Определение коэффициента корреляции на рис. 6.2, *д* невозможно, поскольку переменная X_1 постоянна и, следовательно, в соответствии с формулой (6.1) необходимо осуществить деление на нуль. На рис. 6.2, *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.

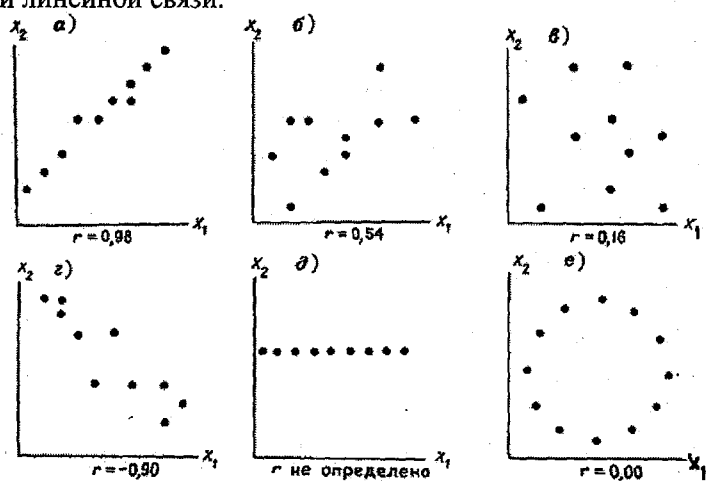


Рис. 6.2. Графики корреляционных полей с различными коэффициентами корреляции между переменными X_1 и X_2.

6.3. Оценка достоверности и значимости коэффициента корреляции

Распределение выборочных коэффициентов корреляции r очень сильно зависит от длины сравниваемых рядов и от самой величины r. При малых величинах 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 средних квадратических погрешностей коэффициентов корреляции используется следующая формула:

$$\sigma_r = (1 - r^2) / n^{1/2}. \qquad (6.3)$$

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:

$$r - \mathbf{t}_{кр}\sigma_r < \mathbf{r} < r + \mathbf{t}_{кр}\sigma_r, \qquad (6.4)$$

где $\mathbf{t}_{кр}$ – критерий Стьюдента при уровне значимости α и числе степеней свободы $\nu = n - 2$.

Оценка значимости коэффициента корреляции осуществляется на основе нулевой гипотезы, которая в этом случае 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:

$$t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}. \qquad (6.5)$$

Затем осуществляется проверка неравенства

$$t > \mathbf{t}_{кр}(\alpha, \nu = n-2).$$

Если данное неравенство 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 незначим, 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 критерия Стьюдента критическое значение коэффициента корреляции, т.е.

$$\mathbf{t}_{кр}\,(1 - \mathbf{r}^2_{кр})^{1/2} / (n-2)^{1/2} = \mathbf{r}_{кр}. \qquad (6.6)$$

Возведем левую и правую части уравнения в квадрат и выполним несложные преобразования

$$\mathbf{t}^2_{кр} - \mathbf{t}^2_{кр}\mathbf{r}^2_{кр} = \mathbf{r}^2_{кр}(n-2)$$

или

$$\mathbf{t}^2_{кр} = \mathbf{r}^2_{кр}(n - 2 + \mathbf{t}^2_{кр}).$$

Отсюда

$$\mathbf{r}_{кр} = \mathbf{t}_{кр}/(n - 2 + \mathbf{t}^2_{кр})^{1/2}. \qquad (6.7)$$

Итак, подставив в формулу (6.7) значения $\mathbf{t}_{кр}$ и *n*, нетрудно вычислить критические значения коэффициента корреляции. Поскольку для длинных выборок при $\alpha = 0{,}05$ $\mathbf{t}_{кр} \approx 2{,}0$, то формулу (6.7) можно упростить:

$$\mathbf{r}_{кр} \approx 2/(n+2)^{1/2}. \qquad (6.8)$$

Преимущество формулы (6.7) перед (6.5) состоит в том, что величину $\mathbf{r}_{кр}$ не нужно пересчитывать как величину $\mathbf{t}_{кр}$ для каждого выборочного значения r.

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:

$$z = 0{,}5 \ln(1+r)/(1-r) = \operatorname{arcth} r, \qquad (6.9)$$

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

$$M(z) = \frac{0{,}5\ln(1+r)}{1-r} + \frac{r}{2(n-1)}$$

и дисперсией

$$D(z) = 1 / (n-3).$$

Отсюда нетрудно видеть, что стандартная погрешность величины z зависит лишь от длины выборки и определяется как

$$\sigma_z = 1 / (n-3)^{1/2}. \qquad (6.10)$$

Доверительные границы для z записываются следующим образом:

$$z - \mathbf{t}_{кр}\sigma_z < \mathbf{z} < z + \mathbf{t}_{кр}\sigma_z. \qquad (6.11)$$

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

$$r = \frac{\exp(2z-1)}{\exp(2z+1)}. \qquad (6.12)$$

Тогда доверительный интервал для величины **r** примет вид:

$$\frac{\exp(2z_1-1)}{\exp(2z_1+1)} < \mathbf{r} < \frac{\exp(2z_2-1)}{\exp(2z_2+1)},$$

где $z_1 = z - \mathbf{t}_{кр}\sigma_z$, $z_2 = z + \mathbf{t}_{кр}\sigma_z$.

Отметим, что рассчитанные таким образом 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) $z = 0{,}5\ln(1 + 0{,}74)/(1 - 0{,}74) = 0{,}95$;

2) $\sigma_z = 1/(50-3)^{1/2} = 1/6{,}86 = 0{,}146$;

3) $(0{,}95 - 2{,}01 \cdot 0{,}146) < \mathbf{z} < (0{,}95 + 2{,}01 \cdot 0{,}146) \rightarrow 0{,}66 < \mathbf{z} < 1{,}24$.

4) Вычислив по формуле (6.9) оценки *r*, окончательно получим $0{,}58 < \mathbf{r} < 0{,}84$. 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

$0{,}74 - 2{,}01 \cdot 0{,}066 < \mathbf{r} < 0{,}74 + 2{,}01 \cdot 0{,}066 \rightarrow 0{,}61 < \mathbf{r} < 0{,}87$.

Сравнение интервальных оценок показывает, что оба доверительных интервала имеют одинаковую ширину, но 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 совершенно 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

$$y_t = \alpha_0 + \sum_{j=1}^{m} \alpha_j x_{t-j} + \varepsilon_t, \qquad (6.13)$$

где m – интервал запаздывания Y по отношению к X; ε_t – случайная компонента.

Как следует из формулы (6.10), при $j = 0$ связь между X и Y синхронная, а при $j \geq 1$ X 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: X влияет на Y в виде H_0: $\alpha_1 = \ldots = \alpha_m = 0$. Как будет показано в главе 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 прежде всего многих методов многомерной статистики, основанных, как правило, 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.

$$T = \begin{vmatrix} T_{11} & T_{12} & \dots & T_{1M} \\ T_{21} & T_{22} & \dots & T_{2M} \\ \dots & \dots & \dots & \dots \\ T_{N1} & T_{N2} & \dots & T_{NM} \end{vmatrix}.$$

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 искомого 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 века. 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 (табл. 6.1). Коэффициенты корреляции для температуры воды составляют верхний треугольник матрицы, а для солености – нижний треугольник.

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

Прежде всего оценим значимость коэффициентов корреляции в табл. 6.1. Воспользуемся для этого формулой (6.7):

$$\mathbf{r}_{кр} = \mathbf{t}_{кр}/(n - 2 + \mathbf{t}^2_{кр})^{1/2}.$$

Величина критерия Стьюдента при $\alpha = 0{,}05$, $\nu = 46$ равна $\mathbf{t}_{кр} = 2{,}02$. Тогда $\mathbf{r}_{кр} = 2/(50)^{1/2} = 0{,}28$. Далее осуществляется проверка неравенства $|r| > \mathbf{r}_{кр}$. Если данное неравенство выполняется, то коэффициент корреляции считается значимым. Незначимые оценки коэффициентов корреляции отмечены в табл. 6.1 чертой снизу.

Как видно из табл. 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 столь 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. Чем выше значения САК, тем интенсивнее зональный перенос. На рис. 6.3 приводится карта изокоррелят между средними годовыми значениями САК и индексом общей циклоничности С, представляющим собой произведе-

ние интенсивности циклонов на их повторяемость и таким образом являющимся интегральным показателем циклонических синоптических вихрей. Значения индексов С 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 $r_{кр} = 0{,}30$.

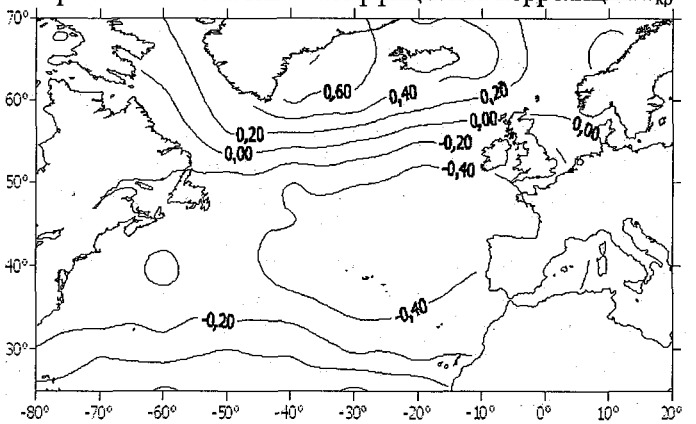


Рис. 6.3. Пространственное распределение коэффициентов корреляции между индексом общей циклоничности *С* и интенсивностью Северо-Атлантического колебания.

Из рис. 6.3 видно, что большая часть акватории Северной Атлантики занята значимой корреляцией САК с индексом *С*. При этом 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.*

Напомним, *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 Спирмэна (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 Спирмэна рассчитывается по формуле:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \qquad (6.14)$$

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:

$$t = \rho[(n-2)/(1-\rho^2)]^{1/2}. \qquad (6.15)$$

Значение коэффициента Спирмэна ρ считается значимым, если $t > \mathbf{t}_{кр}(\alpha,\ \nu = n-2)$.

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

$$\tau = [4Q/n(n-1)] - 1, \qquad (6.16)$$

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

$$\sigma = \sqrt{\frac{2(2n+5)}{9n(n-1)}}.$$

В этом случае t-критерий приобретает вид:

$$t = \frac{|\tau|}{\sigma} = |\tau|\sqrt{\frac{9n(n-1)}{2(2n+5)}}. \qquad (6.17)$$

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 Спирмэна и Кендалла при достаточно большом объеме исходной выборки существует почти функциональная связь

$$\tau \approx (2/3)\rho, \qquad (6.18)$$

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 табл. 6.2.

Таблица 6.2

Порядок расчета коэффициентов ранговой корреляции Спирмэна 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	–

Прежде всего выполним ранжирование значений ледовитости 1-го района в порядке убывания и присвоим им ранг от 1 до 10 (см. табл. 6.2). Теперь определим ранг значений ледовитости 2-го района. Максимальному значению ледовитости 1-го 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:

$$\Sigma d_i^2 = 1 + 1 + 16 + 4 + 9 + 9 + 4 + 16 = 60.$$

Теперь нетрудно рассчитать коэффициент корреляции Спирмэна:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} = 1 - (6 \cdot 60)/(10^3 - 10) = 0{,}64.$$

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:

$$t = \rho[(n - 2)/(1 - \rho^2)]^{1/2} = 0{,}64(8/0{,}59)^{1/2} = 2{,}36.$$

Критическое значение $\mathbf{t}_{кр}(\alpha = 0{,}05,\ \nu = 8) = 2{,}31$. Отсюда видно, что выборочное значение коэффициента корреляции Спирмэна 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 табл. 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 (6.13), находим $\tau = (4 \cdot 33/90) - 1 = 0{,}47$. Для оценки его значимости рассчитываем статистику Стьюдента $t = 0{,}47[(90 \cdot 9/2 \cdot 25)]^{1/2} = 1{,}88$. Сравнивая эту оценку с критической величиной $\mathbf{t}_{кр} = 2{,}31$ видим, что вычисленный коэффициент ранговой корреляции Кендалла оказывается *незначимым*.

Итак, использование коэффициентов корреляции Спирмэна 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 коэффициента Спирмэна приведены в табл. 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 табл. 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. Весь-

ма высокая отрицательная корреляция наблюдается с индексом Южного Колебания (ρ = –0,71), а также с другими параметрами. Это означает, что с заблаговременностью в два года можно построить прогностическую модель вылова ставриды.

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 Спирмэна			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

Сравнение коэффициентов Спирмэна 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 используются *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:

$$r_{cx} = (\bar{x}_1 - \bar{x}_0) p_c q_c / s_x z_p , \qquad (6.19)$$

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 C 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 (6.19) 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 $x_p = \Phi^{-1} \times$ $\times$ ($|p - 0{,}5|$). Затем по величине x_p из таблицы плотности нормального распределения находится значение $z_p = f(x_p)$. Подставляя найденное значение z_p в формулу (6.19), 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 каким-либо образом связаны функционально с третьей переменной X_3, то возникает ложная корреляция.

В этом нетрудно убедиться, если обратиться к следующему примеру. Пусть мы имеем переменные X_1 и X_2, корреляция между которыми отсутствует ($r_{X_1,X_2} = 0$). Сформируем два ряда отношений $Z_1 = X_1/X_3$ и $Z_2 = X_2/X_3$, где X_3 – некоторая третья переменная. 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 при-

близительном 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 $C_3 \approx 2C_1 \approx 2C_2$, то $r_{Z_1,Z_2} = 0{,}8$ и 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 $r_{ложн} = |r_{набл}| - |r_{ист}|$, где $r_{ист}$ – 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 в этом случае 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 $\Delta x_i = x_i - \bar{x}$, где $\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i$. 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, что в результате и привело к росту уровня Мирового океана.

На рис. 6.4 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.

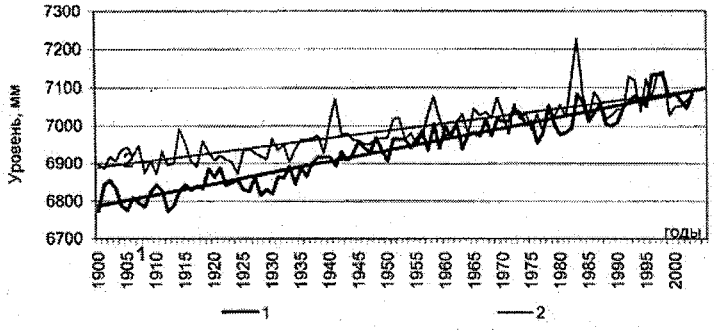


Рис. 6.4. Межгодовой ход морского уровня
на станциях Нью-Йорк (1) и Сан-Франциско (2) с 1901 г.

Корреляция между указанными рядами морского уровня равна $r = 0{,}85$. Естественно полагать, что наличие в каждом из рассматриваемых рядов трендовой компоненты (см. п. 10.2) приводит к появлению эффекта ложной корреляции. Поэтому рассчитаем уравнение линейного тренда (10.11) и вычтем тренд из рядов уровня (рис. 6.5).

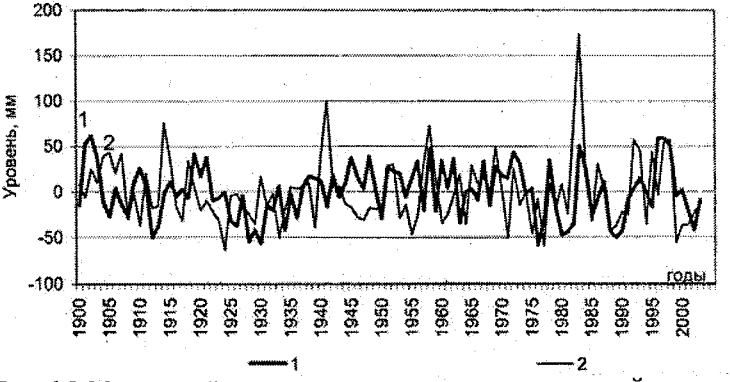


Рис. 6.5. Межгодовой ход морского уровня на станциях Нью-Йорк (1)
и Сан-Франциско (2) после исключения линейного тренда.

Как видно из рис. 6.5, 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 существенно ниже по сравнению с корреляцией между исходными рядами. Итак, получаем, что ложная корреляция оказывается равной $r_{лож} = 0{,}85 - 0{,}24 = 0{,}61$.

Глава 7. ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

7.1. Понятие о методе наименьших квадратов МНК

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 совокупность наблюдений $x_1, ..., x_n$ и $y_1, ..., y_n$, причем между ними существует некоторое приближенное соотношение $y = f(x; a_1, a_2, ..., a_m)$, где $a_1, a_2, ..., a_m$ – неизвестные параметры этой зависимости. В этом случае для отыскания неизвестных коэффициентов может быть использован *метод наименьших квадратов*, суть которого заключается в том, что сумма квадратов отклонений точек от линии связи должна быть наименьшей, т. е.

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - y_{(x)i})^2 = \sum_{i=1}^{n} [y_i - f(x; a_1, a_2, ..., a_m)]^2 = \min, \quad (7.1)$$

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 $n > m$.

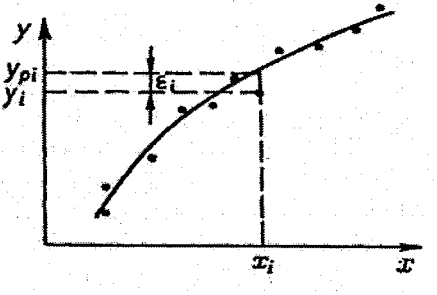


Рис. 7.1. Условие минимизации суммы квадратов отклонений переменной Y в стохастической зависимости между переменными X и Y.

Итак, в соответствии с формулой (7.1) требуется минимизировать сумму квадратов значений ε_i. Как видно из рис. 7.1, остатки ε_i представляют собой 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:

$$\begin{cases} \dfrac{\partial S}{\partial a_1} = -2\sum\limits_{i=1}^{n}[y_i - f(x;a_1,a_2,...,a_m)_i]\left(\dfrac{\partial f}{\partial a_1}\right) = 0, \\ \dfrac{\partial S}{\partial a_2} = -2\sum\limits_{i=1}^{n}[y_i - f(x;a_1,a_2,...,a_m)_i]\left(\dfrac{\partial f}{\partial a_2}\right) = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ \dfrac{\partial S}{\partial a_m} = -2\sum\limits_{i=1}^{n}[y_i - f(x;a_1,a_2,...,a_m)_i]\left(\dfrac{\partial f}{\partial a_m}\right) = 0. \end{cases}$$

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 xOy 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 трех линейных уравнений:

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = -2\sum\limits_{i=1}^{n}[y_i - a_0 - a_1x_i - a_2x_i^2] = 0, \\ \dfrac{\partial S}{\partial a_1} = -2\sum\limits_{i=1}^{n}[y_i - a_0 - a_1x_i - a_2x_i^2]x_i = 0, \\ \dfrac{\partial S}{\partial a_2} = -2\sum\limits_{i=1}^{n}[y_i - a_0 - a_1x_i - a_2x_i^2]x_i^2 = 0. \end{cases}$$

В результате несложных преобразований этой системы получим:

$$\begin{cases} a_0 n + a_1\sum\limits_{i=1}^{n}x_i + a_2\sum\limits_{i=1}^{n}x_i^2 = \sum\limits_{i=1}^{n}y_i, \\ a_0\sum\limits_{i=1}^{n}x_i + a_1\sum\limits_{i=1}^{n}x_i^2 + a_2\sum\limits_{i=1}^{n}x_i^3 = \sum\limits_{i=1}^{n}x_iy_i, \\ a_0\sum\limits_{i=1}^{n}x_i^2 + a_1\sum\limits_{i=1}^{n}x_i^3 + a_2\sum\limits_{i=1}^{n}x_i^4 = \sum\limits_{i=1}^{n}x_i^2y_i. \end{cases} \qquad (7.2)$$

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 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, о, как было показано Фишером, этосвязано 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 следующие: 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.

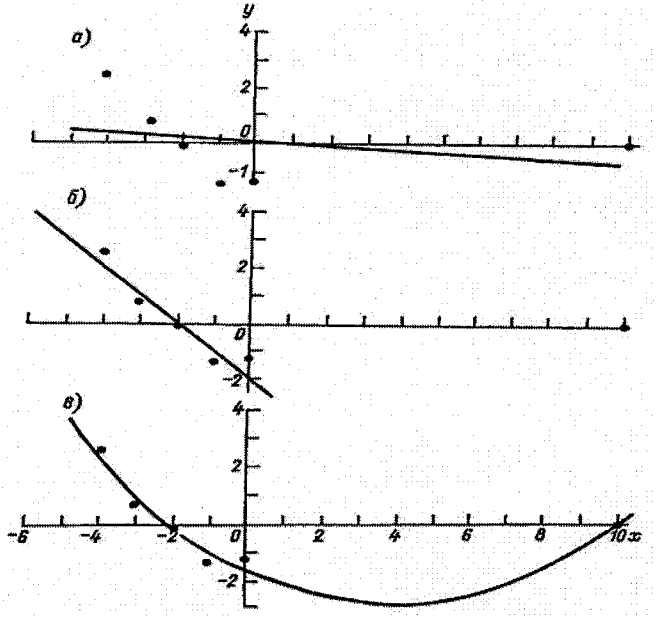


Рис. 7.2. Тестовый пример робастной регрессии.

Поскольку первые два условия достаточно очевидны, то рассмотрим третье условие. Как было указано выше, 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 клас-

сических статистических процедур. На рис. 7.2 представлен тестовый пример регрессионного анализа по шести точкам, значения которых приведены далее в табл. 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 линии и, следовательно, привести к значительному искажению оценок регрессионных параметров. Резко выделяющиеся наблюдения нарушают независимость дисперсии остатков от 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 посредственно-

сти», т.е. к среднему. В дальнейшем под регрессией стали понимать стохастические связи между переменными.

Запишем 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 собой *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 математическое ожидание ($M_\varepsilon = 0$);

2) дисперсия остатков должна быть постоянной (D_ε = const), 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.

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.

$$S = \sum_{1}^{n} [y_i - (a_1 x_i + a_0)]^2 = \min.$$

Отсюда нетрудно получить систему из двух нормальных линейных уравнений:

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = -2\sum\limits_{i=1}^{n} [y_i - a_0 - a_1 x_i] = 0, \\ \dfrac{\partial S}{\partial a_1} = -2\sum\limits_{i=1}^{n} [y_i - a_0 - a_1 x_i] x_i = 0 \end{cases}$$

или, после несложных преобразований,

$$\begin{cases} a_0 n + a_1 \sum x_i = \sum y_i, \\ a_0 \sum x_i + a_1 \sum x_i^2 = \sum x_i y_i. \end{cases} \tag{7.5}$$

Решая (7.5) относительно параметров a_0 и a_1, имеем:

$$a_1 = \frac{\sum^{n} (x_i y_i - \overline{nxy})}{\sum^{n} x_i^2 - n\overline{x^2}}, \tag{7.6}$$

$$a_0 = \overline{y} - a_1 \overline{x}. \tag{7.7}$$

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

$$a_0 = \bar{y} - r\frac{\sigma_y}{\sigma_x}\bar{x}. \qquad (7.9)$$

С учетом выражений (7.8) и (7.9) классическое уравнение регрессии может быть переписано в следующем виде:

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) + \varepsilon. \qquad (7.10)$$

Аналогичным образом может быть представлено и уравнение регрессии x по y:

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) + \varepsilon. \qquad (7.11)$$

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*. Физический смысл параметров a_0 и a_1 становится очевидным, если обратиться к рис. 7.3.

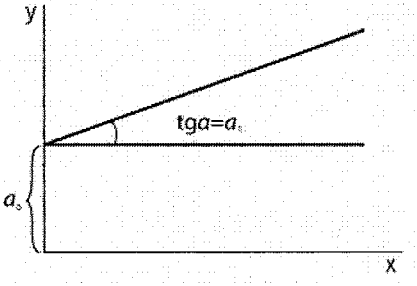


Рис. 7.3. Интерпретация коэффициентов в уравнении линейной регрессии.

Коэффициент регрессии a_1 представляет собой тангенс угла наклона линии регрессии к оси абсцисс (a_1 = tg α), а свободный

член – 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. 7.4). Если $a_0 = 0$, то линия регрессии проходит через начало координат.

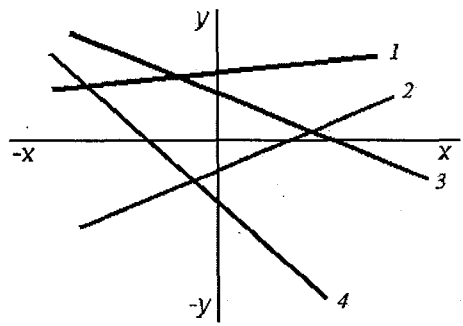


Рис. 7.4. Вид графика линейной регрессии в зависимости от значений параметров: *1* – $a_0 > 0$, $a_1 > 0$; *2* – $a_0 < 0$, $a_1 > 0$; *3* – $a_0 > 0$, $a_1 < 0$; *4* – $a_0 < 0$, $a_1 < 0$.

Нетрудно показать, что линия регрессии должна обязательно проходить через точку с координатами $x = \bar{x}$ и $y = \bar{y}$. Другой такой 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 стандартизованные аналоги. Для этого предварительно производится расчет стандартизованных переменных по формуле $z_{ij} = (x_{ij} - \bar{x_j})/\sigma_j$, т.е. из каждого наблюдения переменной вычитается средняя арифметическая и результат делится на ее стандартное отклонение. Стандартизованная 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 стандартизованное уравнение линейной регрессии:

$$z_y = \beta z_x. \qquad (7.12)$$

Здесь z_y, β, z_x – стандартизованные 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:

$$\beta = a_1(\sigma_x/\sigma_y), \tag{7.13}$$

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

К числу основных критериев качества модели относятся:

– *линейный коэффициент детерминации*, представляющий собой квадрат линейного коэффициента корреляции:

$$r^2 = D_{y(x)} / D_y = 1 - (D_\varepsilon/D_x), \tag{7.14}$$

где $D_{y(x)}$ – дисперсия вычисленных по уравнению регрессии значений функции отклика; D_x – дисперсия фактических значений переменной Y; D_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.

– *среднеквадратическое (стандартное) отклонение модели*:

$$\sigma_{y(x)} = \sqrt{\frac{\sum (y_i - y_{(x)_i})^2}{n-1}}. \tag{7.15}$$

Можно показать, что данная величина для нормально распределенных совокупностей функционально связана с линейным коэффициентом детерминации формулой:

$$\sigma_{y(x)} = \sigma_y (1 - r^2)^{1/2}. \qquad (7.16)$$

Отсюда видно, что чем выше коэффициент корреляции, тем меньшей оказывается стандартная погрешность уравнения регрессии.

– стандартные ошибки коэффициента корреляции и коэффициентов регрессии:

$$\sigma_r = \frac{1 - r^2}{\sqrt{n-1}}; \qquad (7.17)$$

$$\sigma_{a1} = \frac{\sigma_y}{\sigma_x}\sqrt{\frac{1 - r^2}{n-1}}; \qquad (7.18)$$

$$\sigma_{a0} = \frac{\sigma_{y(x)}}{\sqrt{n}} = \frac{\sigma_y \sqrt{1 - r^2}}{\sqrt{n-1}}. \qquad (7.19)$$

Напомним, 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 при построении доверительных интервалов. Для этой цели используется *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 $H_0 : a_j = 0$ при альтернативе $H_1 : a_j \neq 0$, для проверки которой вычисляется $t = | a_j | / \sigma_{bj}$. Далее проверяется неравенство:

$$t > \mathbf{t}_{кр}(\alpha, \nu = n - 2).$$

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 a_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* = 0,032 означает, что рассматриваемый коэффициент регрессии значим на уровне 0,05 и незначим 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_i и y_i 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:

$$D_y = D_{y(x)} + D_\varepsilon.$$

Но так как

$$D_{y(x)} = D(a_0 + a_1 x) = a_1{}^2 D_x = r(D_y/D_x)D_x = r^2 D_y,$$

то получим

$$r^2 = D_{y(x)}/D_y, \quad D_\varepsilon/D_y = 1 - r^2.$$

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 – как *межгрупповая дисперсия*, третий – как *внутригрупповая дисперсия*.

Отношение $r^2 = D_{y(x)}/D_y$ показывает степень ()приближения вычисленных значений $y_{(x)i}$ по уравнению регрессии к фактическим значениям y_i или, другими словами, долю дисперсии функции отклика, описываемой моделью регрессии.

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

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=}^{n}(y_{(x)i} - \bar{y})^2 + \sum_{i=}^{n}(y_{(x)i} - y_i)^2$$

или

$$Q = Q_R + Q_\varepsilon,$$

где Q – общая сумма квадратов отклонений зависимой переменной от средней, а Q_R и Q_ε – соответственно сумма квадратов, обусловленная регрессией и остаточная сумма квадратов, характеризующая случайную ошибку (влияние неучтенных факторов).

В результате оценки указанных сумм квадратов составляется таблица по следующей форме (табл. 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

$$D_{y(x)} = Q_R/(q-1), \quad D_\varepsilon = Q_\varepsilon/(n-q), \; D_y = Q/(n-1).$$

Таблица 7.1

Проверка адекватности линейной регрессионной модели $y = a_0 + a_1 x + \varepsilon$

Источник изменчивости	Сумма квадратов	Число степеней свободы	Критерий Фишера
Линейная регрессия	$Q_R = \sum(y_{(x)i} - \bar{y})^2$	$q-1$	
Отклонение от регрессии	$Q_\varepsilon = \sum(y_{(x)i} - y_i)^2$	$n-q$	
Общая изменчивость	$Q = \sum(y_i - \bar{y})^2$	$n-1$	$F = D_{y(x)}/D_\varepsilon$

При оценке адекватности (значимости) модели составляется нулевая гипотеза о равенстве дисперсий, т.е. $H_0 : D_{y(x)} = D_\varepsilon$ при альтернативе $H_1 : D_{y(x)} \neq D_\varepsilon$. Проверка нулевой гипотезы осуществляется с помощью F-критерия, который в соответствии с табл. 7.1 имеет вид

$$F = D_{y(x)}/D_\varepsilon = Q_R(n-q)/Q_\varepsilon\,(q-1) = Q_R(n-2)/Q_\varepsilon. \quad (7.20)$$

После этого проверяется неравенство

$$F > \mathbf{F}_{\text{кр}}\,(\alpha, \nu_1 = 1, \nu_2 = n-2).$$

Если $F > \mathbf{F}_{\text{кр}}(\alpha, \nu_1, \nu_2)$, то нулевая гипотеза о равенстве дисперсий отвергается и, следовательно, справедливой является альтернативная гипотеза $H_1 : D_{y(x)} \neq D_\varepsilon$. Это означает, что дисперсия, описываемая линейной регрессией, неслучайным образом отличается от дисперсии шума, вследствие чего регрессионную модель следует считать адекватной (значимой). В противном случае у нас есть основания полагать, что модель не является адекватной, т.е. 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 ($Q_В$), изменчивость которого полностью обусловлена зоной формирования стока выше г. Самара. 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.

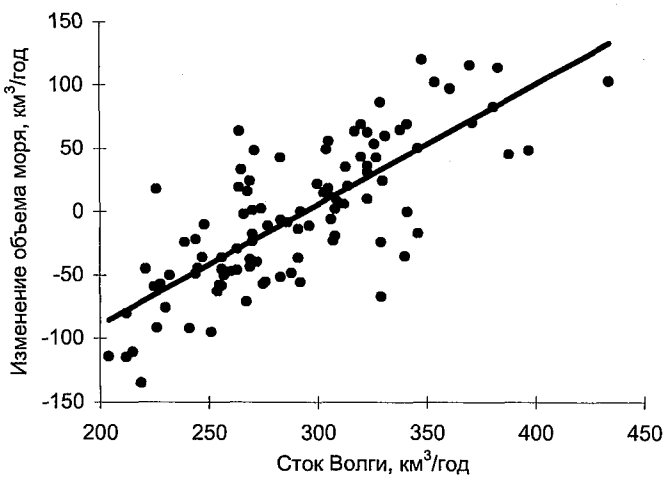


Рис. 7.5. Корреляционное поле между внутригодовыми изменениями объема Каспийского моря (км^3/год) и годовым стоком Волги в г. Самара (км^3/год) за период с 1890 по 1990 г.

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. 7.5. Нетрудно видеть, что между

ними отмечается довольно хорошо выраженная линейная связь. Первичные статистические характеристики ΔV и $Q_В$ приведены в табл. 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 ($Q_В$) является очевидной.

Таблица 7.2

Первичные статистические характеристики изменений объема моря (км^3/год) и стока Волги (км^3/год)

Параметр	Среднее	Стандартное отклонение	X_{max}	X_{min}	R
ΔV	–3,3	56,4	120	–135	255
$Q_В$	237	45	365	147	218

В результате использования МНК получено следующее уравнение регрессии:

$$\Delta V = -246{,}7 + 0{,}0323 Q_B \,. \qquad (7.21)$$

Оценки параметров этого уравнения даются ниже:

– коэффициент корреляции $r = 0{,}79$,

– коэффициент детерминации $r^2 = 0{,}63$,

– среднее квадратическое (стандартное) отклонение модели $\sigma_{y(x)} = 34{,}6$ км^3/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 оценками *p-level*. Для свободного члена *p-level* $= 6{,}8 \cdot 10^{-23}$, для коэффициента регрессии *p-level* $= 5{,}7 \cdot 10^{-23}$, 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.

Адекватность регрессионной модели проверяем по критерию Фишера: $F = D_{y(x)} / D_{\varepsilon} = 166{,}9$. Критическое значение статистики Фишера при $\alpha = 0{,}05$, $\nu_1 = 1$, $\nu_2 = 99$ равно $F_{кр} = 3{,}93$. Отсюда следует, что модель (7.21) полностью адекватна. Сопоставление фактических и вычисленных по модели значений ΔV представлено на рис. 7.6. 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 км^3/год, а максимальное занижение – в 1973 г., достигающее 93,8 км^3/год.

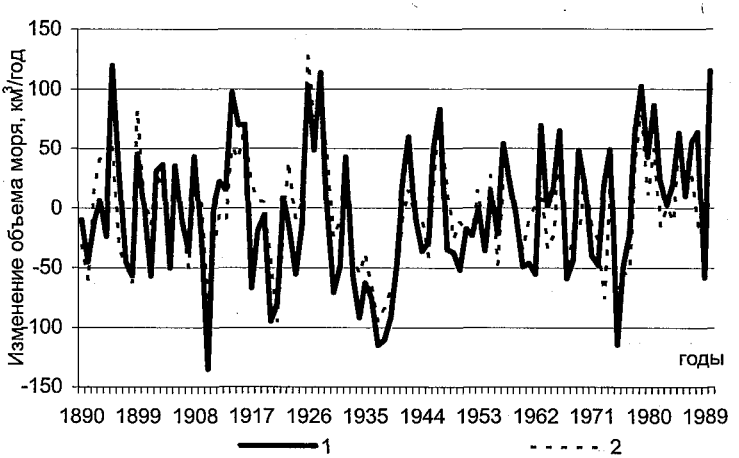


Рис. 7.6. Сопоставление фактических (*1*) и вычисленных по модели (*2*) значений изменений объема вод Каспийского моря за период с 1890 по 1990 г.

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. 7.7, *а*), то это означает адекватность модели. Если полоса рассеяния расширяется (сужается), когда значения x_i или y_i возрастают (рис. 7.7, *б*), то это указывает на непостоянство дисперсии остатков, называемое *гетероскедастичностью*. Если остатки зависят линейно или нелинейно от времени (рис. 7.7, *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. 7.7, *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 выводы.

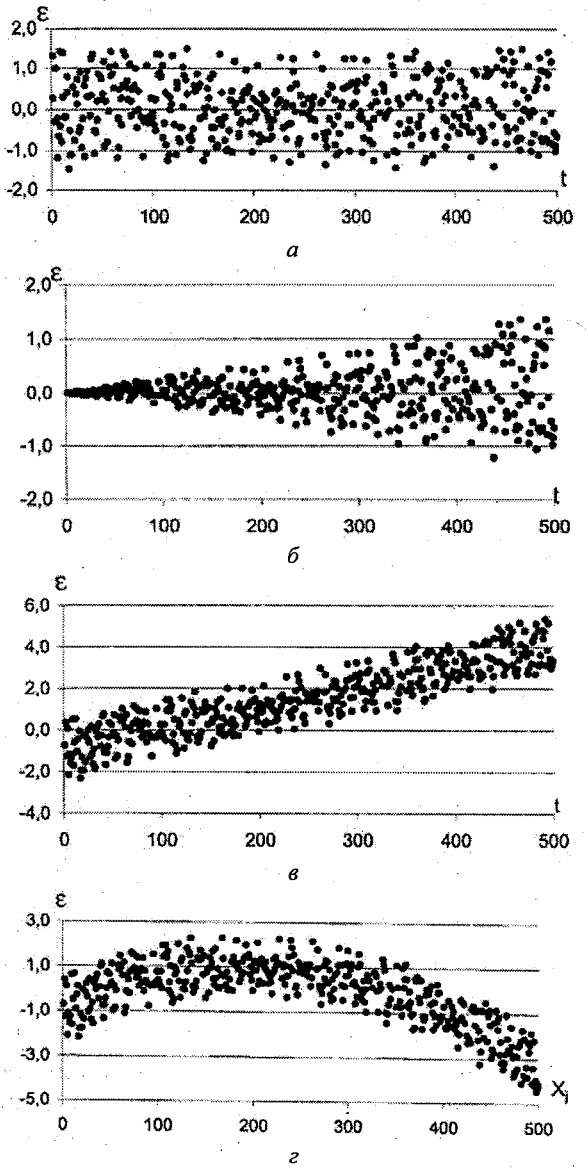


Рис. 7.7. Характеристика модели в зависимости от распределения остатков: *а* – модель адекватна, *б* – гетероскедастичность модели, *в* – наличие в модели линейного тренда, *г* – наличие нелинейной связи функции отклика с переменной *X*.

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:

$$d = \sum(\varepsilon_{i+1} - \varepsilon_i)^2 / \sum \varepsilon_i^2, \qquad (7.22)$$

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 нетрудно видеть, 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 $H_0 : d = 0$. 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 $0 \le d \le 4$. 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 табл. 7.3.

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

Из табл. 7.3 видно, что для двух диапазонов значений статистики d возникает неопределенность в интерпретации результатов, 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.

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. 7.8. 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.

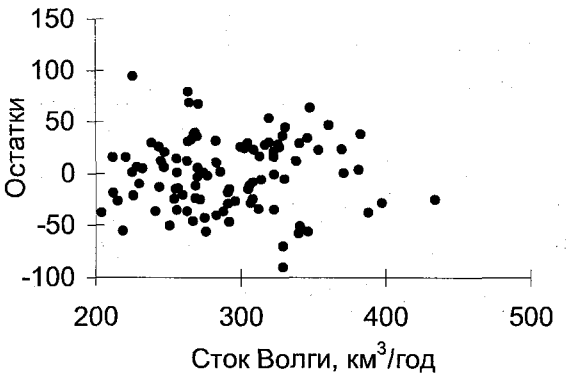


Рис. 7.8. Распределение остатков регрессионной модели (7.21) в зависимости от стока Волги.

7.6. Понятие о робастной регрессии

Как уже отмечалось выше, термин «робастный» означает устойчивость к сравнительно малым отклонениям от принятых предположений. Очевидно, робастная регрессия может служить дополнением к классическому методу наименьших квадратов в случае, если исходная 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 (табл. 7.4), для которых необходимо построить линейную регрессию. На рис. 7.2, *а* дана прямая, построенная с помощью МНК, уравнение которой имеет вид:

$$y_1 = 0{,}068 - 0{,}081x\,. \qquad (7.23)$$

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*):

$$y_2 = -1{,}87 - 0{,}977x. \qquad (7.24)$$

При этом коэффициент детерминации резко возрос, и модель стала значимой даже при самом жестком критерии Фишера. Естественно, стандартная ошибка модели тоже резко уменьшается (табл. 7.5).

Таблица 7.5

Оценки параметров регрессионных моделей

Модель	Параметры регрессионной модели		
	R^2	F	$\sigma_{(x)}$
7.22	0,10	0,4	1,55
7.23	0,91	31,1	0,95
7.24	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*):

$$y_3 = -1{,}74 - 0{,}66x + 0{,}08x^2. \qquad (7.25)$$

Точность данной модели стала лишь чуть-чуть выше (табл. 7.4), 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 ис-

кусственным. 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 $\sigma = 12$.

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.

$$MAD = \text{med}\{|\varepsilon_i - \text{med}(\varepsilon_i)|\}.$$

При этом остатки ε_i предварительно находятся с помощью классического МНК.

Для «засоренных» нормально распределенных выборок П. Хьюбер предлож 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 квадратов остатков. Как следует из графика функции Хьюбера (рис. 7.9), парабола на отрезке $[-h, h]$ продолжается далее двумя прямыми линиями. Значение h, определяющее порог, после которого происходит уменьшение скорости роста ρ, является практически параметром, определяющим степень робастности оценок метода.

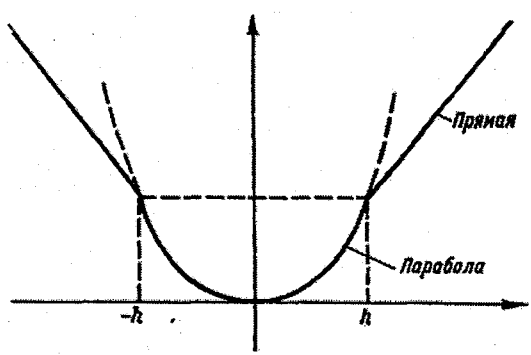


Рис. 7.9. График функции Хьюбера на отрезке $[-h, h]$.

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 Спирмэна $\rho = -0{,}70$). Корреляционное поле между этими переменными приводится на рис. 7.10, 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, тыс. т.

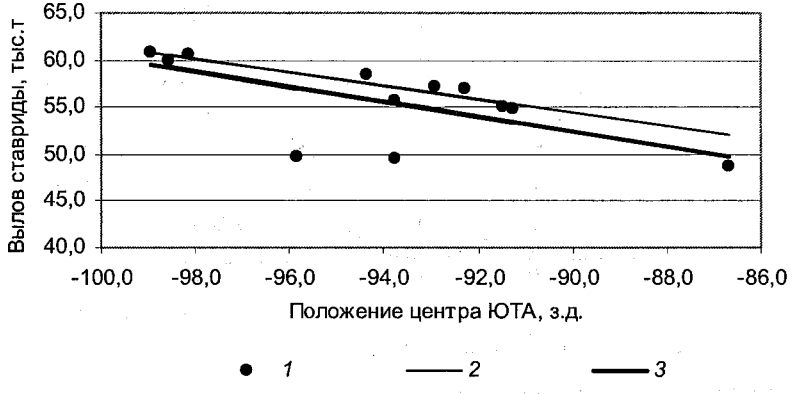


Рис. 7.10. График связи вылова ставриды и смещения ЮТА по долготе:
1 – фактические значения, *2* – робастная регрессия, *3* – линейная регрессия.

Коэффициенты регрессии будем определять классическим методом наименьших квадратов и робастным вариантом – методом

модифицированных весов. Статистические параметры уравнения регрессии даны в табл. 7.6, а сами графики уравнений – на рис. 7.10.

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

Нетрудно видеть, что робастная регрессия не учитывает не две точки, как это казалось визуально, а три точки, вылов в которых составлял меньше $50 \cdot 10^3$ 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 Спирмэна без трех точек возрос до $\rho = -0{,}93$.

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 табл. 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. 7.7 визуально вряд ли может быть признана отскакивающей. И только с помощью робастного подхода это удалось 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. 7.11, на левом графике корреляция между переменными X и Y является довольно высокой и равна $r = 0{,}80$. В то же 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 $r = 0{,}92$.

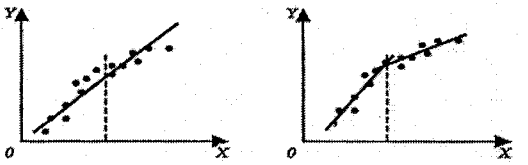


Рис. 7.11. Применение теста Чоу при построении кусочно-линейной регрессии.

Таким образом, 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 его 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

$$[S_0 - (S_1 + S_2)]/(m + 1)$$

определяет оценку уменьшения дисперсии за счет построения двух уравнений регрессии вместо одного. При этом число степеней свободы сократится на $(m + 1)$, поскольку вместо $(m + 1)$ парамет-

ров объединенного уравнения теперь необходимо оценивать $(2m + 2)$ параметров двух регрессий. Следовательно, дробь

$$(S_1 + S_2)/(n - 2m + 2)$$

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:

$$F = \frac{S_0 - S_1 - S_2}{S_1 + S_2} \cdot \frac{n - 2m - 2}{m + 1}. \qquad (7.30)$$

Далее осуществляется проверка неравенства $F > \mathrm{F}_{\text{кр}}(\alpha, \nu_1 = m + 1, \nu_2 = n - 2m - 2)$, где 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:

$$y_i = b_0 + \sum_{j=1}^{m} b_j x_{ij} + \varepsilon_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im} + \varepsilon_i, \qquad (7.31)$$

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 разброс 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 $(y_i, x_{ij}, \ldots, x_{im})$ 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 математическое ожидание ($M_\varepsilon = 0$);

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 $x_1, x_2, \ldots, x_m$ 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. В соответствии с четвертым предположением, единственным источником случайных возмущений значений y_i являются случайные возмущения регрессионных остатков ε_i. Но поскольку по определению ε_i – случайная величина, то соответственно y_i тоже является случайной величиной, причем ее закон распределения соответствует закону распределения ε_i.

Очевидным следствием данного предположения можно считать 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 $x_1, \ldots, x_m$ 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 $\mathbf{X'X}$ и, следовательно, ее детерминант (главный определитель) становится

равным нулю, т.е. $\det(\mathbf{X}'\mathbf{X}) = 0$, 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*.

Заметим, что выше рассмотрены предположения в явном виде, накладываемые на модель МЛР. Однако построение эффективной модели, особенно в прогностическом смысле, вообще говоря, становится бесполезной затеей, если набор исходных предикторов не отвечает определенным условиям. Это означает, что исходные данные должны отвечать ряду требований, к которым 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 различен. 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. Очень важным, особенно при большом числе предикторов, является требование их независимости, напрямую связанное с проблемой мультиколлинеарности, сущность которой будет рассмотрена ниже.

7.9. Вычисление и оценивание параметров множественной линейной регрессии

Коэффициенты регрессии 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.

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{i1} + \ldots + b_m x_{im})]^2 = \min,$$

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

$$(\mathbf{X}'\mathbf{X})\mathbf{B} = \mathbf{X}'\mathbf{Y}, \tag{7.32}$$

где $\mathbf{X}$ – матрица исходных данных; $\mathbf{B}$ и $\mathbf{Y}$ – диагональные матрицы коэффициентов регрессии и функции отклика.

Для решения этой системы умножим (7.32) на матрицу, обратную $(\mathbf{X}'\mathbf{X})$, т.е.

$$(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}). \tag{7.33}$$

Но поскольку произведение в левой части выражения (7.33) представляет единичную матрицу $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}) = \mathbf{I}$, то решение системы нормальных уравнений в матричной форме запишется следующим образом:

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}). \tag{7.34}$$

Стандартизованное уравнение МЛР по аналогии с (7.12) примет вид:

$$z_y = \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_m z_m + \xi = \sum \beta_j z_j. \tag{7.35}$$

Здесь z_y, β_j, z_j – стандартизованные значения функции отклика, коэффициентов регрессии и предикторов соответственно.

Нетрудно видеть, что свободный член в уравнении (7.35) равен нулю. Напомним, что физический смысл стандартизованных коэффициентов регрессии состоит в том, что они показывают относительную роль каждого предиктора в описании изменчивости функции отклика.

Приступая к оцениванию параметров модели МЛР, 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» тре-

бование, состоящее в необходимости выполнения одномерного нормального закона распределения для каждой переменной.

К числу основных критериев качества модели МЛР относятся:

– *множественный коэффициент 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.

$$R = \frac{1}{n\sigma_y\sigma_{y(x)}}\sum_{i=1}^{n}(y_i - \bar{y})(\tilde{y}_i - \bar{y}), \qquad (7.36)$$

где $\tilde{y}_i$ – вычисленные по модели МЛР значения отклика, σ_y – 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 $\mathbf{X'Y}$, 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:

$$R_{ск} = \sqrt{1 - \frac{D_\varepsilon(n-1)}{D_y(n-m)}} = \sqrt{1 - \frac{(1-R^2)(n-1)}{n-m}}. \qquad (7.37)$$

Итак, разность $R - R_{ск}$ – это поправка на положительное смещение величины R. Из формулы (7.37) видно, что для случая одномерной регрессии ($m = 1$) парный коэффициент корреляции является уже несмещенной оценкой. Отметим, что на практике величина $R_{ск}$ используется сравнительно редко. 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 ко-

эффициент $R_{ск}$ 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 корреляции:

$$R^2 = D_{y(x)}/D_y = 1 - (D_\varepsilon/D_x). \qquad (7.38)$$

Отсюда следует, что коэффициент детерминации показывает долю объясненной дисперсии функции отклика. Он функционально связан со стандартизованными 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 стандартизованного 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 новая $m + 1$ переменная линейно зависит от набора 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 ($D_y = D_\varepsilon$) и, следовательно, величина R^2 тоже должна быть равна нулю, определяется по следующей формуле:

$$M(R^2) = m/(n - 1).$$

Отсюда видно, что чем меньше разность $n–m$, тем больше величина R^2 отличается от нуля. По существу, это и означает, что коэффициент множественной корреляции имеет положительное смещение.

