

П. А. ВАЙНОВСКИЙ, В. Н. МАЛИНИН

МЕТОДЫ ОБРАБОТКИ И АНАЛИЗА
ОКЕАНОЛОГИЧЕСКОЙ ИНФОРМАЦИИ
ОДНОМЕРНЫЙ АНАЛИЗ

*Утверждено ученым советом института
в качестве учебного пособия*

Вайновский П. А., Малинин В. Н. Методы обработки и анализа океанологической информации. Одномерный анализ. Учебное пособие. — Л., изд. ЛГМИ, 1991, 136 с.

Излагаются основы методов математической статистики (одномерный анализ) и их использование применительно к обработке и интерпретации океанологических данных.

Предназначено для студентов-океанологов гидрометеорологических вузов и университетов.

Рецензенты: кафедра океанологии Одесского гидрометеорологического института, И. В. Повыкин, канд. техн. наук, ст. науч. сотрудник (Институт озераедения АН СССР).

Научный редактор: проф. Н. П. Смирнов

ВВЕДЕНИЕ

От практической морской гидрометрии к теоретическим вершинам физики и динамики океана лежит широкая область обработки и анализа океанологической информации. Исторически сложилось так, что теория стремится описать физические явления в форме хорошо интерпретируемых функциональных связей, которые понимаются как фундаментальные природные законы. Однако данные натурных наблюдений еще не являются сами по себе информацией в высоком смысле слова, поскольку искомые «зерна фактов» сильно искажены разнообразными помехами и шумами. Поэтому основное содержание дисциплины «Методы обработки и анализа океанологической информации» состоит как раз в том, чтобы дать в руки исследователю инструмент, которым можно отделить «зерна» физически обоснованных закономерностей от «плевел» шумов и ошибок.

Сам по себе океан весьма неудобен для изучения. Современной наукой он понимается как плохо организованная система, насыщенная сложными опосредованными связями и запутанными взаимодействиями в пространстве и времени. Чтобы понять суть океанических процессов, необходимо учесть влияние многих внешних и внутренних факторов, различных по своей природе. В основе современного процесса переработки океанологической информации лежит методология статистико-вероятностного анализа. Применение этих методов позволяет на практике получить достоверные количественные оценки изменчивости океанологических характеристик, интенсивности взаимодействий, направленности причинно-следственных связей. На основе статистико-вероятностной обработки решаются задачи построения карт океанологических полей, восстановления отсутствующих данных наблюдений, фильтрации ошибок, наконец, предсказания тенденций развития океанологических процессов. В результате применения статистических методов исследователь имеет возможность сформулировать и проверить множество гипотез о структуре выявленных колебаний и тем самым вплотную приблизиться к описанию природы наиболее устойчивых закономерностей изменения океанологических характеристик. Теоретической основой для интерпретации результатов обработки океанологических процессов являются общие законы статистической гидромеханики и термодинамики.

Техническая основа успешного решения задач обработки и анализа океанологической информации лежит в применении со-

временных ЭВМ. Причем ЭВМ выступают не только как эффективные средства хранения разнообразных исходных данных. В процессе совершенствования математического обеспечения прикладных вычислений разработано большое количество пакетов программ обработки наблюдений, успешное применение которых существенно ускоряет и упрощает процесс обработки данных. Умелое применение этих программных средств требует от исследователя высокой подготовки и в области программирования, и в области прикладной статистики.

В данном пособии рассмотрены основные принципы обработки и анализа океанологической информации. Последовательность изложения, формулировка важнейших теоретических и прикладных вопросов дана так, как это принято в практике океанологических исследований. Следует также иметь в виду, что данное пособие является лишь первой частью курса «Методы обработки и анализа океанологической информации» и посвящено в основном одномерному анализу экспериментальных данных. Во второй части курса будут рассмотрены методы многомерного статистического анализа океанологической информации.

На протяжении ряда лет этот курс разрабатывался и читался авторами на океанологическом факультете ЛГМИ под руководством проф. Н. П. Смирнова, которому авторы выражают свою искреннюю благодарность за постоянную поддержку в работе. Кроме того, авторы признательны рецензентам за ценные советы и замечания, которые позволили улучшить содержание данного пособия.