– среднеквадратическое (стандартное) отклонение модели;

$$\sigma_{y(x)} = \sqrt{\sum (y_i - \tilde{y}_i)^2 / (n - m - 1)}\,. \qquad (7.40)$$

Можно показать, что данная величина функционально связана с линейным коэффициентом детерминации формулой:

$$\sigma_{y(x)} = \sigma_y \sqrt{1 - R^2}\,. \qquad (7.41)$$

– стандартные ошибки множественного коэффициента корреляции и коэффициентов регрессии:

$$\sigma_R = \frac{1 - R^2}{\sqrt{n - m - 1}}, \qquad (7.42)$$

$$\sigma_{bj} = \frac{\sigma_y}{\sigma_{xj}} \sqrt{\frac{(1 - R^2) D_{yj}}{(n - m - 1) D_{yy}}}\,, \qquad (7.43)$$

где σ_{xj} – стандартное отклонение x_j предиктора; D_{yj} – минор главного определителя (детерминанта) у которого вычеркнута первая строка (y) и j-й 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 $H_0 : R = 0$, $H_0 : |b_j| = 0$. Проверка этой гипотезы осуществляется также с помощью t-критерия:

$$R > t_\alpha \sigma_R, \quad |b_j| > t_\alpha \sigma_{bj}.$$

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 $H_0 : D_{y(x)} = D_\varepsilon$, 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

$$F = Q_R\,(n - m - 1) \,/\, Q_\varepsilon\,(m - 1) = D_{y(x)} \,/\, D_\varepsilon. \qquad (7.44)$$

Вычисленное значение критерия Фишера сравнивается с его табличным (критическим) значением $\mathbf{F}_{кр}(\alpha,\ \nu_1,\ \nu_2)$ при заданном уровне значимости α и степенях свободы $\nu_1 = m$, $\nu_2 = n - 1$. Если выполняется неравенство $F > \mathbf{F}_{кр}$, то нулевая гипотеза о равенстве дисперсий вычисленных значений функции отклика и остатков отвергается и считается, что дисперсия, описываемая моделью МЛР, неслучайным образом отличается от дисперсии ошибок. Это означает, что рассматриваемая модель является адекватной (значимой), т.е. она хорошо соответствует исходным данным функции отклика. Обратный вывод делается, если $F < \mathbf{F}_{кр}$. Заметим также, что критерий Фишера функционально связан с коэффициентом детерминации следующей формулой:

$$F = R^2(n - m - 1) \,/\, [(1 - R^2)m]. \qquad (7.45)$$

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 случайные переменные 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. Но поскольку с физической точки зрения это не совсем понятно, то рассмотрим оценку ρ на примере простейшей модели МЛР:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2}.$$

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 :

$$x_{i1} = b''_0 + b''_1 x_{i2} + e_i.$$

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 ними соответствует частному коэффициенту корреляции:

$$\rho_{yx1} = \frac{1}{n\sigma_\varepsilon \sigma_e} \sum (\varepsilon_i - \bar{\varepsilon})(e_i - \bar{e}). \qquad (7.46)$$

Аналогичным образом может быть рассчитана величина ρ между y_i и X_2. Естественно, если мы имеем набор из m предикторов, то получаем набор из m частных коэффициентов корреляции.

Итак, 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:

$$R^2 = 1 - (1 - \rho^2_{yx1})(1 - \rho^2_{yx2})(1 - \rho^2_{yx3})\ldots(1 - \rho^2_{yxm}). \quad (7.47)$$

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.

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. rank $X < (m + 1)$. Как уже указывалось выше, это означает, что хотя бы одна переменная матрицы **X** может быть выражена линейной функциональной связью через остальные переменные. Вследствие этого матрица $\mathbf{X'X}$ оказывается вырожденной, т.е. ее определитель равен нулю. Так как в данном случае не существует обратной матрицы $(\mathbf{X'X})^{-1}$, то определение коэффициентов регрессии становится невозможным.

Поскольку строгая мультиколлинеарность на практике встречается весьма редко и может быть исключена на первоначальном этапе анализа исходных переменных, то рассмотрим только *реальную мультиколлинеарность*. Суть ее сводится к тому, что если в исходной матрице между большинством 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, 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 ($r_{ij} \geq 0{,}8 - 0{,}9$), то один из дублирующих факторов (X_i или X_j) может быть исключен.

В этом случае объем независимой информации, содержащейся в исходной матрице **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 описании функции отклика (Δ_{xj}). Если выполняется условие $\Delta_{xj} < 2\sigma_R/R$, 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 из которых имеет на один предиктор больше по сравнению с предыдущей моделью. С одной стороны, с включением новой переменной k должно выполняться следующее соотношение: $R_{k+1} \geq R_k$, 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 (n) выборки с включением в модель новой*

переменной повышается качество описания функции отклика, 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 степень мультиколлинеарности модели и, следовательно, ухудшается ее точность. Таким образом, 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

В общем случае построение оптимальной модели МЛР 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 $\mathbf{X}'\mathbf{X}$ 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 наиболее распространенными являются:

– метод включения переменных;
– метод исключения переменных.

Суть *метода включения переменных* заключается в том, что вначале на первом шаге выбирается наиболее коррелированный с функцией отклика предиктор и рассчитываются все параметры модели парной регрессии. 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.

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:

$$F_k = [\rho^2_{yxj}(n-k-2)]/(1-\rho^2_{yxj}). \qquad (7.48)$$

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 происходит проверка адекватности (значимости) модели и сравнение с некоторым пороговым (критическим) значением $\mathbf{F}_{кр}$. Величина $\mathbf{F}_{кр}$ может быть задана самим исследователем. По умолчанию она обычно принимается в ППСП $\mathbf{F}_{кр} = 4{,}0$. Как только величина F_k становится меньше $\mathbf{F}_{кр}$, программа прекращает работу и последний шаг принимается за оптимальную модель регрессии. Заметим, однако, что при этом не все коэффициенты регрессии оказываются значимыми.

Метод исключения переменных реализует обратную процедуру. Вначале строится полная (из m переменных) модель МЛР. Затем из нее исключается наименее значимый (по частному F-критерию)

фактор. 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 $\mathbf{F}_{кр}$. При выполнении условия $F > \mathbf{F}_{кр}$ полученное уравнение МЛР считается оптимальным. По сравнению с методом включения в данной процедуре в некоторых ППСП по умолчанию принимается чуть меньшее значение $\mathbf{F}_{кр}$ ($\mathbf{F}_{кр} = 3{,}9$).

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 и с формальными (вычислительными) аспектами. На наш взгляд, при большом числе переменных в модели предпочтения заслуживает все же первый алгоритм. Действительно, 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, а при совместном включении в модель они адекватно описывают функцию отклика.

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 матрица $(\mathbf{X}'\mathbf{X})^{-1}$ не является ортогональной, т.е. при коррелированности факторов. Это приводит к смещенности значений функции отклика.

Однако главная проблема, на наш взгляд, 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. Прежде всего целесообразно рассчитать полный комплекс (от 1 до m) моделей. И если есть возможность, то 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 ее указанные выше параметры.

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. 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 тесноты этих связей 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 г. 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:

$$y_i^{(1)} = b_0 + b_1 x_{i1} \qquad (R_{(1)}^2,\ \sigma^{(1)}_{y(x)},\ F^{(1)},\ p\text{-}level^{(1)}_{\max})$$

$$y_i^{(2)} = b_0 + b_1 x_{i1} + b_2 x_{i2} \qquad (R_{(2)}^2,\ \sigma^{(2)}_{y(x)}, F^{(2)}, p\text{-}level^{(2)}_{\max})$$

$$\ldots\ldots\ldots\ldots\ldots\ldots \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$y_i^{(8)} = b_0 + b_1 x_{i1} + \ldots + b_6 x_{i6} \qquad (R_{(6)}^2,\ \sigma^{(8)}_{y(x)}, F^{(8)}, p\text{-}level^{(8)}_{\max}).$$

Таблица 7.7

Статистические оценки параметров пошаговых моделей МЛР температуры поверхности океана в области Норвежского течения в зависимости от гидрометеорологических характеристик за 1949–2001 г.

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

Результаты оценивания этих моделей даны в табл. 7.7. Кроме того, графики параметров моделей в зависимости от шага даны на рис. 7.12. 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. Средняя квадратическая 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, она имеет вид:

$$T_w = b_0 + b_1 T_a - b_2 TB - b_3 V + \varepsilon$$

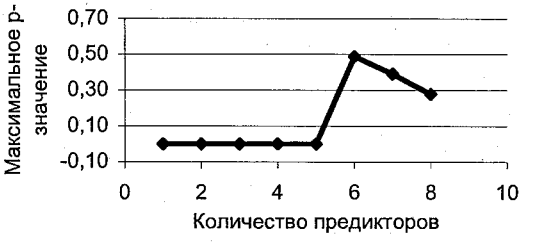
или в стандартизованной форме

$$z_{T_w} = 1{,}35 z_{T_a} - 0{,}54 z_{TB} - 0{,}28 z_V + \varepsilon.$$

а

б

в



г

Рис. 7.12. 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* – $p\text{-}level_{(\max)} = f(m)$.

Из последнего уравнения видно, что вклад температуры воздуха в описании изменчивости T_w более чем в два раза выше вклада *TB* и 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, использовав метод включения переменных, рассчитаем последовательно ряд моделей МЛР, основные параметры для которых приведены в табл. 7.8.

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	$R_{(j)}^2$	$\sigma^{(j)}_{y(x)}$, °C	$F^{(j)}$	$p\text{-}level^{(j)}_{\max}$
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

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 средней квадра-

тической ошибки модели. На пятом шаге она равна 0,07 °С, т.е. становится сравнимой с ошибкой измерения. Все коэффициенты регрессии значимы по критерию Стьюдента. Критерий Фишера во много раз превышает его критическую величину ($\mathbf{F}_{кр}$ = 4,0), а коэффициент детерминации близок к единице.

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 фактически 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

Существуют различные способы установления связи между гидрометеорологическими процессами или явлениями, которая, вообще говоря, может носить любой заранее неизвестный, причем, как правило, нелинейный характер. В общем случае нелинейные модели делятся на два класса:

– модели, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам,

– модели, 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.

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 же 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.

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:

$$E_0 = 6{,}1 \cdot 10^{\frac{7{,}45t}{235+t}}.$$

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 три 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 тогда 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 заключается 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 табл. 8.1. Эти формулы описывают довольно широкий класс нелинейных зависимостей.

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.

Таблица 8.1

Математические функции и их линеаризация

Вид функции	Новые переменные		Уравнение в линейной форме
	y'	x'	
$y = a_0 + a_1/x$	y	$1/x$	$y = a_0 + a_1 x^{-1}$
$y = 1/(a_0 + a_1 x)$	$1/y$	x	$y^{-1} = a_0 + a_1 x$
$y = x/(a_0 + a_1 x)$	x/y	x	$x/y = a_0 + a_1 x$
$y = a_0 a_1^{x}$	$\lg y$	x	$\lg y = \lg a_0 + x \lg a_1$
$y = a_0 e^{a_1 x}$	$\ln y$	x	$\ln y = \ln a_0 + a_1 x$
$y = 1/(a_0 + a_1 e^{-x})$	$1/y$	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	$\lg x$	$y = a_0 + a_1 \lg x$
$y = a_0/(a_1 + x)$	$1/y$	x	$y^{-1} = \frac{a_1}{a_0} + \frac{1}{a_0} x$
$y = a_0 x/(a_1 + x)$	$1/y$	$1/x$	$y^{-1} = \frac{a_1}{a_0} + \frac{1}{a_0} x^{-1}$
$y = a_0 e^{\frac{a_1}{x}}$	$\ln y$	$1/x$	$\ln y = \ln a_0 + a_1 x^{-1}$
$y = a_0 + a_1 x^n$	y	x^n	$y = a_0 + a_1 x^n$

Определение параметров выбранной формулы. В тех случаях, когда определению подлежат два неизвестных параметра эмпирической формулы, которая с помощью выравнивания может быть приведена к линейной зависимости, используется метод наименьших квадратов. Этот метод, суть которого изложена в главе 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 уравнение в линейной форме записывается как

$$\lg y = \lg a_0 + a_1 \lg x,$$

где $y' = \lg y$, $a_0' = \lg a_0$, $a_1' = a_1$, $x' = \lg x$.

Далее составляется система нормальных уравнений:

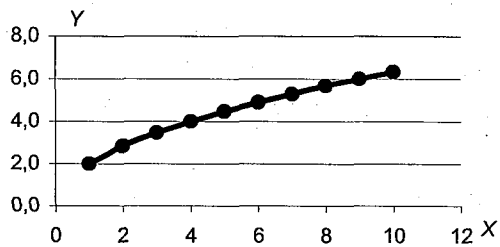
$$a_0' n + a_1' \Sigma x_i' = \Sigma y_i',$$

$$a_0 \Sigma x_i' + a_1' \Sigma x_i'^2 = \Sigma x_i' y_i',$$

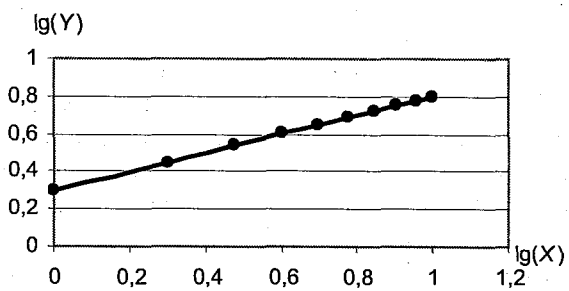
и вычисляются параметры a_0' и a_1':

$$a_1' = \frac{\Sigma(\bar{x}_i' y_i' - n\bar{x}'\bar{y}')}{\Sigma x_i'^2 - n\bar{x}'^2}, \quad a_0' = \bar{y}' - a_1' \bar{x}'.$$

Зная a_1' и a_0', уже не представляет труда найти параметры исходной формулы a_0 и a_1.



а



б

Рис. 8.1. График функции $y = a_0 x^{a_1}$ в исходной (*а*) и в логарифмической (*б*) системе координат.

В качестве иллюстрации сказанному на рис. 8.1 приводится 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 $y = 2x^{0,5}$.

Следует, однако, 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 данном конкретном случае 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

$$\varepsilon_i = y_{(x)i} - y_i.$$

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 $\bar{\varepsilon} = \sum_{i=1}^{n} \varepsilon_i / n = 0$. 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:

$$\eta = \sqrt{\frac{D_{y(x)}}{D_y}} = \sqrt{\frac{\sum (y_{(x)i} - \bar{y})^2}{\sum (y_i - \bar{y})^2}}. \qquad (8.1)$$

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.

Итак, отсюда следует, *что корреляционное отношение служит безразмерной мерой 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 средняя квадратическая ошибка, определяемая как

$$\sigma_{y(x)} = \sqrt{\frac{\sum \varepsilon_i^2}{n-1}} = \sqrt{\frac{\sum (y_i - y_{(x)i})^2}{n-1}}. \qquad (8.2)$$

Нетрудно показать, что данная формула может быть представлена в виде:

$$\sigma_{y(x)} = \sigma_y(1 - \eta^2)^{1/2}. \qquad (8.3)$$

Однако следует помнить, что наиболее действенным способом оценки точности эмпирической формулы является расчет значений функции отклика $y_{(x)i}$ *по независимым данным* (данным, не вошедшим в исходную выборку) и последующее 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 параметра. При этом непосредственное использование МНК оказывается не всегда возможным, ибо далеко

не во всех случаях удается привести эмпирическую формулу к линейному по параметрам виду. 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.

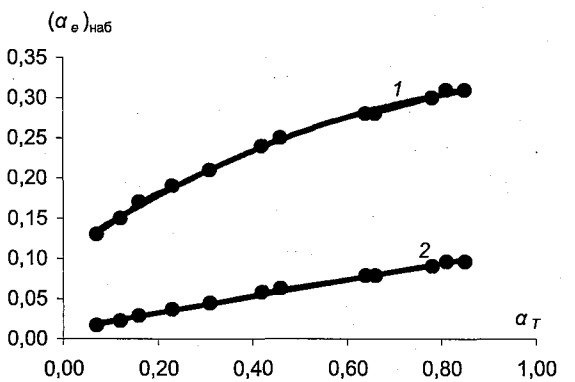


Рис. 8.2. Зависимости безразмерного вертикального профиля влажности (α_e) от вертикального профиля температуры (α_T): $1 - \alpha_e = f(\alpha_T)$, $2 - \alpha_e^2 = f(\alpha_T)$.

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. 8.2 и табл. 8.2). Проверка всех эмпирических формул, приведенных в табл. 8.1, показала, что полного выравнивания нет ни в одном случае.

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. $y' = \alpha_e^2$.

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
$\alpha_T \cdot 10^2$	0,85	0,81	0,66	0,42	0,23	0,12
$(\alpha_e)_{наб}$	0,31	0,31	0,28	0,24	0,19	0,15
$(\alpha_e)_{выч}$	0,31	0,30	0,28	0,24	0,19	0,16

Параметр	VII	VIII	IX	X	XI	XII
$\alpha_T \cdot 10^2$	0,07	0,16	0,31	0,46	0,64	0,78
$(\alpha_e)_{наб}$	0,13	0,17	0,21	0,25	0,28	0,30
$(\alpha_e)_{выч}$	0,14	0,17	0,21	0,24	0,28	0,30

Как следует из графика связи величин α_e^2 и α_T (рис. 8.2, 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.

$$\alpha_e = (0{,}012 + 10\alpha_T)^{1/2}. \qquad (8.4)$$

Предвычисленные по формуле (8.4) значения α_e также представлены в табл. 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:

$$y = a_0 10^{a1x/(a2+x)}. \qquad (8.5)$$

Логарифмируя формулу (8.5), получим:

$$\lg y = \lg a_0 + a_1 x /(a_2 + x).$$

Или после некоторых преобразований

$$[\lg(y/a_0)]^{-1} = (a_2 + x)/a_1 x = (a_2/a_1)x^{-1} + a_1^{-1}.$$

Осуществим теперь замену переменных:

$$y' = [\lg(y/a_0)]^{-1},\ x' = x^{-1},\ a'_1 = a_2/a_1,\ a'_0 = a_1^{-1}.$$

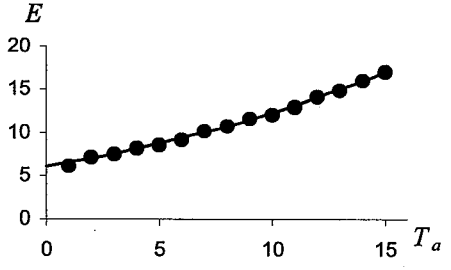


Рис. 8.3. Зависимость насыщающего давления водяного пара от температуры воздуха (формула Магнуса).

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. 8.3). 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 $a_2 = a'_1 / a'_0$. После этого рассчитываем среднюю квадратическую ошибку модели как

$$\sigma_{y(x)} = [\sum \varepsilon_i^2/(n-1)]^{1/2}.$$

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. $\sigma_{y(x)} = \min$.

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 $\Delta = 0{,}1$. Пересчитываем после этого уравнение (8.5). Минимального значения средняя квадратическая ошибка достигает при $a_0 = 8{,}7$ мм/год. Итак, окончательно имеем:

$$[W] = 8{,}7 \cdot 10^{3.83[Ta]/([Ta]-168)}. \qquad (8.6)$$

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. 8.4). Следовательно, $q = m - 1$, где q – число экстремумов функции (8.7).

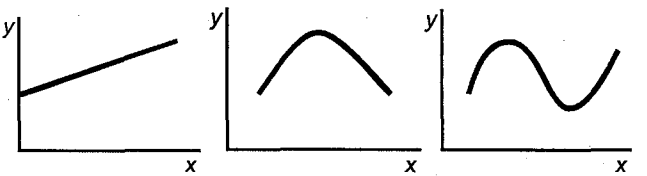


Рис. 8.4. Вид полиномиальной кривой в зависимости от степени полинома.

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:

$$S = \sum_{i=1}^{n} [y_i - (a_0 + a_1 x + ... + a_m x^m)_i]^2 = \min. \qquad (8.8)$$

Приравняв частные производные от S по всем неизвестным параметрам a_0, a_1, ..., a_m к нулю

$$\frac{\partial S}{\partial a_0} = 0, \frac{\partial S}{\partial a_1} = 0, ..., \frac{\partial S}{\partial a_m} = 0,$$

получим систему нормальных уравнений, решив 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,
– средняя квадратическая ошибка модели $\sigma_{y(x)}$,
– средние квадратические 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 $\mathbf{t}_{кр}(\alpha, \nu = n - m - 1)$, где m – степень полинома. Аналогичным образом определяется и критическое значение статистики Фишера $\mathbf{F}_{кр}(\alpha, \nu_1 = m, \nu_2 = n - 1)$.

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:

$$\eta \geq |r| . \qquad (8.9)$$

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 добавле-

ние 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 осуществлена с помощью критерия Стьюдента. T-статистика рассчитывается как

$$t = (\eta^2 - r^2)/\delta_{(\eta 2 - r2)}, \qquad (8.10)$$

где $\delta_{(\eta 2 - r2)}$ – величина ошибки разности $\eta^2 - r^2$.

После этого проверяется неравенство $t > \mathbf{t}_{кр}(\alpha, \nu = n - m - 1)$. Если данное неравенство выполняется, то нулевая гипотеза опровергается 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 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. Напри-

мер, чем меньше степень модели, тем она надежнее. Более подробно задача определения оптимальной степени модели рассматривается ниже в примере 8.4.

Пример 8.4. Один из возможных способов применения полиномиальной регрессии – это аппроксимация вертикальных профилей гидрометеорологических величин. Воспользуемся данными о вертикальном профиле солености, измеренного с помощью гидрозонда через 1 метр до глубины 15 м на гидрологической станции в Кандалакшской губе Белого моря (рис. 8.5). Вертикальный профиль солености является весьма типичным и имеет четыре характерные особенности:

– верхний квазиоднородный слой, который отмечается примерно до горизонта 4 м;

– 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.

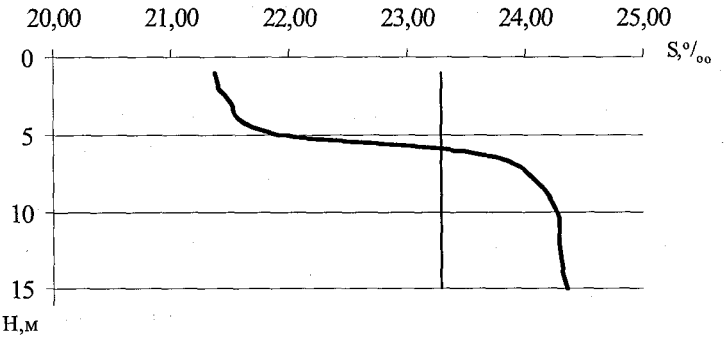


Рис. 8.5. Вертикальный профиль солености на гидрологической станции в Кандалакшской губе Белого моря. Вертикальная линия – среднее значение солености.

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 табл. 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	$7{,}2 \cdot 10^{-6}$
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 табл. 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 $\mathbf{F}_{кр}$. 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 шести уравнений модели. Из табл. 8.3 следует, что значимыми являются всего три уравнения модели: первой, второй и четвертой степени.

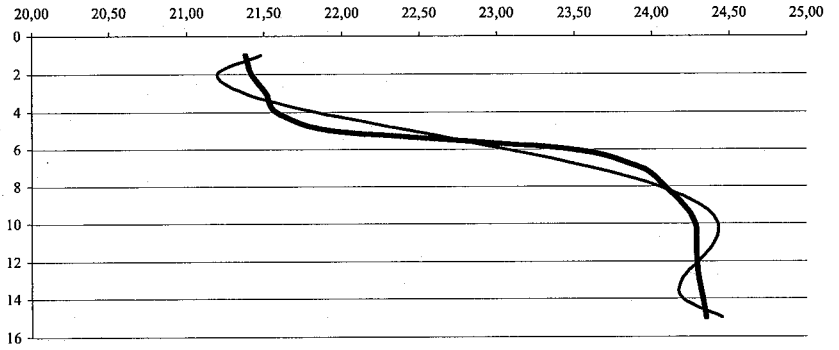
Итак, приняв во внимание противоречивый характер распределения параметров модели (8.7), необходимо выбрать такую степень уравнения, чтобы оно наилучшим (оптимальным) образом аппроксимировало вертикальный профиль солености. Очевидно, если исходить только из чисто формальных условий, то для оптимальной модели ход ее основных параметров должен быть следующим:

– коэффициент детерминации резко возрастает до определенного шага, затем почти не меняется;

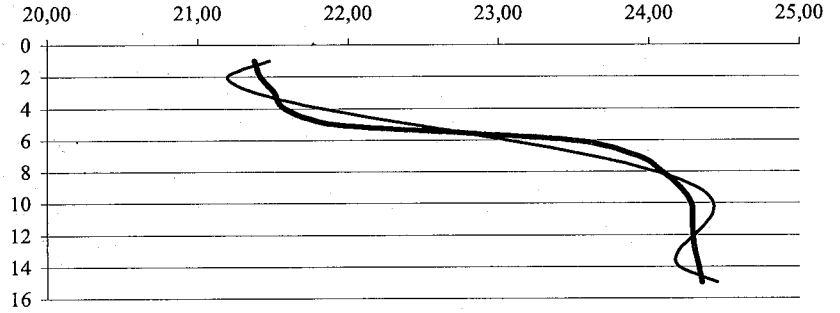
– стандартная ошибка на определенном шаге становится минимальной;

– оценка критерия Фишера больше его критического значения;

– *p-level* меньше заданного уровня значимости (например, $\alpha = 0{,}05$).



а



б

Рис. 8.6. Аппроксимация вертикального профиля солености на гидрологической станции в Кандалакшской губе Белого моря полиномиальной регрессией (*а*) и полиномами Чебышева и 4-го порядка (*б*).

В распределении коэффициента детерминации проявляется два 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. 8.6.

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 удастся производить 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 x, а как некоторую комбинацию многочленов. В результате уравнение (8.7) можно переписать в виде уравнения ортогональной регрессии:

$$y = a_0\varphi_0(x) + a_1\varphi_1(x) + \ldots + a_m\varphi_m(x) + \varepsilon. \qquad (8.11)$$

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:

$$\varphi_{m+1}(x) = \varphi_1(x)\varphi_m(x) - \frac{m^2(n^2 - m^2)}{4(4m^2 - 1)}\varphi_{m-1}(x). \qquad (8.12)$$

Например, многочлен $\varphi_2(x)$ будет иметь вид:

$$\varphi_2(x) = x^2 - (n+1)x + \frac{(n+1)(n+2)}{6}.$$

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. Опустив достаточно громоздкие промежуточные выкладки, приведем сразу окончательные формулы:

$$a_0 = \frac{\Sigma y_i}{n}, \qquad a_1 = \frac{\Sigma y_i \varphi_1(x_i)}{\Sigma \varphi_1^2(x_i)},$$

$$a_2 = \frac{\Sigma y_i \varphi_2(x_i)}{\Sigma \varphi_2^2(x_i)}, \ldots, a_m = \frac{\Sigma y_i \varphi_m(x_i)}{\Sigma \varphi_m^2(x_i)}. \qquad (8.13)$$

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 (см. рис. 8.5). Вначале определялись коэффициенты a_0, a_1, ..., 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 табл. 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).

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

Итак, до $m = 4$ результаты расчета вертикального профиля солености по обеим моделям полностью совпадают. Однако, начиная с $m = 5$, результаты начинают расходиться. Из табл. 8.4 видно, 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 табл. 8.4, то наилучшей следует признать модель с $m = 4$. 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 действительно 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 МНК.

Аналитическая аппроксимация заданного случайного поля может представлять 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 – как отклонение 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 двухмерной полиномиальной регрессии, первые шесть слагаемых – уравнение двухмерной полиномиальной регрессии второго порядка, а все выражение (8.16) представляет уравнение двухмерной 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:

$$\Sigma G = a_0 n + a_1 \Sigma x + a_2 \Sigma y;$$
$$\Sigma xG = a_0 \Sigma x + a_1 \Sigma x^2 + a_2 \Sigma xy;$$
$$\Sigma yG = a_0 \Sigma y + a_1 \Sigma xy + a_2 \Sigma y^2.$$

Аналогичным образом составляются системы нормальных уравнений для поверхностей тренда более высокого порядка. Решением линейного уравнения двухмерной полиномиальной регрессии является семейство прямых линий в системе $x0y$, а уравнения двухмерной полиномиальной регрессии второго порядка – семейство парабол (рис. 8.7). Более сложный характер имеет решение уравнения третьего порядка. Во многом это определяется знаками переменных X и Y последних четырех слагаемых в формуле (8.16).

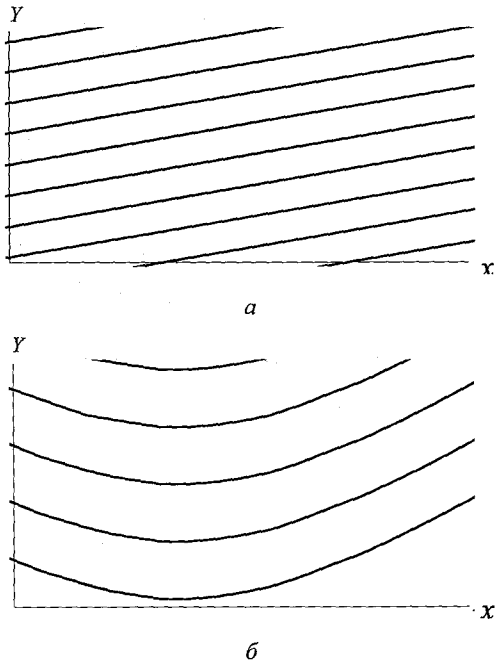


Рис. 8.7. Графический вид решения уравнения двухмерной полиномиальной регрессии:
а – линейное уравнение, *б* – уравнение второго порядка

Для оценки точности аппроксимации рассчитывается дисперсия фактических значений поля:

$$D_T = \frac{1}{n-1}\sum_{i=1}^{n}(G_{Ti} - \overline{G})^2, \qquad (8.17)$$

где n – число исходных точек, и дисперсия вычисленных по уравнению (8.16) значений G_R:

$$D_R = \frac{1}{n-1}\sum_{i=1}^{n}(G_{Ri} - \overline{G})^2 . \qquad (8.18)$$

Отношение D_R/D_T, представляющее собой коэффициент детерминации, дает качество приближения вычисленных значений аппроксимированного поля к его фактическим значениям.

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:

$$c_E = 0{,}85 \cdot 10^{-3} + 0{,}762 \cdot 10^{-4} \cdot U_{10} + 0{,}882 \cdot 10^{-4} \cdot \Delta T_{10} - \\ - 0{,}591 \cdot 10^{-6} \cdot U_{10}^2 - 0{,}11 \cdot 10^{-5} \cdot \Delta T_{10}^2 - 0{,}191 \cdot 10^{-5} \cdot U_{10} \cdot \Delta T_{10}, \quad (8.19)$$

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. п. 5.1).

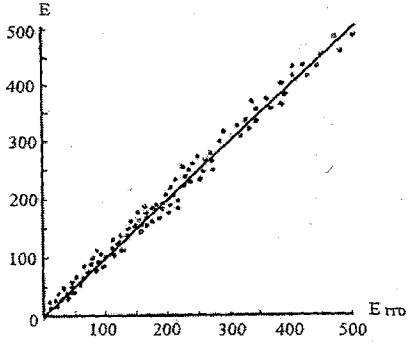


Рис. 8.8. Сравнение среднемесячных величин испарения, рассчитанных по методике ГГО и с использованием формулы (8.16).

Сравнение среднемесячных величин испарения для пяти судов погоды, вычисленных с использованием зависимости (8.19), с аналогичными значениями испарения, рассчитанными по методике ГГО, показало (рис. 8.8.), 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. 8.9, на котором представлено множество 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, касательная к сплайну во внутренней точке P_i обозначена через P_i'. Кубическая сплайн-функция строится по каждой паре точек. В общем виде уравнение кубического сплайна можно записать как

$$P_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3.$$

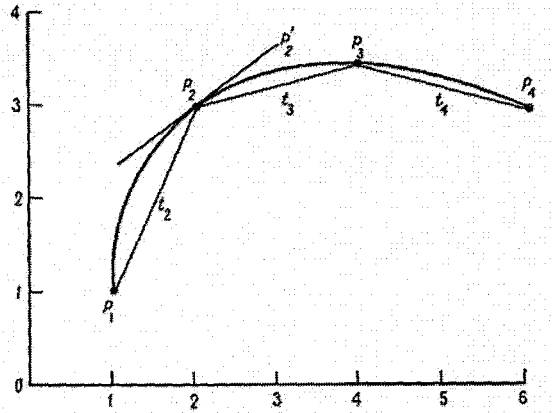


Рис. 8.9. Построение кубической сплайн-функции для четырех точек.

Для определения неизвестных коэффициентов 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 сплай-

ном 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

$$\varphi''(x) = \frac{K_{j-1}(x_j - x)}{h_j} + \frac{K_j(x - x_{j-1})}{h_j},$$

где $h_j = x_j - x_{j-1}$, $j = 2, \ldots, N$.

Проинтегрировав данное равенство дважды на отрезке $x - x_{j-1}$ и определив константы интегрирования, получим аналитическое выражение кубического сплайна:

$$\varphi(x) = \frac{K_{j-1}(x_j - x)^3}{6h_j} + \frac{K_j(x - x_{j-1})^3}{6h_j} + \\ + \frac{x_j - x}{h_j}\left(y_{j-1} - \frac{1}{6}K_{j-1}h_j^2\right) + \frac{x - x_{j-1}}{h_j}\left(y_j - \frac{1}{6}K_j h_j^2\right). \qquad (8.20)$$

На практике вместо (8.20) предпочитают пользоваться обычным кубическим полиномом. Если $x \in \Delta_j = [x_{j-1}, x_j]$, то аналитическое выражение кубического сплайна (8.17) можно переписать в виде

$$y(x) = y_{j-1} + a_{1,j-1}(x - x_{j-1}) + a_{2,j-1}(x - x_{j-1})^2 + a_{3,j-1}(x - x_{j-1})^3, \qquad (8.21)$$

где коэффициенты полинома (8.21) связаны с коэффициентами сплайна (8.20) следующими формулами:

$$a_{1,j-1} = h_j^{-1}(y_j - y_{j-1}) - h_j^{-1}(K_j/6 + K_{j-1}/3),$$
$$a_{2,j-1} = K_{j-1}/2,$$

$$a_{3,j-1} = (K_j - K_{j-1})/\ 6h_j.$$

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:

$$|y_j - f(x_j)| \le \delta,\ j = 1,\ldots, N.$$

Так как кубический сплайн имеет минимальную кривизну, то при построении сглаживающего сплайна естественно потребовать, чтобы он минимизировал характеризующий кривизну интеграл

$$\Phi(\varphi) = \int_a^b [\varphi''(x)]^2 dx \to \min. \qquad (8.22)$$

и при этом удовлетворялись условия

$$|f(x_j) - y_j| \le \delta,\ j = 1, \ldots, N. \qquad (8.23)$$

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:

$$\Phi(\varphi) = \int_a^b [\varphi''(x)]^2 dx + \sum_{j=1}^{N} \frac{[f(x_j) - y_j]^2}{\rho_j}, \qquad (8.24)$$

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 $K_j = \varphi''(x)$. Тогда коэффициенты сглаживающего сплайна можно найти путем решения системы линейных алгебраических уравнений:

$$
\begin{aligned}
& c_1K_1 + b_1K_2 + a_1K_3 = d_1, \\
& b_1K_1 + c_2K_2 + b_2K_3 + a_2K_4 = d_2, \\
& a_{i-2}K_{i-2} + b_{i-1}K_{i-1} + c_iK_i + b_iK_{i+1} + a_iK_{i+2} = d_i, \\
& a_{N-3}K_{N-3} + b_{N-2}K_{N-2} + c_{N-1}K_{N-1} + b_{N-1}K_N = d_{N-1}, \\
& a_{N-2}K_{N-2} + b_{N-1}K_{N-1} + c_NK_N + d_N. \qquad (8.25)
\end{aligned}
$$

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. 8.10 приводятся графики аппроксимации с помощью сглаживающего сплайна функции $y = 0{,}3\,x + 7\sin 5x - 5\cos 8x$ при $\rho = 0{,}1$ (рис. 8.10, *а*) и $\rho = 0{,}005$ (рис 8.10, *б*). Нетрудно видеть, что уменьшение параметра ρ существенно повышает точность аппроксимации данной функции. Однако вопрос заключается в том, действительно ли необходима такая точность или нет. Дело в том, что с повышением точности одновременно аппроксимируются и случайные ошибки экспериментальных данных. Однако 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 $\rho = 0{,}005$.

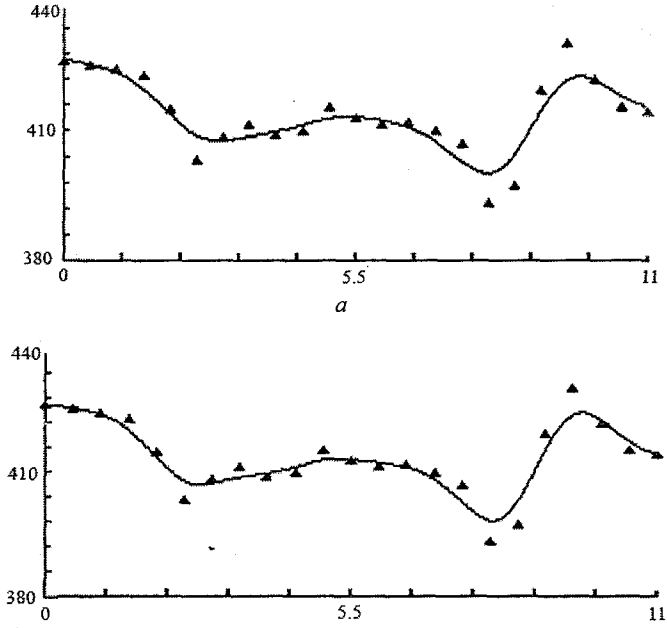


Рис. 8.10. Аппроксимация функции $y = 0{,}3x + 7\sin 5x - 5\cos 8x$ с помощью кубического сглаживающего сплайна при $\rho = 0{,}1$ (*а*) и $\rho = 0{,}005$ (*б*).

Часть 3. АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

Глава 9. ОСНОВНЫЕ ПОНЯТИЯ О СЛУЧАЙНЫХ ПРОЦЕССАХ

9.1. Понятие случайной функции

Вследствие того что природные процессы протекают во времени и пространстве, на практике мы обычно имеем дело не с отдельными случайными величинами, а с их множеством, зависящим от одного или нескольких параметров. Таким образом, приходим к понятию *случайной функции*, представляющей собой обобщение понятия случайной величины. Как известно, случайная величина в процессе опыта принимает одно, заранее неизвестное значение. Такие случайные величины формируются, если комплекс внешних условий, порождающий их, остается постоянным. Однако в действительности этот комплекс обычно изменяется, что приводит к изменению случайной величины в процессе опытов (наблюдений). Подобные случайные величины, изменяющиеся во времени или в пространстве, являются уже случайными функциями. По существу, случайная функция – это совокупность случайных величин.

Итак, *случайной функцией называется такая неслучайная функция, значения которой при каждом значении аргумента представляют случайную величину.*

В качестве примера случайной функции рассмотрим распределение среднемесячных значений температуры поверхности океана за 11 лет (1990–2000 гг.) в районе судна погоды «М» (66° с.ш., 2° в.д.), представленное на рис. 9.1. Естественно, что формирование температуры происходит под воздействием большого числа факторов, составляющих комплекс внешних условий. Очевидно, что эти факторы, сложным образом взаимодействуя друг с другом, оказываются непостоянными во времени. Поэтому каждая годовая серия температуры, похожая на любую другую общими закономерностями (минимальными значениями зимой и максимальными летом), в то же время имеет только ей присущий конкретный вид.

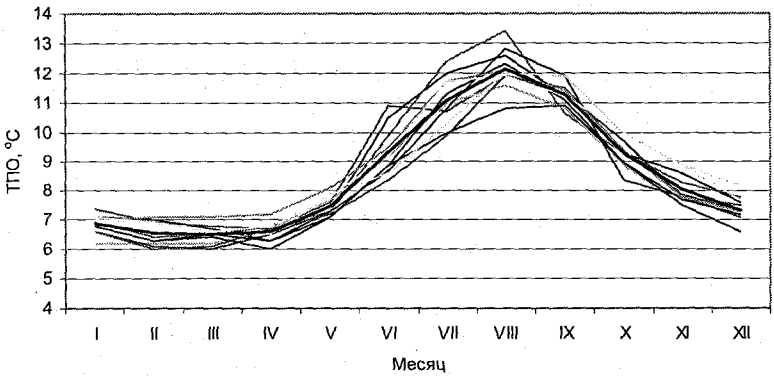


Рис. 9.1. Распределение среднемесячных значений температуры поверхностного слоя воды за 11 лет (1990–2000 гг.) в районе корабля погоды «М».

Поэтому *конкретный вид, который случайная функция X(t) принимает в результате испытаний (наблюдений), называется реализацией случайной функции.* Хотя каждая отдельная реализация носит неслучайный характер (см. рис. 9.1), однако *совокупность всех возможных реализаций образует случайную функцию.*

Фактически всякая функция, встречающаяся на практике, является случайной, поскольку всегда существуют воздействующие на нее, но не поддающиеся учету случайные факторы. Случайные функции, изменяющиеся во времени, называются обычно *случайными процессами*, а случайные функции, изменяющиеся в пространстве, – *случайными полями*. Поэтому случайные процессы представляют собой *временные ряды*, в то время как случайные поля легко представить в виде *пространственных карт*. Например, волнограмма, представляющая собой карту поверхности ветрового волнения в фиксированный момент времени, является реализацией случайной функции широты и долготы.

Зафиксируем теперь момент времени $t = t_j$ и проведем прямую, перпендикулярную оси абсцисс. Эта прямая пересечет каждую реализацию в одной точке. В результате в момент t получим n значений случайной величины, т.е. $x_1(t_j)$, $x_2(t_j)$, ..., $x_n(t_j)$. Набор точек пересечения представляет собой совокупность значений случайной величины, которую называют *сечением* случайной функции, соответствующим значению аргумента $t = t_j$.

Естественно, что число сечений равно длине реализации случайной функции. В рассматриваемом нами случае (рис. 9.1) сечения представляют собой межгодовую изменчивость температуры для каждого из 12 месяцев.

Обычно случайная функция обозначается большими латинскими буквами с указанием аргумента $X(t)$, $Y(t)$ и т.д., а ее реализация – малыми буквами $x_1(t)$, $x_2(t)$, ..., $x_n(t)$, где индекс указывает номер, при котором данная реализация получена. Сечение случайной функции, отвечающее значению аргумента t_j, обозначается как $x_i(t_j)$ или X_i.

Аргумент случайной функции может принимать либо любые вещественные значения в заданном интервале (конечном или бесконечном), либо только определенные дискретные значения. В первом случае $X(t)$ называется *случайным процессом*, во втором – случайным процессом с дискретным временем.

Очевидно, случайная функция может рассматриваться как зависимость от четырех аргументов: трех пространственных координат и времени. Такая случайная функция называется *четырехмерной*. Вследствие сложности ее статистического описания они весьма редко используются в практических расчетах. Наиболее широкое распространение получили *одномерные* и *двухмерные* случайные функции, зависящие соответственно от одного и двух аргументов.

Заметим, что выборочное пространство, связанное с одномерным случайным процессом, дважды бесконечно. Оно простирается для каждого момента времени от $-\infty$ до $+\infty$, причем само время изменяется также от $-\infty$ до $+\infty$. Дважды бесконечное множество функций времени, которые могут быть определены на этом выборочном пространстве, называется *ансамблем*.

Случайная функция, так же как и случайная величина, считается заданной, если известна ее функция распределения. Поскольку совокупность сечений случайной функции $X(t)$ можно рассматривать как систему случайных величин, то ее функция распределения будет представлять функцию распределения $X(t)$, т.е.

$$F(x_1, x_2, ..., x_n) = p(X_1 < x_1, X_2 < x_2, ..., X_n < x_n), \qquad (9.1)$$

где X_1, X_2, ..., X_n – n сечений случайной функции.

Естественно, что функция распределения будет тем полнее характеризовать случайный процесс, чем ближе друг к другу расположены значения аргумента t_j и чем больше число их взято.

Одновременно каждое сечение t_j, представляющее случайную величину, может быть охарактеризовано своей функцией распределения:

$$F_{1j}(x;t_j) = p[X(t_j) < x]. \qquad (9.2)$$

Эта функция называется одномерной функцией распределения случайного процесса. Поскольку она не учитывает взаимной зависимости между различными сечениями, то вводится двухмерная функция распределения для смежных сечений $F(t)$:

$$F_{2j}(x_j, x_{j+1}; t_j, t_{j+1}) = p(X_j < x_j; X_{j+1} < x_{j+1}). \qquad (9.3)$$

Наконец, для n сечений случайного процесса характеристикой служит n-мерная функция распределения, имеющая вид:

$$F_n(x_1, x_2, ..., x_n; t_1, t_2, ..., t_n) = p(X_1 < x_1, X_2 < x_2, ..., X_n < x_n). \qquad (9.4)$$

Итак, случайный процесс считается заданным, если для каждого значения t_j определена одномерная функция распределения (9.2), для каждой пары значений t_j и t_{j+1} определена двухмерная функция распределения (9.3) и для любых n значений t_1, t_2, ..., t_n – n-мерная функция распределения (9.4).

Для непрерывных случайных процессов, каждое сечение которых представляет собой непрерывную случайную величину, можно пользоваться многомерными дифференциальными законами распределения. Если $F_1(x, t)$ имеет частную производную по x

$$\frac{\partial F_1(x,t)}{\partial x} = f_1(x,t), \qquad (9.5)$$

то она называется *одномерной плотностью* распределения или одномерным дифференциальным законом распределения случайного процесса.

Одномерный закон распределения $f_1(x, t)$ есть закон распределения случайной величины – сечения случайного процесса, соответствующего данному значению t. Аналогично определяются многомерные дифференциальные законы распределения случай-

ного процесса. Если существует смешанная частная производная от n-мерной функции распределения

$$\frac{\partial^n F_n(x_1, x_2, ..., x_n; t_1, t_2, ..., t_n)}{\partial x_1 \partial x_2 ... \partial x_n} = f_n(x_1, x_2, ..., x_n, t_1, t_2, ..., t_n), \quad (9.6)$$

то она называется n-мерной плотностью распределения случайного процесса.

Вследствие сложности экспериментального определения многомерных законов распределения, а также их громоздкости и трудности использования при решении прикладных задач, они очень редко применяются на практике. Поэтому в большинстве случаев для описания случайных процессов обычно используются не более чем двухмерные законы распределения.

9.2. Числовые характеристики случайных функций

В отличие от числовых характеристик случайных величин, представляющих собой определенные числа, числовые характеристики случайных функций уже являются некоторыми функциями. Так, под *математическим ожиданием случайной функции $X(t)$ понимается неслучайная функция $m_x(t)$, которая при каждом значении аргумента t равна математическому ожиданию от соответствующего сечения случайной функции*, т. е.

$$m_x(t) = M[X(t)] = \int_{-\infty}^{\infty} x(t) f(x,t) dx, \quad (9.7)$$

где $f(x, t)$ – одномерная плотность вероятности.

На рис. 9.1 математическое ожидание функции $X(t)$ выделено жирной линией. Нетрудно видеть, что геометрически $m_x(t)$ представляет некоторую среднюю кривую из всех возможных реализаций случайной функции. При фиксированном значении аргумента математическое ожидание есть среднее значение сечения, вокруг которого расположены его возможные значения.

Свойства математического ожидания случайной функции во многом идентичны свойствам математического ожидания случайной величины.

Свойство 1. Математическое ожидание неслучайной функции $\varphi(t)$ равно самой неслучайной функции:

$$M[\varphi(t)] = \varphi(t).$$

Свойство 2. Неслучайный множитель $\varphi(t)$ можно вынести за знак математического ожидания:

$$M[\varphi(t)X(t)] = \varphi(t)M[X(t)] = \varphi(t)m_x(t).$$

Свойство 3. Математическое ожидание суммы двух случайных функций равно сумме математических ожиданий слагаемых:

$$M[X(t) + Y(t)] = m_x(t) + m_y(t).$$

Следствие. Для нахождения математического ожидания суммы случайной и неслучайной функций достаточно к математическому ожиданию случайной функции прибавить неслучайную функцию:

$$M[X(t) + \varphi(t)] = m_x(t) + \varphi(t).$$

Дисперсией случайной функции называется неслучайная функция $D_x(t)$, которая для каждого значения аргумента t равна дисперсии соответствующего сечения случайной функции, т.е.

$$D_x(t) = M[\mathring{X}(t)^2] = \int_{-\infty}^{\infty} [x(t) - m_x(t)]^2 f(x,t)dx. \qquad (9.8)$$

Дисперсия характеризует степень рассеяния возможных реализаций (кривых) относительно математического ожидания случайной функции (средней кривой). При фиксированном значении аргумента дисперсия характеризует степень рассеяния возможных значений (ординат) сечения относительно математического ожидания сечения (средней ординаты). Другими характеристиками рассеяния являются среднее квадратическое отклонение и коэффициент вариации, которые определяются как

$$\sigma_x(t) = \sqrt{D_x(t)}, \qquad C_{vx}(t) = \frac{\sigma_x(t)}{m_x(t)}.$$

К свойствам дисперсии относятся:

Свойство 1. Дисперсия неслучайной функции $\varphi(t)$ равна нулю:

$$D[\varphi(t)] = 0.$$

Свойство 2. Дисперсия суммы случайной функции $X(t)$ и неслучайной функции $\varphi(t)$ равна дисперсии случайной функции:

$$D[X(t) + \varphi(t)] = D_x(t).$$

Свойство 3. Дисперсия произведения случайной функции $X(t)$ на неслучайную функцию $\varphi(t)$ равна произведению квадрата неслучайного множителя на дисперсию случайной функции:

$$D[X(t)\varphi(t)] = \varphi^2(t)D_x(t).$$

Заметим, что на практике иногда используется также среднее значение квадрата случайного процесса, дающее представление об его интенсивности и определяемое как

$$\Psi_x^2(t) = \int x^2(t)dt.$$

Нетрудно видеть, что дисперсия случайного процесса равна разности между средним значением квадрата и квадратом среднего значения, т.е. $D_x(t) = \Psi_x^2(t) - m_x^2(t)$.

Итак, математическое ожидание и дисперсия являются характеристиками каждого сечения случайной функции как некоторой случайной величины и отражают лишь внешние свойства случайного процесса. Однако они совершенно не содержат никакой информации о тесноте связи между отдельными сечениями случайной функции и, следовательно, не дают представления об ее внутренней структуре.

Для описания внутренней структуры случайного процесса используется *автокорреляционная функция, которая представляет собой неслучайную функцию* $R_x[(t_1, t_2)]$ *двух независимых фиксированных аргументов* t_1 *и* t_2 *, равную корреляционному моменту сечений этих аргументов*:

$$R_x(t_1, t_2) = M\{[X(t_1) - m_x(t_1)][X(t_2) - m_x(t_2)]\}. \qquad (9.9)$$

Отметим некоторые свойства автокорреляционной функции.

Свойство 1. При равенстве аргументов $t_1 = t_2 = t$, автокорреляционная функция обращается в дисперсию случайной функции:

$$R_x(t, t) = M[X(t) - m_x(t)]^2 = D_x(t), \qquad (9.10)$$

т.е. дисперсия является частным случаем $R_x(t_1, t_2)$. Это означает, что при описании случайной функции, вообще говоря, отпадает необходимость в учете дисперсии как самостоятельной величины.

Свойство 2. Автокорреляционная функция симметрична относительно своих аргументов t_1 и t_2, т.е. не изменяется при их перестановке местами:

$$R_x(t_1, t_2) = R_x(t_2, t_1).$$

Свойство 3. Прибавление к случайной функции $X(t)$ неслучайного слагаемого $\varphi(t)$ не изменяет ее автокорреляционной функции: если $Y(t) = X(t) + \varphi(t)$, то $R_y(t_1, t_2) = R_x(t_1, t_2)$.