РАЗДЕЛ I

ПЕРВИЧНАЯ ОБРАБОТКА ОКЕАНОЛОГИЧЕСКОЙ ИНФОРМАЦИИ

Первичная обработка информации, как правило, является важным этапом большинства фундаментальных и прикладных океанологических исследований. Содержание первичной обработки состоит в оценке принципиально важных свойств изучаемых данных. Прежде всего — это характеристики устойчивости, разброса (изменчивости) вероятности появления отдельных значений. Сумма всех полученных значений является основой как для понимания статистической структуры исследуемых океанологических явлений, так и для их взаимного сравнения. К разделу первичной статистики часто относят и технику статистической проверки гипотез, которая позволяет количественно оценить степень сходства вычислительных статистических оценок изменчивости различных океанологических характеристик и тем самым привести в соответствие наблюдаемые теоретические и экспериментальные данные. Нередко результаты только первичной обработки обеспечивают исследователя исчерпывающей информацией о природе явлений. Примером может быть процесс построения и анализа T -, S -гистограмм, который в океанологии часто именуется «волюмметрическим анализом водных масс». Он позволяет выделить по результатам вычисления первичных статистик, параметров функций распределения и проверки гипотез основные водные массы и дает характеристику процессов их трансформации.

Важное самостоятельное значение первичный статистический анализ несет при выполнении контроля и редактирования океанологической информации в процессе ее накопления. С помощью полученных оценок устойчивости и изменчивости накапливаемого ряда можно проверить принадлежность каждого нового наблюдения к уже накопленной выборке.

Наконец, выводы, сделанные на этапе первичной обработки океанологической информации, являются непременным условием применения более сложных статистических методов анализа океанологических процессов и полей. Теория многих методов прикладной статистики основывается на предположениях стационарности, однородности исходных выборок, их соответствии нормальному закону распределения. Поэтому результаты проверки требуемых статистических гипотез заранее позволяют оценить успешность тех или иных сложных методов анализа и оптимально спланировать всю стратегию обработки океанологической информации.

ГЛАВА I. НАКОПЛЕНИЕ И ХРАНЕНИЕ ОКЕАНОЛОГИЧЕСКОЙ ИНФОРМАЦИИ

Весь сложный процесс разнообразной обработки океанологической информации состоит из четырех основных этапов: сбора информации, ее накопления и хранения, собственно обработки информации, анализа результатов. Практически любая исследовательская океанологическая работа начинается сбором 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.

Перечисленные причины обусловили необходимость создания единой системы сбора, хранения и обработки гидрометеорологической информации в целом, которая и будет объектом рассмотрения в данной главе.

1.1. Океанологическая информация

В общем понимании информация — это любые сведения об объекте исследования. Под океанологической информацией понимают совокупность данных наблюдений и расчетов характеристик состояния океана. Информация об океане предстает перед исследователем в виде большого набора переменных величин, которые изменяются во времени и в пространстве. Чтобы разобраться в особенностях информации, необходимо попытаться ее классифицировать. Прежде всего на практике выделяют группы параметров, характеризующие конкретные океанологические поля: гидрофизические, химические, биологические и др.

В зависимости от структуры наблюдений во времени информация может быть:

- непрерывной, если производится непрерывная регистрация происходящего на различные самописцы;
- дискретной, если наблюдения повторяются многократно через равные или неравные промежутки времени;
- эпизодической, если проводятся единичные, разовые наблюдения.

Эта классификация сохраняет свой смысл и при рассмотрении особенностей распределения наблюдений в пространстве — по вертикали или по горизонтали.

Стремление наиболее точно зафиксировать происходящие в океане сложные явления побуждает исследователей использовать различные способы описания океанологических параметров. Это в первую очередь прослеживается в разнообразии шкал измерений и наблюдений.

Номинальная шкала 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, безразличный, или, например, форма волнения и его связь с ветром позволяют подразделить волны на свободные и вынужденные. Следует отметить, что в номинальной шкале, как правило, фиксируется наличие или отсутствие конкретного 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.

Порядковая шкала — выражает оценку интенсивности явления в дискретном (квантованном) виде. В таком виде задается, например, состояние 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. Наиболее часто количество градаций не превышает 10 или 12 значений.

Метрические шкалы — соответствуют количественному выражению наблюдаемых характеристик в градуционно принятых единицах измерения *СИ* или *СГС*. Метрические шкалы широко используются в практической океанологии для измерения основных гидрофизических характеристик: температуры в градусах, скорости звука в м/с и т. п. Информация в метрической шкале представляется в виде вещественных чисел.

Описание текущего состояния океана имеет ту особенность, что исчерпывающая характеристика его дается, как правило, набором нескольких переменных, выражаемых в различных шкалах. Так, 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 для обмена данными океанологических наблюдений (КН-05), прибрежных наблюдений (КН-02), судовых гидрометнаблюдений (КН-09с). Трудно вычислить и проанализировать статистические характеристики связи для переменных, задаваемых различными шкалами.

Приведенный выше краткий анализ видов океанологической информации убеждает в серьезности проблем обработки данных наблюдений и расчетов. Становится очевидной нетривиальность задач по созданию эффективных путей сбора, хранения и обработки разнородной океанологической информации.

1.2. Единая система сбора, хранения и переработки гидрометеорологической информации

Исследуя особенности изменчивости Мирового океана, нельзя забывать о его неразрывной связи с процессами, происходящими в атмосфере и гидросфере Земли в целом. Именно поэтому реализация системного подхода в смежных науках гидрологии, метеорологии и океанологии позволила разработать и успешно функционировать единой системе переработки всей гидрометеорологической информации.

Всесоюзный научно-исследовательский институт гидрометеорологической информации (г. Обнинск) является центром общесоюзной системы сбора, накопления и обмена всей гидрометеорологической информации. Одновременно ВНИИГМИ является Мировым центром данных по гидрометеорологии. Основными функциями ВНИИГМИ — МЦД являются следующие:

- сбор наиболее полного набора данных в области гидрометеорологии;
- обеспечение полной сохранности накапливаемых данных с соблюдением общепринятых правил их копирования и репродуцирования;
- составление и распространение каталогов имеющихся данных;
- обмен копиями данных с другими мировыми центрами данных (Мельбурн, Вашингтон);
- обеспечение данными всех заинтересованных организаций в СССР и за рубежом.

Осуществление перечисленных функций производится с помощью сети региональных центров данных, созданных на базе территориальных управлений гидрометслужбы (Северное УГМС, Сахалинское УГМС, Приморское УГМС и др.), а также ведущих мореведческих организаций (Института океанологии АН СССР, Государственного океанографического института, Всесоюзного научно-исследовательского института морского рыбного хозяйства и океанографии и бассейновых отделений этих организаций).

Внутри ВНИИГМИ действует специализированный центр океанографических данных (ЦОД), основными задачами которого являются сбор, переработка, хранение и обмен разнообразной океанографической информации.

Функция сбора океанологической информации выполняется во всех перечисленных организациях. Территориальные УГМС собирают текущую информацию о состоянии морей СССР по результатам наблюдений на гидрометстанциях и постах в прибрежной зоне. Мореведческие организации осуществляют сбор информации по итогам проведенных экспедиционных исследований в отдельных районах морей и океанов, которые определены спецификой работ учреждения. Так, Арктический и Антарктический научно-исследовательский институт накапливает данные о режиме Северного Ледовитого и Южного океанов. Тогда как в Балтийском отделении

научно-исследовательского института рыбного хозяйства и океанографии производится сбор данных только по региону Балтийского моря.

Центр океанографических данных собирает воедино информацию, поступающую из территориальных и ведомственных организаций нашей страны. Через ВНИИГМИ как мировой центр данных осуществляется сбор отсутствующих в СССР океанографических данных по различным районам Мирового океана, в первую очередь по Индийскому океану, южным частям Атлантического и Тихого океанов.