Свойство 4. Абсолютная величина автокорреляционной функции не превышает среднего геометрического дисперсий соответствующих сечений:

$$\left|R_x(t_1,t_2)\right| \le \sqrt{D_x(t_1)D_x(t_2)}$$

или

$$\left|R_x(t_1,t_2)\right| \le \sigma_x(t_1)\sigma_x(t_2)\,. \qquad (9.11)$$

Свойство 5. При умножении случайной функции $X(t)$ на неслучайный множитель $\varphi(t)$ ее автокорреляционная функция умножается на произведение $\varphi(t_1)\varphi(t_2)$, т.е.

$$R_y(t_1, t_2) = R_x(t_1, t_2)\varphi(t_1)\varphi(t_2).$$

Следует иметь в виду, что автокорреляционная функция является обычно убывающей функцией времени. Это означает, что по мере увеличения интервала между наблюдениями связь между значениями случайной функции ослабевает.

Отметим, что при *нормальном распределении случайных процессов математическое ожидание и автокорреляционная функция являются их исчерпывающими характеристиками.*

Чтобы иметь возможность сравнивать автокорреляционные функции для случайных функций разной размерности, их обычно нормируют:

$$r(t_1,t_2) = \frac{R(t_1,t_2)}{\sqrt{D_x(t_1)D_x(t_2)}} = \frac{R(t_1,t_2)}{\sigma_x(t_1)\sigma_x(t_2)}\,. \qquad (9.12)$$

Рассчитанная таким образом *нормированная автокорреляционная функция представляет собой безразмерную характеристику линейной связи между сечениями случайной функции.* Из формулы (9.12) следует, что

$$| r(t_1, t_2) | \leq 1. \qquad (9.13)$$

Для каждой пары аргументов t_1, t_2 нормированная автокорреляционная функция равна коэффициенту корреляции соответствующих сечений случайной функции.

На практике довольно часто приходится рассматривать несколько случайных процессов совместно. Естественно, при этом, кроме характеристики каждого случайного процесса, необходимо устанавливать наличие связи между различными процессами. Так, для системы из двух случайных процессов $X(t)$ и $Y(t)$ характеристиками ее являются $m_x(t)$, $m_y(t)$, $R_x(t_1, t_2)$, $R_y(t_1,t_2)$, а также корреляционная функция связи

$$R_{xy}(t_1,t_2) = M\{[X(t_1)-m_x(t_1)][Y(t_2)-m_y(t_2)]\}, \qquad (9.14)$$

которая показывает степень линейной зависимости между сечениями $X(t_1)$ и $Y(t_2)$. При $t_1 = t_2$ данная функция, называемая также *взаимной корреляционной функцией*, будет характеризовать степень линейной зависимости случайных процессов $X(t)$ и $Y(t)$, соответствующих одному и тому же значению аргумента.

В отличие от автокорреляционной функции $R_{xy}(t_1, t_2)$ не является симметричной относительно своих аргументов, однако обладает тем свойством, что не изменяется при одновременной перестановке аргументов и индексов, т.е.

$$R_{xy}(t_1,t_2) = R_{yx}(t_2,t_1).$$

Нормируя $R_{xy}(t_1, t_2)$, получаем безразмерную характеристику тесноты линейной связи между сечениями $X(t_1)$ и $Y(t_2)$, называемую *нормированной взаимной корреляционной функцией*:

$$r_{xy}(t_1,t_2) = \frac{R_{xy}(t_1,t_2)}{\sigma_x(t_1)\sigma_y(t_2)}. \qquad (9.15)$$

При фиксированных значениях t_1 и t_2 данная функция представляет собой коэффициент корреляции сечений $X(t_1)$ и $Y(t_2)$.

Если взаимная корреляционная функция равна нулю при всех значениях своих аргументов t_1 и t_2, то случайные функции $X(t)$ и $Y(t)$ являются некоррелированными (несвязанными).

Следует иметь в виду, что в специальной статистической литературе автокорреляционная и нормированная автокорреляционная функции могут называться соответственно как автоковариационная и автокорреляционная функции, а взаимная корреляционная и нормированная взаимная корреляционная функции соответственно взаимной ковариационной и взаимной корреляционной функциями. Мы будем в дальнейшем пользоваться первыми обозначениями.

9.3. Стационарность случайных процессов

В зависимости от устойчивости вероятностных характеристик случайные процессы подразделяются на стационарные и нестационарные.

Стационарные случайные процессы описывают процессы, которые протекают во времени почти однородно. Это означает, что их вероятностные характеристики с изменением аргумента практически не изменяются и поэтому не зависят от начала отсчета.

Нестационарные случайные процессы описывают процессы, которые протекают во времени так, что их вероятностные характеристики с изменением аргумента существенно изменяют свои значения и, следовательно, зависят от начала отсчета. Поэтому характеристики нестационарного процесса представляют собой функции времени, которые можно определить только осреднением мгновенных значений по ансамблю выборочных функций, формирующих процесс.

Заметим, что исторически понятие «стационарность» применяется в основном в отношении случайных процессов, в то время как применительно к случайным полям более естественным является термин «*однородность*», означающий постоянство их свойств в пространстве.

Различают стационарность в широком и узком смысле.

Случайный процесс X(t) называется стационарным в широком смысле, если он обладает следующими свойствами: его выборочные оценки среднего и дисперсии постоянны во времени и соответствуют математическому ожиданию и генеральной дисперс-

сии, 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.*

Отметим, что величину τ обычно называют сдвигом (лагом). Таким образом, стационарность в широком смысле случайного процесса означает, что

$$\bar{X}_x(t) = \text{const};\ s_x^2(t) = \text{const};\ R_x(t_1, t_2) = f[R_x(\tau)].$$

Заметим, что при более строгом определении стационарности постоянство во времени дисперсии опускается, ибо величина дисперсии представляет собой значение автокорреляционной функции при нулевом сдвиге, т.е. $s_x^2(t) = R_x(\tau = 0)$.

Случайный процесс X(t) называется стационарным в узком смысле, если его многомерные распределения при одновременном прибавлении ко всем аргументам $t_1, t_2, \ldots, t_n$ *одного и того же числа* τ *остаются неизменными*, т. е.

$$f(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n) = f(x_1, x_2, \ldots, x_n; t_1 + \tau, t_2 + \tau, \ldots, t_n + \tau).$$

Стационарность в узком смысле подразумевает знание совместного (многомерного) закона распределения, что вызывает в практических расчетах серьезные затруднения.

Для нормального распределения случайных процессов стационарность в широком смысле эквивалентна строгой стационарности. Во всех остальных случаях стационарность в широком смысле является простейшим частным случаем стационарности в узком смысле. В практических расчетах обычно используется понятие стационарности в широком смысле, которое во многих случаях достаточно точно соответствует стационарности в узком смысле, но в то же время значительно лучше поддается статистическому оцениванию. Отметим, что в гидрометеорологии обычно используется только понятие стационарности в широком смысле.

Следует иметь в виду, что многие нестационарные процессы на некоторых участках, где нет заметных изменений математического ожидания и дисперсии, могут считаться квазистационарными. Строго говоря, выявление стационарности случайного процесса возможно только в результате его физического анализа. К сожалению, это не всегда оказывается реальным. В этих случаях можно использовать статистические методы, ибо большое число неста-

ционарных процессов может быть приведено к стационарному виду путем некоторых преобразований. К числу наиболее простых относится вычисление аномалий случайного процесса, т.е.

$$\Delta X(t) = X(t) - \overline{X}(t),$$

где $\overline{X}(t)$ – среднее арифметическое выборочного случайного процесса, в качестве которого обычно используется норма гидрометеорологической характеристики.

Такой процесс называется *центрированным*. Отметим, что корреляционные функции центрированного и исходного случайных процессов совпадают. Действительно, во многих случаях аномалии гидрометеорологических процессов оказываются стационарными на больших временных интервалах.

Если нестационарность функции вызвана изменяющимся во времени математическим ожиданием, то для приведения ее к стационарному виду можно, например, использовать первые разности:

$$\Delta'_x(t) = X(t_{i+1}) - X(t_i), \quad i = \overline{1,n}.$$

Данная формула позволяет устранить линейный тренд. В том случае, если первые разности случайного процесса еще носят нестационарный характер, то это свидетельствует о существовании нелинейного тренда. Для его устранения можно рассчитать вторые разности:

$$\Delta''_x(t) = \Delta'_x(t_{i+1}) - \Delta'_x(t_i).$$

Как правило, этого оказывается достаточно для приведения функции с изменяющимся во времени математическим ожиданием к стационарному виду. Пример стационарной $X(t)$ и нестационарной $Y(t)$ функций приводится на рис. 9.2. В частности, из рис. 9.2 видно, что для стационарного процесса $X(t)$ характерна неизменность среднего и дисперсии, в то время как процесс $Y(t)$ явно нестационарен по математическому ожиданию и стационарен по дисперсии. Другими словами, для процесса $X(t)$ выборочные средние и дисперсии для всех сечений практически совпадают, в то время как для процесса $Y(t)$ не выполняется равенство оценок выборочных средних значений.

Проверка стационарности требует наличия, по крайней мере, нескольких реализаций процесса, которые далеко не всегда име-

ются в нашем распоряжении. Чаще всего располагают лишь одной реализацией достаточной длины. Предполагая эргодичность процесса, эту реализацию разбивают на отдельные, не обязательно одинаковые по продолжительности интервалы. Затем для каждого из них вычисляются оценки среднего арифметического, дисперсии и автокорреляционной функции.

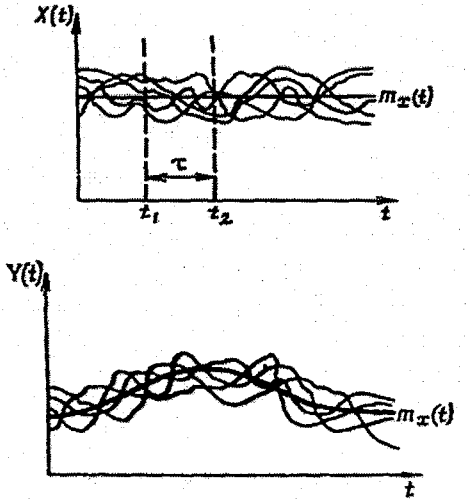


Рис. 9.2. Временной ход реализаций стационарной $X(t)$ и нестационарной $Y(t)$ случайных функций.

После этого, используя критерии статистической проверки гипотез, производится сравнение указанных характеристик. Если расхождение между ними для всех интервалов окажется незначимым, то делается вывод, что данная случайная функция является стационарной. Отметим, что разбиение реализации на отдельные интервалы желательно осуществлять исходя из закономерностей внутренней структуры рассматриваемого процесса, ибо каких-либо формальных критериев для этого нет.

Проверка стационарности по автокорреляционной функции требует ее расчета для нескольких реализаций и последующего сравнения их оценок, отличающихся от нуля значимым образом. Если при этом расхождения между оценками превышают удвоенную среднюю квадратическую ошибку их определения, то процесс следует считать нестационарным.

Пример 9.1. Понятие стационарности является одним из ключевых при анализе случайных процессов. Для проверки стационарности целесообразно воспользоваться рассмотренными выше статистическими гипотезами о равенстве выборочных средних и дисперсий, причем проверку нужно начинать с равенства дисперсий. Однако прежде необходимо построить графики временных рядов и попытаться осуществить физический анализ его колебаний. После этого каждый временной ряд может быть разделен на две не обязательно равные части, для которых выполняется проверка статистических гипотез.

Оценим указанным образом стационарность для температуры поверхности океана (ТПО) в районе юго-восточной части Тихого океана (ЮВТО), являющегося, как известно, важным районом круглогодичного рыбного промысла. Примем ЮВТО в следующих границах: западная – 105° з.д., восточная – 75° з.д., северная – 30° ю.ш., южная – 41° ю.ш.

Для данного района ЮВТО характерно очень плохое покрытие его гидрометеорологическими данными вообще и ТПО в частности. В связи с этим постоянно возникает вопрос о степени репрезентативности тех или иных архивов, содержащих информацию о ТПО и представляющих собой, по существу, некие «черные ящики». Естественно, для этого необходимы реперные данные, т.е. пространственно-временные ряды ТПО, полученные непосредственно по измерениям ее во время экспедиций. Однако пространственная разрозненность и временная нерегулярность таких наблюдений даже за период массовых экспедиционных работ в течение периода с 1979 по 1991 г. не позволили их систематизировать и использовать для сравнения в качестве эталона. Поэтому, очевидно, единственным вариантом оценки степени репрезентативности этих архивов остается тщательный статистический анализ и сравнение архивов друг с другом.

Основным источником данных послужила система CDAS (Climate Data Assimilation System), сведения о которой приведены в табл. 1.1. Поскольку исходные данные заданы в узлах первичной широтно-долготной сетки 1,875×1,875°, то область ЮВТО разделилась на 109 квадратов. Для этих квадратов из Интернета получен архив среднемесячных значений ТПО за период с 1949 по 2003 г.

Визуальный анализ данных показал, что для многих квадратов характерно резкое изменение размаха колебаний. Например, в одном из квадратов (рис. 9.3) межгодовые колебания ТПО до 1980 г. практически не превышали 0,1 °С, и только в последующие годы размах колебаний достиг 0,8 °С. Аналогичный характер колебаний ТПО свойствен и для других квадратов. Ретроспективный физический анализ показал, что такие резкие изменения в дисперсии и средних значениях ТПО вряд ли можно объяснить переломами в ходе метеорологических процессов. Скорее всего, это связано с технологическими особенностями самой системы CDAS. В частности, с использованием новых систем натурных данных (например, ассимиляцией спутниковых карт ТПО).

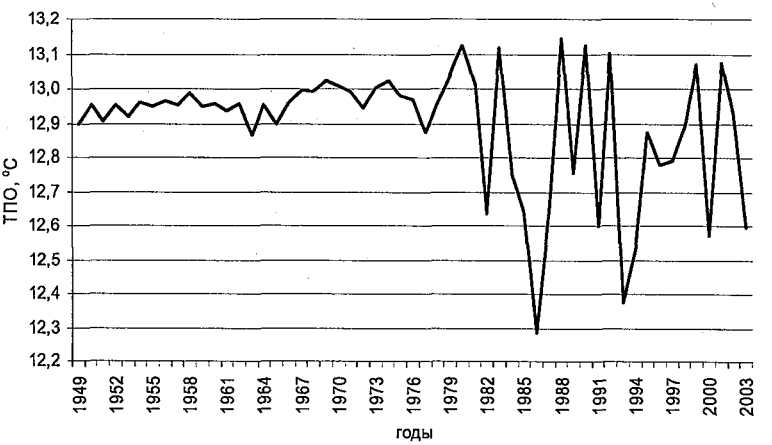


Рис. 9.3. Межгодовой ход температуры поверхности океана в одном из квадратов ЮВТО.

Итак, даже предварительный анализ межгодовых колебаний ТПО свидетельствует об явно выраженной нестационарности ТПО по математическому ожиданию и дисперсии. Естественно, что более строгие выводы можно сделать лишь после соответствующей статистической проверки временных рядов на стационарность с помощью критериев Стьюдента и Фишера.

Для проверки выдвинутых нулевых гипотез временные ряды среднегодовых значений ТПО были разбиты на два относительно однородных промежутка: с 1949 по 1980 г. и с 1981 по 2003 г. Рас-

пределение рассчитанных значений t и F представлено на рис. 9.4 и 9.5. Укажем, что их критические значения при уровне значимости $\alpha = 0{,}05$ равны соответственно $\mathbf{t}_{кр} = 1{,}67$, $\mathbf{F}_{кр} = 1{,}80$.

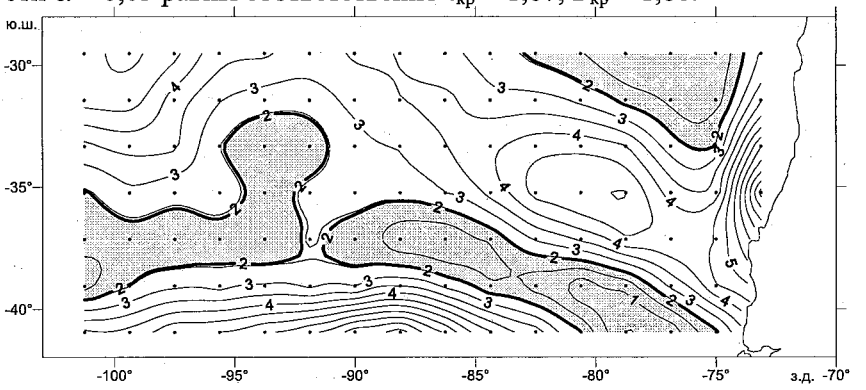


Рис. 9.4. Пространственное распределение критерия Стьюдента для акватории ЮВТО (области значимых величин выделены заливкой).

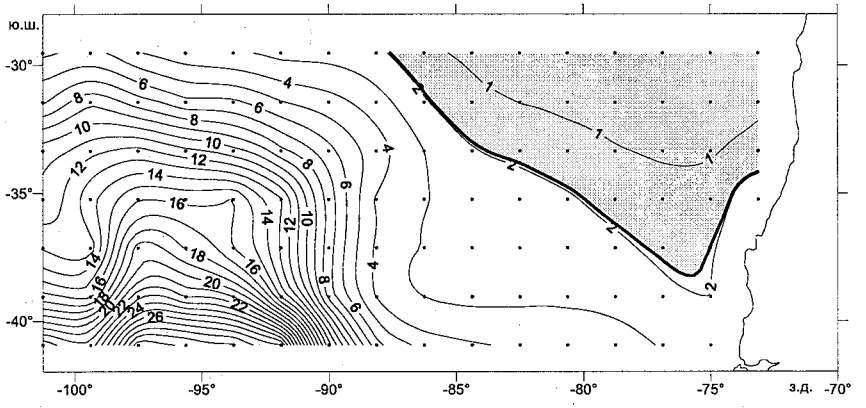


Рис. 9.5. Пространственное распределение критерия Фишера для акватории ЮВТО (области значимых величин выделены заливкой).

Из рис. 9.4 видно, что на большей части акватории $t > \mathbf{t}_{кр}$, т.е. здесь расхождения между выборочными средними являются значимыми. И только два сравнительно небольших языка с малыми значениями t ($t < \mathbf{t}_{кр}$) вдаются с юга и запада навстречу друг другу. Что касается распределения F (рис. 9.5), то легко видеть, что вся акватория делится на два резко неравнозначных района: северо-

восточный, где $F < \mathbf{F}_{кр}$ и остальной, где $F > \mathbf{F}_{кр}$. Заметим, что в юго-западной части акватории оценки F превышают $\mathbf{F}_{кр}$ более чем в десять раз.

Таким образом, достаточно очевидный вывод состоит в том, что межгодовые колебания ТПО в рассматриваемый промежуток времени (с 1949 по 2003 г.) являются нестационарными как по среднему арифметическому, так и по дисперсии. Поэтому, учитывая сомнительный характер данных по ТПО до 1981 г., следует, очевидно, их вообще исключить из последующего статистического анализа.

9.4. Эргодичность стационарных случайных процессов

Очень важным свойством стационарной случайной функции является ее эргодичность. Предположим, мы имеем только одну реализацию случайного процесса большой продолжительности. Тогда математическое ожидание для нее определится по формуле:

$$m_x = \frac{1}{T}\int_0^T x(t)dt, \qquad (9.16)$$

где T – интервал осреднения (длина реализации), а автоковариационная функция как

$$R_x(\tau) = \frac{1}{T-\tau}\int_0^{T-\tau}[x(t)-m_x][x(t+\tau)-m_x]dt. \qquad (9.17)$$

Возникает вопрос, будут ли эти оценки близки к аналогичным значениям, полученным по множеству реализаций?

Если стационарный случайный процесс, для которого статистические характеристики, полученные осреднением по одной реализации, при увеличении интервала осреднения T с вероятностью сколь угодно близкой к единице приближаются к соответствующим характеристикам, полученным осреднением по всему множеству реализаций, то он обладает эргодическим свойством.

Другими словами, *если одна реализация 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, то*

такой стационарный процесс является эргодическим. Стационарный случайный процесс, не обладающий таким свойством, называются *неэргодическим*. На рис. 9.6 представлены своими реализациями две стационарные случайные функции $X(t)$ и $Y(t)$, которые имеют одно и то же математическое ожидание и тот же общий размах колебаний. Однако они заметно отличаются по своей внутренней структуре.

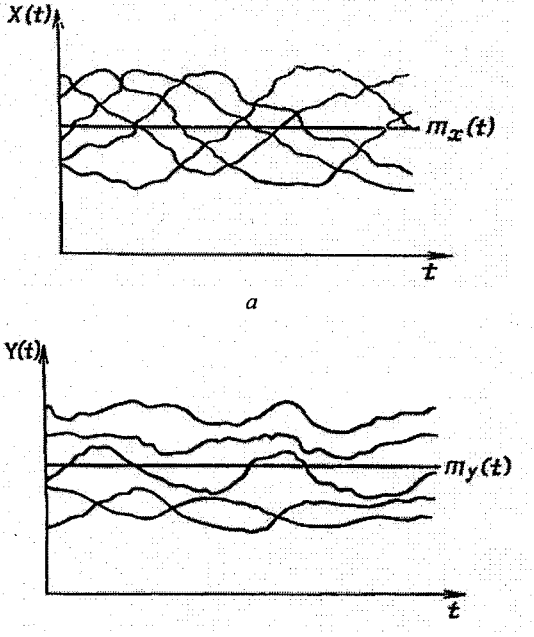


Рис. 9.6. Временной ход реализаций стационарных эргодической $X(t)$ (*а*) и неэргодической $Y(t)$ (*б*) случайной функции

Для случайного процесса $X(t)$ характерно то, что каждая из его реализаций обладает почти одним и тем же средним значением, близким к $m_x(t)$ и почти одним и тем же размахом колебаний, примерно равным величине $D_x(t)$ случайного процесса. Другими словами, для каждой реализации случайного процесса ее выборочные средние и выборочные дисперсии примерно соответствуют друг другу и при этом совпадают с математическим ожиданием и генеральной дисперсией процесса $X(t)$, т.е.

$$\bar{X}_1(t) \approx \bar{X}_2(t) \approx \ldots \approx \bar{X}_n(t) \approx m_x(t),$$

$$s_1^2(t) \approx s_2^2(t) \approx \ldots \approx s_n^2(t) \approx D_x(t).$$

Что касается случайного процесса $Y(t)$, то каждая его реализация имеет свое среднее значение и свой средний размах колебаний, которые могут заметно отличаться от математического ожидания и генеральной дисперсии самого случайного процесса.

Отсюда следует, что оценки соответствующих характеристик случайного процесса $X(t)$, вычисленные по одной достаточно длинной реализации и по множеству реализаций той же общей длины, будут между собой равноточные. Созершенно очевидно, что для случайного процесса $Y(t)$ этого не наблюдается. Поэтому стационарный процесс $X(t)$ является эргодическим, а процесс $Y(t)$ свойством эргодичности не обладает.

При численных расчетах свойство эргодичности имеет большое значение, так как позволяет существенно сократить объемы вычислений. Кроме того, принимая во внимание, что продолжительность многих гидрометеорологических рядов невелика, то эргодичность есть единственная возможность использования аппарата теории случайных функций применительно к временным рядам.

Однако на практике доказать свойство эргодичности очень сложно, поэтому о нем, обычно, судят исходя из физических соображений, связанных с природой изучаемого процесса или же принимают априори. В некоторых случаях в качестве формального признака эргодичности принимается приближение корреляционной функции случайного процесса к нулю при возрастании промежутка τ между сечениями, т.е. $r(\tau) \to 0$ при $\tau \to \infty$.

9.5. Классификация временных рядов

В настоящее время известны различные классификации временных рядов, основанные на разных таксономических признаках. Так, в главе 1 уже была рассмотрена классификация природных процессов по временным масштабам их колебаний. Здесь рассмотрим классификацию временных рядов исходя из их свойств и внутренней структуры. Очевидно, в общем случае все временные ряды могут быть разделены на детерминированные и случайные.

Если значения временного ряда, обусловленные действием одного или, в крайнем случае, нескольких внешних факторов, изменяются по строго определенному, как правило, физическому закону, то такой ряд является детерминированным. Это означает, что значения временного ряда связаны с внешними факторами функциональными зависимостями. Для детерминированных связей свойственно, что каждому значению какой-либо одной переменной соответствует только одно значение другой переменной.

Детерминированные процессы могут быть разделены на *периодические и непериодические*. Первые свидетельствуют о том, что временные ряды подвержены гармоническим колебаниям. Очевидно, *если все основные параметры колебания (амплитуда, период, фаза) остаются строго постоянными во времени, то оно является гармоническим.* Достаточно корректно гармонические колебания описываются разложением Фурье.

Непериодические процессы характеризуют такие изменения во времени, которые протекают по некоторому закону без образования циклов (например, логарифмическому, экспоненциальному и т.п.). Отметим, что детерминированные процессы изучают, как правило, в рамках физических дисциплин.

Естественно, нас больше интересует второй тип временного ряда – случайный, который также может быть представлен как результат действия множества внешних и внутренних факторов. Однако их результирующий эффект на изучаемый процесс вследствие разнообразного, подчас противоположного действия вынуждающих факторов, их изменчивости, наличия между ними прямых и обратных связей, а также ошибок измерений и расчетов, является настолько сложным, что он зачастую не поддается физическому описанию. Поэтому такой временной ряд можно уже рассматривать как случайный.

Как было показано выше, случайные процессы можно разделить на стационарные и нестационарные (рис. 9.7). В свою очередь, стационарные временные ряды делятся на эргодические и неэргодические. В качестве конкретных проявлений эргодического стационарного случайного процесса можно рассматривать модели временного ряда в виде «белого шума» (БШ), «красного шума» (КШ) и циклического колебания (ЦК).

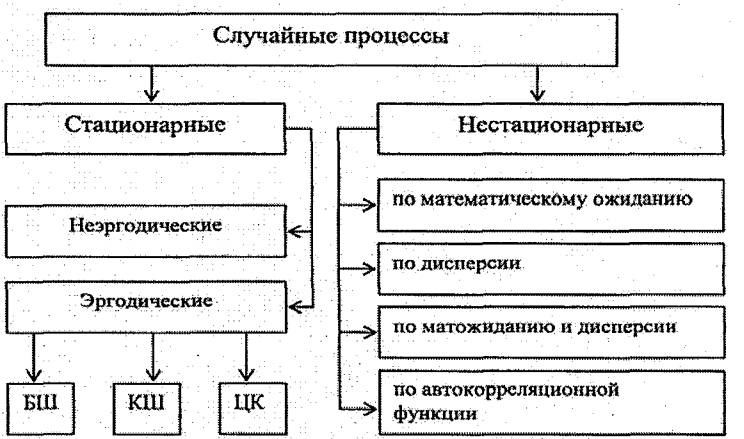


Рис. 9.7. Классификация случайных процессов.

Белый шум – условное название чисто случайного процесса, представляющего собой практически набор случайных чисел, не коррелированных друг с другом. Более строго, белый шум – это дискретный стационарный процесс, случайные значения x_i которого взаимно независимы и одинаково распределены с математическим ожиданием, равным нулю, и постоянной дисперсией. Теоретическая модель белого шума описывается автокорреляционной функцией, равной нулю на всех сдвигах, исключая нулевой сдвиг. Естественно, это не более чем математическая абстракция, ибо на практике такие автокорреляционные функции не встречаются. Поэтому обычно принимается условие незначимости коэффициентов автокорреляции. В этом случае говорят о случайном стационарном процессе, развивающемся по типу модели «белый шум».

Красный шум – условное название процесса, которому свойственна корреляция только между смежными (соседними) значениями временного ряда. Это означает, что во временном ряду присутствует инерционность. Более строго, красный шум – это дискретный стационарный процесс, описывающийся марковской цепью первого порядка. Теоретическая модель красного шума представляет собой автокорреляционную функцию, которая на первом сдвиге имеет значимый коэффициент автокорреляции, а на всех остальных сдвигах она равна нулю. Естественно, для реальных

природных процессов такая автокорреляционная функция не встречается. Поэтому на практике принимается условие незначимости коэффициентов автокорреляции, начиная со сдвига $\tau = 2$, что свидетельствует о случайном процессе, развивающемся по типу модели «красный шум».

Циклическое колебание – такое колебание, параметры которого (период, амплитуда, фаза) испытывают нерегулярные изменения во времени в пределах некоторого диапазона. Пример – годовой ход природных процессов, связанный с годовым ходом притока солнечной радиации. Действительно, амплитуда его в большинстве случаев меняется год от года. Даже период не всегда равен 12 месяцам, а может составлять 11–13 месяцев. Другой пример циклического колебания – солнечная активность, период которой, как известно, составляет 11 лет. Однако период отдельных циклов солнечной активности меняется в широких пределах: от 8 до 15 лет. Также большой изменчивости подвержена и амплитуда колебаний солнечной активности.

Нестационарность случайных процессов обычно проявляется в виде тренда, под которым чаще всего понимают некоторое медленное изменение процесса с периодом, превышающим длину исходной реализации. Отсюда следует, что само существование тренда полностью определяется длиной ряда. При этом возможен тренд по среднему арифметическому, тренд по дисперсии или одновременно тренд по среднему арифметическому и дисперсии, а также по автокорреляционной функции (АФ). Отметим, что последний вид тренда встречается весьма редко. На практике обычно рассматривают частный случай тренда по АФ – тренд по дисперсии. Более подробно тренды будут обсуждены в п. 10.3.

Характерным примером нестационарного процесса с одновременным наличием тренда по среднему и дисперсии является временной ряд типа «*случайного блуждания*». Такой ряд $X(t)$ легко может быть получен из чисто случайного стационарного процесса $Z(t)$ по следующему правилу:

$$x_1 = z_1$$
$$x_2 = z_1 + z_2 = x_1 + z_1$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$x_n = z_1 + z_2 + \ldots + z_n = \Sigma x_{n-1} + z_n.$$

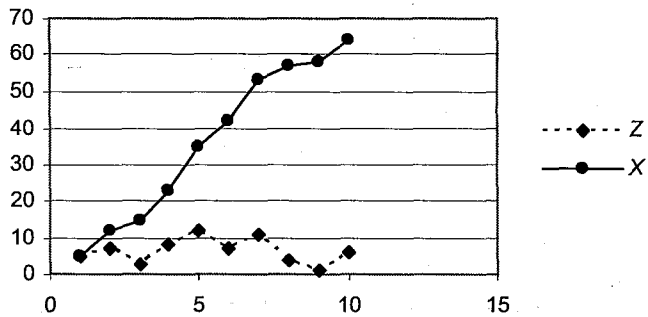


Рис. 9.8. Модель временного ряда в виде «случайного блуждания»

Отсюда мы видим, что ряд $X(t)$ образуется последовательным суммированием случайных не коррелированных друг с другом чисел. На рис. 9.8 приведен график построенного таким образом временного ряда. Действительно, среднее значение его возрастает по прямой $\bar{Z}_t$, причем одновременно увеличивается и амплитуда колебаний ряда $X(t)$. Это и есть нестационарный процесс, называемый случайным блужданием.

Глава 10. МЕТОДЫ АНАЛИЗА ВРЕМЕННЫХ РЯДОВ

10.1. Общая схема исследования временной изменчивости

Если какая-либо случайная величина X измерена в m точках пространства, причем длина каждой реализации составляет n, то получим матрицу размером $m \times n$. Использование методов многомерного статистического анализа применительно к этой матрице представляет основную задачу исследования пространственно-временной изменчивости.

Тогда анализ пространственных полей и анализ временных рядов можно рассматривать как частные случаи исследования пространственно-временной изменчивости. Заметим, что если пространственные поля заданы в узлах регулярной равномерной сетки, то в этом случае к ним применимы практически все методы анализа временных рядов. Однако, учитывая, что гидрометеорологическая сеть станций не является регулярной, анализ пространственных полей обладает целым рядом специфических особенностей.

Что касается временной изменчивости, то для ее исследования прежде всего используется теория случайных функций, аппарат которой в настоящее время хорошо разработан. Естественно, что в зависимости от периода осреднения исходных рядов методы анализа временной изменчивости могут различаться.

Рассмотрим схему исследования крупномасштабных процессов, когда период осреднения исходных данных превышает несколько недель (например, примем τ = 1 мес.), т.е. фильтруются синоптические колебания. В этом случае временной ряд может быть представлен как

$$X(t) = \Phi(t) + Q(t), \qquad (10.1)$$

где $\Phi(t)$ – межгодовая (низкочастотная) изменчивость; $Q(t)$ – внутригодовая (высокочастотная) изменчивость.

Задачей разложения (10.1) является разделение разномасштабных составляющих. Это достигается многими приемами и, в частности, методами фильтрации. Если же необходимо выделить

лишь межгодовую изменчивость, то для этого достаточно усреднить среднемесячные данные за годовые интервалы времени.

Составляющую $Q(t)$ можно выразить как

$$Q(t) = \sum_{i=1}^{k} A_i \cos(\omega_i t + \varphi_i) + \varepsilon(t) = G(t) + \varepsilon(t), \qquad (10.2)$$

где A_i, ω_i, φ_i – амплитуда, частота, фаза i-й гармоники полигармонического ряда сезонного хода; k – число гармоник; $\varepsilon(t)$ – остатки ряда, представляющие собой стационарный случайный процесс с нулевым средним $\bar{\varepsilon} = 0$ и известной корреляционной функцией $r(\tau)$.

Заметим, что представление ряда в виде (10.2) в определенной степени гипотетично, так как не всегда соблюдается требование к стационарности $\varepsilon(t)$ и независимости слагаемых в правой части (10.2).

Анализ ряда $Q(t)$ заключается в том, что вначале методами гармонического анализа или периодограманализа производится оценка годового хода. После исключения годового хода исследуется его остаточная часть $\varepsilon(t)$, если в этом есть необходимость.

С практической точки зрения важным представляется нахождение критерия разделения составляющей $Q(t)$ на гармоническую и случайную части. В предположении их независимости (некоррелированности) можно записать:

$$D(Q) = D(G) + D(\varepsilon), \qquad (10.3)$$

где дисперсия гармонической части представляется как $D(G) = \sum_{i=1}^{k} D(G)_i$, т. е. суммой дисперсий отдельных гармоник.

Рассчитав последовательно гармоники, начиная с первой, будем проверять выполнимость условия:

$$D(G)_i \le D(\varepsilon), \qquad (10.4)$$

где величина $D(\varepsilon)$ определяется из выражения (10.3).

Если условие (10.4) выполняется, то можно считать, что произошло разделение $Q(t)$ на гармоническую и случайную части. Величина $D(\varepsilon)$ характеризует степень «зашумленности» годового хода или, другими словами, степень неопределенности, обусловленную воздействием на рассматриваемый ряд множества случайных

факторов, а также ошибками измерений и расчетов. Заметим, что для большинства гидрометеорологических рядов достаточно первой или, в крайнем случае, второй (полугодовой) гармоники, чтобы выполнялось условие (10.4).

Рассмотрим теперь структуру межгодовой изменчивости $\Phi(t)$. По аналогии с (10.2) ее можно представить в виде следующего разложения:

$$\Phi(t) = T(t) + \sum_{j=1}^{l} A_j \cos(\omega_j t + \varphi_j) + P(t) = T(t) + C(t) + P(t), \quad (10.5)$$

где $T(t)$ – *трендовая составляющая*; A_j, ω_j, φ_j – параметры полигармонического ряда, характеризующего регулярные межгодовые колебания; $C(t)$ – *циклическая компонента*, характеризующая регулярные (циклические) межгодовые колебания; $P(t)$ – *остаточная часть*, характеризующая нерегулярные межгодовые колебания.

При анализе межгодовой изменчивости прежде всего выделяется трендовая составляющая, приемы выделения которой будут рассмотрены в п. 10.3.

После обнаружения и исключения тренда получим некоторый временной ряд $\Phi'(t)$:

$$\Phi'(t) = \Phi(t) - T(t) = C(t) + P(t), \quad (10.6)$$

который обычно уже отвечает условиям стационарности случайного процесса и по существу является его аддитивной моделью. Напомним, что в статистике под *аддитивной* моделью понимается случайный процесс вида:

$$\varphi(t) = \psi(t) + \xi(t), \quad (10.7)$$

где $\psi(t)$ – некоторая квазипериодическая функция; $\xi(t)$ – стационарный случайный процесс.

Мультипликативной моделью называется случайный процесс, представляющий собой произведение данных компонент, т.е.

$$\varphi(t) = \psi(t) x \xi(t). \quad (10.8)$$

Отметим, что исторически модель вида (10.7) трактуется как задача разделения случайного процесса на «сигнал» и «шум». Однако в гидрометеорологии понятия сигнал и шум зачастую весьма условны.

Выделенный ряд $\Phi'(t)$ нетрудно проверить на наличие в нем регулярных колебаний, например с помощью гармонического анализа. Межгодовая изменчивость отличается от внутригодовой тем, что роль регулярных колебаний, как правило, значительно меньше. К причинам, вызывающим регулярные колебания, относятся приливные силы (например, 19-летний цикл) и другие квазипериодические явления (квазидвухлетний цикл и полюсный прилив). Однако соотношение $C(t)/P(t)$ варьирует в широких пределах и априори неизвестно.

Наиболее универсальным методом, позволяющим одновременно оценить квазипериодические и нерегулярные колебания, является спектральный анализ, который дает представление о распределении плотности дисперсии искомого ряда по частотам. Поэтому, оценив суммарную дисперсию квазипериодических колебаний D_c и приняв дисперсию ряда при использовании нормированной спектральной плотности $D_\phi = 1$, получим:

$$\frac{C(t)}{P(t)} = \frac{D_c}{1 - D_c}. \qquad (10.9)$$

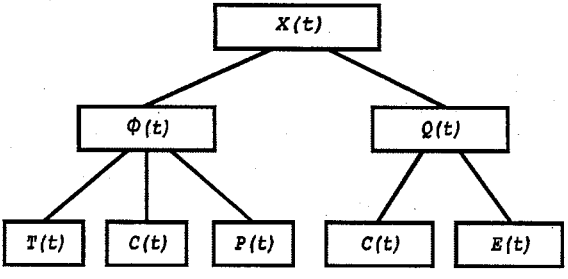


Рис. 10.1. Основные этапы анализа изменчивости временных рядов, интервал дискретизации которых составляет Δt = 1 месяц.

Таким образом, основные этапы исследования временной изменчивости состоят в следующем (рис. 10.1):

1. Разделение временного ряда $X(t)$ на межгодовую и внутригодовую части.

2. Выделение гармоник годового хода на основе гармонического анализа.

3. Оценка вклада случайных колебаний в годовой ход и их анализ.

4. Выделение тренда в межгодовой изменчивости и оценка его вклада в дисперсию исходного ряда.

5. Выделение регулярных межгодовых колебаний и их анализ.

6. Выделение нерегулярных межгодовых колебаний и их анализ.

Разумеется, анализ временных рядов не ограничивается приведенной выше схемой. Поэтому ее можно рассматривать лишь в качестве стандартной процедуры, состоящей из комплекса методов статистического анализа, которые будут обсуждены в следующих разделах.

10.2. Выделение и анализ трендовой компоненты

В разложении (10.5) *под трендовой составляющей понимается некоторое медленное изменение процесса с периодом, превышающим длину исходной реализации.* Отсюда следует, что само существование тренда полностью определяется длиной ряда. При изменении его длины тренд может появляться, исчезать, менять свою интенсивность и форму. Но при этом он не может образовывать циклы, которые, как видно из разложения (10.5), описываются вторым слагаемым.

Отметим, что до настоящего времени существует некоторая путаница в понятии тренда. Следует отличать трендовую компоненту от *тенденции* временного ряда, под которой обычно понимают *главные закономерности в развитии случайного процесса.* Таким образом, в отличие от тренда, тенденция ряда может образовывать циклы. Довольно часто именно долгопериодная изменчивость временного ряда и принимается в качестве его основной тенденции. Кроме того, отсюда следует, что значимый тренд является частным случаем тенденции, но не наоборот.

Очевидно, в некоторых случаях помимо *основного* (главного) тренда целесообразно выделять и *локальные* тренды. Основным является тренд для всего временного ряда. Если же ряд разбить на отдельные характерные отрезки, отличающиеся друг от друга направленностью временных колебаний, то для каждого из них можно построить свои локальные тренды. Пример подобных трендов будет приведен ниже.

В качестве конкретной иллюстрации того, как в зависимости от длины ряда тренд меняет свои характеристики, обратимся к рис. 10.2, на котором приводятся суточные значения температу-

ры воздуха за два годовых интервала (N = 730 сут.) в Санкт-Петербурге. Действительно, если мы рассматриваем весь ряд, то имеем отчетливо выраженный годовой ход температуры, на который накладываются короткопериодные синоптические колебания. Именно годовую гармонику в данном случае можно отождествлять с тенденцией временного ряда. Но поскольку длина исходной реализации позволяет выделить годовой период, то тренд в ряду отсутствует по определению.

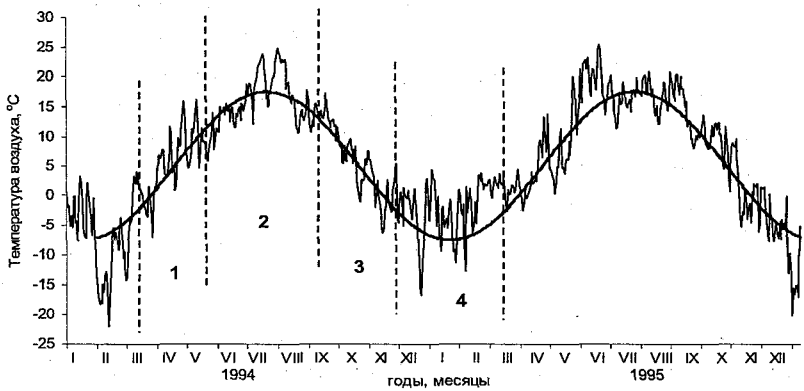


Рис. 10.2. Изменчивость среднесуточных значений температуры воздуха в Санкт-Петербурге за 1994–1995 гг.

В то же время если на рис. 10.2 мы выберем любой интервал, меньший 365 суток, то в этом случае выделить годовой период уже не удается, поэтому можно сделать вывод о наличии для данного интервала трендовой составляющей. При этом тренд является частью годовой гармоники и может быть как линейного, так и нелинейного вида. Из рис. 10.2 следует, что если тренд носит нелинейный характер (например, промежутки времени 2 и 4), то он, очевидно, может быть аппроксимирован полиномом второй степени:

$$T(t) = a_0 + a_1 t + a_2 t^2. \qquad (10.10)$$

По мере уменьшения длины исходной реализации тренд будет претерпевать определенные изменения и для некоторых сравнительно небольших частей этого ряда (интервалы времени 1 и 3 на рис. 10.2) уже вполне достаточным оказывается линейное представление тренда:

$$T(t) = a_0 + a_1 t. \qquad (10.11)$$

Численные значения коэффициентов в этих формулах определяются методом наименьших квадратов. Разумеется, возможны и другие способы аппроксимации тренда. Например, в некоторых случаях нелинейный тренд описывается экспоненциальной зависимостью:

$$T(t) = a_0 + a_1 \exp(a_2 t). \qquad (10.12)$$

Однако использование метода наименьших квадратов возможно только в частном случае при $a_2 = 1$. При оценке тренда наиболее важным представляется оценка его значимости, т.е. насколько существен его вклад в изменчивость случайного процесса. Как и раньше, будем для этой цели использовать критерий Стьюдента. Так, при оценке значимости линейного тренда записывается нулевая гипотеза по отношению к коэффициенту регрессии a_1 и коэффициенту корреляции r:

$$H_0 : |a_1| = 0, \; H_0 : |r| = 0 \qquad (10.13)$$

Для проверки этих гипотез рассчитывается выборочный критерий Стьюдента, причем можно показать, что $t_r = t_{a1}$. Это облегчает оценку значимости тренда. Напомним, что в гл. 6 было показано

$$t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} .$$

Тренд считается значимым, если оценки критерия Стьюдента превышают его критическое значение при заданном уровне значимости, т.е.

$$t > \mathbf{t}_{\text{кр}}(\alpha, \; \nu = n-2). \qquad (10.14)$$

В современных ППСП, как уже указывалось, значимость тренда может быть определена непосредственно по оценке *p-level* коэффициента a_1.

При оценке значимости нелинейного тренда рассчитывается корреляционное отношение η, а затем осуществляется проверка нулевой гипотезы как коэффициента корреляции. По величине коэффициента корреляции и корреляционного отношения легко определить *коэффициент детерминации, показывающий вклад тренда в описание дисперсии функции отклика.*

Другой важной характеристикой тренда является его величина Tr, определяемая в линейном случае как

$$Tr = \frac{(a_0 + a_1 t_n) - (a_0 + a_1 t_1)}{n} = \frac{a_1 (n-1)}{n} \approx a_1, \qquad (10.15)$$

где n – длина ряда.

Отсюда видно, что *величина линейного тренда определяется коэффициентом регрессии.* Аналогичным образом для нелинейного тренда вычисляются его первое и последнее значения по формуле (10.10) и затем разность делится на длину интервала. В результате получаем величину тренда в единицу времени. Следует иметь в виду, что рассчитанный таким образом для нелинейного тренда коэффициент детерминации выше по сравнению с линейным трендом. При этом чем больше «крутизна» тренда, тем больше различия между линейным и нелинейным трендами.

Итак, коэффициент детерминации и величина тренда исчерпывающим образом характеризуют поведение тренда. Однако необходимо помнить, что рассмотренная выше процедура оценивания трендов является параметрической, эффективность которой существенно зависит от того, насколько точно исходный ряд близок к нормальному распределению и от его длины. Действительно, для длинных рядов, даже если исходный ряд не является нормальным, оценка тренда может быть осуществлена рассмотренным выше образом. Для коротких рядов, особенно когда распределение исходных данных неизвестно, эффективность оценивания тренда по формулам (10.10)–(10.15) резко снижается. В этом случае для оценки коэффициента детерминации могут быть использованы непараметрические коэффициенты ранговой корреляции Спирмэна или Кендалла с последующей приближенной оценкой на значимость по критерию Стьюдента или непараметрические критерии серий по числу экстремумов, по числу повышений и понижений значений случайной величины в исходном ряду. Однако непараметрические критерии позволяют лишь предположить присутствие тренда во временном ряду, но не дают его количественных оценок.

Отметим, что приведенные выше формулы (10.10)–(10.12) описывают тренд по среднему арифметическому. Это наиболее часто встречающийся вид трендов в природных процессах, хотя в некоторых случаях может встречаться тренд по дисперсии вре-

менного ряда. Характерный пример такого тренда – развивающийся характер ветрового волнения. Допустим, в штилевую погоду над морем начал дуть ветер с постоянной скоростью и постоянного направления. Естественно, в этих условиях возникает ветровое волнение, которое постепенно будет усиливаться до стадии полностью развитого (установившегося) волнения, когда параметры волн перестают изменяться во времени. Если мы в течение этого периода будем регистрировать, например, высоту морской поверхности (рис. 10.3), то увидим, что амплитуда колебаний уровня плавно изменяется от нуля (штиль) до наибольших значений, соответствующих установившемуся волнению. Огибающая амплитуды свидетельствует о наличии положительного тренда по дисперсии уровня, в то время как математическое ожидание остается постоянным. И, наоборот, отрицательный тренд по дисперсии возникает, когда при развитом волнении стихает ветер и волнение переходит в зыбь, которая постепенно затухает.

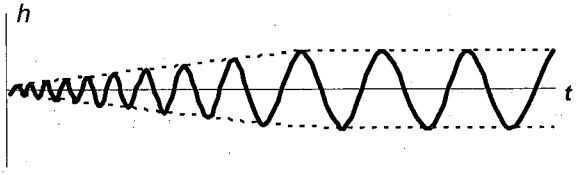


Рис. 10.3. Изменчивость высоты морской поверхности от момента, когда ветер отсутствует, до момента полного развития ветрового волнения.

Итак, исходя из аналитических формул, нетрудно рассчитать трендовую составляющую заданного ряда, определить ее дисперсию и путем последовательного вычитания исключить трендовую составляющую из этого ряда.

Заметим, что в тех случаях, когда аналитическое выражение тренда не представляет для нас интереса, для его исключения из временного ряда могут быть использованы другие приемы. Если, например, тренд аппроксимировать линейным уравнением (10.11), то, обозначая последовательные моменты времени через t_1, t_2, t_3, ..., t_n, исходный временной ряд $\Phi(t)$ можно записать следующим образом:

$$\Phi_1 = a_0 + a_1 + d_1 \qquad (t = 1),$$
$$\Phi_2 = a_0 + 2a_1 + d_2 \qquad (t = 2),$$

$$\Phi_3 = a_0 + 3a_1 + d_3 \quad (t = 3),$$
$$\cdots\cdots\cdots\cdots\cdots$$
$$\Phi_n = a_0 + na_1 + d_n \quad (t = n),$$

где d_1, d_2, d_3, ..., d_n – отклонения величин временного ряда от тренда, т.е. $d(t) = \Phi(t) - T(t)$.

Найдем первые разности:

$$\Delta_1' = \Phi_2 - \Phi_1 = a_1 + (d_2 - d_1),$$
$$\Delta_2' = \Phi_3 - \Phi_2 = a_1 + (d_3 - d_2),$$
$$\cdots\cdots\cdots\cdots\cdots,$$
$$\Delta_{n-1}' = \Phi_n - \Phi_{n-1} = a_1 + (d_n - d_{n-1}).$$

Видно, что во всех разностях присутствует одна и та же константа a_1. Поэтому колебания рассчитанных разностей $\Delta(t)$ зависят только от $d(t)$, т. е. при этом механически исключается тренд.

Если тренд аппроксимируется полиномом второй степени, то в этом случае необходимо рассчитывать уже вторые разности Δ''. Например,

$$\Delta_1'' = \Delta_2' - \Delta_1' = 2a_2 + (d_3 - 2d_2 + d_1),$$
$$\Delta_2'' = \Delta_3' - \Delta_2' = 2a_2 + (d_4 - 2d_3 + d_2).$$

Поскольку во вторые разности входит постоянная величина $2a_2$, то колебания $\Delta''(t)$ зависят только от величин $d(t)$.

Пример 10.1. Обратимся к рис. 10.4, на котором представлен межгодовой ход уровня Каспийского моря в г. Баку по инструментальным наблюдениям за 1900–2004 гг. Нетрудно видеть, что в колебаниях уровня довольно четко выделяются три достаточно длительных стабильных периода, в течение которых изменения уровня оказываются относительно однородными. По аналогии с естественными синоптическими периодами они были названы «естественными климатическими периодами» (ЕКП). Первый ЕКП (1900–1929 гг.) характеризуется стоянием уровня, т.е. его колебания носят случайный характер относительно некоторого мало изменяющегося среднего положения. Второй ЕКП (1930–1977 гг.) сопровождается почти монотонным падением уровня, а третий ЕКП (1978–2004 гг.), наоборот, таким же почти монотонным подъемом уровня.