Хранение океанографических данных осуществляется в активной и пассивной форме. Пассивные данные — это любая океанографическая информация, входящая в гидрометеорологический фонд данного учреждения и не подготовленная для непосредственного ввода в ЭВМ. Это могут быть карты, графики, книжки гидрометеорологических наблюдений, таблицы, фотографии и микрофильмы. В своем большинстве пассивные данные содержат непосредственные результаты наблюдений и измерений. Как правило, они не проконтролированы, вследствие чего могут содержать ошибки и погрешности. Такая форма хранения затрудняет процесс поиска нужной информации, существенно замедляет обмен и массовую обработку данных. Поэтому пассивная форма хранения океанографической информации постепенно уступает место активной.

Активные данные представляют собой материалы наблюдений, занесенные на технические носители и оформленные в соответствии с требованиями подготовки данных для ЭВМ. Современными техническими носителями являются в первую очередь магнитные ленты, пакеты магнитных дисков, гибкие магнитные диски. Данные, представленные в активной форме, позволяют до минимума сократить время ввода информации в ЭВМ и максимально повысить эффективность всего процесса переработки.

Важным достоинством активной формы хранения данных является общедоступность к информации на технических носителях. Запись информации на технические носители осуществляется в соответствии с общими общесоюзными требованиями к представлению гидрометеорологической информации для ЭВМ. Основным носителем океанографической информации в СССР принята на сегодняшний день магнитная лента (МЛ). Использование МЛ позволило унифицировать процедуры обмена океанографической информацией между всеми мореведческими организациями. Подобная унификация облегчила международный обмен океанографической информацией между Мировыми центрами гидрометеорологических данных.

Сбор и накопление океанографической информации служит для целей обработки данных при решении конкретных прикладных и научных задач. Обработка океанографической 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.

Выделим принципиальное отличие перечисленных видов. Оно состоит в том, что первичная обработка и оперативный диагноз осуществляются многократно в заданном ритме с дискретностью, равной дискретности поступления новой информации об океане. Примером подобных видов обработки можно считать 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), обособить сушу и облачность. Затем статистическими методами восстановить недостающие данные и осуществить собственно диагноз, который бы идентифицировал фронтальные зоны, вихри, распределение аномалий температуры и т. п. Тогда как последующие виды обработки производятся эпизодически или даже в единичных случаях.

Выполнение всех видов обработки осуществляется на основе применения разнообразных методов теории вероятностей и математической статистики. Чтобы не нарушать последовательности изложения, будем рассматривать перечисленные виды обработки в соответствующих разделах учебного пособия.

1.3. Понятие Банка Океанологических Данных

Понятие Банка Океанологических Данных (БОД) является центральным в системе сбора, хранения и обработки океанологической информации. По определению, БОД — это комплекс технических средств обработки океанологических данных, состоящий из базы данных (информационной базы), технических средств (ЭВМ, носителей информации, средств ввода-вывода) и программно-математического обеспечения. Общая блок-схема БОД приведена на рис. 1.1.

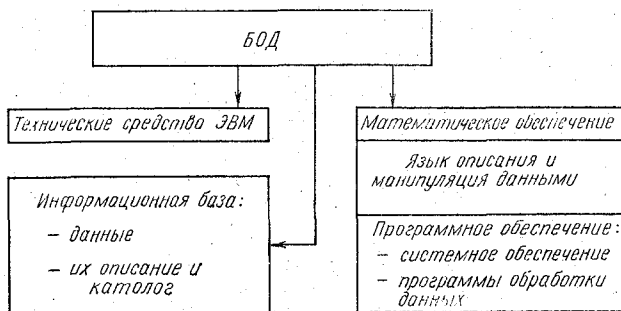


Рис. 1.1. Компоненты банка океанологических данных

Фундаментом БОД является база данных, которая представляет собой совокупность океанологических данных (цифровых массивов и их описаний), хранящихся на техническом носителе в виде строго упорядоченной структуры записей. База данных реорганизует и переупорядочивает исходные ряды или поля наблюдаемых переменных так, чтобы обеспечить наибольшую эффективность функционирования БОД. А первоочередные функции БОД состоят в контроле качества поступающей информации, накоплении новых данных в информационной базе, выполнении конкретных видов стандартной и нестандартной обработки данных.