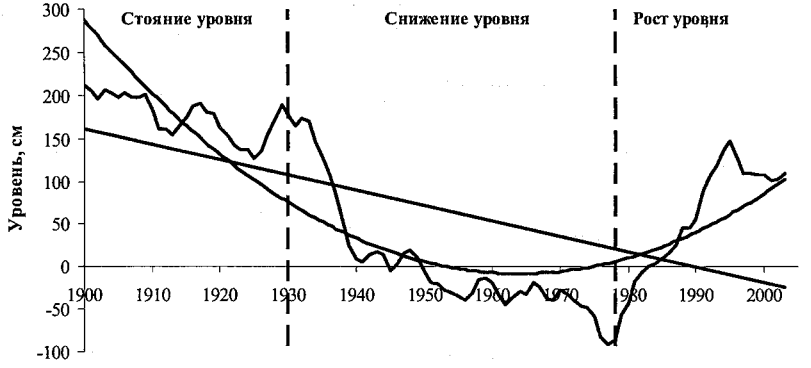


Рис. 10.4. Межгодовой ход уровня Каспийского моря по инструментальным наблюдениям (за нуль принята отметка –28 м БС).

Размах колебаний уровня за весь период времени достигает 3 м (табл. 10.1), причем в течение второго ЕКП он понизился на 2,7 м. Безусловно, такие быстрые однонаправленные колебания уровня привели к огромному экономическому ущербу для прибрежной инфраструктуры, выражаемой в миллиардах долларов. Однако отметим, что подобные колебания уровня нельзя назвать выдающимся событием в истории Каспия. Например, как следует из палеоклиматических данных, в IX в. размах колебаний уровня достигал 7 м.

Нелинейный и линейный тренды уровня моря за 1900–2004 гг. (рис. 10.4) могут быть аппроксимированы следующими выражениями:

$$h = 296{,}77 - 9{,}385t + 0{,}0722t^2\,; \qquad (10.10')$$

$$h = 162{,}76 - 1{,}800t\,. \qquad (10.11')$$

Линейный тренд описывает 33 % дисперсии исходного ряда, а нелинейный – 75 %, причем, естественно, оба тренда являются значимыми. Характеристики линейных локальных трендов для отдельных ЕКП также даются в табл. 10.1. Хотя для периода 1900–1929 гг. линейный тренд является значимым, однако, учитывая, что его величина на порядок меньше трендов для других ЕКП, целесообразно считать, что для первого ЕКП характерно стояние

уровня. Максимальный тренд как по величине, так и по вкладу в дисперсию колебаний уровня наблюдается для третьего ЕКП.

Таблица 10.1

Первичные статистические характеристики для однородных периодов изменений уровня Каспийского моря с 1900 по 2004 г.

Характеристика	1900–1929	1930–1977	1978–2004	1900–2004
Среднее, см/год	175,6	7,5	58,9	68,7
Медиана, см/год	180	–18,5	89	56
Стандартное отклонение, см/год	24,8	68,5	65,7	91,8
Коэффициент эксцесса	–1,01	1,06	–0,73	–1,51
Коэффициент асимметрии	–0,41	1,38	–0,61	0,04
Минимум, см/год	126	–92	–87	–92
Максимум, см/год	211	177	148	211
Величина тренда, мм/год	–2,5	–48	71,5	–17
Коэффициент детерминации тренда	0,465	0,726	0,804	0,326

Пример 10.2. Как известно, в XX столетии отмечалось довольно существенное потепление климата. Наиболее репрезентативным показателем глобального климата является приповерхностная температура воздуха (ПТВ). Поэтому обратимся к рис. 10.5, на котором приводится межгодовой ход аномалий ПТВ для каждого из полушарий за период 1880–2005 гг. При этом аномалии ПТВ отсчитывались от базового периода за 1950–1980 гг. Из рис. 10.5 видно, что в целом отмечается согласованный характер изменений ПТВ в каждом из полушарий. Естественно, что в Северном полушарии не только дисперсия колебаний температуры, но и величина ее тренда несколько выше по сравнению с Южным полушарием.

Можно отметить, что примерно до начала 1940-х годов XX столетия наблюдался сравнительно быстрый рост температуры. Особенно ярко он проявлялся в высоких широтах Северного полушария. Поэтому 20–40-е годы получили название «потепление Арктики». Затем вплоть до середины 70-х годов в Северном полушарии отмечалось практически повсеместное похолодание, характеризуемое значимым отрицательным трендом. В данный период тренд в южном полушарии отсутствовал. И только после этого произошел резкий рост ПТВ, продолжающийся до настоящего

времени. Отметим, что последнее десятилетие XX столетия и 2005 г. считаются самыми теплыми за весь период инструментальных измерений температуры воздуха. Итак, за рассматриваемый период мы можем выделить три интервала, заметно отличающиеся друг от друга характером колебаний температуры и, следовательно, величиной локальных трендов, характеристики которых приведены в табл. 10.2. Нетрудно видеть, что максимальные оценки, как величины тренда, так и его вклада в дисперсию исходных временных рядов, отмечаются в период наиболее сильного потепления, т.е. начиная с 1975 г. и по настоящее время.

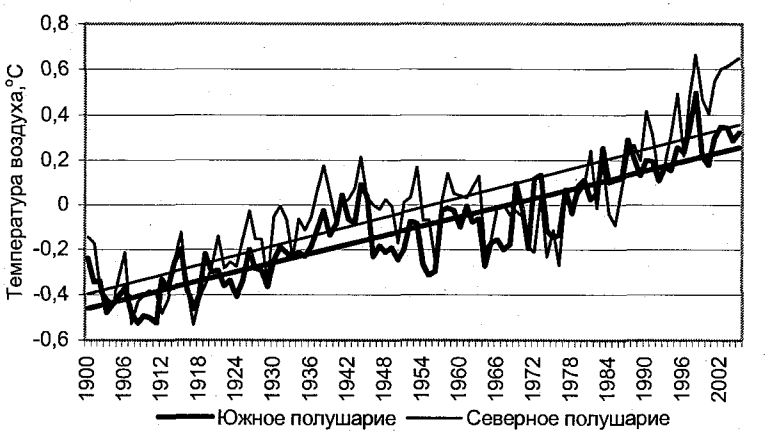


Рис. 10.5. Межгодовой ход приповерхностной температуры воздуха в Северном и Южном полушариях за период 1900–2005 гг.

Таблица 10.2

Оценки линейных трендов ПТВ за различные периоды

Период, годы	Северное полушарие		Южное полушарие		Земной шар	
	R^2	°С / 10 лет	R^2	°С / 10 лет	R^2	°С / 10 лет
1880–2005	0,65	0,060	0,73	0,055	0,72	0,058
1880–1940	0,27	0,045	0,17	0,026	0,25	0,036
1941–1975	0,18	–0,048	0,01	0,012	0,03	–0,018
1976–2005	0,78	0,252	0,62	0,116	0,77	0,184

Поскольку ПТВ в обоих полушариях имеет практически одинаковый линейный тренд, обусловленный глобальным характером потепления климата, то естественно, что между данными времен-

ными рядами отмечается очень высокая корреляция, равная за период 1880–2005 гг. $r = 0{,}90$. Очевидно, что этот тренд можно трактовать как некоторую третью переменную, вызывающую эффект ложной корреляции. Для его исключения вычтем из каждого ряда ПТВ линейный тренд. На рис. 10.6 приводится межгодовой ход ПТВ для каждого из полушарий после исключения трендовых компонент. Нетрудно видеть, что ход ПТВ уже не является таким согласованным, как на рис. 10.5. В этом случае корреляция между данными рядами составляет $r = 0{,}66$. Следовательно, $r_{лож} = 0{,}90 - 0{,}66 = 0{,}24$.

Рис. 10.6. Межгодовой ход приповерхностной температуры воздуха в Северном и Южном полушариях за 1900–2005 гг. после исключения линейного тренда.

10.3. Гармонический анализ

Для описания периодических и квазипериодических процессов, заданных в виде $G(t)$ и $C(t)$ в разложении (10.1), может быть, например, использован гармонический анализ, представляющий собой разложение в ряд Фурье. Весьма важно также, что разложение в ряд Фурье носит более универсальный характер, поскольку позволяет описать любую реализацию случайной функции с помощью конечного числа гармоник.

Итак, если мы имеем функцию $f(t)$, то в интервале $[T, T + 2\pi]$ ее можно представить в виде ряда Фурье:

$$f(t) = \overline{f} + \sum_{n=1}^{\infty} [a_n \cos(n\omega t) + b_n \sin(n\omega t)], \qquad (10.16)$$

где $\overline{f}$ – математическое ожидание функции $f(t)$; ω – частота; a_n и b_n – коэффициенты, определяемые как

$$a_n = \frac{2}{T}\int_0^T f(t)\cos(n\omega t)dt,\ b_n = \frac{2}{T}\int_0^T f(t)\sin(n\omega t)dt. \qquad (10.17)$$

Разложение в ряд Фурье не только точно представляет функцию $f(t)$ при длине ряда $N\to\infty$, но и обеспечивает при фиксированной величине N наименьшую среднюю квадратическую ошибку по сравнению с любым другим представлением функции $f(t)$ в виде тригонометрического ряда по $\sin(n\omega t)$ и $\cos(n\omega t)$ той же длины N.

Слагаемые ряда Фурье

$$U_n = a_n \cos(n\omega t) + b_n \sin(n\omega t),\ n = 1, 2, 3,...$$

называются *гармониками*. Если ввести угол φ_n так, что

$$\cos\varphi_n = \frac{a_n}{\sqrt{a_n^2 + b_n^2}},\ \sin\varphi_n = \frac{b_n}{\sqrt{a_n^2 + b_n^2}},$$

то гармоники можно представить в виде

$$U_n = A_n \cos(n\omega t - \varphi_n),\qquad A_n = \sqrt{a_n^2 + b_n^2}. \qquad (10.18)$$

Величины A_n и φ_n называются соответственно *амплитудой* и *фазой* гармоники. Если физический смысл амплитуды гармоники очевиден ($A = X_{max} - X_{min}$), то фаза представляет временной интервал наступления первого максимума от начала отсчета. Если, например, в годовом ходе температуры воздуха ее максимум наступает в июле, то фаза равна φ = 7 мес.

Заметим, что разложение (10.16) характеризует непрерывный случайный процесс. На практике же обычно имеют дело с дискретными временными рядами. Поэтому разработано несколько численных методов построения дискретного преобразования. К ним, в частности, относятся рекуррентный метод и метод быстрого преобразования Фурье, математическое изложение которых выходит за пределы содержания данной книги. Однако существуют и более простые алгоритмы оценок гармоник. Так, Пановским и Брайером предложена следующая формула для дискретного случайного процесса $X(t)$:

$$X(t) = \bar{x} + \sum_{i=1}^{N/2}\left[a_i \sin\left(\frac{360it}{P}\right) + b_i \cos\left(\frac{360it}{P}\right)\right], \qquad (10.19)$$

где P – полный период, выражаемый в единицах времени; N – длина ряда; i – порядковый номер гармоники; t – время, отсчитываемое от начала периода.

Коэффициенты Фурье здесь рассчитываются как

$$a_i = 0{,}5N\sum [x\sin(360it/P)],\ \ b_i = 0{,}5N\sum [x\cos(360it/P)]. \qquad (10.20)$$

Для удобства вычислений в зависимости от N и P составляется таблица множителей $0{,}5N\sin(360it/P)$ и $0{,}5N\cos(360it/P)$ для всех значений i и t. Для последней гармоники множители равны $N^{-1}\cos(360it/P)$. Значения анализируемого ряда умножаются на соответствующие множители. Их суммы и являются коэффициентами Фурье. Затем синусные и косинусные составляющие i-й гармоники складываются и представляются в виде

$$A_i = \cos[360(t - t_i)/P],$$

где $A_i = (a_i^2 + b_i^2)^{1/2}$; $t_i = (P/360i)\text{arctg}(a_i/b_i)$; $t = (P/360i)\arcsin(a_i/A_i)$, здесь A_i – амплитуда i-й гармоники, а t_i – время наступления максимума i-й гармоники.

В результате разложения в ряд Фурье мы имеем постоянную компоненту $\bar{f}$ и сумму гармоник с частотами, кратными основной частоте ω, постоянными амплитудами A_n и начальными фазами φ_n.

Свойства гармонического разложения:

Свойство 1. Дисперсия гармоники функционально связана с ее амплитудой следующей зависимостью: $\sigma_n^2 = A_n^2/2$;

Свойство 2. Гармоники независимы (некоррелированы) между собой;

Свойство 3. Число рассчитываемых гармоник равно $N/2$;

Свойство 4. Период первой гармоники равен длине исходного ряда, каждой последующей – кратен первой, вследствие чего период последней гармоники равен удвоенной дискретизации ряда.

Пусть, например, длина исходного ряда $N = 48$ месяцев. Тогда первая гармоника имеет период $\tau = 48$ мес., вторая – $\tau = N/2 = 24$ мес., третья – $\tau = N/3 = 16$ мес., последняя – $\tau = N/24 = 2$ мес. При исследовании годового цикла ($N = 12$ мес.) первая гармоника

имеет период, равный длине всего основного периода (12 мес.), вторая – период, равный половине основного (6 мес.), третья – период, равный 1/3 основного (4 мес.). Наконец, последняя шестая гармоника имеет период, равный 2 мес.

Так как гармоники некоррелированы друг с другом, то дисперсия суммы всех гармоник равна сумме дисперсий этих гармоник. Поэтому не составляет большого труда оценить вклад любой гармоники в общую дисперсию процесса, который равен $k_i = A_i^2/2\sigma^2$.

Весьма важным при анализе гармоник представляется отделение значимых гармоник от незначимых. По предложению С.М. Гордеевой для этой цели удобно использовать дисперсионный анализ. Действительно, вклад гармоники k_i можно интерпретировать как коэффициент детерминации, т.е. ту часть дисперсии процесса, которая описывается данной гармоникой. Тогда извлекая корень из k_i, получаем коэффициент корреляции между данной гармоникой и исходным процессом, т.е.

$$r = (k_i)^{1/2} .$$

Значимость величины r, как мы знаем, оценивается по критерию Стьюдента. Определив значимые гармоники и вычтя их суммарную дисперсию из дисперсии ряда, находим дисперсию остатков:

$$D_\varepsilon = D_x - \sum_{i=1}^{p} k_i ,$$

где p – число значимых гармоник.

Очевидно, в первом приближении величину D_ε можно трактовать как дисперсию случайного процесса, близкого к модели белого шума.

Итак, процедуру гармонического анализа удобно разбить на пять этапов:

1) определение средней арифметической величины;

2) оценка коэффициентов Фурье a_i, b_i и построение функции аппроксимации исходных данных;

3) оценка вклада отдельных гармоник в общую дисперсию ряда;

4) интерпретация значимых гармоник;

5) исключение значимых гармоник из исходного ряда с целью анализа остатков.

Гармонический анализ используется для выделения скрытых периодичностей. Однако в природе, вообще говоря, гармонические волны практически не встречаются. Очень близки к строгим гармоникам приливные волны. Действительно, периоды коротких (полусуточных и суточных) приливных волн можно считать практически постоянными. Однако амплитуда даже самых «правильных» полусуточных волн меняется во времени. Причиной этого является полумесячное неравенство приливов. В соответствии с этим неравенством максимальный полусуточный прилив наблюдается в сизигию (момент времени, когда Луна, Земля и Солнце находятся на одной прямой), а минимальный полусуточный прилив отмечается в квадратуру (момент времени, когда Луна, Земля и Солнце составляют угол 90°). В результате во временном ходе уровня моря амплитуда его колебаний плавно изменяется от наибольших значений в сизигию до наименьших в квадратуру.

Кроме того, на формирование амплитуды прилива существенное влияние могут оказывать местные условия, которые либо уменьшают его величину, либо, наоборот, увеличивают. Как известно, максимальные приливы, наблюдающиеся в заливе Фанди (атлантическое побережье Канады), могут достигать 16–18 м. Причиной этого является совпадение по фазе сейш (собственных свободных колебаний уровня) и приливных колебаний, вследствие чего возникает явление резонанса и происходит резкое увеличение амплитуды колебаний.

Кроме приливных колебаний, к явлениям с относительно строгой периодичностью для многих районов земного шара можно отнести годовой и суточный ход основных гидрометеорологических характеристик. В этом случае при исследовании временных рядов, длина которых кратна таким периодичностям, выделенные гармоники будут иметь четкий физический смысл.

Следует иметь в виду, что при использовании гармонического анализа нельзя подходить формально к выделению гармоник. Если с физической точки зрения в данном конкретном временном ряду существование квазипериодических компонент маловероятно, то стоит хорошо подумать, есть ли необходимость в использовании гармонического анализа.

Пример 10.3. Рассмотрим особенности годового хода изменений теплосодержания деятельного слоя океана в десятиградусных широтных зонах Северного полушария. В табл. 10.3 для первых двух гармоник приводятся амплитуды A_i (Вт/м^2), фазы φ_i (месяцы) первого максимума и суммарный вклад этих гармоник (%) в дисперсию исходного процесса.

Из табл. 10.3 видно, что максимальная амплитуда 1-й гармоники годового хода изменений теплосодержания океана наблюдается в зоне 40–30° с.ш. и быстро уменьшается по направлению к экватору и полюсу от этой зоны. Время наступления максимума изменений теплосодержания океана очень слабо меняется с широтой и почти везде (исключая зону 20–10° с.ш.) отмечается в летний период, когда происходит интенсивное прогревание верхних слоев океана за счет солнечной радиации.

Таблица 10.3

Гармонический анализ изменений теплосодержания деятельного слоя океана в десятиградусных широтных зонах (Вт/м^2) Северного полушария

Параметр	90–80	80–70	70–60	60–50	50–40	40–30	30–20	20–10	10–0
A_1, Вт/м^2	31	37	27	58	86	109	66	50	16
A_2, Вт/м^2	6	16	6	2	10	0	19	25	11
φ_1, месяц	4,9	6,5	5,8	5,7	6,0	5,8	5,7	2,9	4,2
φ_2, месяц	3,9	4,4	0,4	0,5	3,6	5,1	3,1	3,9	2,6
%	54	78	90	97	99	100	99	92	60

Полугодовая гармоника играет существенную роль в сезонном ходе изменений теплосодержания только в низких широтах, причем в экваториальной зоне (10–0° с.ш.) ее амплитуда сравнима с амплитудой 1-й гармоники. Сумма 1-й и 2-й гармоник с высокой точностью описывает годовой ход изменений теплосодержания океана в пределах 10–70° с.ш., т.е. в большей части акватории океана Северного полушария. В экваториальной зоне и в Северном Ледовитом океане уже заметна роль непериодических (случайных) колебаний в формировании годового хода изменений теплосодержания океана.

Пример 10.4. Оценим наличие скрытых периодичностей в колебаниях морского уровня в Онежском заливе Белого моря в районе о. Большой Соловецкий, наблюдения над которым выполнены студентами-океанологами 3 курса РГГМУ в рамках производст-

венной практики. Измерения уровня осуществлялись с 5 по 31 июля 2005 г. с дискретностью 2 часа. Временной ход уровня приведен на рис. 10.7. Горизонтальными черточками отмечены осредненные за сутки значения уровня. Среднее значение уровня за истекший период равно 135,9 см, дисперсия – 514,1 см^2, размах колебаний – 107 см. Из рис. 10.7 видно, что наиболее характерной чертой колебаний морского уровня является наличие значительного полусуточного прилива. Второй пик принадлежит довольно слабому суточному приливу.

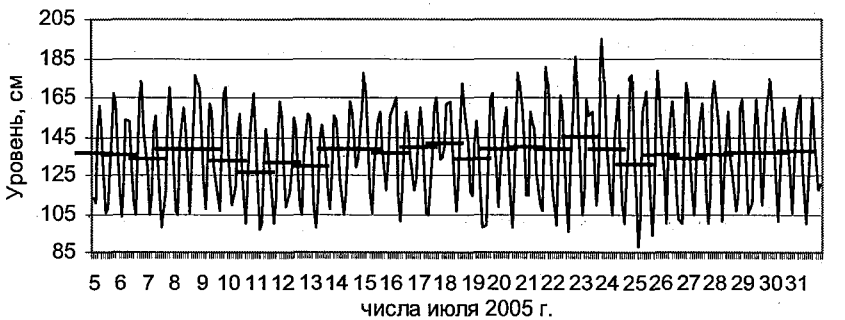


Рис. 10.7. Временная изменчивость уровня моря в июле 2005 г. в Онежском заливе Белого моря. Прямые линии – среднесуточные значения уровня.

Рассчитаем амплитуды гармоник временного ряда, общее число которых равно $N = 162$, и нанесем их на график, который называется периодограммой (рис. 10.8). Из этого рисунка видно, что действительно наиболее мощной гармоникой является главная лунная полусуточная волна M_2 (τ = 12,4 ч). Характеристики ее приводятся в табл. 10.4. Амплитуда волны M_2 = 26,1 см, а вклад в дисперсию исходного процесса достигает 66 %. Отсюда коэффициент корреляции этой волны с исходным процессом $r = (0{,}661)^{1/2} =$ = 0,813. Рассчитаем теперь парную корреляцию непосредственно между данной гармоникой и исходным рядом уровня. Получаем r = 0,812 и R^2 = 0,660, т.е. имеем практически точное соответствие с результатами периодограммы.

Суточный прилив составляет лунно-солнечная деклинационная волна K_1. Ее амплитуда почти в 4 раза меньше волны M_2. Практически в 4 раза меньше и коэффициент корреляции (r = 0,21) с исходным рядом. Однако он значим, так как критическое значе-

ние коэффициента корреляции при уровне значимости $\alpha = 0{,}05$ равно $\mathbf{r}_{кр} = 0{,}11$.

Таблица 10.4

Гармонический анализ морского уровня в районе о. Большой Соловецкий

Гармоника	Период, ч	Амплитуда, см	Дисперсия, см^2	Коэффициент детерминации	Коэффициент корреляции
Полусуточная волна	12,4	26,1	340.1	0,661	0,813
Суточная волна	24,0	6,6	21,7	0,042	0.21
Сейша	72	3,3	5,4	0,011	0,10

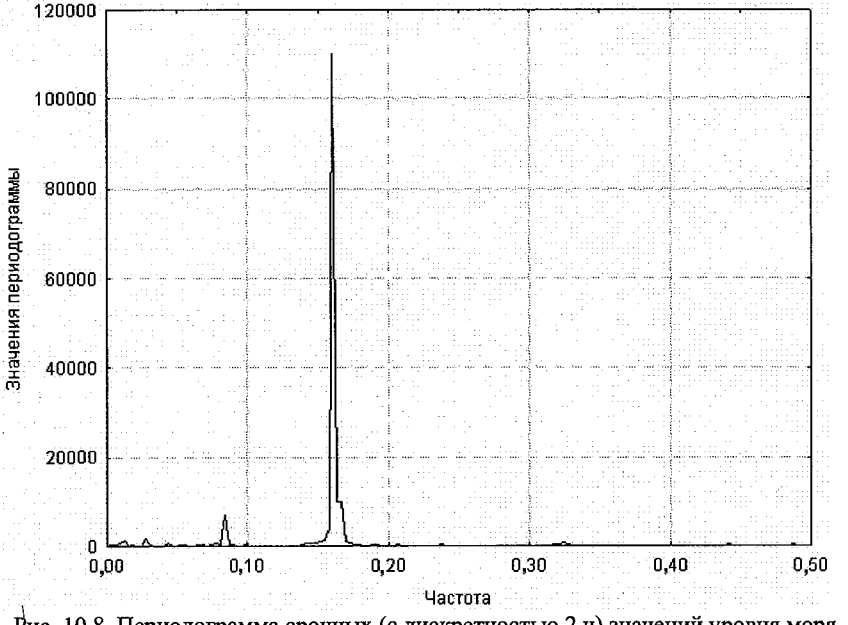


Рис. 10.8. Периодограмма срочных (с дискретностью 2 ч) значений уровня моря за период с 5 по 31 июля 2005 г. ($N = 314$).

Следующая по величине гармоника, которая практически даже не видна на рис. 10.8, соответствует периоду $\tau = 72{,}0$ ч, а ее вклад в дисперсию исходного ряда равен $R^2 = 5{,}4/514{,}1 = 0{,}011$, что соответствует величине $r = 0{,}10$. Очевидно, ее можно интерпретировать как крупномасштабную сейшу, обусловленную изменениями синоптических процессов в данном районе и, прежде всего, ветровым режимом. Данная гармоника незначима на уровне $\alpha =$

=0,05, но значима при $\alpha = 0{,}10$ ($r_{кр} = 0{,}09$). Если исключить эти гармоники из исходного ряда, то получим чисто случайный процесс, близкий к модели «шума белого». Дисперсия его равна 147 см^2 или 28,6 % от исходной дисперсии.

10.4. Автокорреляционный анализ

Знание автокорреляционных функций (АФ) позволяет решить широкий круг задач, связанных с исследованием и прогнозированием изменчивости гидрометеорологических условий. К ним относятся: выделение скрытых периодичностей, в том числе регулярных межгодовых колебаний; вычисление оценки спектральной плотности; оценка степени связности ряда; долгосрочный прогноз на основе методов экстраполяции АФ.

Для одной реализации эргодического стационарного процесса АФ показывает степень линейной зависимости значений процесса от предшествующих значений, относящихся к различному сдвигу τ. Таким образом, АФ характеризует внутреннюю структуру процесса, его динамику во временной области.

Коэффициентом автокорреляции называют коэффициент корреляции между значениями данного ряда и его же значениями, относящимися к некоторому сдвигу τ. Следовательно, АФ представляет последовательность коэффициентов автокорреляции, начиная с $\tau = 0$.

Для непрерывного стационарного процесса бесконечной продолжительности нормированная автокорреляционная функция имеет вид:

$$r(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} X(t)X(t+\tau)d\tau . \qquad (10.21)$$

Переходя в (10.21) к дискретному стационарному процессу конечной продолжительности, получаем:

$$r(\tau) = \frac{\sum^{N-\tau} (x_i - \bar{x})(x_{i+\tau} - \bar{x})}{(N-\tau)\sigma^2}, \qquad (10.22)$$

где N – длина реализации, которая предполагается весьма значительной.

Помимо нормированной автокорреляционной функции, в статистике часто используют понятие автокорреляционной (автоковариационной) функции, также характеризующей внутреннюю структуру процесса и определяемой по формуле:

$$R(\tau) = \frac{\sum_{}^{N-\tau} (x_i - \bar{x})(x_{i+\tau} - \bar{x})}{N - \tau}. \qquad (10.23)$$

Из формул (10.22) и (10.23) следует, что

$$\frac{R(\tau)}{r(\tau)} = \sigma^2, \qquad (10.24)$$

т.е. нормированная автокорреляционная и автокорреляционная функции отличаются друг от друга на величину дисперсии.

Перечислим свойства автокорреляционной функции для типичных гидрометеорологических процессов:

Свойство 1. АФ – четная функция, т.е. $r(-\tau) = r(\tau)$, что позволяет сократить процесс вычислений, рассчитывая $r(\tau)$ только для положительных сдвигов.

Свойство 2. АФ – убывающая функция, т.е. $r(\tau = 0) \geq r(\tau)$, причем при $\tau \to \infty$ $r(\tau) \to 0$.

Свойство 3. Так как при $\tau = 0$ $r = 1$, то дисперсия АФ $\sigma^2 = 1$.

Свойство 4. Если в процессе присутствует периодическая составляющая, то ее период сохраняется в АФ.

Для характеристики временного сдвига, на котором происходит затухание АФ, в практике анализа стационарных случайных процессов вводится понятие интервала корреляции. *Интервал корреляции* является мерой протяженности линейной связи между ординатами процесса.

Максимальный (абсолютный) интервал корреляции определяется следующим выражением:

$$\tau_{\text{кор}}^{(1)} = \int_0^{\infty} |r(\tau)|\, d\tau. \qquad (10.25)$$

Ординаты процесса, разделенные интервалом, большим $\tau_{\text{кор}}^{(1)}$, можно считать некоррелированными. Поэтому экстраполяция их на время, большее $\tau_{\text{кор}}^{(1)}$, не имеет смысла. Очевидно, в качестве приближенной оценки максимального интервала корреляции целе-

сообразно принимать сдвиг τ_i, который соответствует последнему значимому коэффициенту автокорреляции.

Существует и другое определение интервала корреляции:

$$\tau_{\text{кор}}^{(2)} = \int_0^{\infty} |r(\tau)|^2 \, d\tau, \tag{10.26}$$

которое, например, применяется при характеристике эффективности оценок АФ. Величина $\tau_{\text{кор}}^{(2)}$, как правило, не превышает временного сдвига, соответствующего первому пересечению АФ оси τ. Величина $\tau_{\text{кор}}^{(2)}$ часто называется *радиусом корреляции*. Приближенно радиус корреляции может быть определен по первому пересечению АФ оси τ.

Задача определения интервалов корреляции существенно облегчается, если известно аналитическое выражение, аппроксимирующее АФ исследуемого процесса. Широкий класс физических, в том числе океанологических, процессов описывается АФ вида:

$$r(\tau) = e^{-\alpha|\tau|}, \tag{10.27}$$

$$r(\tau) = e^{-\alpha|\tau|} \cos \beta\tau, \tag{10.28}$$

где α – коэффициент затухания; β – частота колебания АФ.

Подставив (10.27) в (10.25) и (10.26), имеем

$$\tau_{\text{кор}}^{(1)} = \frac{1}{\alpha}, \qquad \tau_{\text{кор}}^{(2)} = \frac{1}{2\alpha}.$$

Таким образом, если АФ выражается экспоненциальной зависимостью, то $\tau_{\text{кор}}^{(1)} = 2\tau_{\text{кор}}^{(2)}$.

Подставив (10.28) в (10.25) и (10.26), получим

$$\tau_{\text{кор}}^{(1)} = \frac{1}{\alpha(1+\mu^2)}\left[1 + \mu \exp\left(\frac{2\pi}{\mu}\right) + \frac{\mu}{\exp\left(\frac{2\pi}{\mu} - 1\right)}\left(\exp\left(-\frac{\pi}{2\mu}\right) + 2\exp\left(\frac{\pi}{2\mu}\right) + \exp\left(\frac{3\pi}{2\mu}\right)\right)\right], \tag{10.29}$$

$$\tau_{\text{кор}}^{(2)} = \frac{2\alpha^2 + \beta^2}{4\alpha(\alpha^2 + \beta^2)}, \qquad (10.30)$$

где $\mu = \beta / \alpha$.

Наиболее точным методом определения параметров α и β является метод наименьших квадратов. Однако во многих случаях можно ограничиться более простыми способами, основанными на определении параметров аппроксимации по нескольким характерным точкам АФ. Например, при аппроксимации АФ зависимостью (10.28), имеем:

$$\alpha = \ln\cos\frac{\beta\tau_2}{r(\tau_\Delta)}, \qquad \beta = \frac{\pi}{2\tau_1},$$

где τ_1 – временной сдвиг, соответствующий первому пересечению АФ оси τ; τ_2 – временной сдвиг, на котором отмечается ближайший к τ_1 экстремум АФ.

Рассмотрим вопрос о необходимой продолжительности наблюдений для надежной оценки АФ, если в исследуемом ряду имеется цикличность. С этой целью представим временной ряд в виде разложения Фурье:

$$X(t) = \sum_{k=1} A_k \sin(\omega_k t + \varphi_k) \cdot$$

Задавшись величиной максимального сдвига АФ τ_m, равной периоду самой низкочастотной гармоники, присутствующей в процессе $\tau_m = T_1$, можно получить соотношение между продолжительностью наблюдений N, максимальным сдвигом τ_m и относительной средней квадратической ошибкой σ:

$$N \geq \tau_m\left(1 + \frac{1}{2\pi\sigma}\right). \qquad (10.31)$$

Из (10.31) следует, например, что для вычисления ординат АФ с погрешностью 3 % от общей дисперсии процесса необходимо, чтобы

$$N \approx (5\text{–}6)\tau_m,$$

а с погрешностью 2 %

$$N \approx 9\tau_m.$$

Таким образом, надежное определение оценок АФ возможно лишь для тех составляющих процесса, период которых в 5–10 раз меньше продолжительности наблюдений. Если во временном ряду имеются более низкочастотные компоненты, то при помощи фильтрации их предварительно следует исключать из реализации.

При расчетах АФ для случайного процесса с не очень большой длиной ряда проверить условие стационарности весьма сложно. Понятно, что статистические характеристики ($\bar{x}$, σ), рассчитанные при различных величинах сдвига τ по формуле (10.22) для короткого ряда, когда нет уверенности в выполнении условия его стационарности, могут заметно отличаться друг от друга. В этом случае значения АФ целесообразно рассчитывать по формуле:

$$r(\tau)=\frac{\sum_{i=1}^{N-\tau}(x_i-\bar{x}_{1,N-\tau})(x_{i+\tau}-\bar{x}_{\tau+1,N})}{(N-\tau-1)\sigma_{1,N-\tau}\sigma_{\tau+1,N}}. \qquad (10.32)$$

Один из способов оценки АФ заключается в оценке значимости коэффициентов автокорреляции на основе нулевой гипотезы, т. е. $H_0: |r(\tau)| = 0$, при $\tau \neq 0$. Коэффициент автокорреляции окажется значимым, если будет выполняться следующее условие:

$$|r(\tau)|/\sigma_{r(\tau)} > \mathbf{t}_{кр},$$

где $\mathbf{t}_{кр}$ – критерий Стьюдента с N–τ–1 степенями свободы при заданном уровне значимости α; $\sigma_{r(\tau)}$ – стандартное отклонение ординат АФ, рассчитанное как

$$\sigma_{r(\tau)}=\frac{1-r^2(\tau)}{\sqrt{N-\tau-1}}. \qquad (10.33)$$

Если при данном τ ордината $r(\tau)$ выходит за пределы величины $t_\alpha\sigma_{r(\tau)}$, т.е. отличается от нуля более чем это можно приписать случайной вариации, то нулевая гипотеза отвергается, и отличие ординаты от нуля считается значимым. Недостаток этого критерия в том, что он не учитывает наличие во временном ряду циклических колебаний.

Для построения доверительных интервалов используется тот факт, что при малых величинах $r(\tau)$ отклонение коэффициентов автокорреляции от их истинного значения распределяется по нор-

мальному закону. Тогда для длинных временных рядов при уровне значимости $\alpha = 0{,}05$ ($t_\alpha \approx 2$) доверительные границы для оценки значений АФ могут быть найдены как

$$r_{\text{н}}(\tau_i) = r(\tau_i) - \frac{2}{\sqrt{N-\tau_i}}[1 - r^2(\tau_i)];$$

$$r_{\text{в}}(\tau_i) = r(\tau_i) + \frac{2}{\sqrt{N-\tau_i}}[1 - r^2(\tau_i)],$$

где $r_{\text{н}}(\tau_i)$, $r_{\text{в}}(\tau_i)$ – соответственно нижняя и верхняя доверительные границы.

10.5. Автокорреляционные функции различных временных рядов

Большое разнообразие временной изменчивости гидрометеорологических процессов заключено между двумя эталонами: изменчивостью детерминированных процессов и изменчивостью чисто случайного процесса типа «белого шума». При этом во многих случаях гидрометеорологические характеристики не соответствуют какому-либо одному виду процесса, а представляют собой смесь нескольких составляющих изменчивости. Поэтому практический анализ вычисленных выборочных автокорреляционных оценок состоит в том, чтобы провести их сравнение с известными теоретическими автокорреляционными функциями. Вследствие этого становится возможным определение типа случайного процесса.

Рассмотрим вначале стандартные графики нормированных автокорреляционных функций временных рядов для конкретных видов моделей гидрометеорологических процессов. Исходя из классификации случайных процессов (см. п. 9.6), в качестве таковых следует рассматривать модели временного ряда в виде «белого шума», «красного шума» и циклического колебания. Как уже указывалось, теоретическая модель белого шума описывается автокорреляционной функцией, равной нулю на всех сдвигах, исключая сдвиг $\tau = 0$. График такой АФ приведен на рис. 10.9. Следует иметь в виду, что теоретическая модель белого шума не более чем *математическая абстракция*, поэтому на практике такие автокорреляционные функции не встречаются. Поэтому целесообразно от теоретической модели перейти к такому случайному стационарному процессу, о котором можно говорить, что он развивается

по типу модели «белый шум». В этом случае автоматически принимается условие, что коэффициенты автокорреляции являются незначимыми на всех сдвигах, исключая нулевой (см. рис. 10.9). Подобные автокорреляционные функции являются типичными для многих гидрометеорологических процессов. Конкретных примеров можно привести множество. Например, пространственно-временная изменчивость параметров ветрового волнения при различных периодах осреднения, исключая зыбь, как правило, близка к белому шуму.

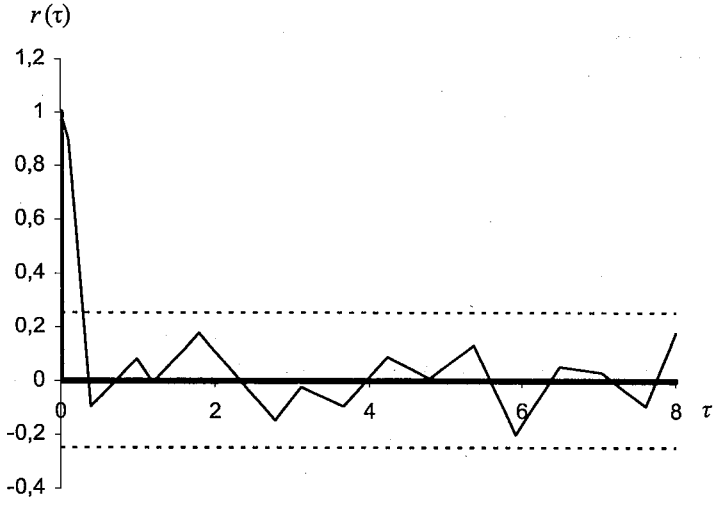


Рис. 10.9. График автокорреляционной функции теоретической (толстая линия) и реальной (тонкая линия) моделей случайного процесса «белый шум».

Теоретическая модель «красный шум» описывается автокорреляционной функцией, которая на первом сдвиге имеет значимый коэффициент автокорреляции, а на всех остальных сдвигах она равна нулю (рис. 10.10). Как и в предыдущем случае, переходим от теоретической модели к случайному стационарному процессу, который развивается по типу модели «красный шум». Тогда принимаем условие, что на первом сдвиге коэффициент автокорреляции является значимым, а на всех остальных он отклоняется от нуля случайным образом (см. рис. 10.10). Данная автокорреляционная функция может быть аппроксимирована экспоненциальной фор-

мулой вида (10.27). Для многих районов Мирового океана межгодовая изменчивость температуры довольно близко соответствует модели красного шума. Как правило, очень хорошо соответствует этой модели и межгодовая изменчивость стока крупных рек.

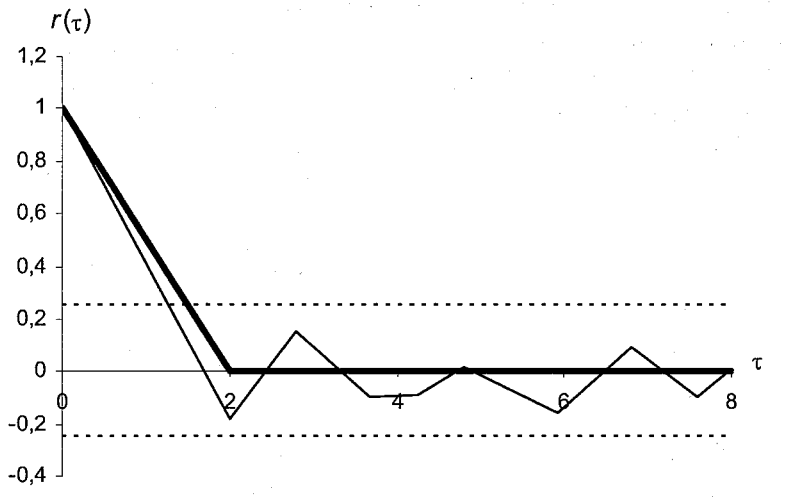
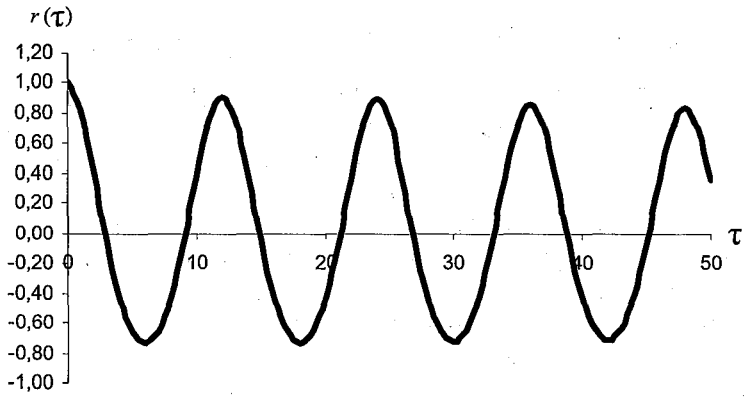


Рис. 10.10. График автокорреляционной функции теоретической (толстая линия) и реальной (тонкая линия) моделей случайного процесса «красный шум».

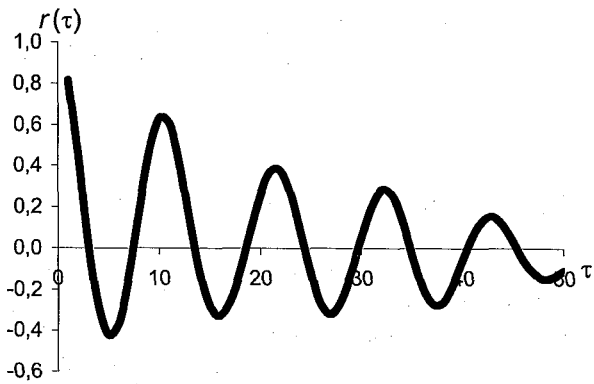
Автокорреляционная функция циклического колебания представляет собой затухающую косинусоиду, степень затухания которой зависит от характера нерегулярных изменений параметров данного колебания, т.е. от его амплитуды и периода. Поэтому если АФ сезонного хода большинства гидрометеорологических процессов вне экваториальной зоны выражается в виде слабо затухающей косинусоиды (рис. 10.11, *а*), то АФ межгодовой изменчивости солнечной активности уже представляет более быстро затухающую косинусоиду (рис. 10.11, *б*). На практике данный вид АФ можно аппроксимировать формулой вида (10.28).

По существу, циклическое колебание с почти постоянным периодом (например, годовой или суточный циклы) можно рассматривать как совокупность (смесь) двух процессов: квазигармонического колебания и белого шума. Как указывалось выше, для гармонического колебания, представляющего чисто детерминированный процесс, все основные параметры колебания (амплитуда, пе-

риод, фаза) остаются строго постоянными во времени. Поэтому автокорреляционная функция такого процесса полностью с ним совпадает и имеет вид гармонического косинусоидального незатухающего колебания, изменяющегося в пределах от $r(\tau) = -1$ до $r(\tau) = 1$ с периодом τ_0 (рис. 10.12). Строго говоря, в чистом виде гармонические колебания в природе не встречаются. Наиболее близкими к ним являются приливы.



а



б

Рис. 10.11. График автокорреляционной функции циклического колебания.
а – среднемесячные значения ТПО в районе судна погоды «М»,
б – средние годовые значения солнечной активности (чисел Вольфа) за 1750–1996 гг.

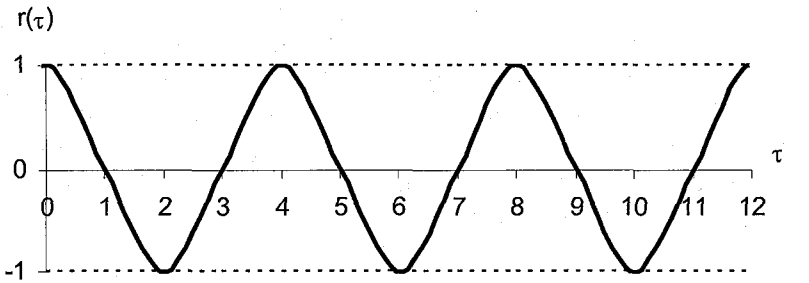


Рис. 10.12. График автокорреляционной функции гармонического колебания.

Кроме того, довольно часто в межгодовой изменчивости реальных гидрометеорологических характеристик отмечается наложение случайной изменчивости в виде белого шума на трендовую компоненту. Автокорреляционная функция такого процесса имеет значительный радиус корреляции, который тем больше, чем выше инерционность исходного процесса. После первого пересечения через нуль коэффициенты автокорреляции становятся незначимыми вплоть до точки отсечения АФ (рис. 10.13). Радиус корреляции, как видно из рис. 10.13, равен 5 годам.

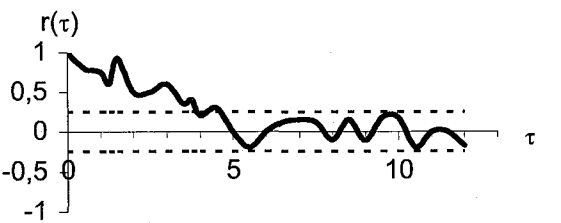


Рис. 10.13. График автокорреляционной функции случайного процесса, состоящего из совокупности белого шума и трендовой компоненты.

Пример 10.5. Проанализируем автокорреляционную функцию вертикального градиента температуры в термоклине по годовому ряду среднесуточных наблюдений в пункте, находящемся в районе станции погоды «Танго» *T* ($\varphi = 29^\circ$ с.ш., $\lambda = 135^\circ$ в.д.), рассчитанную Григоркиной, Губером и Фуксом. Исходный ряд центрировался относительно среднегодового значения. Затем вычислялись значения АФ до $\tau = 60$ сут. (рис. 10.14). Одновременно были рассчитаны также нижняя и верхняя доверительные границы.

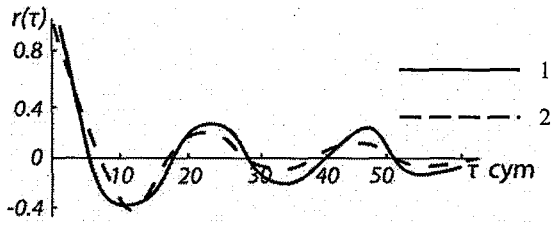


Рис. 10.14. Автокорреляционная функция колебаний вертикального градиента температуры в термоклине (по Григоркиной, Губеру, Фуксу).
1 – АФ после исключения низкочастотной составляющей;
2 – аппроксимация АФ зависимостью (10.28).

Как видно из рис. 10.14, график АФ представляет собой затухающую косинусоиду. Это свидетельствует о том, что в исходном процессе присутствует квазипериодическая (циклическая) составляющая, период которой, определенный по коррелограмме, близок к 20–22 сут. Для аппроксимации АФ может быть использована формула вида (10.28). После нахождения численных значений параметров α ($\alpha \approx 0{,}07$ 1/сут.) и β ($\beta \approx 0{,}28$ рад/сут.) имеем:

$$r(\tau) = e^{-0{,}07|\tau|}\cos 0{,}28\tau.$$

Сравнение экспериментальной и аппроксимированной кривых АФ показало, что между ними наблюдается удовлетворительное соответствие. Это значительно облегчает задачу определения интервалов корреляции. Использовав найденные параметры α и β, нетрудно получить $\tau_{кор}^{(1)} \approx 20$ сут., $\tau_{кор}^{(2)} \approx 4$ сут.

Заметим, что интервал $\tau_{кор}^{(1)}$, определенный по числу пересечений реализацией процесса нулевого уровня, оказался равным 22 сут.

Таким образом, линейная зависимость вертикального градиента температуры от предшествующей ситуации в данном пункте достаточно велика на интервале, не превышающем 4 сут. Через 20–22 сут. значения процесса практически не коррелированны.

10.6. Понятие о взаимнокорреляционной функции

При описании гидрометеорологических процессов часто приходится иметь дело не с одной, а с двумя или несколькими функциями времени. В этом случае возникает задача определения статистической связи между различными случайными функциями.

Мерой линейной взаимосвязи двух временных реализаций $X(t)$ и $Y(t)$ является нормированная взаимнокорреляционная функция (ВКФ), которая применительно к непрерывным процессам бесконечной продолжительности определяется как

$$r_{xy}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} X(t)Y(t+\tau)dt. \qquad (10.34)$$

Функция $r_{xy}(\tau)$ характеризует степень коррелированности (взаимосвязности) ординаты реализации $X(t)$, взятой в момент времени t_1, и ординаты реализации $Y(t)$, взятой в момент времени t_2, причем $t_2 - t_1 = \tau$. Величина $r_{xy}(\tau)$ – всегда действительная функция, которая может принимать как положительные, так и отрицательные значения. В отличие от АФ она не обязательно имеет максимум при $\tau = 0$ и не обязательно является четной функцией.

В общем случае ВКФ несимметрична, т.е.

$$r_{xy}(\tau) \neq r_{xy}(-\tau),$$

поэтому вычисление $r_{xy}(\tau)$ осуществляется как для положительных, так и для отрицательных значений аргумента. Но объем вычислений может быть сокращен на основе следующего свойства ВКФ:

$$\left.\begin{aligned} r_{xy}(\tau) &= r_{yx}(-\tau) \\ r_{yx}(\tau) &= r_{xy}(-\tau) \end{aligned}\right\}.$$

Функция $r_{xy}(\tau)$ характеризует степень зависимости при опережении $X(t)$ относительно $Y(t)$; функция $r_{yx}(\tau)$ – при запаздывании $X(t)$ по отношению к $Y(t)$.

При переходе от (10.34) к дискретным процессам большой продолжительности формулы для вычисления ВКФ приобретают вид:

$$r_{xy}(\tau) = \frac{\sum^{N-\tau} (x_i - \bar{x})(y_{i+\tau} - \bar{y})}{(N-\tau)\sigma_x\sigma_y}, \qquad (10.35)$$

$$r_{yx}(\tau) = \frac{\sum^{N-\tau} (y_i - \bar{y})(x_{i-\tau} - \bar{x})}{(N-\tau)\sigma_x\sigma_y}. \qquad (10.36)$$

Естественно, что при небольших значениях N на точности вычисления ВКФ начинают уже сказываться изменения статистиче-

ских характеристик: средней арифметической и стандартного отклонения. Учет их изменений производится аналогично (10.32).

Нормированная взаимная корреляционная функция связана с взаимной корреляционной функцией $R_{xy}(\tau)$ следующим образом:

$$r_{xy}(\tau)=\frac{R_{xy}(\tau)}{\sigma_x(0)\sigma_y(0)}, \qquad r_{yx}(\tau)=\frac{R_{yx}(\tau)}{\sigma_x(0)\sigma_y(0)}, \qquad (10.37)$$

где произведение $\sigma_x(0)\sigma_y(0)$ при нулевом сдвиге представляет *взаимную дисперсию* переменных X и Y.

Приведем два неравенства, которые показывают, как абсолютные значения взаимной корреляционной функции связаны со значениями этих функций при нулевом сдвиге:

$$|R_{xy}(\tau)|^2 \le R_x(0)R_y(0)$$

$$|R_{xy}(\tau)| \le 0{,}5[R_x(0)+R_y(0)].$$

Отсюда следует, что если $R_{xy}(\tau)=0$, то функции $X(t)$ и $Y(t)$ являются некоррелированными.

Перечислим свойства взаимной корреляционной функции.

Свойство 1. При одновременной перестановке индексов и аргументов взаимная корреляционная функция не изменяется:

$$R_{xy}(t_1, t_2) = R_{yx}(t_2, t_1).$$

Свойство 2. Прибавление к случайным функциям $X(t)$ и $Y(t)$ неслучайных слагаемых $\varphi(t)$ и $\psi(t)$, не изменяет их взаимной корреляционной функции. Так, если

$$X_1(t) = X(t) + \varphi(t) \text{ и } Y_1(t) = Y(t) + \psi(t),$$

то

$$R_{x1y1}(t_1, t_2) = R_{xy}(t_1, t_2).$$

Свойство 3. При умножении случайных функций $X(t)$ и $Y(t)$ на неслучайные множители $\varphi(t)$ и $\psi(t)$, взаимная корреляционная функция умножается на произведение $\varphi(t_1)\psi(t_2)$. Так, если

$$X_1(t) = X(t)\varphi(t) \text{ и } Y_1(t) = Y(t)\psi(t),$$

то

$$R_{x1y1}(t_1, t_2) = R_{xy}(t_1, t_2)\varphi(t_1)\psi(t_2).$$

Свойство 4. Абсолютная величина взаимной корреляционной функции двух случайных функций не превышает среднего геометрического их дисперсий:

$$|R_{xy}(t_1, t_2)| \le [D_x(t_1)D_y(t_2)]^{1/2}.$$

Следствие. Абсолютная величина нормированной взаимной корреляционной функции двух случайных функций не превышает единицы:

$$|r_{xy}(t_1, t_2)| \le 1.$$

Взаимнокорреляционная функция (ВКФ) является более информативной по сравнению с АФ в том смысле, что она дает возможность получить также разность фаз процессов $X(t)$ и $Y(t)$. Временной сдвиг τ_i, соответствующий максимуму функции взаимной корреляции, определяет среднюю разность фаз исследуемых процессов. Этот сдвиг называют иногда *оптимальным сдвигом*.

Симметрия ВКФ относительно нулевого сдвига $\tau = 0$, т.е. максимум ВКФ при $\tau = 0$ означает, что процессы протекают синфазно. Асимметрия ВКФ (максимум при $\tau \neq 0$) свидетельствует о том, что процессы протекают с некоторой разностью фаз, соответствующей τ_i. Кроме того, если исследуемые ряды содержат основное колебание с одной и той же частотой, то их функция взаимной корреляции содержит основное колебание с той же частотой.

По абсолютной величине ВКФ судят о степени взаимосвязи процессов, а по ее знаку – об их прямой или обратной зависимости. Так, если два процесса протекают в известной степени однородно, то независимо от того, обусловлена ли эта однородность взаимным влиянием $X(t)$ и $Y(t)$ или зависимостью обоих процессов от некоторого третьего, в любом случае будут иметь место одинаково направленные отклонения от средней величины.

Взаимнокорреляционный анализ широко используется при решении многих задач, связанных с изучением взаимосвязей океанологических процессов во временной области. К их числу относятся:

1. Определение величины и знака взаимосвязи двух стационарных процессов.

2. Определение средней разности фаз по сдвигу.

3. Установление временных масштабов, в которых статистическая связь процессов является наиболее сильной.

4. Выделение периодических составляющих, присущих обоим статистическим рядам.

5. Выяснение возможной заблаговременности прогноза и выбор оптимальных предикторов.

Пример 10.6. Как известно, изменение в развитии многих атмосферных процессов происходит раньше, чем в океане. В частности, это относится к тепловым процессам. Действительно, вследствие инерционности океана наступление экстремумов сезонного хода в компонентах притока солнечной радиации к океану опережает годовой ход температуры поверхностного слоя воды в зависимости от географических условий на 1–3 месяца. В качестве конкретного примера обратимся к району Канарского апвеллинга, расположенному вдоль берега Африки между широтами 15 и 25° с.ш. и долготой 20° з.д. Его замечательной особенностью является то, что он обладает исключительно высокой биологической и промысловой продуктивностью вод и поэтому служит областью интенсивного рыбного промысла.

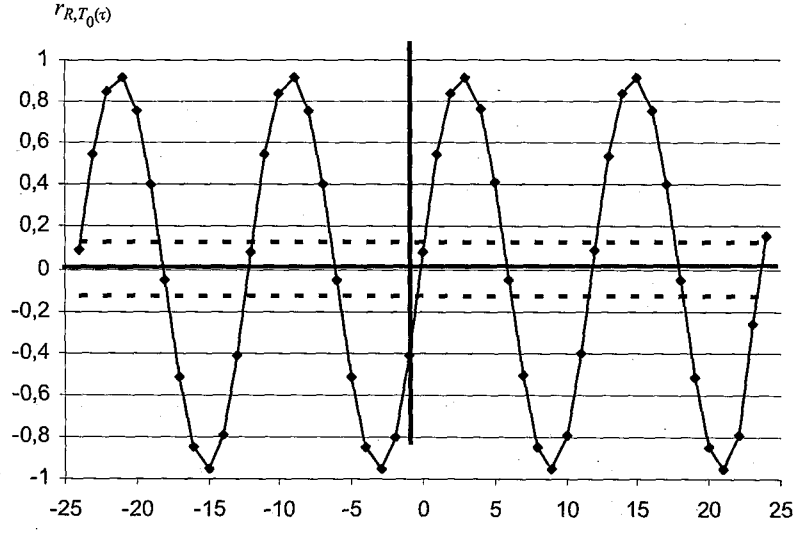


Рис. 10.15. Функция взаимной корреляции между среднемесячными значениями радиационного баланса R и температурой поверхности океана T_0 за период 1981–2000 гг. ($n = 240$) в районе Канарского апвеллинга до сдвига $\tau_{max} = 25$ мес.

Для среднемесячных значений результирующего радиационного притока тепла (радиационного баланса R) и температуры поверхности океана (T_0) за период 1981–2000 гг. ($n = 240$) рассчитана нормированная взаимнокорреляционная функция до максимального сдвига $\tau_{max} = 25$ (рис. 10.15). Нетрудно видеть, что указанные процессы протекают асинфазно. Действительно, значения R опережают значения T_0 на 3 мес. Коэффициент взаимной корреляции при $\tau = 3$ мес. равен $r = 0{,}92$. Таким образом, сдвиг 3 мес. – это период адаптации температуры поверхности океана, её естественная реакция на климатический сигнал, внешнее возмущение. Кроме того, отметим, что через 9 мес. связь возобновляется, но уже с противоположным знаком. Естественно, поскольку оба случайных процесса содержат колебание с одной и той же частотой, то их функция взаимной корреляции также содержит основное колебание с той же частотой. В данном случае это годовой период, равный 12 мес.

Выше отмечалось, что взаимнокорреляционный анализ может быть использован для прогноза гидрометеорологических характеристик. В данном случае сдвиг 3 мес. представляет собой естественную (природную) заблаговременность прогноза T_0, что может быть положено в основу построения физико-статистической модели прогноза. Кроме того, другим важным прогностическим сдвигом является полупериод (9 месяцев) между сезонными изменениями T_0 и R, причем при сдвиге $\tau = 9$ мес. коэффициент взаимной корреляции даже несколько выше ($r = -0{,}95$).

10.7. Авторегрессионные модели временных рядов

При моделировании и прогнозе стационарных случайных процессов, особенно при наличии в них значительной инерционности, целесообразно использовать так называемые параметрические модели, которые включают малое число параметров, зависящих исключительно от предшествующих значений процесса. Наибольшее распространение получили *модели авторегрессии* (АР), *модели скользящего среднего* (СС) и *смешанные модели авторегрессии – скользящего среднего* (АРСС).

В модели АР текущее значение случайного процесса выражается как конечная линейная комбинация предыдущих его значений и случайного импульса. Пусть $X(t)$, $X(t-1)$, $X(t-2)$,... есть значения стационарного случайного процесса $X(t)$ в равноотстоящие моменты $t, t-1, t-2, \ldots$, а $Z(t)$ – белый шум. Обозначим через $X^{\circ}(t) = X(t) - m_x$ центрированный случайный процесс. Тогда выражение

$$X^{\circ}(t) = \alpha_1 X^{\circ}(t-1) + \alpha_2 X^{\circ}(t-2) + \ldots + \alpha_p X^{\circ}(t-p) + Z(t) \quad (10.38)$$

называется процессом авторегрессии порядка p.

Данная модель содержит $p + 2$ параметров: m_x; $\alpha_1, \alpha_2, .., \alpha_p$; σ^2_z, где σ^2_z – дисперсия белого шума $Z(t)$. Нетрудно видеть, что при $p = 1$ мы получаем марковский процесс первого порядка или другими словами «красный шум». При $p = 0$ модель АР вырождается в «белый шум».

В модели СС текущее значение случайного процесса $X(t)$ выражается через предыдущие значения $Z(t-1)$, $Z(t-2)$,... белого шума $Z(t)$. Тогда выражение

$$X^{\circ}(t) = Z(t) + \beta_1 Z(t-1) + \beta_2 Z(t-2) + \ldots + \beta_q Z(t-q) \quad (10.39)$$

называется процессом скользящего среднего порядка q. Данная модель содержит $q + 2$ параметров: m_x; $\beta_1, \beta_2, \ldots, \beta_q$; σ^2_z.

Наконец, в тех случаях, когда использование этих моделей не приводит к желаемой точности в описании случайного процесса, целесообразно объединить в одной модели авторегрессию и скользящее среднее. В результате приходим к комбинированной модели АРСС, т.е.

$$X^{\circ}(t) = \alpha_1 X^{\circ}(t-1) + \ldots + \alpha_p X^{\circ}(t-p) + Z(t) + \beta_1 Z(t-1) + \ldots + \beta_q Z(t-q). \quad (10.40)$$

Эта модель содержит уже $p + q + 2$ параметров: m_x; $\alpha_1, \alpha_2, \ldots, \alpha_p$; $\beta_1, \beta_2, \ldots, \beta_q$; σ^2_z.

Заметим, что данная модель имеет физический смысл, поскольку она является дискретным аналогом линейного дифференциального уравнения, используемого для описания линейных систем. Отсюда следует, что эта модель представляет собой временной ряд в виде выходного сигнала линейной системы, на вход которой подается белый шум.

Моделям (10.38)–(10.40) свойственна сравнительно высокая мощность и экономичность, которая заключается в том, что имеющиеся наблюдения расходуются на оценивание сравнительно малого числа параметров. Однако они требуют более полной априорной информации о физической сущности рассматриваемых процессов.

Построение параметрической модели включает в себя три этапа:

1) идентификация моделей,

2) оценивание параметров моделей,

3) проверка точности построенных моделей.

Рассмотрим вкратце каждый из этапов.

Идентификация моделей. Суть первого этапа состоит в выборе типа модели, который заключается в том, чтобы из общего класса линейных моделей отобрать одну из трех моделей (АР, СС или АРСС), наилучшим образом описывающую исходный случайный стационарный процесс. Основой метода идентификации является сравнение теоретической автокорреляционной функции с эмпирической, полученной в результате обработки натурных данных. Действительно, если теоретические автокорреляционные функции для рассматриваемых нами моделей известны, то в результате сравнения их с выборочными автокорреляционными функциями можно выбрать наилучшую модель. Относительно просто это сделать для моделей первого и второго порядка. Так, автокорреляционные функции для данных моделей первого порядка имеют вид:

– модель авторегрессии:

$$R(\tau) = \sigma^2 \alpha_1^{\tau}, \tag{10.41}$$

– модель скользящего среднего:

$$R(\tau) = \sigma_z^2 \sum_{i=0}^{q-\tau} \beta_i \beta_{j+\tau}, \tag{10.42}$$

– модель авторегрессии – скользящего среднего:

$$R(\tau) = \frac{\sigma_z^2 (1-\alpha_1\beta_1)(\alpha_1-\beta_1)}{1-\alpha_1^2}. \tag{10.43}$$

Из формулы (10.42) видно, что автокорреляционная функция процесса скользящего среднего обрывается при $\tau = q$. Заметим, что на практике довольно редко используются модели более высокого порядка, чем второй.

Оценивание параметров моделей. На данном этапе решаются две следующие задачи: нахождение оптимального порядка выбранной модели и оценка неизвестных коэффициентов модели.

Заметим, что универсального метода нахождения порядка модели не существует. Считается, что наиболее точным методом оценивания параметров является метод максимального правдоподобия. Вследствие его сложности рассмотрим более простые методы. На наш взгляд, вполне удовлетворительные результаты могут быть достигнуты при минимизации дисперсии остатков ($D_{\varepsilon M}$) выбранной параметрической модели, т.е.

$$D_{\varepsilon M} = \rightarrow \min.$$

Нахождение экстремума $D_{\varepsilon M}$ может быть осуществлено пошаговым путем, т.е. последовательно выполняется их расчет, начиная с модели первого порядка.

Кроме того, для оценки порядка модели могут быть использованы те или иные критерии. К ним относятся критерии Акаике, Парзена, Шварца–Риссанена и др. Приведем в качестве примера критерий Акаике (FRE)

$$FPE(M) = \frac{N+M+1}{N-M-1}\left(\frac{P_M}{2(N-M)}\right).$$

Следует иметь в виду, что указанные критерии применимы преимущественно в идеализированных условиях, т.е. для очень длинных выборок. К тому же полученные с их помощью оценки порядка модели могут противоречить друг другу.

Проверка точности построенных моделей. После оценки всех параметров выбранной модели необходимо проверить, насколько хорошо она согласуется с данными наблюдений, т.е., по существу, оценить ее адекватность. С этой целью может быть использован критерий Фишера точно таким же образом, как при проверке адекватности, например регрессионных моделей (см. п. 7.4). Кроме того, осуществляется анализ остатков построенной модели по схеме, рассмотренной в п. 7.5.

10.8. Понятие о цепях Маркова

Рассмотрим случайные процессы с дискретным временем и конечным множеством состояний. При этом под *состоянием* будем иметь в виду определенные классы (градации, типы, группы и т.п.), на которые могут быть подразделены гидрометеорологические процессы или явления. Так, многие характеристики нетрудно разбить на 3 градации: норма, выше нормы, ниже нормы. Легко выделяется также естественное множество состояний, такие как типы атмосферной циркуляции, типы ледового режима и т.д.

Последовательность событий $A_i^{(t)}$ *называют цепью Маркова k-го 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* $t-k+1$, т.е.

$$P\left\{A_j^{(t+1)}\middle|A_i^{(t)},A_{i_1}^{(t-1)},A_{i_2}^{(t-2)},\ldots\right\}=P\left\{A_j^{(t+1)}\middle|A_i^{(t)},\ldots,A_{i_{k-1}}^{(t-k+1)}\right\}. \qquad (10.44)$$

Отсюда следует, что для цепи Маркова первого порядка, называемой *простой марковской цепью*, для каждого момента времени t справедливо равенство:

$$P\left\{A_j^{(t+1)}\middle|A_i^{(t)},A_{i_1}^{(t-1)},A_{i_2}^{(t-2)},\ldots\right\}=P\left\{A_j^{(t+1)}\middle|A_i^{(t)}\right\}. \qquad (10.45)$$

Это означает, что для *простой цепи Маркова* 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. Для такой случайной последовательности имеется автокорреляционная связь только между соседними членами ряда. Если вероятность системы в настоящий момент зависит от некоторого множества состояний системы в предшествующие моменты времени, то имеем *сложную цепь Маркова*. Это означает, что существует автокорреляционная связь между различными сечениями рассматриваемого процесса, начиная со сдвига $\tau = 1$.

В общем случае, свойство случайного процесса зависеть только от ближайшего прошлого называют *марковским*. Простейшим объектом, обладающим таким свойством, является *однородная*

простая цепь Маркова, у которой условная вероятность не зависит от времени:

$$P\left\{A_j^{(t+1)}\middle|A_i^{(t)}\right\} = P\left\{A_j^{(t+1+\Delta t)}\middle|A_i^{(t+\Delta t)}\right\}. \qquad (10.46)$$

Другими словами, однородной называют такую цепь Маркова, когда условная вероятность перехода из состояния i в состояние j не зависит от номера испытания. Эту условную вероятность перехода за один шаг из i-го состояния в j-е обозначают P_{ij}. Переходные вероятности P_{ij} образуют матрицу перехода системы за один шаг:

$$\pi_1 = \begin{pmatrix} P_{11}, & P_{12}, & ..., & P_{1k} \\ P_{21}, & P_{22}, & ..., & P_{1k} \\ ... & ... & ..., & ... \\ P_{k1}, & P_{k2}, & ..., & P_{kk} \end{pmatrix}. \qquad (10.47)$$

Отметим, что в обозначении переходной вероятности P_{ij} первый индекс указывает номер предшествующего, а второй – номер последующего состояния. Тогда P_{11} означает вероятность «перехода» из первого состояния в первое, а P_{12} – вероятность перехода из первого состояния во второе.

Можно показать, что матрица перехода за два шага из i-го состояния в j-е равна произведению двух матриц π_1 перехода за один шаг, т.е.

$$\pi_2 = \pi_1 \cdot \pi_1 = \pi_1^2. \qquad (10.48)$$

Тогда при переходе за n шагов имеем следующую матрицу перехода:

$$\pi_n = \pi_1^n. \qquad (10.49)$$

Итак, зная матрицу перехода за один шаг, нетрудно найти вероятности перехода между состояниями за любое количество шагов.

Матрицы перехода обладают тем свойством, что все их элементы неотрицательны, а суммы вероятностей по любой строке равны единице, т.е.

$$\sum_{j=1}^{k} P_{ij} = 1 \qquad (i = 1,2,\ldots,k).$$

Иногда матрицы, обладающие таким свойством, называются *стохастическими*.

Зная матрицу перехода за n шагов и начальное положение A_i, можно найти вероятность на шаге n в состоянии A_j. Обозначим через $\mathbf{e}_i$ вектор-строку, состоящую из нулей и единиц, которая находится на i-м месте. Тогда для каждого i можно получить в матричной форме следующее соотношение:

$$e_i \pi_{(n-1)} \pi_1 = e_i \pi_n = e_i \pi_1^i, \qquad (10.50)$$

которое определяет распределение цепи Маркова по состояниям A_j через n шагов после выхода из состояния A_i. При этом если для некоторого n все элементы матрицы $\pi_1^{\,i}$ положительны, то вероятность находиться в состоянии A_j, для цепи Маркова при $n \to \infty$ не зависит от начального состояния A_i и удовлетворяет уравнению $\mathbf{p} = \mathbf{p}\,\pi_1$, где $\mathbf{p}$ – вектор-строка с неотрицательными элементами $p_j\left(\sum_{j=1}^{k} p_j = 1\right)$.

Вектор $\mathbf{p}$ называют *предельным распределением*, смысл которого состоит в следующем. При $n \to \infty$ цепь Маркова входит в устойчивый режим, характеризующийся следующими свойствами:

1) среднее время пребывания в состоянии A_j равно $\mathbf{p}_j T$, где T – достаточно длительный промежуток времени;

2) среднее время возвращения в состояние A_j равно $1/\mathbf{p}_j$.

Пример 10.7. В зоне переменного увлажнения межгодовая изменчивость речного стока практически полностью определяется количеством выпавших осадков. Поскольку чередование засушливых и дождливых лет обычно крайне нерегулярно и не поддается оценке стандартными статистическими методами, то воспользуемся для этой цели построением цепи Маркова. Прежде всего была выбрана одна из рек в бассейне Дона, межгодовую изменчивость стока которой было решено разделить на четыре устойчивых состояния: экстремально низкий сток, сток ниже нормы, сток выше нормы, экстремально высокий сток. Из анализа экспериментальных данных было установлено, что за первой градацией (экстремально низкий сток) никогда не следует четвертая (экстремально высокий сток), а за четвертой – первая. Все остальные переходы возможны, причем:

– из первой градации можно попасть в каждую из средних градаций вдвое чаще, чем опять в первую. Следовательно, вероятность переходов из первой градации составляет $P_{11} = 0{,}2$, $P_{12} = 0{,}4$, $P_{13} = 0{,}4$, $P_{14} = 0$;

– из четвертой градации переходы во вторую и третью бывают в четыре и пять раз чаще, чем возвращение в четвертую градацию, поэтому $P_{41} = 0$, $P_{42} = 0{,}4$, $P_{43} = 0{,}5$, $P_{44} = 0{,}1$;

– из второй градации переход в другие градации может быть только реже: в первую – в два раза, в третью – на 25 %, в четвертую – в четыре раза, чем переход во вторую. В результате имеем $P_{21} = 0{,}2$, $P_{22} = 0{,}4$, $P_{23} = 0{,}3$, $P_{24} = 0{,}1$;

– из третьей градации переход во вторую столь же вероятен, как и возвращение в третью, а переходы в первую и четвертую градации случаются в четыре раза реже, поэтому $P_{31} = 0{,}1$, $P_{32} = 0{,}4$, $P_{33} = 0{,}4$, $P_{34} = 0{,}1$.

Таким образом, матрица вероятностей переходов для речного стока имеет вид:

$$\pi_1 = \begin{pmatrix} 0{,}2 & 0{,}4 & 0{,}4 & 0 \\ 0{,}2 & 0{,}4 & 0{,}3 & 0{,}1 \\ 0{,}1 & 0{,}4 & 0{,}4 & 0{,}1 \\ 0 & 0{,}4 & 0{,}5 & 0{,}1 \end{pmatrix}. \qquad (10.51)$$

Действительно, как мы видим из формулы (10.51), сумма вероятностей события в каждой строке равна 1. Найдем теперь среднее время между засухами, определяющими экстремально низкий сток, и избыточным увлажнением, формирующим экстремально высокий сток. Для этого определим предельное распределение для цепи Маркова с матрицей вероятностей переходов (10.51) . Очень засушливые и дождливые годы, как правило, не повторяются, поэтому $P_{14} = P_{41} = 0$. Нетрудно проверить, что все элементы матрицы $\boldsymbol{\pi}_1^2$ положительны.

Предельные вероятности засушливых и дождливых лет составляют 0,15 и 0,08. Учитывая, что периодичность возвращения системы в состояние A_j равна $1/\mathbf{p}_j$, получаем, что периодичность засушливых лет в среднем составляет 6–7 лет, а дождливых – 12–13 лет.

Глава 11. СПЕКТРАЛЬНЫЙ АНАЛИЗ

11.1. Понятие о спектральной плотности

Понятие спектра нашло широкое применение в различных областях науки, особенно в физике и радиотехнике. Так, например, оптический спектр показывает вклад волн различной частоты или длины в энергию источника света. В гидрометеорологии понятие спектра широко используется для выделения скрытых периодичностей временного ряда, для исследования закономерностей его частотной структуры, при моделировании и прогнозе стационарных процессов.

Спектральное представление случайного стационарного процесса является обобщением гармонического анализа периодических функций на случайные процессы. Большой вклад в развитие этого метода внесли Винер, Колмогоров, Хинчин, Яглом и другие исследователи.

Как было показано А.Я. Хинчиным, если применять преобразование Фурье не к реализации случайного процесса, а к его автокорреляционной функции, которая является строго затухающей, то получается математически корректное выражение для пары преобразований Фурье:

$$S_x(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} R(\tau) e^{-i\omega\tau} d\tau, \qquad (11.1)$$

$$R_x(\tau) = \int_{-\infty}^{\infty} S(\omega) e^{i\omega\tau} d\omega \ , \qquad (11.2)$$

где $S_x(\omega)$ – *спектральная плотность* (энергетический спектр, плотность спектра мощности и т. п.); $e^{-i\omega\tau}$ – показательная функция мнимого аргумента.

Формула (11.1) – прямое преобразование Фурье, формула (11.2) – обратное преобразование Фурье.

Физический смысл $S(\omega)$ состоит в том, что она означает *плотность дисперсии стационарной случайной функции на данной частоте*, т. е.

$$S(\omega_k) = \sigma^2/\omega_k.$$

Сказанное можно пояснить следующим образом. Если исследуемый процесс представляет собой гармоническую функцию времени с частотой ω и амплитудой A, то в этом случае дисперсия процесса равна $A^2/2$. В более общем случае, когда процесс представляет собой совокупность (смесь) нескольких гармоник с частотами ω_i и амплитудами A_i, дисперсия такого процесса запишется выражением:

$$\sigma^2 = \frac{1}{2}\sum_{i=1}^{n} A_i^2 , \qquad (11.3)$$

где n – число гармоник.

Такое разложение общей дисперсии процесса на ее отдельные составляющие, соответствующие выделенным частотам, и оценка их значимости представляют собой сущность спектрального анализа.

Полагая в формуле (11.2) $\tau = 0$, получаем выражение для дисперсии случайной функции:

$$D_x = R_x(0) = \int_{-\infty}^{\infty} S_x(\omega) d\omega. \qquad (11.4)$$

Довольно часто вместо спектральной плотности в расчетах используется *нормированная спектральная плотность*, отражающая распределение плотности дисперсии по частотам в долях единицы, т.е.

$$s_x(\omega) = \frac{S_x(\omega)}{\int_{-\infty}^{\infty} S_x(\omega) d\omega} = \frac{S_x(\omega)}{D_x}$$

или

$$s_x(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} r_x(\tau) e^{-i\omega\tau} d\tau . \qquad (11.5)$$

Естественно, нормированная корреляционная функция и нормированная спектральная плотность также являются взаимными преобразованиями Фурье. Заметим, что спектральная плотность – величина размерная, ее размерность соответствует квадрату случайной величины. В то же время нормированная спектральная плотность уже представляет безразмерную величину.

Основные свойства спектральной плотности (СП):

Свойство 1. СП – четная функция, т.е. $S_x(\omega) = S_x(-\omega)$.

Свойство 2. СП является неотрицательной величиной, т.е. $S_x(\omega) \geq 0$.

Использовав формулу Эйлера $e^{-i\omega\tau} = \cos\omega\tau - i\sin\omega\tau$ и учтя четность $S_x(\omega)$ и $R_x(\tau)$, для вещественного случайного процесса можно записать:

$$R_x(\tau) = 2\int_0^\infty S_x(\omega)\cos\omega\tau d\omega; \qquad (11.6)$$

$$S_x(\omega) = \frac{1}{\pi}\int_0^\infty R_x(\tau)\cos\omega\tau d\tau. \qquad (11.7)$$

Аналогичные формулы могут быть записаны для $s_x(\omega)$ и $r_x(\tau)$. Если нанести значения $S_x(\omega)$ на график, ось абсцисс которого представляют частоты ω, то получим кривую спектральной плотности. При этом сам график обычно называется *спектрограммой*. Так как $D_x = R_x(0)$, то дисперсия есть удвоенная площадь, ограниченная кривой СП при $\omega \geq 0$, или площадь, ограниченная кривой СП на всем интервале $(-\infty, \infty)$. Для кривой нормированной спектральной плотности площадь, лежащая под ней на интервале $(-\infty,\infty)$, равна единице.

11.2. Аналитическое оценивание спектральной плотности

Аналитическое вычисление СП возможно, если известно соответствующее истинное выражение автокорреляционной функции (АФ). Рассмотрим спектральные плотности для некоторых простых вариантов задания АФ.

1. Нормированная автокорреляционная функция имеет вид $r(\tau) = e^{-\alpha|\tau|}$ при $\alpha > 0$, где α – *коэффициент затухания*. График этой функции приведен на рис. 11.1, *а* для трех значений $\alpha = 0{,}5$; 1; 3. Нетрудно видеть, что АФ представляет собой монотонно затухающую положительно определенную кривую, убывание которой зависит от величины α. Это означает, что корреляционная связь между сечениями случайной функции $X(t)$ и $X(t + \tau)$ при одном и том же интервале τ уменьшается с ростом α. Интервал корреляции, равный $\tau^{(1)}_{\text{кор}} = \alpha^{-1}$, характеризует скорость затухания

корреляционной связи. Подставив выражение для АФ в формулу (11.5), получим:

$$s(\omega) = \frac{1}{\pi}\frac{\alpha}{\omega^2 + \alpha^2}. \qquad (11.8)$$

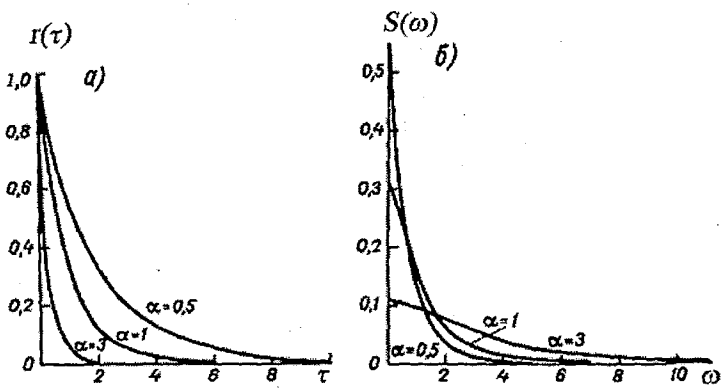


Рис. 11.1. График автокорреляционной функции $r(\tau) = e^{-\alpha|\tau|}$ (*а*) и ее спектральной плотности (*б*).

Спектр вида (11.8) характерен для случайных процессов, интенсивность которых монотонно убывает с увеличением частоты (рис. 11.1, *б*), причем чем меньше α, тем быстрее убывает кривая СП. Это означает, что превалирующее значение в спектре имеют малые частоты. Случайный процесс такого типа называется *узкополосным,* так как его энергия сосредоточена в узкой полосе частот. Для больших значений α СП с увеличением частоты убывает уже весьма медленно. Такой процесс называется *широкополосным.* Максимум спектральной плотности (11.8) находится на частоте $\omega = 0$ и равен $1/\pi\alpha$. При $\alpha > \omega$ кривая спектральной плотности выравнивается и переходит в прямую линию, параллельную оси частот и находящуюся от нее на расстоянии $1/\pi\alpha$.

Стационарный случайный процесс, для которого СП постоянна во всем диапазоне частот $s_x(\omega) = \text{const}$, представляет «белый шум» по аналогии с белым светом, у которого спектральный состав примерно однороден.

2. Нормированная автокорреляционная функция задана по затухающему косинусоидальному типу $r(\tau) = e^{-\alpha|\tau|}\cos\beta\tau$ при $\alpha > 0$,

$\beta > 0$, где β – частота колебания АФ. График этой функции представлен на рис. 11.2, *а* для трех случаев: кривая *I* – $\alpha = 0{,}5$, $\beta = 2$, кривая *II* – $\alpha = 1$, $\beta = 1$, кривая *III* – $\alpha = 2$, $\beta = 0{,}5$. Из данного рисунка видно, что при малой величине отношения α/β (кривая I, $\alpha/\beta = 0{,}25$) график АФ близок к гармоническим колебаниям частоты ω. Если подставить данное выражение АФ в формулу (11.5), то после некоторых преобразований получим:

$$s_x(\omega) = \frac{\alpha(\alpha^2 + \beta^2 + \omega^2)}{\pi\left[(\alpha^2 + \beta^2 + \omega^2)^2 - 4\omega^2\beta^2\right]}. \qquad (11.9)$$

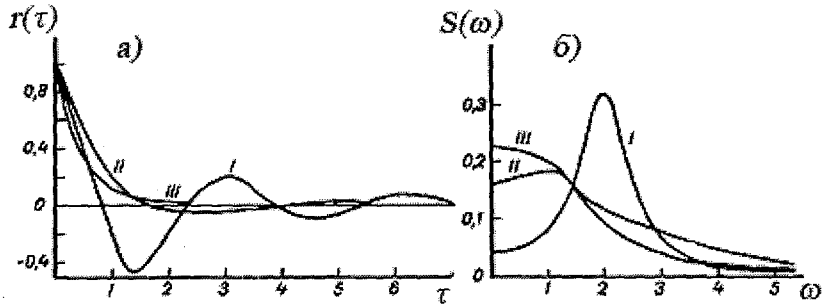


Рис. 11.2. График автокорреляционной функции $r(\tau) = e^{-\alpha|\tau|}\cos\beta\tau$ (*а*) и ее спектральной плотности (*б*).

Спектр вида (11.9) существенно зависит от отношения α/β. При малой величине отношения $\alpha/\beta = 0{,}25$ спектральная плотность имеет ярко выраженный максимум на частоте $\omega = \beta$ (рис. 11.2, *б*). При $\alpha \to 0$ максимум сужается и в предельном случае при $\alpha = 0$ спектр представляет прямую линию бесконечной длины, параллельную оси ординат.

Процесс, характеризуемый таким спектром, является гармоническим колебанием бесконечной длительности. С увеличением α/β ускоряется затухание корреляционной функции, вследствие чего максимум СП становится более размытым. При больших значениях α/β (кривая *III*, $\alpha/\beta = 4$) корреляционная функция отличается от нуля только при малых значениях τ. Поэтому СП с ростом частоты изменяется медленно, оставаясь на значительном диапазоне близкой к начальному значению $s_x(\omega) = 0$.

К сожалению, во многих случаях не представляется возможным получить аналитические оценки СП, поскольку истинные оценки АФ, как правило, либо вообще неизвестны, либо имеют настолько сложный вид, что не поддаются интегрированию.

11.3. Понятие о частотной весовой функции

Поскольку на практике мы имеем дело в основном с выборочными оценками АФ, то соответственно оценки СП также являются выборочными, т.е.

$$S^*_x(\omega) = \frac{1}{\pi}\int_0^T R^*_x(\tau)\cos\omega\tau d\tau, \qquad (11.10)$$

где $S^*_x(\omega)$ – выборочная оценка СП; T – конечная реализация случайного процесса.

Можно показать, что выборочная оценка $S^*_x(\omega)$ является несмещенной, но оказывается несостоятельной, поскольку дисперсия этой оценки не стремится к нулю при стремлении T к бесконечности. Поэтому возникает задача устранения несостоятельности оценок $S^*_x(\omega)$, которая в общем случае не имеет аналитического решения и обычно сводится к каким-либо методам оптимизации. Вследствие этого возникает необходимость в построении оптимальных оценок СП, к которым, например, приводят методы, основанные на предварительном сглаживании оценок АФ.

Проанализируем выборочную функцию $R^*_x(\tau)$, равную истинному значению $R_x(\tau)$ при $|\tau| \le \tau_m$ и равную нулю при $|\tau| > \tau_m$, где τ_m – *точка отсечения* АФ. Другими словами, точка отсечения АФ – это максимальный сдвиг τ, до которого осуществляется расчет АФ. Функцию $R^*_x(\tau)$ можно рассматривать как произведение истинной функции $R_x(\tau)$ на некоторую функцию $\lambda(\tau)$, называемую *частотной весовой функцией* (временным или корреляционным окном):

$$R^*_x(\tau) = \lambda(\tau)R_x(\tau). \qquad (11.11)$$

При этом идеальная частотная весовая функция должна обладать следующим свойством:

$$\lambda(\tau) = \begin{cases} 1 & \text{при} \quad |\tau| \le \tau_m \\ 0 & \text{при} \quad |\tau| > \tau_m. \end{cases} \qquad (11.12)$$

Подставив (11.10) в (11.11), получим:

$$S_x^*(\omega) = \frac{1}{\pi}\int_0^T \lambda(\tau) R_x(\tau)\cos\omega\tau d\tau . \qquad (11.13)$$

Нетрудно показать, что выборочная оценка СП может быть представлена как произведение истинных оценок СП на спектр весовой функции, т.е.

$$S^*_x(\omega) = S_{\lambda(\tau)} S_x(\omega), \qquad (11.14)$$

где спектр функции $\lambda(\tau)$, называемый иногда спектральным окном, определяется по следующей формуле:

$$S_{\lambda(\tau)} = \frac{1}{2\pi}\int_{-\tau_m}^{\tau_m} e^{-i\omega\tau} d\tau = \frac{\sin\omega\tau_m}{\pi\omega} . \qquad (11.15)$$

Итак, смысл функции $\lambda(\tau)$ состоит в том, что с ее помощью осуществляется сглаживание оценки АФ. При этом способ сглаживания определяется только выбором функции $\lambda(\tau)$. В связи с этим возникает задача такого подбора $\lambda(\tau)$, чтобы выборочные оценки СП были бы наиболее близкими к истинным оценкам СП. К сожалению, данная задача осложняется тем, что истинный вид АФ в формуле (11.13) обычно неизвестен. Поэтому на практике выборочную оценку СП приходится осуществлять по выборочной оценке АФ:

$$S_x^*(\omega) = \frac{1}{2\pi}\int_{-\tau_m}^{\tau_m} \lambda(\tau) R_x^*(\tau)\cos\omega\tau d\tau . \qquad (11.16)$$

В формуле (11.16) частотная функция $\lambda(\tau)$ и точка отсечения (среза) АФ τ_m подбираются таким образом, чтобы удовлетворять некоторому выбранному критерию оптимальности. В качестве критерия используется, например, средняя квадратическая ошибка $S^*_x(\omega)$, характеризующая разброс оценок СП около их математического ожидания. Однако осуществить это довольно сложно уже по той причине, что истинные значения СП, как уже упоминалось выше, обычно неизвестны.

Другим осложняющим обстоятельством является то, что формула (11.12) характеризует идеальную частотную функцию. К сожалению, в действительности добиться выполнения этого условия практически не удается. В настоящее время известно значительное число различных видов весовых функций, получивших применение в практических расчетах. В гидрометеорологии, например, до-

вольно часто используются функции Парзена, Бартлетта, Тьюки, Хэмминга.

– функция Бартлетта:

$$\lambda(\tau)=\begin{cases}1-|\tau|/\tau_m & \text{при} \quad |\tau|\le\tau_m \\ 0 & \text{при} \quad |\tau|>\tau_m.\end{cases} \qquad (11.17)$$

– функция Хемминга:

$$\lambda(\tau)=\begin{cases}0{,}54+0{,}46\cos\pi\tau/\tau_m & \text{при} \quad 0\le\tau\le\tau_m \\ 0 & \text{при} \quad \tau>\tau_m.\end{cases} \qquad (11.18)$$

– функция Парзена:

$$\lambda(\tau)=\begin{cases}1-\dfrac{6\tau^2}{\tau_m}\left(1-\dfrac{\tau}{\tau_m}\right) & \text{при} \quad 0\le\tau<\dfrac{\tau_m}{2} \\ 2\left(1-\dfrac{\tau}{\tau_m}\right)^3 & \text{при} \quad \dfrac{\tau_m}{2}\le\tau\le\tau_m \\ 0 & \text{при} \quad \tau>\tau_m.\end{cases} \qquad (11.19)$$

Существует также целый ряд других весовых функций, сведения о которых можно найти в специальной литературе.

Как видно из приведенных формул, значения $\lambda(\tau)$, а следовательно, и оценки СП определяются выбором τ_m. При этом в зависимости от выбора величины τ_m будет происходить смещение оценок СП при малых значениях τ_m и существенное увеличение дисперсии оценок при больших значениях τ_m. Очевидно, при малых значениях τ_m в формуле (11.18) выборочные оценки АФ не очень заметно отличаются от истинных оценок АФ. При $|\tau| > \tau_m$ принимается условие $R^*_x(\tau) = 0$, хотя в действительности АФ может существенно отличаться от нуля. В результате происходит появление систематической ошибки, вызывающей смещение оценки $S^*_x(\omega)$.

Увеличение τ_m приводит к уменьшению этой систематической ошибки, однако при этом одновременно происходит увеличение различий между выборочными и истинными оценками АФ. Вследствие этого повышается дисперсия оценок $S^*_x(\omega)$, что особенно

заметно при уменьшении длины исходного ряда. В результате возникает очевидное противоречие, обусловленное выбором величины τ_m при неизменной длине выборки. Так, с уменьшением τ_m возрастает достоверность (эффективность) оценки спектра, но увеличивается его сглаженность и, следовательно, уменьшается подробность описания исходного процесса. Наоборот, с увеличением τ_m увеличивается количество пиков, т. е. возрастает степень подробности описания исходного процесса, однако при этом уменьшается эффективность оценок спектра. Таким образом, выигрывая в одном, мы всегда проигрываем в другом.

Характерной особенностью практически всех частотных весовых функций является то, что вследствие сглаживания автокорреляционных оценок (весовая функция Хэмминга может приводить даже к появлению отрицательных ординат спектра) происходит искажение оценок спектральной плотности, которое в некоторых случаях может быть весьма существенным. Но поскольку истинные оценки спектра, как правило, неизвестны, то очень трудно определить и возможные погрешности за счет введения весовой функции.

Из результатов численных расчетов следует, что оценки спектральной плотности значительно больше зависят от величины τ_m, чем от вида корреляционного окна. Однако выбор τ_m, вообще говоря, представляет собой не такую уж простую задачу. Естественно, что τ_m должна зависеть от длины реализации случайного процесса. Для получения надежных ошибок $s(\omega)$ обычно принимается $\tau_m = (0{,}1 - 0{,}3)N$. Но поскольку заранее нельзя указать оптимальное значение τ_m, то на практике этот вопрос решается экспериментальным путем, т.е. заданием различных значений τ_m и сравнением полученных спектров. Это позволяет определить, какие особенности спектра являются закономерными, т.е. присущими исследуемому процессу.

11.4. Численное оценивание спектральной плотности

В настоящее время известно довольно много численных методов оценки СП. Рассмотрим здесь только наиболее часто используемые в практических расчетах.

1. Прежде всего, это метод, предложенный Блэкманом и Тьюки (1958), который заключается в последовательном оценивании автокорреляционной функции и преобразовании Фурье достоверной части коррелограммы с какой-либо из весовых функций, обеспечивающей получение статистически значимых оценок СП. Данный метод считается традиционным.

Поскольку в большинстве случаев автокорреляционная функция реальных гидрометеорологических процессов очень трудно поддается точной аппроксимации, то в практических расчетах интеграл (11.3) обычно заменяют суммированием с некоторым верхним пределом m, называемым длиной реализации спектра или *максимальным сдвигом*:

$$s(\omega) = \frac{1}{\pi} \sum_{\tau=1}^{m} r(\tau) \cos(\omega\tau) \Delta\tau . \qquad (11.20)$$

Существует целый ряд численных методов аппроксимации выражения (11.20). Например, официальное признание получил метод, рекомендованный Всемирной метеорологической организацией, согласно которому

$$s(i) = \frac{r_0}{m} + \frac{2}{m} \sum_{\tau=1}^{m-1} \left[r_\tau \cos\left(\frac{360}{2m} i\tau \right) \right] + (-1)^m \frac{r_m}{m}, \qquad (11.21)$$

где i – номер гармоники (i = 1, 2, ..., n); m – максимальный сдвиг; r_0 и r_m – значения $r(\tau)$ при $\tau = 0$ и $\tau = m$.

Значения s_0 и s_m нужно уменьшить в 2 раза. Заметим, что период колебаний и номер гармоники связаны между собой соотношением $\theta = 2m/i$. Отсюда период первой гармоники принимается равным $2m$, второй – m, третий – $2m/3$ и т. д.

При оценках спектральной плотности приходится считаться с тем, что каждая ее ордината рассчитывается со значительной ошибкой. Это обусловлено, с одной стороны, ошибками расчета АФ, в том числе выбором точки ее среза, а с другой – неопределенностью выбора частотной весовой функции, ее ошибками и ошибками численной реализации спектра.

Заметим, что должна существовать взаимосвязь между m и τ_m, причем $m \leq \tau_m$. В численных расчетах обычно принимается $m = \tau_m$. В этом случае, задавая лишь один параметр (более удобным представляется τ_m), нетрудно рассчитать оценки спектральной плотности с использованием любой частотной функции.

Следует иметь в виду, что кривые СП, рассчитанные по рядам наблюдений, обычно имеют много пиков, из которых большая часть является чисто случайной и должна быть исключена из анализа. Поэтому весьма важной является задача выделения достоверных пиков.

В качестве первого приближения при построении доверительных интервалов для значений спектральной плотности используется метод Тьюки. Им было показано, что для случайной выборки, подчиняющейся нормальному закону, распределение выборочных спектральных оценок по отношению к спектру соответствующей генеральной совокупности должно соответствовать распределению χ^2, деленному на число степеней свободы ν. Величина ν определяется как

$$\nu = \frac{2N - m/2}{m}. \qquad (11.22)$$

Таблица 11.1

Доверительные границы χ^2/ν оценок спектральной плотности

Число степеней свободы, ν	5 %	10 %	90 %	95 %
2	2,99	2,30	0,10	0,05
3	2,60	2,08	0,20	0,12
4	2,37	1,94	0,26	0,18
5	2,21	1,85	0,32	0,23
6	2,10	1,77	0,37	0,27
8	1,94	1,68	0,44	0,34
10	1,88	1,60	0,49	0,39
12	1,75	1,53	0,53	0,43
15	1,66	1,48	0,57	0,48
20	1,51	1,42	0,62	0,54
30	1,46	1,34	0,69	0,62
50	1,34	1,26	0,75	0,69
100	1,22	1,18	0,82	0,77
200	1,16	1,14	0,87	0,83

Для нахождения χ^2/ν составлена специальная таблица, входными параметрами для которой являются уровень значимости и число степеней свободы (табл. 11.1). Доверительные интервалы составляются относительно среднего спектра

$$S_0(\omega) = \sum_{\omega=0,01}^{2\pi} \frac{S(\omega)}{N} \qquad (11.23)$$

по следующей формуле:

$$S_0(\omega)-\frac{\chi^2}{\nu}<S(\omega)<S_0(\omega)+\frac{\chi^2}{\nu}. \qquad (11.24)$$

В качестве уровня значимости используют обычно $\alpha = 5$ % и $\alpha = 10$ %. Как показывают результаты расчетов, чем выше оценка СП и чем длиннее исходный ряд, тем более узким становится доверительный интервал.

Кроме того, целесообразно дополнительно на график наносить оценки СП белого и красного шумов. Как известно, белый шум нормированной СП определяется по формуле:

$$s_{\text{бш}}(\omega) = \Delta t/2\pi. \qquad (11.25)$$

Для СП, рассчитанной по $R(\tau)$, необходимо данное выражение умножать на величину дисперсии.

Кривая СП для красного шума вычисляется по формуле:

$$s_{\text{кш}}(\omega)=s_0(\omega)\frac{1-r_1^2}{1+r_1^2-2r_1\cos(\pi/m)}, \qquad (11.26)$$

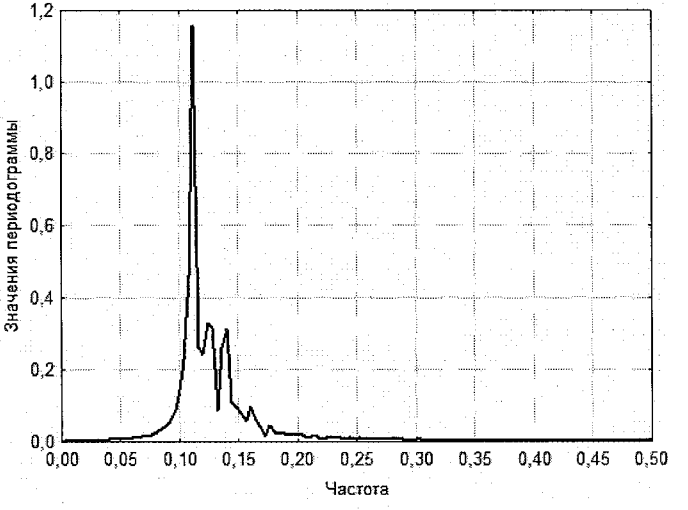
где r_1 – значение корреляционной функции при сдвиге $\tau = 1$.

Как видно из формулы (11.26), кривая СП представляет затухающую экспоненту.

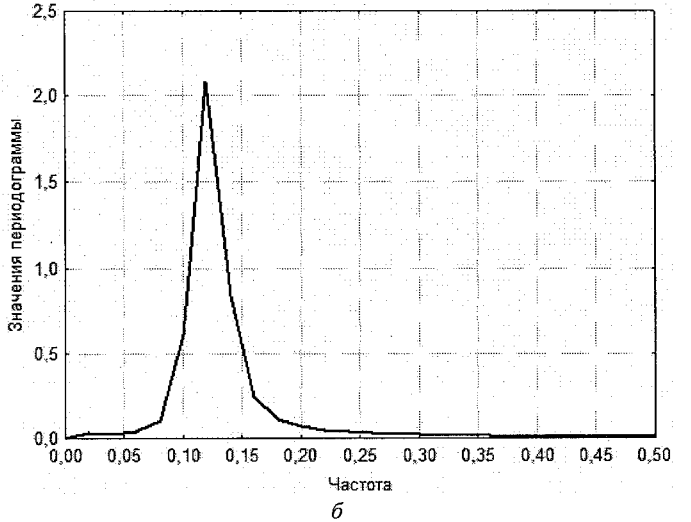
Поведение выборочной кривой СП относительно теоретических оценок позволит дать более полную интерпретацию частотной структуры исследуемого случайного процесса.

На рис. 11.3 представлены графики СП, рассчитанные для вертикальных смещений морского уровня в северо-восточной части Черного моря вблизи г. Геленджик, исходные данные для которого получены с помощью измерений на волномерном буе с дискретностью 0,78 с. Всего длина исходного ряда составляет N = =1500 значений уровня. Оценки СП основаны на двух реализациях АФ: первая – $\tau_m = 250$ (рис. 11.3, *а*), вторая – $\tau_m = 50$ (рис.11, *б*). Как видно из рис. 11.3, *а*, спектрограмма характеризуется ярко выраженным пиком на частоте $\omega = 0{,}112$ цикл/0,78 с, что соответствует периоду $\tau = (1/0{,}112)0{,}78 = 6{,}96$ с, а также еще 2 значительно более мелкими пиками. С уменьшением длины АКФ, как было указано выше, должна возрастать достоверность оценки спектра и увеличиваться его сглаженность. Действительно, из рис. 11.3, *б*

видно, что кривая СП является более сглаженной по сравнению с рис. 11.3, *а*. При этом значение СП основного колебания уровня возросло более чем в полтора раза, но период его почти не изменился ($\tau = 6{,}50$ с).



а



б

Рис. 11.3. Спектральная плотность уровня моря, рассчитанная по волнограмме с дискретностью 0,78 с ($N = 1500$) при различной длине АФ (τ_m): *а* – $\tau_{m=}250$, *б* – $\tau_m = 50$.