Поэтому задаваемая форма базы данных должна так организовать логическую структуру записи информации, чтобы реализовать быстроту и свободу доступа к информации в условиях данного типа ЭВМ и сформулированного круга задач, решаемых с помощью БОД. Кроме собственно массивов наблюдений в базе данных обязательно хранится файл описаний, в котором определяется принятая логическая структура записей, иерархия их взаимного расположения. Здесь же хранится каталог имеющихся в наличии данных, где зафиксирована длительность наблюдений, их видовое разнообразие и местоположение на техническом носителе.

Функционирование базы данных невозможно без соответствующего программно-математического обеспечения, в котором выделяется отдельно системное обеспечение и комплекс программ обработки данных. Системное обеспечение представляет собой набор программ, с помощью которых отдельные составные части Банка Океанологических Данных организуются в единое целое. Набор программ обработки данных составляется таким образом, чтобы в рамках данного БОД рационально выполнять все виды обработки поступающей информации (см. разд. 1.2).

Выделим основные группы программ обработки данных.

1. Программы предварительной подготовки информации обеспечивают сортировку данных, их преобразование к удобному для вычислений виду; выполняют контроль качества данных; осуществляют наполнение базы данных.

2. Программы первичного статистического анализа основываются на вычислительных алгоритмах классической статистики и выполняют операции оценивания статистических моментов, эмпирических функций распределения и критериев проверки статистических гипотез.

3. Программы корреляционно-регрессионного анализа реализуют алгоритмы вычисления корреляционных и спектральных функций, построения уравнений регрессии и ряд более сложных методов изучения взаимосвязей между исследуемыми переменными.

4. Программы океанологических вычислений являются специфическими, направленными на решение конкретных задач при использовании БОД на практике. Это могут быть, например, программы расчета условной удельной плотности воды, скорости звука,

гидростатического давления, скоростей дрейфового и геострофического течения, параметров вертикальной устойчивости, коэффициентов обмена и другие.

Вся совокупность программно-математических средств является системой управления базами данных (СУБД). Для решения океанологических задач нередко разрабатываются специализированные СУБД. Такие СУБД направлены на эффективное решение отдельных наиболее важных океанологических задач. Например, на обработку результатов спутниковых наблюдений океана, на обработку наблюдений океанологических зондов и т. п. Сложная структура исходной информации и многообразие возможных задач ее обработки, пока не позволили создать единую общую СУБД для океанографии.

Наконец, комплекс технических средств, состоящий из ЭВМ, технических носителей и периферийных устройств ввода-вывода информации, является важнейшей составной частью материального обеспечения любого БОД. Именно прогресс в области производства средств вычислительной техники создает условия для усовершенствования БОД.

Основными требованиями, предъявляемыми к комплексу технических средств, следует считать быстрдействие вычислительных операций и процесса обмена информацией, объем памяти процессора и разнообразие видов представления обрабатываемой информации. Последнее — особенно важно для практической океанологии, где большое количество наблюдений до сих пор представляется в виде фотографий карт и графиков. Наличие устройств обработки видеoinформации позволяет в десятки раз ускорить процесс функционирования БОД и существенно повысить качество информации.

Комплекс технических средств наиболее крупных БОД формируется на базе Единой Системы ЭВМ. Как правило, это ЭВМ типа ЕС-1022, ЕС-1033, ЕС-1045, которые обладают высокими техническими характеристиками и богатым арсеналом периферийных устройств. Высокая эффективность функционирования БОД достигается применением сети дисплейных станций, на которых в режиме диалога человек — ЭВМ осуществляется формулировка рабочего задания и контроль процесса обработки.

Процесс организации БОД представляется весьма сложным, трудоемким и длительным. При создании БОД учитывается, что высокая эффективность его практического применения достигается при точном соответствии структуры банка тем задачам, которые будут решаться на его основе. Ведущими областями применения банков данных в океанологии следует считать:

- исследования фундаментальных проблем океана;
- решение прикладных задач обеспечения морского флота;
- изучение сырьевой базы океана (пищевых, минеральных, энергетических ресурсов).