2. Вторая вычислительная схема оценивания СП предложена Кули и Тьюки в 1965 г., которая заключается в предварительном вычислении *периодограммы* и последующем ее сглаживании с помощью той или иной весовой функции. Периодограмма рассчитывается непосредственно по исходной реализации эргодического стационарного процесса $X(t)$, заданного на промежутке $[0, T]$ следующим образом:

$$S_x^*(\varpi) = \frac{1}{2\pi T}\left|\int_0^T X(t)\mathrm{e}^{-i\varpi\tau}dt\right|^2. \qquad (11.27)$$

Рассчитанная таким образом периодограмма представляет собой выборочную спектральную плотность, которая получается преобразованием Фурье самой реализации исходного процесса $X(t)$. Доказано, что выборочная оценка $S^*_x(\omega)$ относится к классу несмещенных оценок, но при этом она не является состоятельной, поскольку ее дисперсия не стремится к нулю при стремлении T к бесконечности. При больших значениях T периодограмма носит крайне нерегулярный характер, причем ее значения могут резко меняться даже при малых изменениях длины ряда. Для получения состоятельной оценки периодограммы необходимо произвести ее сглаживание с помощью частотной весовой функции:

$$S_x^*(\omega) = \int_{-\infty}^{\infty} S_1(\omega)Q(\omega-\omega_1)d\omega_1 = S_{\lambda(\tau)}S_x(\omega), \qquad (11.28)$$

где $Q(\omega)$ – спектр частотной функции $\lambda(\tau)$.

Нетрудно видеть, что выборочные оценки СП, рассчитанные по автокорреляционной функции и периодограмме исходного процесса, полностью совпадают. Широкое распространение данной вычислительной схемы стало возможным благодаря предложенному Кули и Тьюки методу дискретного преобразования Фурье, который сводит к минимуму необходимое число операций и обеспечивает наиболее высокую точность. Эта схема, названная *быстрым преобразованием Фурье,* особенно эффективна при анализе очень длинных временных рядов, превышающих несколько сотен или даже тысячу значений. В большинстве ППСП при оценивании СП применяется данный метод.

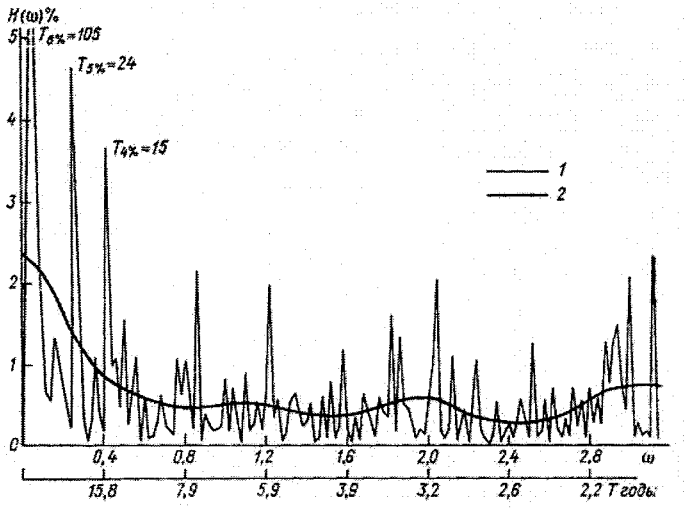


Рис. 11.4. Периодограмма (*1*) и оценки спектральной плотности (*2*) среднегодовых значений температуры воздуха в Центральной Англии за период 1969–1973 гг. по И.И. Поляку.

На рис. 11.4 приводятся периодограмма и оценки спектральной плотности среднегодовых значений температуры воздуха в Центральной Англии за 315 лет (1669–1973 гг.), рассчитанные И.И. Поляком. Оценки периодограммы даны в процентах от оценки дисперсии ряда. Нетрудно видеть, что периодограмма содержит довольно много флуктуаций, имеющих различную мощность на отдельных частотных интервалах. Наиболее мощные из них находятся в низкочастотной части и имеют периоды 105 лет (8 %), 24 года (5 %) и 15 лет (4 %). Вклад в дисперсию остальных флуктуаций еще меньше. Естественно, прежде всего возникает вопрос об устойчивости выделенных флуктуаций во времени. Длина данного временного ряда позволяет осуществить такую проверку. Как оказалось, при уменьшении длины ряда происходит почти полная перестройка внутренней структуры ряда температуры. Только период 15 лет является относительно устойчивым. Поэтому можно заключить, что отдельные всплески на периодограмме носят в основном случайный характер, их появление и исчезновение во времени обусловлены выборочной изменчивостью изучаемых реализаций. Что касается кривой СП, то она имеет весьма сглаженный

вид, обусловленный тем, что используемый интервал предварительной фильтрации временного ряда на основе регрессионного фильтра составлял 35 лет.

На практике вместо оценок СП можно дополнительно воспользоваться *спектральной функцией*, которая представляет собой функцию, характеризующую интегральную долю дисперсии, приходящейся на этот интервал частот, т.е.

$$i_x(\omega) = \int_{-\infty}^{\omega} S_x(\omega')d\omega' = 2\int_{0}^{\omega} S_x(\omega')d\omega'. \qquad (11.29)$$

Отсюда легко перейти к нормированной спектральной функции:

$$i_x(\omega) = \int_{-\infty}^{\omega} s_x(\omega')d\omega' = 2\int_{0}^{\omega} s_x(\omega')d\omega'. \qquad (11.30)$$

Как следует из формулы (11.30)

$$i_x(0) = 0, \quad i_x(\infty) = 2\int_{0}^{\omega} s_x(\omega')d\omega' = 1 \qquad (11.31)$$

Таким образом, каждой гармонике $a_k\cos\omega_k\tau$ разложения Фурье корреляционной функции соответствует скачок величины a_k на графике нормированной спектральной функции. Если величину (11.29) умножить на 100, то она будет представлять долю дисперсии, приходящейся на частоты $|\omega'| \le \omega$, в процентах от общей дисперсии σ^2. Нормированная спектральная функция белого шума имеет вид:

$$i_x(\omega) = \Delta t\omega/\pi \quad (0 \le \omega \le \pi/\Delta t). \qquad (11.32)$$

Если функцию (11.32) нанести на график, то она будет представлять прямую, соединяющую точки (0,0) и ($\pi/\Delta t$, 1). Чем больше оценки периодограммы, наносимые на этот график, отклоняются от теоретической прямой белого шума, тем больше вероятность того, что анализируемая реализация случайного процесса не соответствует модели белого шума. При этом степень отклонения от модели белого шума определяется с помощью критерия значимости Колмогорова–Смирнова. С этой целью вокруг теоретической прямой строится двусторонняя критическая область шириной $\pm\lambda/(n/2-1)^{1/2}$, где величина λ легко может быть определена из табл. 4.3 по числу степеней свободы.

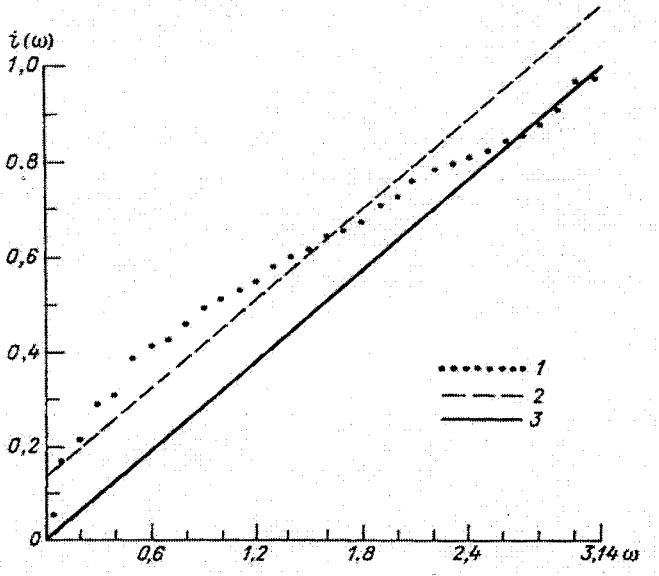


Рис. 11.5. График спектральной функции (*1*) среднегодовых значений температуры воздуха в Центральной Англии за период 1969–1973 гг. и 99 %-ный доверительный интервал (*2*) для белого шума (*3*) по И.И. Поляку.

Из графика спектральной функции, полученной по тем же наблюдениям среднегодовых значений температуры воздуха в Центральной Англии за 315 лет (рис. 11.5), видно, что рассчитанные значения спектра выходят за границы 99 %-ного доверительного интервала для белого шума в области низких частот, т.е. данный ряд температуры не является полностью случайным. Отметим, что анализ других временных рядов температуры воздуха также свидетельствует о наличии в них низкочастотной изменчивости, очевидно, обусловленной трендовой компонентой. Однако основная доля изменчивости температуры обусловлена белым шумом.

3. В последнее время все большее распространение при оценивании СП получает *метод наибольшей энтропии,* предложенный, очевидно, впервые Бургом в 1967 г. Суть метода заключается в том, что между удельной энтропией и автокорреляционной функцией нормально распределенного стационарного случайного процесса, заданного на конечном промежутке, существует функциональная зависимость. Это позволяет получить оценки СП «ме-

тодом наибольшей энтропии» (МНЭ), которые таким образом выражают максимум неопределенности по отношению к отсутствующей информации и вместе с тем соответствуют уже имеющейся информации. Было установлено, что МНЭ-оценки СП функционально связаны с параметрической моделью случайного процесса, заданной в виде модели авторегрессии порядка p следующим образом:

$$S(\omega)_{МНЭ} = \frac{2\sigma_{\alpha}^2}{\left[1 - \sum_{j=1}^{P} \alpha_j \exp(-2\pi\omega ij)\right]^2}. \qquad (11.33)$$

Здесь σ^2_{α} – дисперсия белого шума в модели АР (10.38); α_j – коэффициенты авторегрессии; p – порядок модели АР.

Отсюда следует, что МНЭ-оценка СП сводится к подгонке к исходному временному ряду процесса авторегрессии, определению порядка модели p, ее коэффициентов α_j и дисперсии белого шума в формуле (10.38).

Для расчета коэффициентов авторегрессии используются численные схемы Юла–Уокера, Бурга, Бокса–Дженкинса и др. Так, в соответствии со схемой Юла–Уокера

$$a_{p,p} = \frac{R(p) - \sum_{\tau=1}^{p-1} a_{p-1,\tau} R(p-\tau)}{1 - \sum_{\tau=1}^{p-1} a_{p,\tau} R(\tau)}, \qquad (11.34)$$

где $R(\tau)$ – коэффициенты автокорреляции, рассчитанные по традиционной формуле (10.23).

Первый коэффициент авторегрессии принимается равным выборочному коэффициенту автокорреляции с единичным сдвигом, т.е. $a_{1,1} = R(1)$. Для оценки порядка моделей авторегрессии могут применяться критерии Акаике, Парзена и других, приведенные в п. 10.7.

Достоинством данного метода является возможность оценки СП по сравнительно коротким реализациям. При этом разрешающая способность, т.е. способность разделять близкие по частоте пики СП, превышает разрешающую способность таких методов

оценивания СП, как сглаживание периодограммы и максимального правдоподобия. Однако «расплатой» за это является жесткое ограничение временного ряда на стационарность и нормальность, а также неопределенность в выборе порядка авторегрессионной модели. При изменении ее порядка оценки спектра могут существенно меняться.

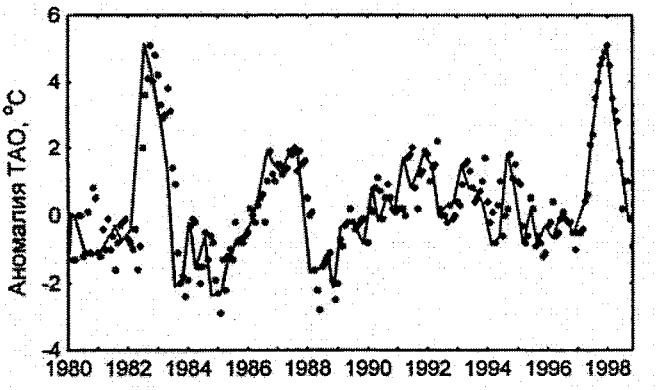


Рис. 11.6. Временной ряд среднемесячных аномалий температуры поверхности океана, полученный на буйковой станции в пункте с координатами 110° з.д. и 0°с.ш. за 1980–1999 гг.

В качестве примера приведем МНЭ-оценки СП, рассчитанные С.Г Добровольским для временного ряда среднемесячных аномалий ТПО, полученного на буйковой станции в пункте с координатами 110° з.д. и 0°с.ш. за 1980–1999 гг. (рис. 11.6) Данный ряд относится к числу наиболее важных индексов, характеризующих явление Эль-Ниньо. Действительно, как видно из рис. 11.6, отчетливо выделяется несколько пиков значительных положительных аномалий ТПО: 1982–1983 гг., 1987–1988 гг. и 1998 г., соответствующих развитию явления Эль-Ниньо. На рис. 11.7 приведены МНЭ-оценки СП для данного ряда с разрешением 3 месяца Мы видим, что изменение порядка модели довольно существенно изменяет характер СП. Появляются хорошо различимые пики спектральной плотности. Один из них составляет 50 месяцев, а другой – 19 месяцев. Однако оба они не являются статистически значимыми, ибо оказываются меньше ширины 95 %-ного доверительного интервала относительно кривой красного шума.

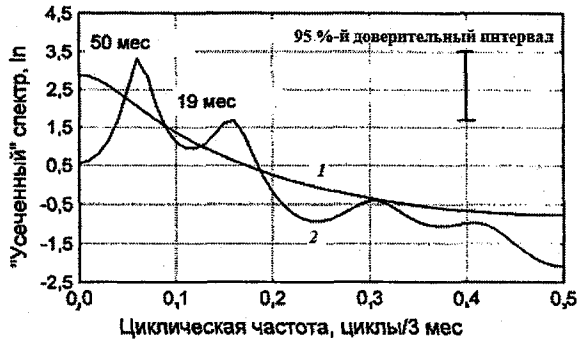


Рис. 11.7. МНЭ-оценки спектральной плотности для временного ряда среднемесячных аномалий температуры поверхности океана с разрешением 3 месяца. Кривая *1* соответствует модели авторегрессии 1-го порядка («красный шум»), кривая *2* – модели 8-го порядка.

МНЭ-анализ СП для временного ряда январских значений индекса Южного колебания за 112 лет, выполненный С.Г. Добровольским (рис. 11.8), также показал, что какие-либо отчетливо выраженные пики СП, выходящие за пределы доверительного интервала, отсутствуют. Таким образом, если судить о межгодовой изменчивости явления «Южное колебание – Эль-Ниньо» только по результатам МНЭ-оценок СП, то она не имеет каких-либо циклических колебаний и представляет собой чисто стохастический процесс.

Рис. 11.8. МНЭ-анализ спектральной плотности для временного ряда январских значений индекса Южного колебания за 112 лет.
1 – модель «белый шум», *2* – модель авторегрессии 1-го порядка, *3* – модель авторегрессии 5-го порядка, *4* – модель авторегрессии 7-го порядка.

Разумеется, есть и другие методы оценки СП. Например, метод наибольшего правдоподобия. Но вследствие его довольно сложной вычислительной схемы он на практике используется довольно редко.

11.5. Виды спектральной плотности временных рядов

В п. 10.5 были рассмотрены стандартные графики нормированных автокорреляционных функций временных рядов для различных типов гидрометеорологических процессов. Поэтому целесообразно теперь обратиться к анализу выборочных оценок спектральной плотности, вычисленных по этим автокорреляционным функциям. Выборочная оценка нормированной СП теоретической модели белого шума, как уже отмечалось выше, представляет собой прямую линию, параллельную оси ординат и находящуюся от нее на расстоянии $1/\pi\alpha$ (рис. 11.9). Это означает, что плотность дисперсии постоянна для всех частот. Поскольку на практике такие спектральные плотности не встречаются, то переходим от теоретической модели к случайному стационарному процессу, развивающемуся по типу модели «белый шум». В этом случае мы получаем такую кривую СП, которая может быть аппроксимирована прямой линией.

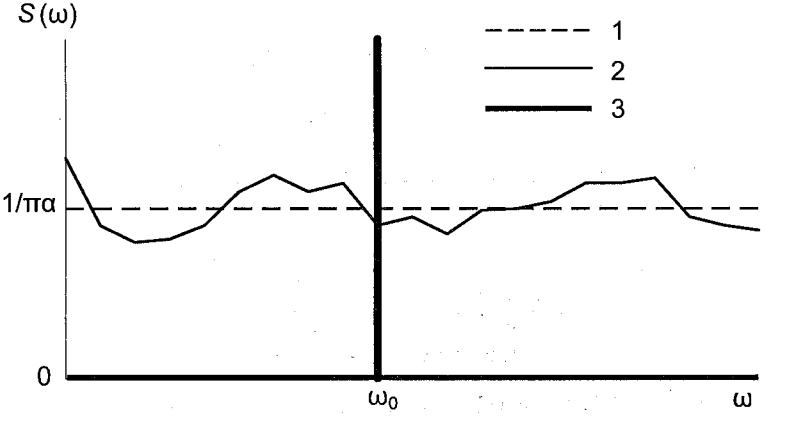


Рис. 11.9. График спектральной плотности теоретической (*1*) и реальной (*2*) моделей «белого шума» и гармонического колебания (*3*).

Поскольку противоположным модели белого шума является гармоническое колебание, которое представляет собой уже детерминированный процесс, то его спектральная плотность имеет вид прямой линии, перпендикулярной оси частот и устремленной в бесконечность (рис. 11.9). Естественно, что подобный вид спектральной плотности для гидрометеорологических процессов невозможен. Поэтому даже выборочные оценки спектральной плотности короткопериодных приливных колебаний уровня представляют очень острый пик с узкой шириной.

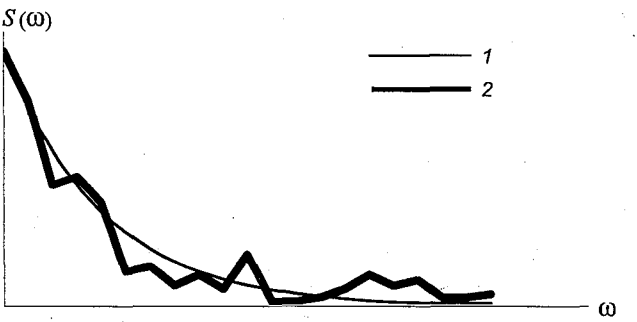


Рис. 11.10. График спектральной плотности теоретической (*1*) и реальной (*2*) моделей «красного шума».

Для теоретической модели красного шума кривая спектральной плотности имеет вид убывающей экспоненты, причем степень ее «вогнутости» зависит от коэффициента затухания α автокорреляционной функции. Чем меньше величина α, тем быстрее приближается СП к оси частот (рис. 11.10). Переходя от теоретической модели к случайному стационарному процессу, который развивается по типу модели «красный шум», получаем кривую спектральной плотности с множеством мелких пиков, которая аппроксимируется убывающей экспонентой.

Спектральная плотность циклического колебания представляет собой острый пик, соответствующий основной частоте этого колебания. Естественно, чем «сильнее» циклическое колебание, тем острее пик СП. В предельном случае оно стремится к гармоническому колебанию, т.е. к прямой линии. Для слабовыраженного циклического колебания пик спектральной плотности становится размытым (рис. 11.11).

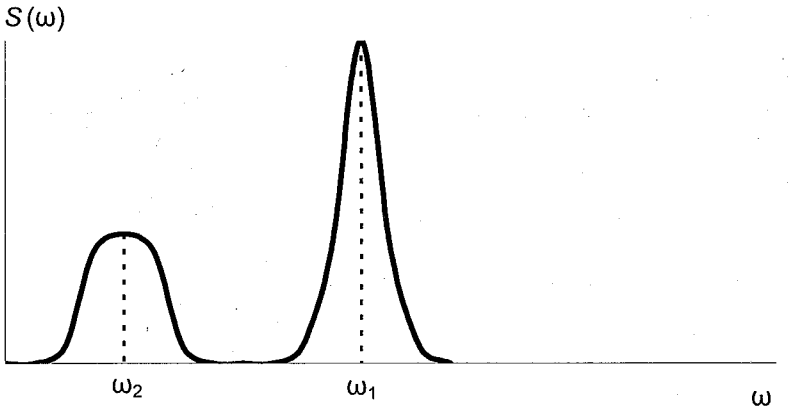


Рис. 11.11. График спектральной плотности циклического колебания.

Как уже отмечалось выше, во многих случаях гидрометеорологические характеристики представляют собой совокупность нескольких элементарных составляющих изменчивости. В частности, для межгодовой изменчивости реальных гидрометеорологических величин характерно наложение случайной изменчивости в виде белого шума на трендовую компоненту. Спектральная плотность трендовой компоненты выражается в виде плавной кривой в области низких частот, затухающей по мере увеличения частоты колебаний, которая постепенно переходит в случайно меняющуюся кривую, приблизительно параллельную оси абсцисс (рис. 11.12). Пожалуй, единственное существенное отличие ее от спектральной плотности в виде модели красного шума состоит в том, что затухание кривой спектральной плотности происходит медленнее модели красного шума.

Довольно часто реальная межгодовая изменчивость гидрометеорологических процессов выражается как совокупность тренда, белого шума и циклического колебания. В этом случае выборочная спектральная плотность имеет вид плавной кривой, переходящей в ненулевой спектр на более высоких частотах, на фоне которого выделяется «горб», соответствующий основной частоте циклического колебания (рис. 11.12).

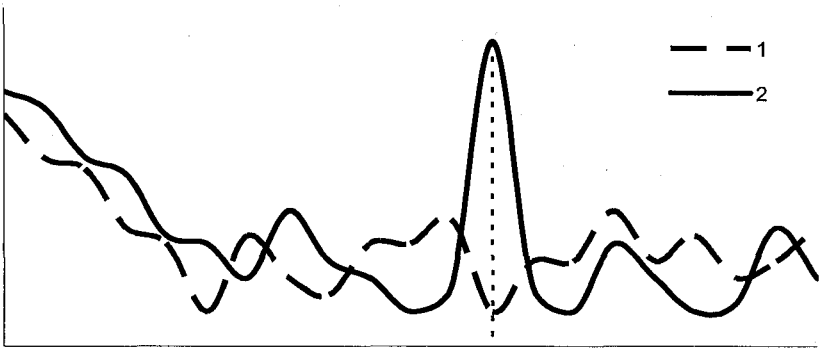


Рис. 11.12. График спектральной плотности совокупности нескольких элементарных составляющих случайного процесса.
1 – случайный процесс в виде белого шума и трендовой компоненты,
2 – случайный процесс в виде белого шума, тренда и циклического колебания.

11.6. Понятие о взаимной спектральной плотности

Нормированная взаимная спектральная плотность (ВСП) связана с взаимной корреляционной функцией (ВКФ) двух стационарных случайных процессов $X(t)$ и $Y(t)$ прямым преобразованием Фурье, т.е.

$$s_{xy}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} r_{xy}(\tau) e^{-i\omega\tau} d\tau . \qquad (11.35)$$

Естественно, что взаимная корреляционная функция выражается через ВСП с помощью обратного преобразования Фурье:

$$r_{xy}(\tau) = \int_{-\infty}^{\infty} s_{xy}(\omega) e^{i\omega\tau} d\omega . \qquad (11.36)$$

Как уже отмечалось выше, ВКФ не обладает свойством четности. Обозначим через $r^{+}_{xy}(\tau)$ четную часть, а через $r^{-}_{xy}(\tau)$ нечетную часть ВКФ. Тогда мы можем записать:

$$r_{xy}(\tau) = r^{+}_{xy}(\tau) + r^{-}_{xy}(\tau).$$

В общем случае ВСП представляет собой комплексную величину, состоящую из вещественной части, называемой коспектром $C_{xy}(\omega)$, и мнимой части – квадратурного спектра $Q_{xy}(\omega)$, т.е.

$$s_{xy}(\omega) = C_{xy}(\omega) + iQ_{xy}(\omega). \qquad (11.37)$$

Вещественная часть ВСП (косинус-спектр) находится как косинус преобразования Фурье четной части функции взаимной корреляции:

$$C_{xy}(\omega) = \frac{1}{\pi}\int_0^{\infty} r^{+}_{xy}(\tau)\cos(\omega\tau)d\tau, \qquad (11.38)$$

где $r^{+}_{xy}(\tau) = [r_{xy}(\tau) + r_{xy}(-\tau)]/2$.

Так как $r^{+}_{xy}(\tau)$ – четная функция, то коспектр является четной функцией и его называют четной (симметричной) частью ВСП.

Операцией (11.38) осуществляется сглаживание эффекта асимметрии функции взаимной корреляции $r_{xy}(\tau)$. Симметрия $r^{+}_{xy}(\tau)$ относительно нулевого сдвига означает, что разность фаз между процессами равна нулю, т. е. процессы происходят синхронно. Коспектр характеризует вклад энергии колебаний различных частот в общую взаимную корреляцию при нулевом сдвиге двух временных рядов, т. е. является мерой взаимной энергии двух процессов.

Мнимая часть взаимного спектра находится как синус-преобразование Фурье нечетной части взаимной корреляции:

$$Q_{xy}(\omega) = \frac{1}{\pi}\int_0^{\infty} r^{-}_{xy}(\tau)\sin(\omega\tau)d\tau, \qquad (11.39)$$

где $r^{-}_{xy}(\tau) = [r_{xy}(\tau) - r_{xy}(-\tau)]/2$.

Данным преобразованием осуществляется усиление эффекта асимметрии функции $r_{xy}(\tau)$. Поскольку $r^{-}_{xy}(\tau)$ является нечетной функцией $[r^{-}_{xy}(\tau) = -r^{-}_{xy}(-\tau)]$, то асимметрия функции $r^{-}_{xy}(\tau)$ проявляется в смещении максимума $r^{-}_{xy}(\tau)$ на некоторый сдвиг $\tau \neq 0$. Это означает, что процессы имеют некоторую разность фаз, т.е. происходят несинхронно.

Квадратурный спектр характеризует распределение по частотам энергии несинхронного взаимодействия. Другими словами, он характеризует вклад в общую взаимную корреляцию пары случай-

ных процессов содержащихся в них гармоник различной частоты при сдвиге фаз этих гармоник на четверть периода T.

Подставляя (11.37) в (11.36) и полагая $\tau = 0$, получаем выражение для взаимной дисперсии случайных процессов:

$$r_{xy}(0) = \int_{-\infty}^{\infty} C_{xy}(\omega) d\omega. \qquad (11.40)$$

При частотном представлении взаимной энергии процессов полезно сравнивать взаимную энергию на фиксированной частоте с энергиями каждого из процессов на той же частоте путем вычисления отношения:

$$F^2(\omega) = \frac{|s_{xy}^2(\omega)|}{s_x(\omega) s_y(\omega)} = \frac{C_{xy}^2(\omega) + Q_{xy}^2(\omega)}{s_x(\omega) s_y(\omega)}. \qquad (11.41)$$

Функция $F^2(\omega)$, называемая *когерентностью*, характеризует линейную статистическую связь спектральных компонент одинаковой частоты и по смыслу аналогична коэффициенту детерминации, но в отличие от него зависит от частоты. Заметим, что по своей сути функция когерентности аналогична квадрату нормированной взаимной корреляционной функции $r_{xy}(\tau)$, определяемой по формуле (10.28).

Вследствие того что всегда справедливо неравенство

$$C_{xy}^2(\omega) + Q_{xy}^2(\omega) \le s_x(\omega) s_y(\omega), \qquad (11.42)$$

называемое *неравенством когерентности*, то величина $F^2(\omega)$ изменяется в пределах от 0 до 1. Отметим, что когерентность служит мерой устойчивости разности фаз. Если разность фаз двух процессов постоянна, то $F^2(\omega) = 1$, если разность фаз неустойчива, то должно выполняться условие $F^2(\omega) \to 0$.

Из (11.41) следует, что при $Q_{xy}(\omega_i) = 0$, $C_{xy}(\omega_i) \neq 0$ разность фаз колебаний на частоте ω_i должна быть равна нулю, так как взаимосвязь процессов будет существовать только за счет их синхронного взаимодействия. При $C_{xy}(\omega_i) = 0$, $Q_{xy}(\omega_i) \neq 0$ разность фаз спектральных компонент на частоте ω_i равна 90°, так как взаимосвязь процессов на частоте ω_i, имеет место лишь за счет энергии несинхронного взаимодействия. Во всех остальных случаях, т.е. при $C_{xy}(\omega_i) \neq 0$, $Q_{xy}(\omega_i) \neq 0$ разность фаз спектральных компонент фиксированной частоты вычисляется по формуле:

$$\varphi_y(\omega) - \varphi_x(\omega) = \theta_{xy}(\omega) = \operatorname{arctg}\frac{Q_{xy}(\omega)}{C_{xy}(\omega)}. \qquad (11.43)$$

Исходя из геометрической интерпретации ВСП, величину $\theta_{xy}(\omega)$ называют *фазовым спектром*. Он характеризует отставание по фазе процесса $Y(t)$ от процесса $X(t)$ при условии, что Q_{xy} считают положительным от 0 до 180° и отрицательным от 180 до 360°. Тогда в соответствии с геометрической интерпретацией ВСП величину $S_{xy}(\omega)$ можно назвать *амплитудным спектром,* который представляет модуль ВСП и определяется следующим выражением:

$$|S_{xy}(\omega)| = \sqrt{C_{xy}^2(\omega) + Q_{xy}^2(\omega)}. \qquad (11.44)$$

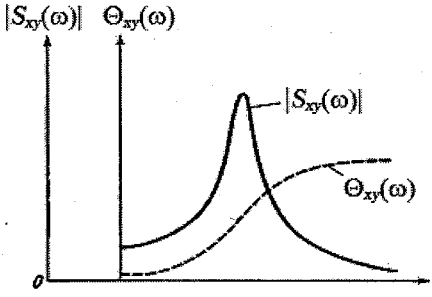


Рис. 11.13. Типовой график взаимной спектральной плотности.

Типовой график взаимной спектральной плотности приведен на рис. 11.13. Нетрудно видеть, что он состоит из двух кривых, одна из которых является амплитудным спектром, а другая – фазовым спектром.

11.7. Фильтрация временных рядов

В спектральном анализе, а также при решении некоторых других задач, связанных с выделением тех или иных колебаний, часто возникает задача фильтрации исходных рядов. Эта операция позволяет значительно уменьшить роль случайных ошибок и подавить те колебания, которые не представляют интереса для решаемой задачи.

Фильтрация осуществляется с помощью следующего преобразования:

$$\tilde{X}(t) = \int_{-\infty}^{\infty} h(\tau) X(t+\tau) d\tau, \qquad (11.45)$$

где $h(\tau)$ – весовая функция фильтра.

Любая фильтрация предусматривает изменение амплитуд гармоник процесса $X(t)$. Функция, определяющая характер изменения амплитуд исходного процесса при прохождении ряда через фильтр, называется *частотной характеристикой фильтра*. При условии, что начальные фазы гармоник после фильтрации должны остаться неизменными, частотная характеристика $c(\omega)$ будет иметь вид

$$c(\omega) = \int_{-\infty}^{\infty} h(\tau) \cos(2\pi\omega\tau) d\tau. \qquad (11.46)$$

Как следует из выражения (10.46), весовая функция $h(\tau)$ такого четного фильтра связана с частотной характеристикой обратным преобразованием Фурье:

$$h(\tau) = \int_{-\infty}^{\infty} c(\omega) \cos(2\pi\omega\tau) d\omega. \qquad (11.47)$$

Таким образом, для фильтрации процесса $X(t)$ следует вначале задать частотную характеристику $c(\omega)$, а затем из выражения (11.47) найти весовую функцию и выполнить преобразование (11.45).

Задача фильтрации решается наилучшим образом, если гармоники с частотами $\omega_1 \le \omega \le \omega_2$ на выходе фильтра сохраняются неизменными, а амплитуды всех остальных гармоник обратятся в нуль. Частотная характеристика такого идеального фильтра имеет вид:

$$c(\omega) = \begin{cases} 1 & \text{при} \quad \omega_1 \le \omega \le \omega_2 \\ 0 & \text{при} \quad \omega < \omega_1, \omega > \omega_2 \end{cases}. \qquad (11.48)$$

Подставив характеристику (11.48) в уравнение (11.47), можно найти весовую функцию идеального фильтра

$$h(\tau) = \frac{\cos\pi\tau(\omega_1+\omega_2)\sin\pi\tau(\omega_2-\omega_1)}{\pi\tau}. \qquad (11.49)$$

Величина спектра после фильтрации определяется как

$$\hat{S}(\omega) = c(\omega)\tilde{S}(\omega).$$

Однако частотная характеристика (10.48) может быть реализована в том случае, если весовая функция (10.49), а значит, и сам процесс $X(t)$, заданы на бесконечном интервале. Поскольку на практике интервал задания функции $X(t)$ всегда конечен, то, следовательно, также должен быть конечен интервал задания весовой функции $h(\tau)$. В этом случае амплитуды гармоник, лежащих в диапазоне $\omega_1 \le \omega \le \omega_2$, умножаются на величины, отличные от единицы, а амплитуды гармоник с частотами вне этого диапазона – на величины, отличные от нуля.

Поскольку реализация идеального фильтра на ограниченном интервале в виде (11.48) невозможна, то задачей фильтрации является нахождение таких $c(\omega)$ и $h(\tau)$, которые бы обеспечили выделение гармоник в заданном диапазоне оптимальным образом, т.е. с минимальными искажениями.

Следует иметь в виду, что решение данной задачи возможно лишь при условии, что спектр исходного процесса известен или есть физические основания предполагать наличие определенных гармоник в исследуемом интервале частот. Последнее обстоятельство может в значительной степени способствовать правильному выбору как вида фильтра, так и интервалов его задания.

Перейдя к сумме по конечному интервалу $(2M + 1)$, выражение (11.45) запишем так:

$$\tilde{X}_i = \sum_{j=-M}^{M} h_j X_{i+j} \,. \qquad (11.50)$$

В зависимости от выделяемого диапазона частот фильтры подразделяются на *низкочастотные* или сглаживающие ($0 \le \omega \le \omega_2$), *высокочастотные* ($\omega_1 \le \omega \le \infty$) и *полосовые* ($\omega_1 \le \omega \le \omega_2$).

Первые применяются в случае необходимости исследования долгопериодных колебаний, тенденции или тренда в исходном ряду, а также для подавления случайных колебаний, связанных с ошибками измерений.

Высокочастотные фильтры позволяют стационировать ряд и исследовать высокочастотные колебания до определенного задаваемого предела частоты. Полосовые фильтры применяются для изучения колебаний в определенной полосе частот.

Фильтров, различающихся своей весовой функцией, предложено в настоящее время большое количество. К их числу, напри-

мер, относятся: фильтр «скользящее среднее», треугольный фильтр (фильтр Бартлетта), фильтр Тьюки, нормальный фильтр и др.

Простейшим и, очевидно, наиболее широко применяемым в гидрометеорологии видом фильтрации является *скользящее осреднение*. Весовая функция этого фильтра имеет вид:

$$h_j = \begin{cases} \dfrac{1}{2M+1} & \text{при} \quad |j| \le M \\ 0 & \text{при} \quad |j| > M \end{cases}, \tag{11.51}$$

где $(2M+1)$ – *интервал сглаживания*.

Общая формула для отфильтрованного ряда при низкочастотной фильтрации может быть записана как

$$\tilde{X}_i = \frac{1}{2M+1} \sum_{j=-M}^{M} X_{i+j} . \tag{11.52}$$

При использовании скользящего осреднения в качестве высокочастотного фильтра общее выражение для $\tilde{X}_i$ приобретает вид

$$\tilde{X}_i = X_i - \frac{1}{2M+1} \sum_{j=-M}^{M} X_{i+j} . \tag{11.53}$$

И, наконец, для полосового фильтра общую формулу для ряда X_i можно представить как

$$\tilde{X}_i = \frac{1}{2M_1+1} \sum_{j=-M_1}^{M_1} X_{i+j} - \frac{1}{2M_2+1} \sum_{j=-M_2}^{M_2} X_{i+j} . \tag{11.54}$$

где $M_2 > M_1$.

Частотная характеристика высокочастотного фильтра, определяемая выражением

$$c_i(\omega) = \frac{\sin \pi\omega(2M+1)}{\pi\omega(2M+1)},$$

приводится на рис. 11.14, *а*.

Частотную характеристику низкочастотного фильтра $c_h(\omega)$ можно получить через частотную характеристику сглаживающего фильтра (рис. 11.14, *б*):

$$c_h(\omega) = 1 - c_i(\omega) .$$

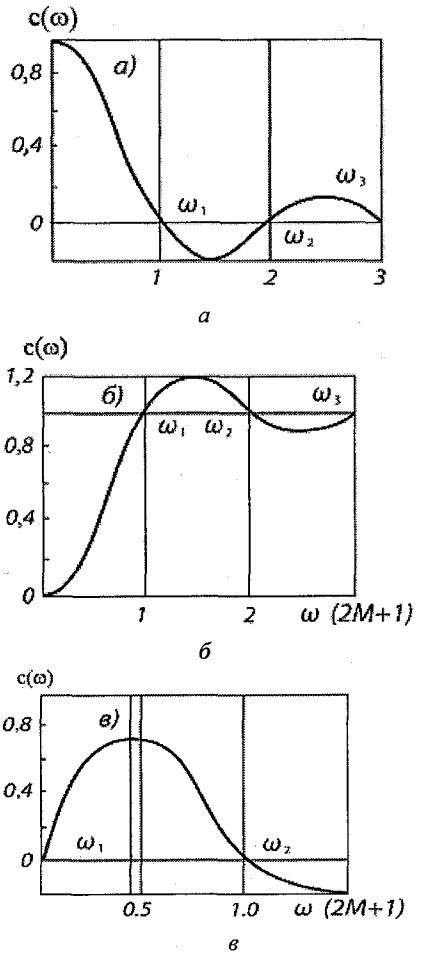


Рис. 11.14. Частотная характеристика различных фильтров:
а – высокочастотный фильтр; *б* – низкочастотный фильтр; *в* – полосовой фильтр.

Наконец, частотная характеристика полосового фильтра может быть представлена как разность частотных характеристик двух сглаживающих фильтров (рис. 11.14, *в*):

$$c(\omega) = c_t(\omega, M_1) - c_t(\omega, M_2).$$

Рассмотрим, например, частотную характеристику сглаживающего фильтра. Как видно из рис. 11.15, величина $c_t(\omega)$ на уча-

стке главного лепестка (от $\omega = 0$ до первого перехода кривой $c(\omega)$ через нуль) быстро убывает. Однако значения частотной характеристики на боковых лепестках (участках между соседними переходами $c(\omega)$ через нуль) довольно велики. Так, экстремум первого бокового лепестка составляет более 0,2. К тому же гармоники с частотами от $p/(2M + 1)$ до $(p + 1)/(2M + 1)$, где $p = 1, 3, 5, ...$, в процессе $X(t)$ на выходе фильтра меняют фазу на 180°.

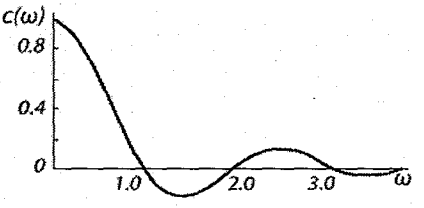


Рис. 11.15. Частотная характеристика фильтра скользящего среднего.

В общем случае можно указать три следующих свойства фильтрации рядов:

1) длина отфильтрованного ряда сокращается по сравнению с исходной на $2M$ ординат. Это накладывает ограничение на выбор интервала задания весовой функции, так как должно быть соблюдено условие $2(N - M) < 1/\omega_m$, где ω_m – наиболее низкая частота в исходном ряду;

2) диапазон исследуемых частот ограничен сверху частотой $\omega_N = 0{,}5$ цикла на интервал дискретности. Поэтому, располагая рядом с числом членов $2N + 1$, можно фильтровать частоты в диапазоне

$$\frac{1}{2(N-M)} < \omega \le \frac{1}{2}$$

или с периодами в диапазоне $2 \le p < 2(N - M)$, где p выражается через число интервалов дискретности;

3) средние арифметические значения рядов X_i и $\tilde{X}_i$ связаны между собой соотношением:

$$\frac{\sum_{i}^{N} \tilde{X}_i}{\sum_{i}^{N} X_i} = \sum_{j=-M}^{M} h_j .$$

Поэтому низкочастотные фильтры строят обычно таким образом, чтобы сумма их весов равнялась единице. Для высокочастотных и полосовых фильтров сумма весов равна нулю, поэтому при их использовании среднее арифметическое значение полученного ряда равно нулю.

Итак, очевидными недостатками фильтра «скользящее среднее» являются:

– уменьшение длины исходного ряда, которое может быть заметным при значительном интервале сглаживания;

– дисперсия сглаженного ряда всегда меньше дисперсии исходного ряда;

– наличие сдвига по фазе между сглаженным и исходным процессами.

Использование предварительной фильтрации ряда при расчете спектра приводит к искажению его вида практически при любом типе фильтра. Особенно значительны искажения при полосовых фильтрах, так как при этом происходит изменение амплитуд не только вне заданной полосы пропускания, но и внутри нее. Но в силу того, что истинные оценки СП, как правило, неизвестны, то и исправление выборочных значений СП оказывается очень сложным делом. Кроме того, статистическая значимость полученных спектральных оценок с предварительной фильтрацией ряда всегда будет завышена, особенно при узкой полосе пропускания частот и малой длине ряда.

Пример 11.1. На рис. 11.16 представлен межгодовой ход фактических значений температуры воздуха в Санкт-Петербурге за период 1880–1996 гг. Кроме того, на этом же рисунке нанесены скользящие средние температуры воздуха с интервалами сглаживания СС = $2M + 1 = 7$ и СС = 21. Обычно низкочастотное осреднение применяют, как уже указывалось выше, с целью подавить случайные ошибки и мелкомасштабные флуктуации. Действительно, из рис. 11.16 отчетливо видно, что уже 7-летние скользящие средние существенно искажают структуру временного ряда температуры. Так, стандартное отклонение исходных данных равно $\sigma = 1{,}11$ °С, а отфильтрованного ряда при СС = 7 $\sigma = 0{,}50$ °С. Следовательно, дисперсия отфильтрованного ряда занижается более чем в четыре раза. Еще больше уменьшается дисперсия ряда

при скользящем осреднении СС = 21. Так как в этом случае σ = =0,29 °С, то дисперсия отфильтрованного ряда занижается уже более чем на порядок (в 14 раз).

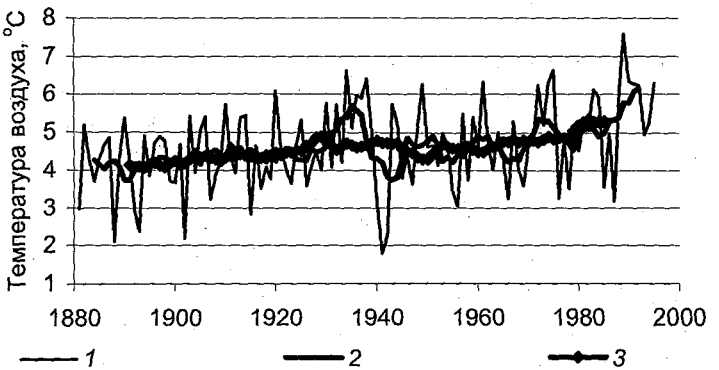


Рис. 11.16. Межгодовой ход температуры воздуха в Санкт-Петербурге. *1*– фактические значения, *2* – скользящие средние температуры воздуха с интервалом сглаживания $(2M + 1) = 7$, *3* – скользящие средние температуры воздуха с интервалом сглаживания $(2M + 1) = 21$.

Очевидно, использование скользящего осреднения целесообразно при анализе тенденции изменения временного ряда. Из данного рисунка нетрудно видеть, что температура воздуха в Санкт-Петербурге имеет довольно устойчивую тенденцию к повышению. При этом тенденция не является полностью монотонной. Так, примерно до начала 30-годов XX в. отмечался устойчивый рост температуры, затем до начала 70-х годов она практически не изменялась, а после этого опять начался ее достаточно быстрый рост, который продолжается до настоящего времени.

Пример 11.2. При анализе межгодовой изменчивости гидрометеорологических характеристик многими исследователями обычно предварительно выполняется скользящее осреднение исходных рядов с целью сглаживания случайных ошибок и мелкомасштабных флуктуаций. Наиболее часто применяется интервал сглаживания СС = 5. Воспользуемся для этой цели данными из примера 7.4. В табл. 11.2 приводятся оценки корреляции между средними годовыми значениями различных гидрометеорологических величин в районе Норвежского течения за 1949–2001 гг.

(верхний треугольник) и их скользящими 5-летними средними (нижний треугольник). Отметим, что значимые коэффициенты корреляции на уровне $\alpha = 0{,}05$ ($r_{кр} = 0{,}26$) выделены полужирным шрифтом.

Таблица 11.2

Корреляционная матрица между средними годовыми значениями различных гидрометеорологических характеристик в районе Норвежского течения за 1949–2001 гг. (верхний треугольник) и их скользящими 5-летними средними (нижний треугольник)

Хар-ка	T_w	T_A	U	V	P	*Prec*	*Cloud*	R	*TB*
T_w	1	**0,84**	0,12	**0,30**	–0,19	0,07	–0,03	**0,55**	**0,28**
T_a	**0,88**	1	0,25	**0,59**	–0,22	0,06	0,05	**0,77**	**0,67**
U	**0,30**	**0,39**	1	**0,29**	**–0,28**	**0,65**	**0,27**	0,19	0,04
V	**0,28**	**0,46**	**0,72**	1	**–0,27**	0,00	–0,07	**0,45**	**0,43**
P	–0,25	**–0,38**	**–0,55**	**–0,60**	1	**–0,43**	–0,17	0,08	–0,06
Prec	0,23	0,18	**0,32**	–0,05	–0,17	1	**0,78**	0,05	0,04
Cloud	0,05	0,01	–0,21	**–0,49**	0,19	**0,80**	1	0,10	**0,30**
R	**0,72**	**0,82**	0,25	0,20	–0,19	**0,34**	0,22	1	**0,69**
TB	**0,29**	**0,58**	–0,17	–0,07	–0,06	0,24	**0,46**	**0,67**	1

Прежде всего, обратившись к анализу исходных гидрометеорологических характеристик (верхний треугольник), отметим, что довольно высокая корреляция между некоторыми переменными является очевидной. Например, между температурами поверхности океана и воздуха (T_w и T_a), между облачностью и осадками (*Prec* и *Cloud*), между радиационным и тепловым балансом (R и *TB*). Именно между этими характеристиками корреляция максимальна. Менее очевидной, например, является корреляция между радиационным балансом поверхности океана и температурой воздуха и совсем она не очевидна между радиационным балансом и меридиональной компонентой скорости ветра. Всего значимых коэффициентов корреляции 18.

Обратимся теперь к анализу стохастических связей между 5-летними скользящими средними гидрометеорологических характеристик (нижний треугольник). Если само число значимых коэффициентов корреляции практически не изменилось (19), то вот оценки многих из них изменились весьма заметно. При этом почти постоянными остались оценки коэффициентов корреляции между теми переменными, связь для которых является физически обу-

словленной. Так, очень мало изменились коэффициенты корреляции $\Delta r(T_w, T_a) = 0{,}84–0{,}88$, $\Delta r(Prec, Cloud) = 0{,}78–0{,}80$, $\Delta r(R, TB) = 0{,}67–0{,}69$. В то же время связь между переменными, когда связь между ними не очевидна, обычно приводит к завышению оценок коэффициентов корреляции, причем в некоторых случаях это завышение весьма значительно. Например, корреляция между P и V увеличилась от $r = –0{,}27$ для исходных данных до $r = –0{,}60$ для скользящих средних. Еще более значительным оказывается увеличение корреляции для компонент скорости ветра U и V (соответственно $r = 0{,}29$ и $r = 0{,}72$). Правда, в некоторых случаях связь занижается. Например, корреляция между TB и V составляет $r = 0{,}43$ для исходных данных и $r = –0{,}07$ для скользящих средних.

Итак, не вызывает сомнения, что скользящее осреднение может приводить к заметному искажению стохастической связи между переменными, особенно в том случае, когда она является неочевидной.

Часть 4. АНАЛИЗ СЛУЧАЙНЫХ ПОЛЕЙ

Глава 12. СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ И СВОЙСТВА СЛУЧАЙНОГО ПОЛЯ

12.1. Первичные характеристики случайного поля

Как следует из теории случайных функций, *случайное поле – это случайная функция, изменяющаяся в пространстве.* В общем виде случайное поле можно представить как $U(x, y, z, t)$, где x, y, z – пространственные координаты, t – время. Такое поле называется четырехмерным. Если случайное поле не зависит от времени $U(x, y, z)$, то имеем трехмерное поле и т.д. Классическим примером двухмерного поля служит пространственная карта.

Учитывая векторный характер пространственных координат, случайное поле можно рассматривать в виде $U(\rho)$, где ρ – k-мерный вектор (k = 2–4). По аналогии со случайными процессами, случайное поле представляется как совокупность всех его реализаций или как совокупность всех его сечений. При этом сечением случайного поля будем называть случайную величину, получающуюся при фиксированных значениях всех аргументов, т.е. при фиксированном значении вектора ρ. Тогда единичной реализацией случайного поля будет являться неслучайное поле, полученное в результате наблюдений или опытов.

В этом случае простой заменой времени t на ρ все формулы по оценке вероятностных характеристик случайных процессов могут быть распространены на случайные поля. Однако существует два важных ограничения: во-первых, случайное поле должно быть задано в узлах регулярной сети точек таким образом, чтобы шаг пространственной дискретизации по всем направлениям был одинаковым; во-вторых, если случайный процесс является скалярным, то случайное поле – векторным.

Отметим, что важной особенностью случайных полей является то, что они носят, как правило, нерегулярный характер. Действительно, сеть стационарных метеорологических, гидрологических и иных станций на территории суши находится преимущест-

венно в населенных пунктах, никак не связанных с узлами географической сетки.

По аналогии со случайными процессами для исчерпывающего описания случайного поля необходимо знание n-мерной функции распределения

$$F_u(u_1, u_2,\ldots, u_n;\ \rho_1, \rho_2,\ldots, \rho_n) = P(U_1 < u_1,\ U_2 < u_2,\ \ldots,\ U_n < u_n) \quad (12.1)$$

и n-мерной плотности распределения

$$f_n(u_1, u_2,\ \ldots,\ u_n;\ \rho_1, \rho_2,\ \ldots,\ \rho_n) =$$

$$= \partial^n F_n(u_1, u_2,\ldots, u_n; \rho_1, \rho_2,\ \ldots,\ \rho_n)/\partial u_1 \partial u_2 \ldots \partial u_n \quad (12.2)$$

Поскольку на практике определить n-мерные функции распределения или плотности распределения чрезвычайно сложно, то обычно ограничиваются анализом моментов распределения. Так, момент первого порядка

$$m_1(\rho) = M[U(\rho)] = m_u(\rho) \quad (12.3)$$

называется математическим ожиданием случайного поля. Его выборочной характеристикой является пространственное среднее, обычно обозначаемое как $<x>$.