Каждое из перечисленных направлений весьма специфично в силу особенностей по сути решаемых задач, отличий в методике обработки и анализа информации. Поэтому до создания нового БОД специалистом-океанологом должны быть сформулированы основные требования к будущему банку. Перечислим наиболее важные из них.

1. Нужно определить цель создания БОД и конкретные задачи, им решаемые.

2. Необходимо указать объем (количество наблюдений и признаков) и вид (шкалу наблюдений) хранимой информации; определить правила отбора информации в базу данных.

3. Необходимо сформулировать предполагаемую технологию обработки информации, указать требуемые методики вычислений.

4. Нужно задать вид представления, хранения и передачи результатов обработки исходной информации.

Строгое соблюдение этих требований должно обеспечить организацию надежных и эффективных банков океанологических данных.

1.4. Характеристика современных банков океанографических данных

Вследствие сложной физической структуры гидрометеорологической информации и большого числа решаемых с ее помощью задач возникает необходимость в разработке целой иерархии различных БОД. В системе организаций Госкомитета по гидрометеорологии с 1982 г. установлена иерархия режимно-справочных банков гидрометеорологических данных, которая состоит из двух основных уровней (рис. 1.2). Верхний уровень представляет глобальный БОД, нижний уровень отведен региональным и специализированным банкам данных.

Региональный банк океанографических данных (РБОД) содержит материалы комплексных океанографических наблюдений, выполненных в конкретном районе Мирового океана. Выбор границ района определяется принятым географическим делением океанов и морей и соответствующими задачами обработки информации. В СССР сформированы региональные архивы прежде всего по морям, омывающим территорию страны (Балтийскому, Черному, Азовскому), а также по акваториям отдельных океанов (Атлантическому, Тихому, Индийскому и Северному Ледовитому). Размещение региональных БОД производится в конкретных организациях, которые являются ведущими в процессе изучения отдельных акваторий Мирового океана.

Специализированный банк океанографических данных (СБОД) содержит специальную информацию об океане, подготовленную целевым образом. Подобные БОД создаются для накопления и переработки, как правило, узкого круга океанологических характеристик с целью решения ограниченного количества конкретных

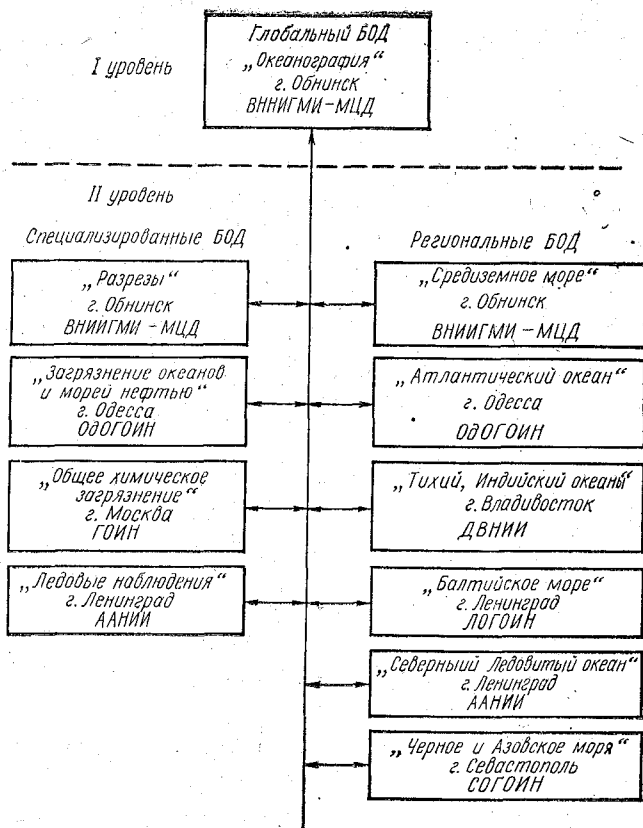


Рис. 1.2. Взаимосвязь режимно-справочных банков океанологических данных

проблем. Поэтому специализированные БОД, не претендуя на комплексность имеющейся информации, содержат данные, которые способствуют углубленному изучению конкретных свойств океана.