Отклонение случайного поля от его математического ожидания называют центрированным случайным полем:

$$U^*(\rho) = U(\rho) - m_u(\rho).$$

Дисперсией случайного поля называется одноточечный центральный момент второго порядка:

$$\mu_2(\rho) = M\{[U(\rho) - m_u(\rho)]^2\} = D_u(\rho). \quad (12.4)$$

Двухточечный центральный момент второго порядка

$$\mu_{1,1}(\rho_1,\rho_2) = M\{[U(\rho_1) - m_u(\rho_1)][U(\rho_2) - m_u(\rho_2)]\} = R_u(\rho_1,\rho_2) \quad (12.5)$$

называется пространственной (кросс) корреляционной функцией случайного поля.

При равенстве векторных аргументов $\rho_1 = \rho_2 = \rho$ $t_1 = t_2 = t$, кросскорреляционная функция обращается в дисперсию случайного поля

$$R_u(\rho, \rho) = D_u(\rho).$$

Чтобы иметь возможность сравнивать кросскорреляционные функции для случайных полей разной размерности, их обычно нормируют:

$$r_u(\rho_1, \rho_2) = R_u(\rho_1, \rho_2) / [D_u(\rho_1)D_u(\rho_2)]^{1/2}. \qquad (12.6)$$

Нормированная кросскорреляционная функция для каждой фиксированной пары точек ρ_1,ρ_2 представляет собой коэффициент корреляции между сечениями случайного поля, соответствующими этим точкам.

Отметим, что кросскорреляционная функция в общем случае характеризует связь между значениями случайного поля в двух различных точках пространства в различные моменты времени. Поэтому, фиксируя либо пространство, либо время, мы получаем совершенно различные корреляционные функции. Например, при фиксировании пространственных координат точек поля изучается временная изменчивость в данной конкретной точке путем расчета автокорреляционной функции. Наоборот, фиксируя моменты времени, можно изучать пространственную структуру случайного поля и, в частности, вычислять пространственные корреляционные функции. Выборочный коэффициент пространственной линейной корреляции, характеризующий меру линейной связи точек пространства в два момента времени, определяется как

$$r_u = \Sigma(x_{ip} - \langle x_p \rangle)(x_{iq} - \langle x_q \rangle)/n\sigma_p\sigma_q . \qquad (12.7)$$

где $\langle x_p \rangle$ и $\langle x_q \rangle$ – пространственные средние, а σ_p и σ_q – стандартные отклонения параметра x_i в моменты времени p и q; n – число точек.

Все свойства коэффициента пространственной корреляции совпадают со свойствами коэффициента парной корреляции, перечисленными в главе 6. Аналогичным образом также оцениваются его стандартная ошибка и значимость.

12.2. Однородность и изотропность случайного поля

Как было показано выше, принципиальное отличие случайного процесса от случайного поля заключается в том, что случайный процесс является скалярным, а случайное поле – векторным. Данное обстоятельство следует иметь в виду при анализе свойств слу-

чайного поля. Понятие стационарности, используемое для случайных процессов, эквивалентно понятию однородности случайного поля. Случайное поле называют *однородным*, если все n-мерные законы распределения не изменяются при переносе системы точек $\rho_1, \rho_2, ..., \rho_n$ на один и тот же вектор, т.е. если функция распределения не изменяется при замене сечения, соответствующего точкам $\rho_1, \rho_2, ..., \rho_n$, сечениями, которые соответствуют точкам $\rho_1 + \rho_0, \rho_2 + \rho_0, ..., \rho_n + \rho_0$ при любом векторе ρ_0.

Указанные условия однородности соответствуют условию строгой однородности (однородности в узком смысле). *Полем однородным в широком смысле называют такое поле, для которого математическое ожидание является постоянной величиной, а корреляционная функция зависит только от одного аргумента – разности векторов* $l = \rho_2 - \rho_1$.

Вследствие векторного характера случайного поля дополнительно вводится понятие изотропности. Случайное однородное поле называется изотропным, если все его n-мерные законы распределения не изменяются при всевозможных вращениях системы точек $N_1(\rho_1), N_2(\rho_2), ..., N_n(\rho_n)$ вокруг любой оси, проходящей через начало координат, и при зеркальном их отражении относительно любой плоскости, также проходящей через начало координат.

Таким образом, для однородного изотропного поля n-мерные плотности распределения $f_N(u_1, u_2, ..., u_n; \rho_1, \rho_2, ..., \rho_n)$ не изменяются при параллельном переносе, вращении и зеркальном отображении системы точек $N_1(\rho_1), N_2(\rho_2), ..., N_n(\rho_n)$. При этом корреляционная функция $R_U(\rho_1, \rho_2)$ должна принимать одни и те же значения для любой пары точек $N_1(\rho_1), N_2(\rho_2)$, для которых одинаков модуль разности $l = |\rho_2 - \rho_1|$, поскольку такие пары точек всегда могут быть совмещены друг с другом с помощью параллельного переноса, вращения и зеркального отражения. Следовательно, корреляционная функция однородного и изотропного поля является функцией одного скалярного аргумента l, представляющего собой расстояние между точками $N_1(\rho_1)$ и $N_2(\rho_2)$.

Поэтому характеристикой однородности и изотропности могут служить концентрические окружности – изокорреляты, относительно любой заданной точки $N_1(\rho_1), N_2(\rho_2), \ldots, N_n(\rho_n)$. Пример однородного и изотропного поля приведен на рис. 12.1.

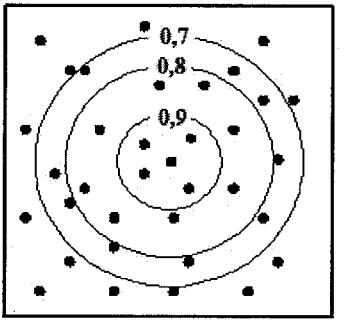


Рис. 12.1. Распределение изокорреляг для однородного и изотропного случайного поля.

Для однородного и изотропного поля математическое ожидание есть величина постоянная, т.е. $M_U(l) = M_U$, а корреляционная функция является функцией только одного скалярного аргумента l, т.е.

$$R_U(l) = R_U(\rho_1, \rho_2);$$

$$l = |\rho_2 - \rho_1| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}\ .$$

Многочисленные исследования гидрометеорологических полей указывают на существенное различие изменений гидрометеорологических характеристик в горизонтальном и вертикальном направлениях, поэтому при изучении статистической структуры крупномасштабных полей допустимой можно считать изотропность и однородность лишь применительно к двухмерному горизонтальному полю. При этом однородность принимается относительно центрированного случайного поля, т.е. рассматривается поле отклонений от его математического ожидания. Само математическое ожидание обычно нельзя считать постоянным.

В предположении об эргодичности однородного и изотропного поля, которое обычно не доказывается, а принимается априори, математическое ожидание и корреляционную функцию можно на-

ходить осреднением по одной реализации, взятой по одной большой пространственной области. В этом случае для трехмерной области математическое ожидание (пространственное среднее) вычисляется следующим образом:

$$m_U = \frac{1}{v} \iiint_{(D)} u(x, y, z) dxdydz, \qquad (12.8)$$

где D – поверхность, стягивающая объем v,
а для плоской области

$$m_u = \frac{1}{S} \iint_{(D)} u(x, y) dxdy \cdot \qquad (12.9)$$

где S – площадь плоской области D.

Корреляционная функция для трехмерного поля определяется по следующей формуле:

$$R_U(l) = \frac{1}{v} \iiint_{(D)} [u(x, y, z) - m_u] \times [u(x + \Delta x, y + \Delta y, z + \Delta z) - m_u] dxdydz \cdot \qquad (12.10)$$

Нетрудно видеть, если $R_U(l)$ разделить на значения стандартного отклонения x, y, z, то получим нормированную кросскорреляционную функцию.

Таким образом, однородное изотропное поле обладает эргодическим свойством, если математическое ожидание и корреляционная функция, полученные осреднением по одной реализации по формулам (12.8) и (12.10) при безграничном увеличении диаметра области, могут быть с вероятностью, сколь угодно близкой к единице, приближены к соответствующим характеристикам, полученным осреднением по всему множеству реализаций. Поскольку, как правило, на практике мы не имеем множества реализаций, то эргодичность представляет собой важнейшее условие обработки пространственных полей.

Отметим, что существует важное отличие пространственной корреляционной функции от автокорреляционной. Установлено, что в отличие от временных рядов пространственная корреляционная функция при стягивании области в точку стремится не к единице, а к величине

$$r_u(0) = 1/(1 + \eta^2),$$

где $\eta^2 = \delta^2/\sigma^2$ – мера случайных погрешностей в исходных данных в отдельной точке; δ^2 – дисперсия случайных погрешностей; σ^2 – дисперсия поля.

При этом ошибки измерений неотделимы от отклонений, вызванных локальными нарушениями однородности поля. При некоторых упрощающих предположениях, используя равенство

$$\eta^2 = [1 - r_u(0)]/\, r_u(0),$$

получаем расчетную формулу для оценки дисперсии случайных погрешностей

$$\delta^2 = [1 - r_u(0)]\sigma^2/\, r_u(0).$$

На практике оценка $r_u(0)$ осуществляется экстраполяцией эмпирической пространственной корреляционной функции в точку $l = 0$.

В некоторых случаях для однородного изотропного поля наряду с корреляционной функцией используется структурная функция

$$B_u(l) = M\{[U(\rho + l) - U(\rho)]^2\}. \qquad (12.11)$$

По аналогии со случайным процессом структурная функция случайного поля однозначно определяется через ее корреляционную функцию:

$$B_u(l) = 2[R_u(0) - R_u(l)]. \qquad (12.12)$$

Естественно, что существует и обратное преобразование, выражающее зависимость корреляционной функции от структурной функции:

$$R_u(l) = 0{,}5[B_u(\infty) - B_u(l)]. \qquad (12.13)$$

Отсюда следует, что при анализе статистической структуры случайных полей в принципе безразлично, какую из этих функций использовать. Отметим, что в практических приложениях структурная функция используется довольно часто. Кроме того, следует иметь в виду, что все приведенные выше формулы для статистических характеристик генеральной совокупности полностью справедливы и для их выборочных аналогов.

К основным задачам анализа случайных полей относится, прежде всего, изучение их статистической структуры, которая включает определение статистических характеристик (пространственное среднее, корреляционная функция, дисперсия и т.д.). Кроме того, важной задачей анализа случайных полей является объектив-

ный анализ, т.е. перевод данных из нерегулярной сети точек в регулярную.

К другим задачам относятся:

– построение пространственных изолиний или карт;

– сглаживание и фильтрация исходных данных;

– восстановление полей с приведением их к одному моменту времени и т.д.

12.3. Анализ схем размещения точек на карте

Известный статистик и геолог Дж. Девис сказал, что «в науках о Земле карты играют ту же роль, что и ноты в музыке, будучи компактным и эффективным средством выражения зависимостей и различных деталей».

Как уже отмечалось выше, карта – это двухмерное представление некоторой пространственной области. Зависимости, изучаемые на карте, почти всегда изображаются с помощью точек, которые представляют собой некоторые числовые оценки случайной величины, непосредственно измеренные или рассчитанные тем или иным способом.

Существующие схемы расположения точек на карте можно условно разделить на три категории: равномерные, регулярные и случайные. Схема расположения точек называется *равномерной*, если плотность точек в любой подобласти равна плотности точек во всех других подобластях (рис. 12.2). Как видно из рис. 12.2, на каждый квадрат приходится 1–2 точки. При неравномерном размещении точек плотность их размещения в пространстве существенно различна. Другими словами, в одних подобластях точек может быть густо, а в других – пусто.

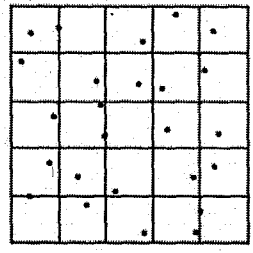


Рис. 12.2. Схема равномерного размещения точек на карте.

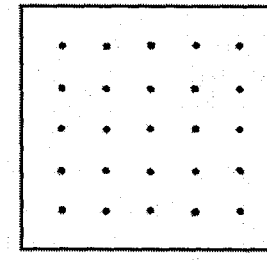


Рис. 12.3. Схема регулярной сети точек на карте в декартовой системе координат.

Схема называется *регулярной*, если точки образуют какой-либо вид сети. Это значит, что расстояния между любыми точками i и j подчиняются определенному закону, причем данный закон не обязательно должен быть одинаков по осям абсцисс и ординат. Принимая, например, равенство расстояний по обеим осям, получаем регулярную сеть в декартовой системе координат (рис. 12.3). Если расстояние постоянно по оси ординат, но меняется вдоль оси абсцисс, то имеем географическую сеть точек, образованную пересечение широт и долгот. При изучении крупномасштабных гидрометеорологических процессов географическая сетка используется наиболее часто. Значение регулярной сети точек, как при вероятностном описании полей, так и при численном моделировании гидрометеорологических процессов, трудно переоценить. Однако реальная сеть наблюдений таковой обычно не является.

Схема называется *случайной*, если 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. Это означает, что их распределение не поддается каким-либо закономерностям.

Таким образом, по аналогии с классификацией зависимостей (см. п. 6.1) регулярную схему точек мы можем считать детерминированной, а равномерную – стохастической, которая в предельных случаях может вырождаться в регулярную или в случайную схему. Например, если из регулярной сети точек случайным образом исключить ряд точек, то она превращается в равномерную схему.

Естественно, в реальных условиях большинство гидрометеорологических станций наблюдений занимают некоторое промежуточное положение между равномерной и случайной схемами расположения точек. Регулярная сеть точек наблюдений обычно выдерживается только при проведении специально спланированных натурных экспериментов. Например, даже одним судном может быть выполнена гидрологическая съемка акватории моря в узлах регулярной сети точек.

Если регулярная сеть точек на карте видна сразу и вряд ли нуждается в дополнительной проверке, то, вообще говоря, не так просто отличить равномерную схему расположения точек от неравномерной или от случайной. Для этого используется довольно много различных способов и критериев, описание которых можно

найти в специальной литературе. Отметим, что требование равномерного размещения измерительной сети является весьма важным при решении многих задач. Даже такой простой, как оценка пространственной средней. Действительно, при крайне неравномерной сети точек, как будет показано ниже, задача оценки пространственной средней становится уже нетривиальной.

Степень равномерности размещения точек на карте может быть проверена в рамках статистической проверки гипотез. С этой целью вся карта делится на ряд одинаковых подобластей (например, квадратов). При этом число их должно быть не меньше 5, а в каждой из подобластей желательно, чтобы число точек было также не меньше 5.

Если каждая подобласть содержит примерно одно и то же число точек, то естественно считать, что система расположения точек является равномерной. Проверка гипотезы равномерности осуществляется с помощью критерия Пирсона χ^2, который теоретически не зависит от формы или ориентировки подобластей. Ожидаемое («теоретическое») число точек для каждой подобласти будет равно:

$$E = n/k, \tag{12.14}$$

где n – общее число точек наблюдения; k – число подобластей.

Далее рассчитывается критерий Пирсона по формуле:

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i , \tag{12.15}$$

где O_i – фактическое число точек для каждой подобласти.

После этого проверяется неравенство $\chi^2 > \chi^2_{кр}(\alpha, \nu = k–2)$. Если данное неравенство выполняется, то нулевая гипотеза отвергается и делается вывод о неравномерном (случайном) распределении точек. Если данное неравенство не выполняется, то у нас есть основания полагать, что распределение точек в пространстве равномерное. Очевидный недостаток такого подхода состоит в произвольном разбиении области на подобласти. Существует вероятность того, что при варьировании числа подобластей далеко не всегда результаты будут совпадать.

Очень простыми критериями проверки сети точек на случайность являются непараметрические тесты по числу экстремумов

(смены знака) или числу скачков. Эти тесты являются универсальными в том смысле, что значения исследуемой характеристики в точках могут быть представлены в выборке в любом произвольном порядке.

Как известно, экстремумом называется точка, в которой производная функции имеет разрыв. Другими словами, это точка, в которой знак приращения рассматриваемой характеристики изменяется на противоположный: с убывания на возрастание или наоборот. Это означает, что должно выполняться одно из следующих неравенств:

$$x_{i-1} < x_i > x_{i+1}$$

или

$$x_{i-1} > x_i < x_{i+1}.$$

В первом случае x_i – максимум, во втором – минимум. Естественно полагать, что число смены знаков в случайной выборке должно зависеть только от ее объема n. Установлено, что если $n > 10$, то статистическое распределение числа экстремумов близко к нормальному с математическим ожиданием

$$M(\xi) = (2n - 4)/3 \qquad (12.16)$$

и дисперсией

$$D(\xi) = (16n - 29)/90. \qquad (12.17)$$

Проверка гипотезы о случайном характере распределения точек основана на сравнении фактического значения числа точек экстремумов $\Sigma\xi$, полученного по исследуемой карте, с теоретическим его значением $M(\xi)$, рассчитанным по формуле (12.16). Далее определяется фактическое нормированное число экстремумов как

$$Z = [\Sigma\xi - M(\xi)]/[D(\xi)]^{1/2}, \qquad (12.18)$$

которое сравнивается с нормированным значением нормального закона распределения $\boldsymbol{Z_{кр}}$ при заданном уровне значимости α. Если $Z > \boldsymbol{Z_{кр}}$, то гипотеза о случайном характере точек отвергается и делается вывод, что их распределение применительно к рассматриваемой характеристике обладает определенными закономерностями, в том числе оно может носить равномерный характер. Если $Z < \boldsymbol{Z_{кр}}$, то значения $\Sigma\xi$ и $M(\xi)$ не существенно отличаются друг от друга, и можно полагать, что рассматриваемая выборка является совершенно случайной.

Суть непараметрического теста проверки сети точек на случайность по числу скачков или по числу повышений (понижений) ряда сводится к следующему. Пусть у нас имеется выборка x_1, x_2, …, x_n. Переход $x_{i-1} < x$ будем называть повышением и обозначать знаком (+), а переход $x > x_{i-1}$ – понижением и обозначать знаком (–). Установлено, что общее число повышений (понижений) значений случайной величины в выборке при ее объеме более $n > 10$ распределено асимптотически нормально с математическим ожиданием

$$m_+ = m_- = n/2$$

и дисперсией

$$D_+ = D_- = (n + 1)/12.$$

Определив число повышений (n_+) и понижений (n_-) ряда, рассчитаем нормированную (стандартизованную) величину числа понижений (повышений):

$$t_+ = (n_+ - m_+)/ (D_+)^{1/2},\ t_- = (n_- - m_-)/ (D_-)^{1/2}. \qquad (12.19)$$

Далее сравним t_+ или t_- со значениями нормированных ординат таблицы нормального закона распределения. Если вероятность рассчитанных значений t_+ или t_- по таблице окажется меньше заданного уровня значимости, то гипотеза о случайном характере распределения точек опровергается и считается, что оно имеет устойчивую тенденцию к закономерным изменениям в пространстве.

Пример 12.1. На рис. 12.4 представлено распределение 123 артезианских скважин в одном из районов США. Вся область была разделена на 12 равных квадратов и для каждого из них было подсчитано число скважин (табл. 12.1). Затем рассчитывался критерий Пирсона, который оказался равным $\chi^2 = 15{,}2$. Критическое значение данного критерия при $\alpha = 0{,}05$ и $\nu = 10$ составляет $\chi^2_{кр} = 18{,}3$. Поскольку $\chi^2 < \chi^2_{кр}$, то делаем вывод, что сеть скважин размещена по территории довольно равномерно.

Таблица 12.1

Размещение скважин по 12 квадратам карты

Хар-ка	1	2	3	4	5	6	7	8	9	10	11	12
Число точек	10	5	5	11	12	6	12	16	15	9	14	8
$\frac{(O-E)^2}{E}$	0,00	2,60	2,60	0,06	0,32	1,73	0,32	3,30	2,26	0,14	1,42	0,48

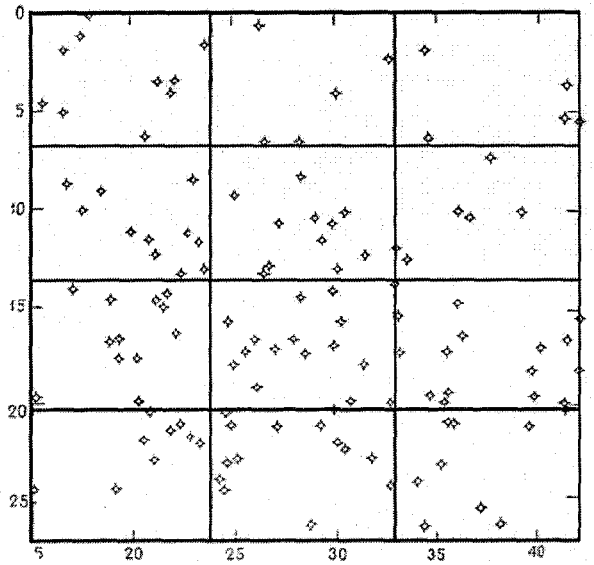


Рис. 12.4. Пространственное распределение 123 артезианских скважин, пробуренных в одном из районов США (по Девису, 1968).

12.4. Понятие о регионализированной переменной

В общем случае любая характеристика, заданная в некоторой пространственной области, может быть представлена суммой трех компонент:

1) дрифтом, описывающим поверхность тренда в определенном направлении,

2) регионализированной переменной, обладающей пространственной корреляцией,

3) случайной переменной, пространственно бессвязной и аналогичной процессу «белого шума» в модели временного ряда.

Нетрудно видеть, что подобное представление довольно хорошо соответствует разложению временного ряда в виде (10.5). Дрифт считается детерминированной переменной и оценивается как полиномиальная функция от пространственных координат в линейном или квадратичном приближении, т.е. эквивалентен тренду временного ряда. Регионализированная переменная занимает некоторое промежуточное положение между полностью случайной

и полностью детерминированной переменными. В отличие от случайной компоненты, регионализированная переменная непрерывна от точки к точке, однако ее изменения настолько сложны, что они не могут быть описаны какой-либо детерминированной функцией. В то же время регионализированная переменная имеет пространственную корреляцию на коротких расстояниях. Если точки расположены далеко друг от друга, то они становятся уже статистически независимыми. Степень пространственной непрерывности регионализованной переменной может быть выражена вариограммой.

Важнейшей характеристикой регионализированной переменной является величина полудисперсии, служащая мерой степени пространственной зависимости между ее отдельными значениями вдоль заданного направления. Для простоты положим, что значения рассматриваемой характеристики равномерно расположены в пространстве вдоль прямых линий, т.е. образуют равномерную сетку. Приняв постоянное расстояние вдоль прямой равным Δ, полудисперсия может быть вычислена для расстояний, кратных Δ, как

$$\gamma_h = (2n)^{-1}\sum_{i=1}^{n-k}(x_i - x_{i+h})^2, \qquad (12.20)$$

где x_i – значение регионализированной переменной, взятой в точке i; x_{i+h} – ее значение, взятое через h интервалов; n – число точек.

Если вычислить значения полудисперсии для различных интервалов h и нанести результаты на график, то полученная кривая, отражающая зависимость γ_h от расстояния, будет называться полувариограммой (рис. 12.5). Очевидно, что полувариограмма является аналогом коррелограммы. Но в отличие от автокорреляционной функции, равной 1 при сдвиге $\tau = 0$, при равным нулю расстоянии значение рассматриваемой характеристики сравнивается с самим собой. В результате все разности равны нулю и величина γ также обращается в нуль.

С увеличением расстояния Δh сравниваемые точки становятся слабее связанными друг с другом, вследствие чего значения γ_h увеличиваются. Если предположить, что на некотором расстоянии сравниваемые точки уже никак не связаны друг с другом, то их квадраты разностей будут равны по величине дисперсии относи-

тельно среднего значения. Полудисперсия более не растет, и полувариограмма начинает колебаться возле некоторого значения, называемого *порогом*. Расстояние, на котором полудисперсия приближается к дисперсии, называется *размахом* (рангом) регионализированной переменной.

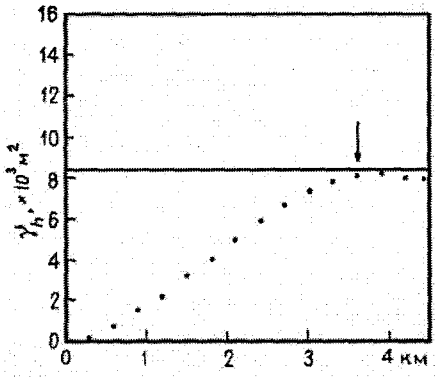


Рис. 12.5. Схематическое представление полувариограммы (по Девису, 1968).

Для произвольной точки в пространстве можно построить симметричный интервал, равный двойной величине размаха. Если регионализированная переменная однородна (всюду имеет одно и то же среднее значение), то любое положение точки вне этого интервала совершенно независимо от центральной точки. И только внутри интервала значения регионализированной переменной в любой точке определенным образом связаны с центральной точкой. Следовательно, они могут быть использованы для оценки центральной точки. Например, в виде весов, которые должны быть приписаны каждому измерению.

Можно показать, что между полудисперсией и другими статистиками (автоковариацией и автокорреляцией) имеется определенная взаимосвязь. Для однородной регионализированной переменной полудисперсия для расстояния Δh равна разности между дисперсией и автоковариацией для того же расстояния, т.е.

$$\gamma_h = R_u(0) - R_u(h). \qquad (12.21)$$

Для стандартизованной однородной регионализированной переменной полувариограмма будет зеркальным отражением автокорреляционной функции (рис. 12.6).

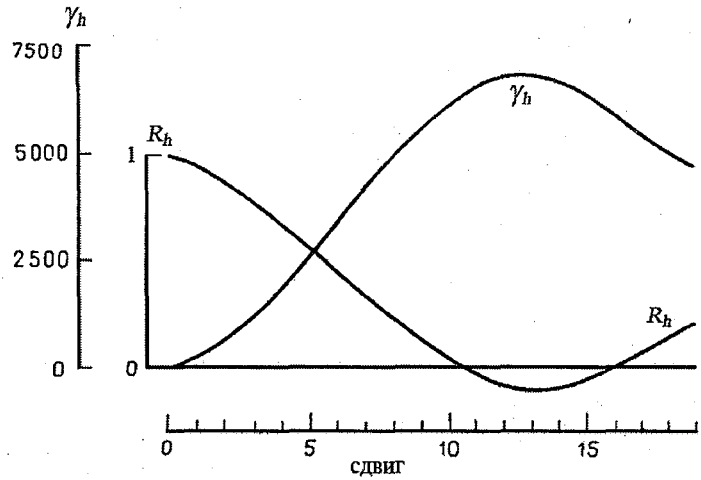


Рис. 12.6. Соотношение между полудисперсией и автокорреляцией для стандартизованной однородной регионализированной переменной.

Для аппроксимации вариограммы используются следующие модели:

1. Сферическая модель:

$$\gamma_h = \sigma^2[(3h/2a) - (h^3/2a^3)] \quad \text{для } h < a,$$
$$\gamma_h = \sigma^2 \quad \text{для } h > a, \qquad (12.22)$$

где σ^2 – дисперсия регионализированной переменной; a – размах регионализированной переменной.

2. Экспоненциальная модель:

$$\gamma_h = \sigma^2(1 - e^{-h/a}). \qquad (12.23)$$

3. Кусочно-линейная модель:

$$\gamma_h = \alpha h \qquad h < a,$$
$$\gamma_h = \sigma^2 \qquad h \geq a, \qquad (12.24)$$

где α – угловой коэффициент.

Заметим, что параметр a, входящий в сферическую и экспоненциальную модели, довольно часто интерпретируется как радиус пространственной корреляции исходных данных.

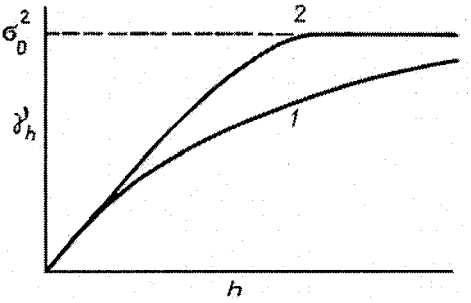


Рис. 12.7. Сравнение экспоненциальной и сферической моделей полувариограммы. *1* – экспоненциальная модель, *2* – сферическая модель.

На рис. 12.7 приводится сравнение первых двух моделей. Нетрудно видеть, что сферическая модель довольно быстро достигает некоторого *порога* (критического расстояния), в то время как экспоненциальная модель никогда не достигает своего предельного значения, а лишь приближается к нему асимптотически. Следовательно, полудисперсия второй модели всегда меньше, чем первой. Отметим, что вблизи начала координат обе модели совпадают и график их имеет вид прямой линии, описываемой кусочно-линейной моделью. Поэтому для расстояний, значительно меньших порога, последняя модель может служить уже хорошим приближением.

Глава 13. МЕТОДЫ АНАЛИЗА СЛУЧАЙНЫХ ПОЛЕЙ

13.1. Построение и анализ карт

Одной из важных практических задач является построение карт в изолиниях. Как известно, изолинии на карте связывают точки равных значений, вследствие чего пространство между двумя последовательными изолиниями содержит только точки, значения которых попадают внутрь интервала, определенного этими изолиниями. Карты служат обобщением эмпирической информации или результатов теоретических расчетов. Понятно, что практическая значимость их не вызывает сомнения. В то же время карты, особенно объединенные какой-либо тематической идеей, могут представлять значительную научную ценность. В частности, «Атлас океанов», содержащий несколько тысяч самых разнообразных карт о физических, химических, метеорологических, геоформологических, биологических и иных полях в океанах и в атмосфере над ними, опубликованный в начале 1980-х годов, еще на многие десятилетия останется фундаментальным научным трудом.

Построение карт включает в себя несколько этапов:

– оценка репрезентативности данных;

– выбор гипсометрической основы и масштаба карты;

– выбор интервала между изолиниями;

– проведение изолиний.

Оценка репрезентативности данных, по существу, означает их первичный анализ, который рассмотрен в первой части учебника. Дополнительно оценивается степень репрезентативности самих станций, что особенно важно при построении метеорологических карт. Речь идет о характере ландшафта, где расположена станция. Например, при картировании элементов климата вряд ли можно пользоваться данными станции, находящейся в пределах крупного города. Известно, что города служат дополнительными очагами тепла. Так, количество тепла, выделяемое крупными промышленными городами, по некоторым оценкам может даже превышать естественный приток от солнца. Особенно значительными различия в температуре в городе и вне его пределов оказываются в зим-

ний период. Например, в Москве эти различия зимой в отдельные дни могут достигать 8–10 °С, среднемесячные значения составляют порядка 4 °С, а средние годовые температуры внутри города выше более чем на 1 °С.

Приведем другой пример. Если станция расположена в узкой долине, то в зависимости от преобладающего направления ветра ее данные (характеристики ветра, осадков, температуры и др.) могут существенно отличаться от аналогичных данных, измеренных вне этой долины. Следовательно, данные указанной станции отражают локальные (мезомасштабные) условия колебаний метеорологических характеристик и при построении климатических карт они могут внести серьезные искажения.

При построении океанологических карт или метеорологических карт над океанами основные проблемы заключаются в катастрофическом отсутствии данных. Кроме того, при построении метеорологических карт следует принимать во внимание искажение характеристик вблизи берегов крупных островов и материков. Известно, что, например, количество осадков, выпадающих в прибрежной зоне, обычно существенно выше, чем над открытой акваторией.

Что касается выбора масштаба карты, то для этого можно использовать следующее соотношение:

$$M = \delta/\delta_k, \tag{13.1}$$

где δ_k – предельно допустимый градиент на карте в единицах измерения картируемой величины на 1 мм расстояния; δ – наибольший пространственный градиент на местности, выраженный в тех же единицах.

Примем в качестве допустимой ошибки, порождаемой картографическим процессом, 1/3 средней квадратической ошибки σ картируемой величины и будем считать суммарную среднюю квадратическую погрешность, обусловленную погрешностями оформления издательских оригиналов, неточности снятия данных с карты и др., близкой к 0,6 мм. Тогда формула (13.1) примет вид:

$$M = \delta/0{,}55\sigma. \tag{13.2}$$

Выбор интервала между изолиниями определяется исходя из формальных и неформальных соображений. Интервал между изо-

линиями, с одной стороны, должен обеспечивать необходимую точность карты, а с другой – должен быть увязан с избранным масштабом карты. При большом интервале между изолиниями (недостаточном количестве изолиний) возникает необходимость сложной графической интерполяции при снятии значений в каждой заданной точке карты, что приводит к потере точности. При избыточном количестве изолиний они в отдельных местах карты могут сливаться, вследствие чего создается неверное впечатление о большой точности карты.

При выборе интервала следует исходить из того, что изолинии должны проводиться так, чтобы погрешности, возникающие при интерполяции между ними, не сказывались на точности получаемых результатов. Поэтому интервал выбирается с таким расчетом, чтобы колебания картируемой характеристики по отдельным станциям укладывались между изолиниями, не превышая двойной средней квадратической ошибки его средних значений.

Проведение изолиний может осуществляться вручную, как это делалось ранее, или с помощью ЭВМ. Например, в течение многих десятилетий синоптические карты погоды непосредственно строились синоптиками. Затем эта работа была «поручена» ЭВМ. Сравнение синоптических карт, полученных автоматизированным и ручным способами, показало, что в некоторых случаях благодаря опыту и знаниям синоптик может провести изолинии на неосвещенной измерениями территории более правильно, чем ЭВМ. В то же время синоптик может и серьезно ошибиться. Самое интересное заключается в том, что в результате сравнения нескольких сотен синоптических карт, построенных ЭВМ и вручную, оказалось, что усредненные карты почти совпали друг с другом. Это означает, что ЭВМ вела себя как «усредненный» синоптик.

Естественно, что построение изолиний ЭВМ происходит на основе каких-либо геометрических соотношений, причем далеко не всегда они вытекают из строгих теоретических закономерностей, а могут основываться на соображениях здравого смысла и практического опыта. Поэтому те допущения, которые заложены в расчетную схему, переносятся на результаты картирования.

Кроме того, любая расчетная схема должна предусматривать непрерывность картируемой характеристики всюду в пределах данного пространства, однозначность ее в любой точке и коррели-

рованность на расстоянии, большем типичного расстояния между исходными точками. Последнее условие означает, что в соседних точках картируемая характеристика должна иметь близкие значения, вследствие чего построенные изолинии будут носить более закономерный и плавный характер. И, наоборот, при малой коррелированности значения в соседних точках начинают сильно различаться, поэтому проведение изолиний в значительной степени носит уже случайный характер и имеет много локальных, зачастую плохо интерпретируемых экстремумов.

По-видимому, универсального алгоритма построения изолиний, пригодного для любых условий, просто не существует. В зависимости от характера исходных данных, их точности, размещения точек на карте, рельефа местности и некоторых других факторов результаты картирования могут существенно различаться. Достаточно просто выполнять построение изолиний в рамках статистико-графического пакета «Серфер», одна из последних версий которого (8-я версия) содержит 12 различных методов интерполяции и проведения изолиний. Причем определить априори, какой из них лучше практически невозможно.

13.2. Пространственное осреднение гидрометеорологических данных

На первый взгляд, задача оценки пространственной средней относится к числу рутинных операций. Однако это не совсем так. Поскольку случайные поля обычно заданы на нерегулярной сети точек, то задача их пространственного осреднения в отличие от временного осреднения носит уже нетривиальный характер. Действительно, далеко не всегда простое арифметическое осреднение позволяет получить корректную оценку пространственной средней. Предположим, нам требуется выполнить осреднение значений осадков для предгорной части Кавказа. На равнинной части этой территории осадкомерная сеть станций существенно более плотная, чем в предгорной части. Однако хорошо известно, что вследствие топографического эффекта в предгорной части осадков выпадает заметно больше, чем на равнине. По существу, это означает не выполнимость условий однородности и изотропности. Поэтому простое арифметическое осреднение по существенно неравномерной сети приведет к явному занижению пространственной средней.

Можно привести также и другой пример, связанный с определением величины осадков над водными акваториями. Известно, что над открытыми водными акваториями осадков выпадает существенно меньше, чем над крупными островами и прибрежными районами. Поэтому пространственное осреднение осадков, например по акватории Каспийского моря, по данным островных и береговых станций неминуемо приводит к существенным искажениям осредненной величины осадков.

Весьма важной данная задача оказывается также в связи с внедрением в практику спутниковых и радиолокационных наблюдений. Как известно, получаемые с их помощью гидрометеорологические характеристики уже изначально являются осредненными по пространству. Для идентификации дистанционных параметров приходится решать обратные задачи. Простейшая из них – это сравнение пространственных средних, полученных по наземной сети станций и дистанционными методами.

Если случайные поля имеют ансамбль (множество) реализаций, то это позволяет решить задачу оценки пространственной средней при условии однородности и изотропности на основе известной статистической структуры поля методом оптимальной интерполяции.

Однако во многих случаях производить осреднение приходится по единственной реализации и тогда решение данной задачи с помощью оптимальной интерполяции уже невозможно. В настоящее время известно значительное число методов, реализующих процедуру пространственного осреднения исходных данных, однако почти все они могут быть использованы при выполнении условий однородности, изотропности и эргодичности случайного поля, принимаемых обычно априори.

Как известно, среднее арифметическое для некоторой пространственной области S определяется следующим образом:

$$< f >= \frac{1}{S} \iint\limits_{(S)} f dx dy \,. \qquad (13.3)$$

При достаточно густом равномерном распределении точек обычно ограничиваются пространственным осреднением вида:

$$< f >= \frac{1}{n} \sum_{i=1}^{n} f_i , \qquad (13.4)$$

где n – число точек.

Естественно, как уже отмечалось выше, при неравномерном распределении точек и невыполнении условия однородности и изотропности данная формула может приводить к серьезным погрешностям. Например, использовав данные по среднемесячным осадкам в выделенных узлах на рис. 13.1, получим по формуле (13.4) $<P>$ = 52 мм/мес. Если же мы предварительно проведем изолинии осадков, например, с помощью рассмотренного ниже метода конечных элементов, а затем снимем их значения в узлах регулярной квадратной сетки и выполним осреднение, то имеем $<P>$ = 60 мм/мес. Различие между указанными оценками осадков является значимым при уровне значимости $\alpha = 0{,}05$.

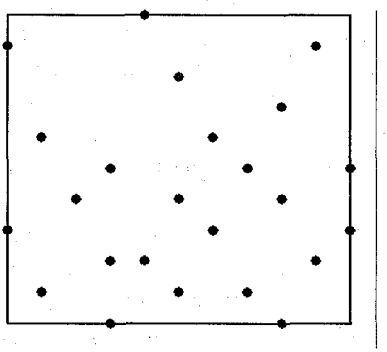


Рис. 13.1. Пространственное распределение осадкомерной сети станций.

К числу наиболее простых методов осреднения относится *метод квадратов*. Суть его состоит в следующем. Вся рассматриваемая территория (акватория) делится на определенное число квадратов (рис. 13.2) таким образом, чтобы число пустых квадратов было бы минимальным. Для каждого квадрата путем простого арифметического осреднения находится локальная (частная) пространственная средняя. Для «пустых» квадратов, т.е. там, где отсутствуют данные наблюдений, для определения средней могут привлекаться данные соседних квадратов. После этого путем арифметического осреднения частных средних определяется средняя по всей территории. Если число точек в квадратах существен-

но различается, то для повышения точности следует вводить весовые множители, учитывающие их число.

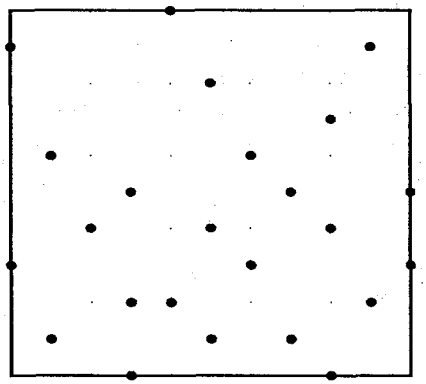


Рис. 13.2. Осреднение по площади методом квадратов.
Точки – пункты наблюдений.

Примем, что рассматриваемая область на рис. 13.1 разбита на M квадратов, в каждом из которых находится n точек наблюдения. Общее число точек составляет N. Тогда общая пространственная средняя рассчитывается как

$$< f > = \sum_{i=1}^{N} f_i p_i ,$$

где $p_i = 1/n_i M$, причем $N = M\sum n_i$.

Безусловным достоинством метода квадратов является его простота. К очевидным недостаткам относятся неоднозначность разбиения площади на квадраты и неравномерное освещение квадратов данными.

Заметим, что данный принцип был использован при обработке гидрометеорологических наблюдений попутных (коммерческих) судов в Северной Атлантике. Как известно, 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. В ВНИГМИ-МЦД бывшего СССР эти данные обрабатывались. При этом вся акватория Северной Атлантики была разделена на пятиградусные «квадраты», границами которых являлись широты и долготы, кратные пяти. По координатам судна гидрометеорологические данные относились к тому или иному

квадрату. Накопленные за месяц наблюдения усреднялись и затем по прошествии годового интервала времени публиковались атласы. Таким образом, начиная с 1977 г. в течение более чем 10 лет была издана серия атласов и справочников с характеристиками крупномасштабного взаимодействия океана и атмосферы, которые затем активно использовались в научных и практических целях.

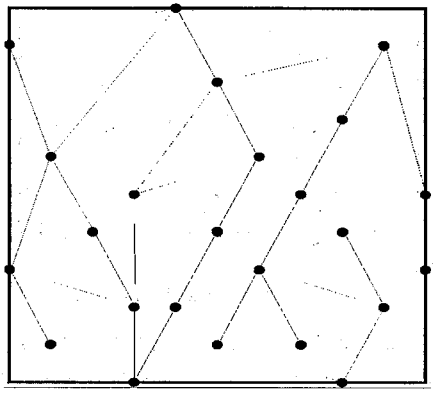


Рис. 13.3. Осреднение по площади методом треугольников.
Точки – пункты наблюдений.

Другой достаточно простой способ осреднения – это *метод треугольников*. В соответствии с этим методом рассматриваемая область разбивается на треугольники, в вершинах каждого из которых расположены станции (рис. 13.3). Полагая линейный характер изменения элемента между вершинами треугольника, находим значение его в центре тяжести треугольника, который представляет среднее арифметическое от данных в вершинах треугольника:

$$f_i = \frac{1}{3}\sum_{j=1}^{3} f_j \,. \qquad (13.5)$$

Отметим, что центр тяжести находится на пересечении медиан треугольника. После этого производится осреднение с весами, пропорциональными площадям треугольников (всего m треугольников):

$$< f >= \frac{1}{S}\sum_{i=1}^{m} f_i S_i \,, \qquad (13.6)$$

где S_i – площадь треугольника; S – площадь пространственной области осреднения.

Данный метод имеет те же достоинства и недостатки, что и метод квадратов. Сравнение оценок пространственной средней по обоим методам показывает, что, как правило, расхождения невелики.

В качестве обобщения данного метода можно рассматривать *метод конечных элементов*, в соответствии с которым вначале производится разбиение всей площади на треугольники; затем внутри треугольника может быть произведена интерполяция в любую точку треугольника по данным из вершин (из окружающих станций). Интерполяционная формула при этом имеет вид:

$$f = N_i f_i + N_j f_j + N_k f_k , \qquad (13.7)$$

где $N_i = [(x_j y_k - x_k y_j) + (y_j - y_k)x + (x_k - x_j)y] / 2S,$

$N_j = [(x_k y_i - x_i y_k) + (y_k - y_i)x + (x_i - x_k)y] / 2S,$

$N_k = [(x_i y_j - x_j y_i) + (y_i - y_j)x + (x_j - x_i)y] / 2S.$

Здесь i, j, k – вершины треугольника; x, y – координаты интерполируемого узла; S – площадь треугольника.

Данная интерполяционная формула обеспечивает гладкость поля при переходе от одной подобласти к другой. Поэтому она легко может быть использована для получения данных в регулярной сетке. Недостаток – определенный произвол в разбиении области на треугольники. Следует помнить, что точность любого способа разбиения на треугольники зависит от соотношения сторон. Наибольшая точность достигается для равносторонних треугольников. Используя формулу (13.7), осуществляем интерполяцию в узлы регулярной сетки, а далее выполняем обычное арифметическое осреднение. На наш взгляд, именно предварительная интерполяция в узлы регулярной сетки с последующим арифметическим осреднением является наиболее оптимальным вариантом нахождения пространственной средней.

Пример 13.1. Рассмотрим модельный пример. По изначально густой сети станций в узлы регулярной квадратной сетки интерполировались значения осадков. Затем случайным образом сеть была разрежена (см. рис. 13.1). Далее данная территория была разбита

на 25 квадратов и 35 треугольников. После этого рассчитывалась указанными выше методами пространственная средняя осадков. В результате были получены следующие оценки:

1. Метод квадратов: $<P> = 55$ мм/мес.
2. Метод треугольников: $<P> = 57$ мм/мес.
3. Метод конечных элементов: $<P> = 60$ мм/мес.
4. Среднее арифметическое по 24 точкам: $<P> = 52$ мм/мес.
5. Среднее арифметическое по всем узлам квадрата: $<P> = 62$ мм/мес.

Итак, практически все методы осреднения несколько занизили величину пространственной средней. Естественно, возникает вопрос, насколько значимы эти расхождения? Использование критерия Стьюдента показало, что значимыми при $\alpha = 0{,}05$ являются расхождения только между 1 и 5, 3 и 4, 4 и 5 методами осреднения. На наш взгляд, наиболее точными представляются результаты осреднения методом конечных элементов.

13.3. Построение и анализ кросскорреляционной функции

Как уже отмечалось выше, кросскорреляционная функция принципиально отличается от автокорреляционной и чисто пространственной корреляционной функций, ибо синтезирует в себе их свойства, вместе взятые. Действительно, АФ рассчитывается по временной реализации характеристики в 1 точке поля, а пространственная корреляционная функция – по реализации случайного поля в заданный момент времени.

Естественно, что построение кросскорреляционной функции сочетает в себе элементы расчета той и другой функций. При расчете кросскорреляционной функции весьма важным является выполнимость условия однородности и изотропности случайного поля, а также нормального закона распределения для временных реализаций рассматриваемой характеристики в каждой точке пространства. Последнее условие связано с параметрическим характером коэффициента корреляции.

До построения кросскорреляционной функции необходимо проверить выполнимость условия однородности и изотропности случайного поля. Самый простой способ – это анализ изокоррелят,

построенных относительно различных пунктов – центров корреляции. Если изокорреляты представляют собой концентрические окружности, то рассматриваемое поле можно считать однородным и изотропным. Естественно данный вывод будет справедлив только относительно той точки, для которой строилось поле изокоррелят. В принципе, при переходе к другой точке пространства все расчеты необходимо повторять. И таких «повторов» может оказаться много. Кроме того, для точек вблизи края карты построение изокоррелят становится вообще невозможным.

Поэтому в ГГО был разработан метод, который позволяет строить лишь одну карту изокоррелят для всего поля. В соответствии с ним одновременно выполняется осреднение коэффициентов корреляции как по градациям расстояния, так и в зависимости от направления между станциями, которое отсчитывается от параллели или меридиана. После этого строится единое поле изокоррелят. Таким образом, осуществляется учет векторного характера поля. Другое преимущество указанного метода состоит в том, что он сглаживает несущественные мелкомасштабные детали, в результате чего изолинии носят более закономерный характер.

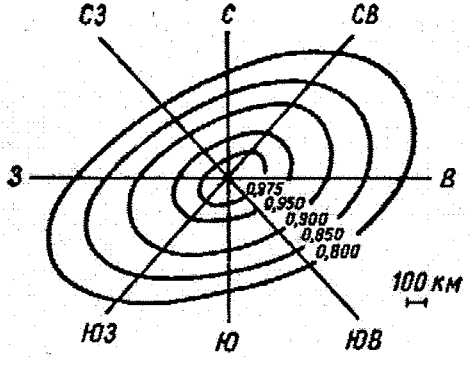


Рис. 13.4. Карта изокоррелят средней месячной температуры воздуха в феврале на Европейской территории России (ЕТР).

На рис. 13.4 приводится карта изокоррелят средней февральской температуры воздуха для центральной части Европейской территории России, полученная по данным 172 метеорологических станций. Нетрудно видеть, что поле изокоррелят имеет вид эллипсов, ориентированных вдоль преобладающего переноса воздуш-

ных масс, т.е. поле февральской температуры является анизотропным. Правда, на малых расстояниях анизотропия проявляется слабо и только с их увеличением она усиливается. Отметим, что в поле аномалий температуры воздуха анизотропия уже практически не выражена.

После анализа поля изокоррелят переходят к построению кросскорреляционной функции, процедура расчета которой обычно сводится к следующему. Для всех пар точек рассчитываются выборочные оценки коэффициента корреляции. Далее составляется пространственный вариационный ряд в порядке возрастания расстояния между станциями. Естественно, чем длиннее полученный ряд, тем надежнее будут результаты оценки. Вариационный ряд коэффициентов корреляции разбивается на градации, например по формуле Стерджесса $k \approx 1 + 3{,}32\lg n$. Затем определяется ширина градаций. Заметим, что данная формула служит только ориентиром выбора числа градаций, которое обычно корректируется так, чтобы получить более «круглое» значение ширины градации. После этого для каждой градации осуществляется осреднение коэффициентов корреляции. Далее строится график зависимости средних коэффициентов корреляции от расстояния. Для полученной таким образом корреляционной функции подбирается та или иная аппроксимационная формула.

Пример 13.2. Обратимся к рис. 13.5, на котором приводятся графики кросскорреляционных функций температуры воздуха для лета Европейской территории России при различных периодах осреднения, построенных в предположении однородности и изотропности рассматриваемого поля. Нетрудно видеть, что с увеличением периода осреднения пространственная связность температуры воздуха возрастает. При всех масштабах осреднения высокая корреляция ($r(l) > 0{,}6$) сохраняется на довольно значительных расстояниях. Так, для суточного периода $r(l) > 0{,}6$ соответствует расстоянию $l \approx 500$ км, а для месяца – $l \approx 900$–1000 км. Вне зависимости от периода осреднения все корреляционные функции могут быть аппроксимированы выражением вида:

$$r(l) = r(0)\exp(-l/l_0)^{\alpha},$$

где $r(0)$ – значение корреляции, получаемое экстраполяцией корреляционной функции $r(l)$ на нулевой сдвиг $l = 0$; l_0 – радиус кор-

реляции, за который принимается расстояние, на котором $r(0)$ убывает в e раз; α – подгоночный коэффициент.

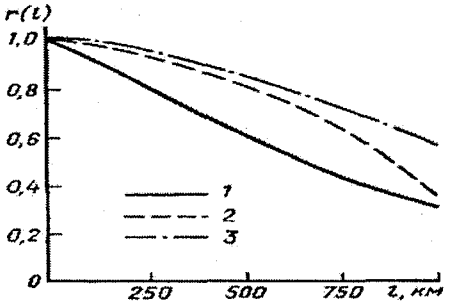


Рис. 13.5. Кросскорреляционная функция температуры воздуха для лета на ЕТР при различных периодах осреднения:

1 – сутки, *2* – декада, *3* – месяц.

Отметим, что в отличие от автокорреляционной функции величина $r(0)$ обычно меньше единицы. Например, для месячных значений температуры воздуха в соответствии с рис. 13.5 величина $r(0) = 0{,}99$.

13.4. Понятие объективного анализа

Как уже отмечалось выше, в узком смысле, объективный анализ (ОА) – это процедура перевода данных из нерегулярной сети точек в регулярную сеть. Однако в метеорологии объективный анализ обычно понимается более широко и включает в себя комплекс автоматических методов анализа информации в целях численного моделирования и прогноза. При этом основными задачами ОА являются:

1) исключение грубых ошибок наблюдений,
2) автоматизация сбора и подготовки информации,
3) интерполяция данных в узлы регулярной сетки,
4) согласование полей характеристик,
5) построение пространственных карт,
6) построение четырехмерных полей данных.

Задачи ОА в океанологии и экологии отличаются от метеорологических, так как, по существу, отсутствует постоянная сеть станций мониторинга. В связи с этим ОА в этих науках понимается только в узком смысле.

В общем случае восстановление поля любого элемента производится в четырехмерном пространстве. Если обозначить через r радиус-вектор точки такого пространства, то тогда интерполяция элемента g в точку r_0 производится по известным значениям в точках $r_1, r_2, ..., r_n$ по следующей формуле:

$$g^*(r_0) = F\{g(r_1), g(r_2), ..., g(r_n)\}. \qquad (13.8)$$

Вид функции F определяется способом интерполяции, причем в некоторых случаях она может зависеть от взаимного расположения точек $r_1, r_2, ..., r_n$. Естественно, что результат интерполяции в точке r_0 обычно отличается от его исходного значения $g(r_0)$, вследствие чего возникает ошибка интерполяций

$$\delta(r_0) = g^*(r_0) - g(r_0).$$

Обычно функцию интерполяции представляют в виде линейной функции своих аргументов:

$$g^*(r_0) = \sum_{i=1}^{n} a_i g(r_i), \qquad (13.9)$$

где a_i – веса, связанные с системой расположения точек в пространстве.

Данное представление называется линейной интерполяцией, поскольку описывает характеристику в виде некоторой ее линейной комбинации. Вследствие своей простоты методы линейной интерполяции получили довольно широкое распространение на практике.

Полиномиальная интерполяция. Одним из первых методов ОА, получивших распространение в США в 50-е годы прошлого столетия при численном анализе метеорологических полей, был метод полиномиальной интерполяции. Суть его заключается в следующем. В предположении однородности и изотропности исходное поле принимается как полиномиальная функция от пространственных координат. Тогда разложение поля по некоторой базисной системе координат можно представить как

$$f = \sum_{k=1}^{m} b_k \varphi(r), \qquad (13.10)$$

где $\varphi(r)$ – известная функция координат; b_k – коэффициенты разложения, которые подлежат определению.

Если функция координат аппроксимируется, например, полиномом второй степени, то имеем:

$$f(r_0) = f(x, y) = b_1 + b_2 x_i + b_3 y_i + b_4 x_i^2 + b_5 y_i^2 + b_6 x_i y_i, \qquad (13.11)$$

где x_i, y_i – координаты i-й станции.

Отсюда видно, чтобы получить систему из шести уравнений с шестью неизвестными, должно быть известно значение элемента не менее чем в шести точках. Решая систему линейных уравнений каким-либо численным методом, получаем оценки неизвестных коэффициентов b_1, b_2,..., b_6. Это позволяет вычислить значение рассматриваемого элемента в заданном узле интерполяции. Аналогичным образом для каждого узла регулярной сети точек составляется своя система уравнений. Естественно, что данный метод может быть применен только в случае густой и достаточно равномерной сети станций наблюдений.

Недостатки:

1. Недостаточная точность интерполяции при неравномерном распределении точек в пространстве.

2. В случае изменения системы точек или отсутствия наблюдений на некоторых станциях необходим пересчет коэффициентов.

Метод полиномиальной регрессии. Данный метод является обобщением предыдущего. Он практически полностью дублирует двухмерную полиномиальную регрессию (п. 8.5). Запишем его до второго порядка:

$$G(x, y) = \sum\sum a_{ij} x^i y^j + \varepsilon =$$

$$= a_0 + a_{10} x + a_{01} y + a_{20} x^2 + a_{02} y^2 + a_{11} xy + \varepsilon. \qquad (13.12)$$

Неизвестные коэффициенты в данной формуле a_{ij} могут определяться разными способами. Укажем здесь способ, несколько отличный от приведенного в п. 8.5. Если осуществить замену переменных: $z_1 = x^2$, $z_2 = y^2$, $z_3 = xy$, то получим уравнение классической множественной линейной регрессии:

$$G(x, y) = a_0 + a_{10}x + a_{01}y + a_{20}z_1 + a_{02}z_2 + a_{11}z_3 + \varepsilon.$$

С помощью МНК нетрудно определить коэффициенты регрессии. После этого задавая значения пространственных координат x и y в узлах регулярной сетки, рассчитываются оценки параметра G.

Рассчитав коэффициенты статистической регрессии, затем осуществляется интерполяция в любую точку области. Отличие ее от метода математической полиномиальной интерполяции состоит в том, что вычисляется лишь одно уравнение, с помощью которого можно сразу же вычислить значения интерполируемой величины для всей пространственной области. Кроме того, данный метод позволяет провести детальную статистическую оценку точности результатов интерполяции. Наконец, с помощью использования пошаговых алгоритмов можно оптимизировать получение «наилучшего» уравнения регрессии.

Пример 13.3. Рассмотрим задачу объективного анализа температуры поверхности океана в Северной Атлантике. На рис. 13.6 представлено поле ТПО в узлах географической сетки 5° широты на 10° долготы для сентября 1991 г. Общая длина выборки составила $N = 64$ точки. Нетрудно видеть, что в распределении изолиний ТПО достаточно четко прослеживается широтная зональность, т.е. постепенное уменьшение ТПО от низких широт к высоким. Это означает, что для интерполяции можно применить метод двухмерной полиномиальной регрессии.

Разобьем акваторию Северной Атлантики на две примерно равноценные части. Для одной из них (крестики на рис. 13.7) будем считать, что значения ТПО в ней известны. А для другой части (точки на рис. 13.7), наоборот, значения ТПО неизвестны. Таким образом, первая – это зависимая выборка, длина которой составляет $n_1 = 33$ значения ТПО, а другая – независимая выборка длиной $n_2 = 31$. Присвоим западной долготе знак минус и рассчитаем для зависимой выборки уравнение второй степени:

$$\text{ТПО} = 8{,}869 + 0{,}418\varphi - 0{,}638\lambda - 0{,}007\varphi^2 - 0{,}004\lambda^2 + 0{,}010\varphi\lambda.$$

Стандартизованное уравнение регрессии будет иметь вид:

$$Z_{\text{ТПО}} = 0{,}89z_1 - 1{,}74z_2 - 1{,}11z_3 - 0{,}78z_4 + 1{,}15z_5.$$

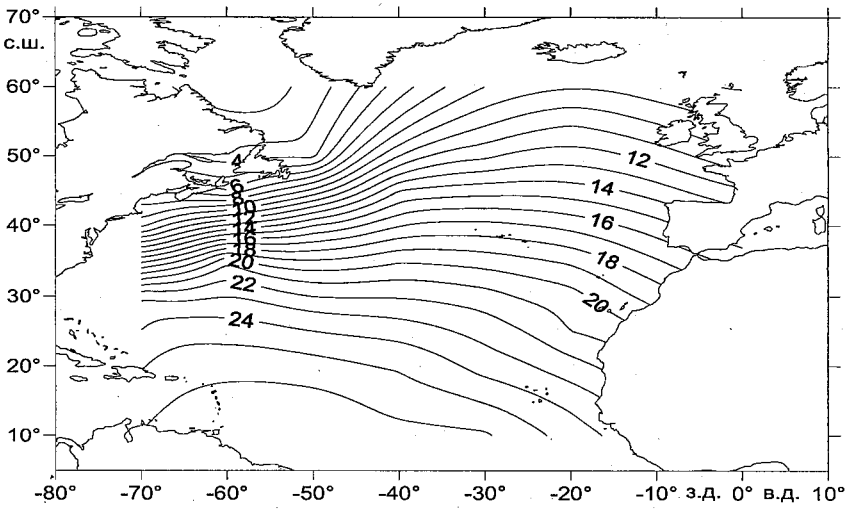


Рис. 13.6. Пространственное распределение температуры поверхности океана в сентябре 1991 г. на акватории Северной Атлантики

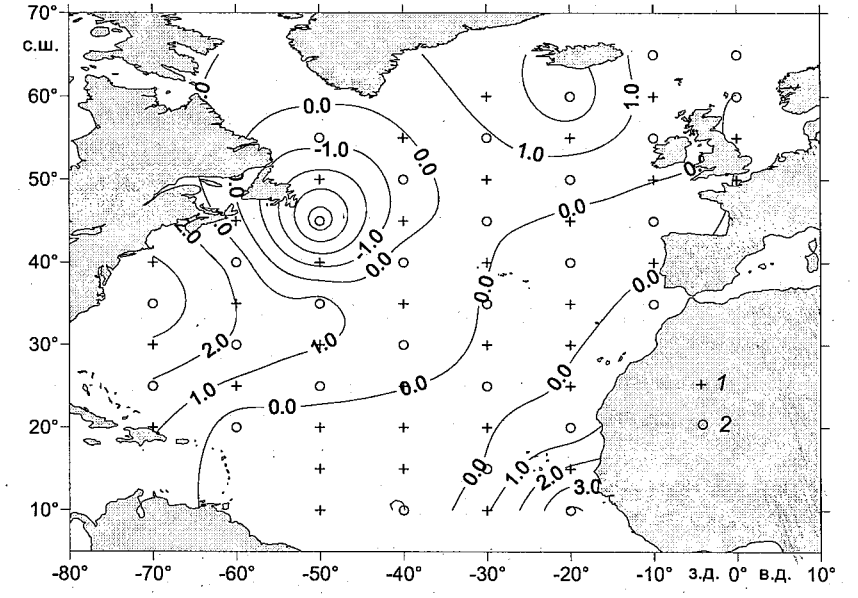


Рис. 13.7. Пространственное распределение ошибок восстановления температуры поверхности океана (°С) в сентябре 1991 г. на акватории Северной Атлантики: *1* – точки зависимой выборки; *2* – точки независимой выборки.

В этом уравнении z_1 соответствует переменной φ, z_2 – переменной λ и т.д. Итак, мы видим, что наибольший вклад в описание изменчивости функции отклика принадлежит λ, следующий – произведению φλ и т.д. Коэффициент детерминации данной модели равен $R^2 = 0{,}94$, критерий Фишера $F = 84{,}9$, стандартная ошибка модели $\sigma_{y(x)} = 1{,}85$ °С, а максимальное значение критерия *p-level*, которое принадлежит свободному члену, составляет $p\text{-}level_{(b0)} = 0{,}06$. Следовательно, модель регрессии в статистическом смысле является очень «хорошей». Даже довольно большая ошибка $\sigma_{y(x)}$ по отношению к стандартному отклонению ТПО зависимой выборки ($\sigma_{ТПО} = 6{,}95$ °С) оказывается малой величиной. Следовательно, можно ожидать, что точность интерполяции для независимой выборки должна быть весьма высокой.

Действительно, для большинства сеточных узлов (24 из 31) ошибка интерполяции оказалась меньше средней квадратической ошибки, которая для независимой выборки равна $\sigma_{y(x)} = 1{,}60$ °С, т.е. даже ниже, чем для зависимой выборки. Пространственное распределение ошибок также приводится на рис. 13.7. Как и следовало ожидать, максимальные ошибки интерполяции отмечаются в наиболее мощной фронтальной зоне, находящейся в дельте Гольфстрима, где на это течение натекает холодное Лабрадорское течение. Ошибка интерполяции в точке 45° с.ш. и 50° з.д. превышает 4 °С. Отметим, что данный очаг максимальных ошибок является стационарным, т.е. он существует во все сезоны и для любого года. Наличие его связано с тем, что гидрометеорологические процессы здесь носят подсеточный характер. Другими словами, масштаб их существенно меньше по сравнению с масштабом заданной географической сетки. Вследствие этого избежать больших ошибок интерполяции при заданной географической сетке вряд ли возможно, к каким бы совершенным методам мы не прибегали. Второй очаг существенных ошибок находится на крайнем юго-востоке вблизи побережья Африки в зоне действия Гвинейского течения. Скорее всего, его существование носит локальный характер, поскольку для других карт ТПО он проявляется далеко не всегда. Итак, можно считать, что в целом интерполяция ТПО для сентября 1991 г. методом полиномиальной регрессии является успешной.

13.5. Использование статистического пакета «Сёрфер» (Serfer) для объективного анализа полей

Одним из наиболее простых и доступных способов двухмерной интерполяции является использование пакета «Сёрфер», который позволяет производить интерполяцию данных, заданных в произвольной сети точек в узлы регулярной сетки и осуществлять проведение изолиний. При этом число используемых алгоритмов зависит от версии пакета. Так, в седьмой версии таких алгоритмов 9, а в восьмой – 12. Естественно, что получаемые результаты могут очень сильно зависеть от метода интерполяции. Поэтому очень кратко перечислим характерные особенности методов интерполяции восьмой версии пакета «Сёрфер».

1. Метод обратных расстояний (Inverse Distance to a Power) позволяет производить последовательную интерполяцию данных от одной точки к другой относительно узлов сетки. В результате происходит сглаживание истинных значений, и чем дальше удалено значение от узла сетки, тем больше оно искажается. С помощью данного метода осуществляется взвешивание расстояния, так что влияние одной точки относительно другой уменьшается с увеличением расстояния от узла сетки. Вычисляя узел сетки, всем точкам присваиваются веса. Когда конкретное наблюдение совпадает с узлом сетки, расстояние между наблюдением и узлом сетки равно 0,0, поэтому такому наблюдению присваивается вес равный 1,0 (в данном случае значение является узлом сетки), в то время как всем другим наблюдениям назначаются веса 0,0. Поэтому метод имеет тенденцию интерполировать данные таким образом, что вокруг узла образуются концентрические изолинии. Метод не экстраполирует значения вне диапазона данных.

2. Метод крайкинга (Kriging) – один из наиболее гибких методов, который может использоваться для любого набора данных, причем он наиболее эффективен для линейно изменяющихся данных. С его помощью осуществляется построение визуально привлекательных карт даже для весьма нерегулярных полей данных. Метод крайкинга позволяет связать экстремальные значения между собой, а не изолировать их контурами от остальных значений. Применяется для создания точной сетки данных. Имеет оптималь-

ные статистические свойства и может экстраполировать значения вне диапазона данных.

3. Метод минимальной кривизны (Minimum Curvature) формирует плавные изолинии, но может создать искажение (завышать, занижать) значений в областях, где данные отсутствуют. Искажение данных не превышает величины крайних значений ряда (максимального и минимального). С помощью метода интерполируются данные с минимальным искривлением, что способствует сохранению структуры данных. В основе метода лежит взвешенный метод наименьших квадратов. Данные могут экстраполироваться за пределами заданных значений.

4. Модифицированный метод Шеппарда (Modified Shepard's Method) подобен методу обратных расстояний. Используется метод наименьших квадратов, что предотвращает появление замкнутых изолиний и позволяет объединять соседние значения при минимальном расстоянии между данными. Можно экстраполировать значения вне диапазона данных.

5. Метод дальнего соседа (Natural Neighbor) позволяет хорошо интерполировать только значения с полным набором данных, без пропусков в ряду, не экстраполируются пропуски в данных. В результате интерполяции вся область данных разбивается на многоугольники. Изолинии проводятся с учетом взвешенного среднего числа соседних наблюдений через центры многоугольников.

6. Метод ближайшего соседа (Nearest Neighbor) позволяет приспособить данные наблюдений к узлам регулярной сетки. Данный метод используется для полного набора данных.

7. Метод полиномиальной регрессии (Polynomial Regression) отображает основные, ярко выраженные тенденции в ряду, не учитывая внутренние локальные особенности. Позволяет заполнить пропуски в наблюдениях.

8. Метод радиальных базисных функций (Radial Basis Functions) наиболее гибкий метод, результаты подобны методу Крайкинга, подходит для большого набора данных. В методе определяется оптимальный набор весов так, чтобы значения ряда интерполировались в узел сетки.

9. Метод линейной интерполяции (Triangulation with Linear Interpolation). Триангуляция, т.е. разбиение на треугольники с линейной интерполяцией, эффективно для небольшого набора дан-

ных. Создаются хорошие треугольные образы между точками данных. Не интерполируются данные вне диапазона данных. Алгоритм метода позволяет строить треугольники, вершинами которых служат исходные точки, причем узлы сетки оказываются внутри каждого треугольника. Кроме того, точки связываются так, что грани треугольников не пересекаются между собой. Лучшие результаты получаются при относительно равномерном распределении точек по области сетки.

10. Метод скользящего среднего (Moving Average) применим к очень большим рядам наблюдений (> 1000 наблюдений). Метод извлекает промежуточные, скрытые тенденции в ряду.

11. Метод меры данных (Data Metrics) используется, чтобы получить информацию о данных в узлах сетки.

12. Локальный полиномиальный метод (Local Polynomial) применимым к монотонным, линейно изменяющимся наборам данных, без резких перепадов в структуре ряда. В основе алгоритма лежит метод взвешенных наименьших квадратов, который стремиться присвоить интерполированным данным значения в узлах сетки. Вычислительная скорость метода не зависит от размера данных.

Безусловно, каждый из перечисленных методов обладает как определенными достоинствами, так и недостатками. Априори нельзя сказать, какой из них лучше, а какой хуже. Просто в зависимости от специфики исходных данных лучшие (худшие) результаты интерполяции могут относиться к любому из них. Указанные методы являются формальными в том смысле, что не учитывают специфические особенности исходного поля. Практически все они дают наилучший результат для нормально распределенных пространственных данных. Если же распределение их отличается от нормального, то желательно предварительное приведение их к нормальному виду.

Кроме того, многими методами интерполяции принимается равномерное изменение рассматриваемой характеристики от точки к точке. В действительности, это выполняется далеко не всегда. Так, на территории суши она может быть подвержена влиянию локальных условий: особенностям орографии, степени залесенности, наличию водоемов и т.п. В океане локальное влияние на гидрометеорологические характеристики оказывают острова, побережья, фронтальные зоны, холодные (теплые) течения и др.

Таким образом, в зависимости от характера исходных данных результаты интерполяции в каждом конкретном случае могут быть совершенно различными. В тех случаях когда существует возможность проверки результатов интерполяции, обязательно следует использовать все методы и по минимуму средней квадратической ошибки находить лучший вариант.

Если плотность точек (станций) в пространстве существенно различна, то это может привести к завышению роли точек, образующих сгущения. В этом случае может быть применена так называемая декластеризация. Суть ее состоит в том, что точкам, находящимся в области разреженной плотности сети, присваивается больший вес, а точкам в области сгущений – меньший вес. При этом вся область интерполяции разбивается на ряд элементарных ячеек, а затем на нее накладывается регулярная сетка. Вес принимается обычно обратно пропорциональным числу точечных измерений каждой из ячеек сетки.

Наконец, необходимо принимать во внимание дополнительные моменты. Как известно, расположение сети исходных станций (точек) обычно задается в географической системе координат. Но поскольку декартовая система координат удобнее в практических расчетах, то для пространственных областей не очень больших размеров целесообразно использовать именно ее. С этой целью при переходе от географической системы координат к декартовой могут быть применены следующие формулы:

$$x = a\cos\varphi(\lambda - \lambda_{min}), \quad y = a(\varphi - \varphi_{min}),$$

где x и y – декартовы координаты в километрах, причем ось x направлена на восток, а ось y – на север; φ – широта; λ – долгота точки в градусах; φ_{min} и λ_{min} – минимальные из значений широты и долготы для точек контура; $\bar{\varphi}$ – средняя широта, определяемая как полусумма минимальной и максимальной широт контура; a – средняя длина градуса меридиана ($a = 111{,}2$ км).

Для пространственных областей больших размеров следует уже учитывать изменения длины одного градуса по широте.

Кроме того, представляется удобным перейти от исходных координат, задаваемых в километрах, к безразмерным координатам, которые можно задать следующим образом:

$$x_{без} = (x - x_{min})/Lx, \quad y_{без} = (y - y_{min})/Ly,$$

где x_{min} и y_{min} – минимальные из значений абсцисс и ординат точек контура; $Lx = x_{max} - x_{min}$ – поперечник контура в направлении оси абсцисс (x_{max} – максимальное из значений абсцисс точек контура) аналогично $Ly = y_{max} - y_{min}$.

В безразмерных координатах на площади осреднения значения абсцисс изменяются в пределах от 0 до 1, ординат – от 0 до $l = Ly/Lx$.

Пример 13.4. Корректный статистический анализ невозможен, если в матрице исходных данных имеются пропуски. Восстановление исходных данных в тех случаях, когда они отсутствуют, представляет собой необходимый и одновременно самый сложный этап первичного анализа информации. При единичных пропусках их заполнение, вообще говоря, является стандартной операцией и не представляет особых затруднений. Однако с увеличением числа пропусков задача их заполнения становится все более сложной. Причем в общем случае данная задача не имеет единственного решения, ибо в зависимости от количества пропусков, точности данных, их изменчивости и других факторов могут применяться различные методы восстановления.

Но все же принципиальная сложность заключается в том, что мы не можем непосредственно оценить точность восстановления. В то же время некорректность в заполнении пропущенных данных приводит, как правило, к искаженным статистическим выводам. Было бы прекрасно, если бы можно было найти такой способ заполнения пропусков, при котором анализ полученных полных данных оказывался бы корректным. Особенно опасно искажение структуры исходных рядов при прогнозе гидрометеорологических процессов и явлений.

Хотя по своей сути задача восстановления носит явно неформальный характер, тем не менее ее решение должно осуществляться на основе формальных методов. Поэтому рассмотрим задачу восстановления исходных данных на основе алгоритмов пакета «Сёрфер». Отметим, что восстановление пропусков временного ряда, по существу, тождественно задаче объективного анализа пространственных полей.

Воспользуемся для этого среднемесячными данными по температуре поверхностной воды на ст. Малые Кармакулы с 1977 по 2003 г., т.е. длина ряда составляет $n = 27$. Число пропусков данных на станции для каждого месяца составляет от 6 до 7 % или от 22 до 26 % случаев (табл. 13.1). При этом целиком отсутствуют данные за 1978, 1982, 1983, 1988, 1992 и 1995 гг. Учитывая сравнительно высокий процент отсутствующих данных, достаточно очевидно, что любое восстановление неизбежно приведет к искажению структуры временного ряда. Поэтому речь идет лишь о нахождении такого метода, который обеспечивал бы минимальное искажение структуры данных.

Таблица 13.1

Первичные статистические характеристики температуры воды на станции Малые Кармакулы (72°23′с.ш.; 52°44′в.д.) с 1977 по 2003 г. ($n = 27$)

Месяц	Число пропусков	Среднее	Медиана	X_{min}	X_{max}	Стандартное отклонение
I	6	–1,8	–1,8	–1,9	–1,7	–
II	6	–1,8	–1,8	–1,9	–1,7	–
III	6	–1,8	–1,8	–1,9	–1,7	–
IV	6	–1,8	–1,8	–1,9	–1,4	–
V	6	–1,7	–1,7	–1,8	–1,5	–
VI	6	–0,3	–0,7	–1,5	2,3	0,98
VII	6	5,2	4,7	1,2	8,8	1,84
VIII	6	6,8	6,9	4,6	9,7	1,42
IX	6	4,5	4,3	2,3	6,2	1,11
X	7	0,9	0,8	–0,8	3,6	1,23
XI	7	–1,3	–1,5	–1,8	–0,4	0,45
XII	7	–1,7	–1,8	–1,8	–1,5	–

Из табл. 13.1 отчетливо видно, что зимой (с декабря по май) температура поверхности моря (ТПМ) практически постоянна. Очевидно, заполнение пропусков в данных для этих месяцев может быть осуществлено подстановкой средних значений. Следовательно, в интерполяции нуждаются лишь оценки температуры воды за период с июня по ноябрь. Учитывая, что максимальная изменчивость температуры отмечается в июле и августе, естественно ожидать наибольшей ошибки восстановления именно в эти месяцы. Кроме того, анализ первичных статистических характеристик за отдельные месяцы свидетельствует о том, что исходные данные

имеют резко асимметричное распределение. Естественно, это тоже затрудняет задачу восстановления данных.

Поскольку априори неизвестно, какой из методов пакета «Сёрфер» наилучшим образом отвечает задаче восстановления данных температуры воды, то будем использовать все восемь методов его седьмой версии.

Суть интерполяции заключалась в следующем. По осям координат (x – месяцы, y – годы) размещаются исходные данные. Затем выбирается метод интерполяции и осуществляется оценка значений ТПМ в пустых узлах сетки. После этого остается только оценить точность восстановленных значений температуры. Но это как раз самое сложное. Поэтому вначале были отсеяны те методы, которые заведомо искажали сезонный ход температуры и завышали ее крайние (X_{max} и X_{min}) значения. В результате были исключены три метода (обратных расстояний, полиномиальной регрессии и радиальных базисных функций). В табл. 13.2 приводятся сезонные изменения температуры по фактическим и проинтерполированным данным для отсутствующих месяцев для ст. Малые Кармакулы. Очевидно, можно полагать, что восстановленные за шесть лет значения температуры должны быть близки к фактическим оценкам.

Таблица 13.2

Сравнение сезонного хода температуры воды на ст. Малые Кармакулы с восстановленными за 6–7 лет различными методами

Месяц	Фактические значения	Метод				
		крайкинга	минимальной кривизны	Шеппарда	ближайшего соседа	линейной интерполяции
I	–1,82	–1,85	–1,77	–1,82	–1,83	–1,82
II	–1,82	–1,85	–1,82	–1,83	–1,83	–1,82
III	–1,82	–1,88	–1,87	–1,85	–1,83	–1,82
IV	–1,79	–1,84	–1,85	–1,91	–1,80	–1,80
V	–1,72	–1,39	–1,55	–1,85	–1,73	–1,71
VI	–0,27	0,56	0,17	0,20	–0,33	–0,44
VII	5,23	4.71	5,09	6,08	5,19	5,33
VIII	6,82	6,58	7,08	8.66	7,71	7,39
IX	4,42	4,80	4,95	5,68	4,81	4,84
X	0,95	1,56	1,30	1,25	0,74	1,04
XI	–1,33	–0,70	–1,22	–1,40	–1,40	–1,27
XII	–1,75	–1,68	–2,46	–1,86	–1,79	–1,74
σ	–	1,43	1,11	2,46	1,00	**0,74**

Поскольку визуально определить, какой из методов оказывается наилучшим, весьма сложно, то в качестве критерия «близости» будем использовать среднее квадратическое отклонение восстановленных значений ТПМ от их фактических значений σ. Из табл. 13.2 видно, что наименьшее значение величины σ дает метод линейной интерполяции (σ = 0,74), вторым по точности является метод ближайшего соседа.

Естественно, полученные оценки σ еще не дают полной уверенности в расстановке приоритетов среди методов интерполяции, ибо мы не знаем, насколько точно осредненные за семь лет данные должны соответствовать осредненным значениям температуры за более длительный период времени. Поэтому мы провели следующий модельный эксперимент. Из матрицы исходных данных X случайным образом исключалась часть фактических значений ТПМ и затем эти пропуски заполнялись тем или иным методом. При этом в качестве меры точности опять будем использовать среднее квадратическое отклонение проинтерполированных значений от истинных. Из исходного ряда температуры на ст. Малые Кармакулы было исключено произвольным образом семь лет (1980, 1981, 1987, 1991, 1994, 1998, 2002 гг.).

После восстановления искусственных пропусков различными методами интерполяции оценивалась средняя квадратическая ошибка для всего периода в целом (n = 84) и для каждого месяца в отдельности (n = 7). Результаты расчетов приведены в табл. 13.3. Наименьшую ошибку восстановления температуры за семилетний период опять дает метод линейной интерполяции. Вторым по точности является метод минимальной кривизны.

Итак, в обоих случаях линейная интерполяция дает более точные результаты, поэтому примем ее за основу. Этот алгоритм реализует оптимальную триангуляцию Деланая. Главный недостаток метода – разрывы на границах. Поэтому, если исходные данные отсутствуют на границах, их нужно восстанавливать другими методами.

Заметим, что если, исходя из табл. 13.3, для каждого месяца выбирать свой собственный наилучший метод восстановления, то в этом случае средняя квадратическая ошибка в целом за весь период окажется равной σ = 0,63 °С, т.е. она становится заметно ни-

же аналогичной ошибки метода линейной интерполяции. Но все равно она достаточно высокая. Поэтому еще раз подчеркнем, что в связи с высокими ошибками интерполяции возможны искажения внутренней структуры ряда, что необходимо учитывать при дальнейшем статистическом анализе температуры.

Таблица 13.3

Оценки средней квадратической ошибки восстановления температуры воды (°С) для ст. Малые Кармакулы различными методами, в

Месяц	Метод				
	крайкинга	минимальной кривизны	Шепарда	ближайшего соседа	линейной интерполяции
I	0,06	0,09	0,03	**0,00**	0,03
II	0,06	**0,03**	0,05	0,04	0,04
III	0,09	0,09	0,12	0,16	**0,05**
IV	**0,07**	0,09	0,17	0,12	0,09
V	0,55	0,26	0,40	1,92	**0,07**
VI	1,82	1,36	1,36	1,35	**0,82**
VII	**1,02**	1,21	1,44	1,82	1,47
VIII	1,59	**1,37**	1,87	1,66	1,44
IX	**0,93**	0,94	1,13	1,24	1,19
X	1,19	**1,03**	1,26	1,57	1,08
XI	0,65	**0,17**	0,27	0,22	0,30
XII	0,19	0,88	0,17	0,14	**0,01**
Весь период	0,91	0,82	0,94	1,00	**0,79**

Примечание. Выделены минимальные ошибки восстановления.

Дополнительная проверка точности интерполяции может быть осуществлена с помощью гармонического анализа. Рассмотрим данный подход на примере августовских значений ТПМ за 1977–2003 гг. С довольно высокой точностью (90 % дисперсии от исходного процесса) временной ряд может быть аппроксимирован семью наиболее значимыми гармониками. На рис. 13.8 приводится исходный ряд, в котором квадратиками отмечены значения ТПМ, полученные методом линейной интерполяции для 6 указанных выше лет с пропусками, а также аппроксимация его с помощью 7 гармоник. Нетрудно видеть, что только в 1988 г. расхождение в оценке ТПМ двух разных методов интерполяции составляет 0,42 °С, в остальные годы оно существенно меньше. Таким образом, это подтверждает достаточно высокую точность интерполяции значений ТПМ на ст. Малые Кармакулы.

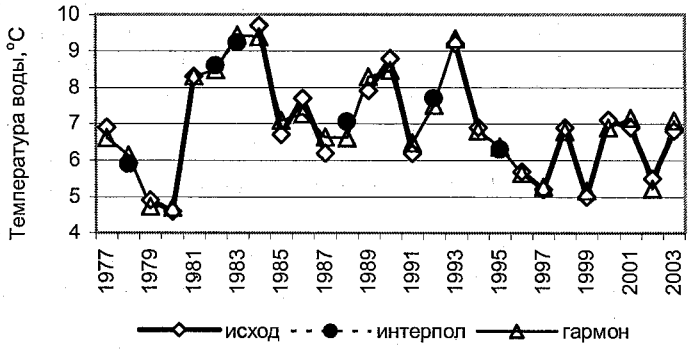


Рис. 13.8. Межгодовой ход температуры воды в августе на ст. Малые Кармакулы

13.6. Понятие о крайкинге

Одним из наиболее теоретически разработанных в статистическом отношении методов двухмерной интерполяции является метод крайкинга, названный в честь южноафриканского горного инженера и пионера применения статистических методов в геологии Д. Г. Крайга (Krige). В литературе этот метод называется также кригингом. На наш взгляд, первое название его является более точным.

Метод крайкинга можно использовать для построения карт в изолиниях, но, в отличие от обычных алгоритмов оконтуривания, он имеет статистически оптимальные свойства. В частности, он имеет два больших преимущества перед обычными процедурами оценивания, которые используются при построении карт в изолиниях. Одно заключается в том, что оценки процедур в среднем имеют наименьшую возможную ошибку, а также обеспечивают явное выражение величины этой ошибки. Другое сводится к использованию информации из полувариограммы для нахождения оптимального множества весов с целью оценки поверхности в точках, отличных от точек опробования. Так как полувариограмма является функцией расстояния, то веса изменяются в соответствии с географическим положением точек опробования.

В общем случае можно выделить три вида крайкинга:

1) точечный (простой) крайкинг (simple kriging),

2) ординарный крайкинг (ordinary kriging),

3) универсальный крайкинг (universal kriging).

Точечный крайкинг. Точечный крайкинг – простейшая форма крайкинга, в котором наблюдения состоят из измерений, взятых в безразмерных точках, и оценки проводятся в других местах, которые сами также являются безразмерными точками. Для упрощения задачи принимается, что картируемая переменная статистически стационарна или свободна от дрифта. Значение в точке, не принадлежащей выборке, может быть оценено как взвешенное среднее известных наблюдений, причем не выдвигается жесткое условие равенства единице суммы весовых коэффициентов. Система линейных уравнений простого крайкинга для нахождения весов интерполяции λ_β может быть представлена в виде:

$$\sum_{\beta} \lambda_\beta \gamma_{\alpha\beta} = \gamma_{\alpha 0}, \qquad (13.13)$$

где $\gamma_{\alpha\beta}$ – значение вариограммы для пар точек наблюдений; $\gamma_{\alpha 0}$ – значение вариограммы для узла интерполяции; β – количество измерений.

Тогда значение в сеточном узле определяется как линейная комбинация наблюдений:

$$Z^* = \Sigma \lambda_\beta Z_\beta + \lambda_0 , \qquad (13.14)$$

где $\lambda_0 = m(1 - \Sigma\lambda_\beta) = m\lambda_m$; m – среднее значение внутри заданной пространственной области.

В этом случае средняя квадратическая ошибка интерполяции будет определяться как

$$\sigma^2(\varepsilon) = R_u(0) - \Sigma \lambda_\beta R_{u\beta 0}, \qquad (13.15)$$

где $R_u(0)$ – дисперсия пространственного поля.

Естественно, что при «хорошем» качестве интерполяции последнее слагаемое в правой части выражения (13.14) должно стремиться к нулю. Отклонения суммы весовых коэффициентов от единицы может служить критерием качества интерполяции в конкретном узле. Чем больше сумма коэффициентов отклоняется от единицы, тем хуже интерполяция в заданный узел. Кроме того, для оценки точности используется обычная средняя квадратическая ошибка интерполяции. Если все точки расположены далеко от интерполи-

руемого узла, то погрешность метода будет практически равна средней квадратической ошибке внутри области интерполяции.

Ординарный крайкинг. Недостатков, связанных с изменениями среднего значения внутри области интерполяции, лишен метод ординарного крайкинга:

$$\Sigma\lambda_\beta\gamma_{\alpha\beta} + \mu = \gamma_{\alpha 0}, \tag{13.16}$$

$$\Sigma\lambda_\beta = 1, \tag{13.17}$$

$$Z^* = \Sigma\lambda_\beta Z_\beta, \tag{13.18}$$

$$\sigma^2(\varepsilon) = R_u(0) - \Sigma\lambda_\beta R_{u\beta 0} - \mu, \tag{13.19}$$

где μ – подгоночный параметр, вводимый в исходную систему весов крайкинга, чтобы выполнялось условие (13.17).

Универсальный крайкинг. Хотя ординарный крайкинг более точен, чем простой крайкинг, тем не менее, он не учитывает наличие пространственного тренда в исходных данных. В этом случае может быть применен метод универсального крайкинга. Такой подход оправдан, когда интерполяция проводится для больших пространственных областей с явно выраженной тенденцией к повышению (понижению) рассматриваемой характеристики вдоль какого-либо направления. Однако следует помнить, что универсальный крайкинг очень чувствителен к неравномерной сети станций. При использовании его в областях с редкой сетью наблюдений могут возникать серьезные ошибки, связанные с экстраполяцией тренда на большие расстояния.

Универсальный крайкинг можно считать состоящим из трех операций: первая – оценка и устранение дрифта; затем стационарные остатки крайгируются с целью получения необходимых оценок. Наконец, оцененные остатки комбинируются с дрифтом с целью получения истинной поверхности. Оценка дрифта может быть осуществлена, например, с помощью полинома первой или второй степени (13.12).

Словарь основных статистических терминов

Термин	Определение
Амплитуда гармоники	Разность между максимальным и минимальным значениями гармоники
Анализ объективный	Процедура перевода данных из нерегулярной сети точек в регулярную сетку
Величина случайная	Переменная величина, которая в результате испытания (измерения) в одинаковых условиях может принимать то или иное заранее неизвестное значение
Величина случайная количественная	Величина, выражаемая в метрической шкале
Величина случайная ординальная	Величина, соответствующая порядковой (ординальной) шкале
Величина случайная номинальная	Величина, соответствующая номинальной шкале
Величина случайная стандартизованная	Величина, полученная преобразованием $t=(x-m_x)/\sigma_x$, 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
Гармоника	Слагаемое в разложении Фурье
Генеральная совокупность	Весь мыслимо возможный набор случайной величины
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
Дециль	Квантиль, соответствующий одной из вероятностей 0,10; 0,20; …; 0,90
Доверительный интервал	Область значений случайной величины внутри доверительных границ
Дисперсия генеральная (выборочная)	Мера изменчивости случайной величины в генеральной (выборочной) совокупности, имеющая размерность ее квадрата
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
Корреллограмма	График автокорреляционной функции

Корреляционное отношение	Безразмерная мера нелинейной связи двух случайных величин
Корреляционное поле	График двух случайных величин в декартовой системе координат
Корреляция ранговая	Линейная стохастическая связь между порядковыми переменными
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 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 эксцесса	Характеристика крутости кривой плотности распределения случайной величины
Коэффициент регрессии	Коэффициент пропорциональности между зависимой и независимой переменными
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 всех оставшихся переменных
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
Модель авторегрессионная	Конечная линейная комбинация предыдущих его значений и случайного импульса
Мода	Наиболее часто встречающаяся в данном статистическом ряду величина

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 строгая	Корреляционная матрица, в которой хотя бы одна переменная описывается линейной функциональной связью через остальные переменные
Мультиколлинеарность реальная	Корреляционная матрица, в которой между большинством исходных переменных отмечается высокая коррелированность
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 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
Оценка робастная	Оценка, которая является устойчивой к существенным отклонениям в значениях данных
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 случайная	Погрешность, которая при испытаниях в одинаковых условиях меняется произвольным образом
Погрешность систематическая	Погрешность, изменяющаяся по определенному закону
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

Полигон распределения	Ломаная линия, соединяющая частоты вариационного ряда
Преобразование Фишера	Функциональное преобразование вида $z = 0{,}5 \ln(1 + r)/(1 - r)$, 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 $t_1, t_2, \ldots, t_n$ 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 случайной величины	Порядковый номер значения признака ранжированного ряда
Ранг матрицы	Наибольший порядок ее отличного от нуля минора, совпадающий с максимальным числом линейно независимых столбцов матрицы
Реализация случайной функции	Конкретный вид, который случайная функция принимает в результате испытаний наблюдений
Регрессионный анализ	Построение эмпирических зависимостей (моделей) и оценивание их параметров
Регрессия парная линейная	Уравнение линейной зависимости между двумя случайными переменными
Регрессия множественная линейная	Уравнение, описывающее линейную зависимость функции отклика от множества факторов
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 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

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 порядковая	Каждый член вариационного ряда
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
Точка отсечения АФ	Максимальный сдвиг, до которого осуществляется расчет автокорреляционной функции
Тренд	Медленное изменение случайного процесса с периодом, превышающим длину исходной реализации
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	Неслучайная функция, значения которой при каждом значении аргумента представляют случайную величину

Функция частотная весовая	Функция, которая уравнивает выборочные и истинные оценки автокорреляционной функции
Функция спектральная	Функция, характеризующую интегральную долю дисперсии, приходящейся на некоторый интервал частот
Характеристика фильтра частотная	Функция, определяющая характер изменения амплитуд случайного процесса при прохождении ряда через фильтр
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
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
Шум белый	Случайный процесс, представляющий набор случайных чисел, не коррелированных друг с другом
Шум красный	Случайный процесс, которому свойственна корреляция только между смежными (соседними) значениями временного ряда

ЛИТЕРАТУРА

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.: изд. РГГМИ, 1991. – 136 с.
3. *Вайновский П.А., Малинин В.Н.* Методы обработки и анализа океанологической информации. Ч. 2. Многомерный анализ.– СПб.: изд. РГМИ, 1992. – 96 с.
4. *Григоркина Р.Г.. Губер П.К., Фукс В.Р.* Прикладные методы корреляционного и спектрального анализа крупномасштабных океанологических процессов. – Л.: Изд-во 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-ИНФРА-М, 2004. – 462 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. *Дженкинс Г., Ваттс Д.* Спектральный анализ и его приложения. Вып. 1. – М.: Мир, 1971. – 316 с., Вып. 2. – М.: Мир, 1972. – 287 с.
9. *Добровольский С.Г.* Климатические изменения в системе «гидросфера-атмосфера». – М.: Геос, 2002. – 231 с.
10. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. Кн. 1, 2. – М.: Финансы и статистика, 1986. – 366 с., 1987. – 351 с.
11. *Исаев А.А.* Статистика в метеорологии и климатологии. – М.: Изд-во МГУ, 1988. – 245 с.
12. *Казакевич Д.Л.* Основы теории случайных функций в задачах гидрометеорологии. – Л.: Гидрометеоиздат, 1989. – 230 с.
13. *Кремер Н.Ш.* Теория вероятностей и математическая статистика. – М.: ЮНИТИ, 2003. – 543 с.
14. *Львовский Е.Н.* Статистические методы построения эмпирических формул. – М.: Высшая школа, 1982. – 224 с.
15. *Макарова Н.В., Трофимец В.Я.* Статистика в Excel. – М.: Финансы и статистика, 2002. – 365 с.
16. *Малинин В.Н., Гордеева С.М.* Физико-статистический метод прогноза океанологических характеристик. – Мурманск, Изд-во ПИНРО, 2003. – 164 с.
17. *Мюллер П., Нойман П., Шторм Р.* Таблицы по математической статистике. – М.: Финансы и статистика, 1982. – 272 с.
18. *Носач В.В.* Решение задач аппроксимации с помощью персональных компьютеров. – М.: МИКАП, 1994. – 382 с.
19. *Пановский Г.А., Брайер Г.В.* Статистические методы в метеорологии. – Л.: Гидрометеоиздат, 1967. – 242 с.

20. *Поляк И.И.* Методы анализа случайных процессов и полей в климатологии. – Л.: Гидрометеоиздат, 1979. – 255 с.
21. *Пузаченко Ю.Г.* Математические методы в экологических и географических исследованиях. – М.: Академия, 2004. – 407 с.
22. *Рожков В.А.* Теория и методы статистического оценивания вероятностных характеристик случайных величин и функций с гидрометеорологическими приложениями. Кн. 1. – СПб.: Гидрометеоиздат, 2001. – 340 с.
23. *Рожков В.А.* Теория и методы статистического оценивания вероятностных характеристик случайных величин и функций с гидрометеорологическими приложениями. Кн. 2. – СПб.: Гидрометеоиздат, 2002. – 440 с.
24. *Смирнов Н.П., Вайновский П.А., Титов Ю.Э.* Статистический диагноз и прогноз океанологических процессов. – СПб.: Гидрометеоиздат, 1992. – 199 с.
25. *Тьюки Дж.* Анализ результатов наблюдений. Разведочный анализ. – М.: Мир, 1981. – 693 с.
26. *Тюрин Ю.Н., Макаров А.А.* Статистический анализ данных на компьютере. – М.: ИНФА-М, 1998. – 528 с.
27. *Хьюбер П.* Робастность в статистике. – М.: Мир, 1984. – 303 с.
28. *Шелутко В.А.* Численные методы в гидрологии. – Л.: Гидрометеоиздат, 1991. – 238 с.

Функция Лапласа

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,0000	0,32	0,1255	0,64	0,2389	0,96	0,3315
0,01	0,0040	0,33	0,1293	0,65	0,2422	0,97	0,3340
0,02	0,0080	0,34	0,1331	0,66	0,2454	0,98	0,3365
0,03	0,0120	0,35	0,1368	0,67	0,2486	0,99	0,3389
0,04	0,0160	0,36	0,1406	0,68	0,2517	1,00	0,3413
0,05	0,0199	0,37	0,1443	0,69	0,2549	1,01	0,3438
0,06	0,0239	0,38	0,1480	0,70	0,2580	1,02	0,3461
0,07	0,0279	0,39	0,1517	0,71	0,2611	1,03	0,3485
0,08	0,0319	0,40	0,1554	0,72	0,2642	1,04	0,3508
0,09	0,0359	0,41	0,1591	0,73	0,2673	1,05	0,3531
0,10	0,0398	0,42	0,1628	0,74	0,2703	1,06	0,3554
0,11	0,0438	0,43	0,1664	0,75	0,2734	1,07	0,3577
0,12	0,0478	0,44	0,1700	0,76	0,2764	1,08	0,3599
0,13	0,0517	0,45	0,1736	0,77	0,2794	1,09	0,3621
0,14	0,0557	0,46	0,1772	0,78	0,2823	1,10	0,3643
0,15	0,0596	0,47	0,1808	0,79	0,2852	1,11	0,3665
0,16	0,0636	0,48	0,1844	0,80	0,2881	1,12	0,3686
0,17	0,0675	0,49	0,1879	0,81	0,2910	1,13	0,3708
0,18	0,0714	0,50	0,1915	0,82	0,2939	1,14	0,3729
0,19	0,0753	0,51	0,1950	0,83	0,2967	1,15	0,3749
0,20	0,0793	0,52	0,1985	0,84	0,2995	1,16	0,3770
0,21	0,0832	0,53	0,2019	0,85	0,3023	1,17	0,3790
0,22	0,0871	0,54	0,2054	0,86	0,3051	1,18	0,3810
0,23	0,0910	0,55	0,2088	0,87	0,3078	1,19	0,3830
0,24	0,0948	0,56	0,2123	0,88	0,3106	1,20	0,3849
0,25	0,0987	0,57	0,2157	0,89	0,3133	1,21	0,3869
0,26	0,1026	0,58	0,2190	0,90	0,3159	1,22	0,3883
0,27	0,1064	0,59	0,2224	0,91	0,3186	1,23	0,3907
0,28	0,1103	0,60	0,2257	0,92	0,3212	1,24	0,3925
0,29	0,1141	0,61	0,2291	0,93	0,3238	1,25	0,3944
0,30	0,1179	0,62	0,2324	0,94	0,3264		
0,31	0,1217	0,63	0,2357	0,95	0,3289		

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
1,26	0,3962	1,59	0,4441	1,92	0,4726	2,50	0,4938
1,27	0,398	1,60	0,4452	1,93	0,4732	2,52	0,4941
1,28	0,3997	1,61	0,4463	1,94	0,4738	2,54	0,4945
1,29	0,4015	1,62	0,4474	1,95	0,4744	2,56	0,4948
1,30	0,4032	1,63	0,4484	1,96	0,475	2,58	0,4951
1,31	0,4049	1,64	0,4495	1,97	0,4756	2,60	0,4953
1,32	0,4066	1,65	0,4505	1,98	0,4761	2,62	0,4956
1,33	0,4082	1,66	0,4515	1,99	0,4767	2,64	0,4959
1,34	0,4099	1,67	0,4525	2,00	0,4772	2,66	0,4961
1,35	0,4115	1,68	0,4535	2,02	0,4783	2,68	0,4963
1,36	0,4131	1,69	0,4545	2,04	0,4793	2,70	0,4965
1,37	0,4147	1,70	0,4554	2,06	0,4803	2,72	0,4967
1,38	0,4162	1,71	0,4564	2,08	0,4812	2,74	0,4969
1,39	0,4177	1,72	0,4573	2,10	0,4821	2,76	0,4971
1,40	0,4192	1,73	0,4582	2,12	0,483	2,78	0,4973
1,41	0,4207	1,74	0,4591	2,14	0,4838	2,80	0,4974
1,42	0,4222	1,75	0,4599	2,16	0,4846	2,82	0,4976
1,43	0,4236	1,76	0,4608	2,18	0,4854	2,84	0,4977
1,44	0,43	1,77	0,4616	2,20	0,4861	2,86	0,4979
1,45	0,4265	1,78	0,4625	2,22	0,4868	2,88	0,498
1,46	0,4279	1,79	0,4633	2,24	0,4875	2,90	0,4981
1,47	0,4292	1,80	0,4641	2,26	0,4881	2,92	0,4982
1,48	0,4306	1,81	0,4649	2,28	0,4887	2,94	0,4984
1,49	0,4319	1,82	0,4656	2,30	0,4893	2,96	0,4985
1,50	0,4332	1,83	0,4664	2,32	0,4898	2,98	0,4986
1,51	0,4345	1,84	0,4671	2,34	0,4904	3,00	0,49865
1,52	0,4357	1,85	0,4678	2,36	0,4909	3,20	0,49931
1,53	0,437	1,86	0,4686	2,38	0,4913	3,40	0,49966
1,54	0,4382	1,87	0,4693	2,40	0,4918	3,60	0,499841
1,55	0,4394	1,88	0,4699	2,42	0,4922	3,80	0,499928
1,56	0,4406	1,89	0,4706	2,44	0,4927	4,00	0,499968
1,57	0,4418	1,90	0,4713	2,46	0,4931	4,50	0,499997
1,58	0,4429	1,91	0,4719	2,48	0,4934	5,00	0,499997

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

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)					

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

СОДЕРЖАНИЕ

CONTENTS

Учебное издание

Малинин Валерий Николаевич

СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ГИДРОМЕТЕОРОЛОГИЧЕСКОЙ ИНФОРМАЦИИ

Учебник

Редактор Л.В. Ковель
Компьютерная верстка Н.И. Афанасьевой

ЛР № 020309 от 30.12.96.

Подписано в печать 29.10.08. Формат 60 × 90 1/16. Гарнитура Times New Roman.
Бумага офсетная. Печать офсетная. Усл.-печ. л. 25,4. Тираж 300 экз. Заказ № 50/08.
РГГМУ, 195196, Санкт-Петербург, Малоохтинский пр., 98.
ЗАО «НПП «Система», 195112, Санкт-Петербург, Малоохтинский пр., 80/2.