Примером СБОД может служить банк океанографических данных по загрязнению акватории Мирового океана, в котором накоплена информация по общему уровню загрязнения поверхности воды, а также по концентрациям отдельных загрязняющих веществ. Другой пример — БОД характеристик льда в Северном Ледовитом океане, созданный в ААНИИ в рамках автоматизированной ледово-информационной системы. В этом СБОД собраны материалы о количестве и формах льда, о его разрушенности, степени торошения и т. п.

Особый интерес представляют те СБОД, которые организованы на основе привлечения материалов наблюдений по крупным (международным) исследовательским программам. Так, во ВНИИГМИ-МЦД создан СБОД по программе «Разрезы», куда

включены наблюдения в энергоактивных зонах Мирового океана, проведенные в 1980-е годы. Направленность программы «Разрезы» на исследование механизмов короткопериодных колебаний климата обусловила включение в СБОД прежде всего параметров термодинамического взаимодействия атмосферы и океана.

Все рассмотренные специализированные и региональные БОД объединены в Глобальном Банке Данных «Океанография», где обобщаются одновременно функции и региональных, и специализированных БОД. В базе данных Глобального БОД содержится вся информация о состоянии океана и атмосферы над ним за всю историю наблюдений. База данных Глобального БОД пополняется, как правило, в первую очередь и включает в себя все материалы океанографических исследований, вплоть до самых современных результатов экспедиционных наблюдений. Тогда как базы данных в Региональных и специализированных БОД представляют собой, как правило, целевые выборки из Глобального БОД.

Общий объем наблюдений, хранящихся в Глобальном БОД, достигает величин 10^7 — 10^8 . Особенности распределения данных наблюдений по акватории Мирового океана можно проследить на примере хотя бы анализа батометрических данных (табл. 1.1). Как видно из табл. 1.1, количество батометрических станций явно недостаточно даже для анализа фонового распределения гидрофизических полей: в среднем на 1000 км^2 акватории океана приходится менее 1 наблюдения за всю историю исследования океано-

Таблица 1.1

Распределение батометрических станций в Мировом океане за всю историю наблюдений (по данным ВНИИГМИ-МЦД на 1985 г.)

Акватория океана	Общее число наблюдений (тыс.)	Число наблюдений		% от общего числа наблюдений
		на площадь в 1000 км^2	на одну трапецию $5^\circ \times 5^\circ$	
Атлантический	191,7	2,20	469	62,9
Северная часть	162,4	3,98	873	53,3
Южная часть	29,3	0,63	131	9,6
Тихий	101,0	0,60	137	33,1
Северная часть	82,6	1,08	279	27,1
Южная часть	18,4	0,20	42	6,0
Индийский	12,1	0,16	34	4,0
Северная часть	4,0	0,34	83	1,3
Южная часть	8,1	0,13	27	2,7
Мировой	304,8	0,92	204	100
Северное полушарие	249,0	1,92	470	81,7
Южное полушарие	55,8	0,28	58	18,3

графии. Естественно, что хуже всего освещена наблюдениями акватория океанов в южном полушарии, где сосредоточено всего 18,3 % от общего числа батометрических станций. Наименее освещен данными наблюдений Индийский океан, где одна батометрическая станция приходится на 10 000 км².

В качестве примера эффективной организации БОД рассмотрим особенности БОД по Балтийскому морю. Накопленные в БОД «Балтийское море» данные ориентированы прежде всего на решение задач по исследованию сезонной и межгодовой изменчивости гидрологических и гидрохимических процессов. Основными частями БОД являются база данных, программное обеспечение и комплекс технических средств.

База данных организована в соответствии с решаемыми задачами так, чтобы пользователь имел возможность эффективного доступа ко всем накапливаемым гидрометпараметрам (рис. 1.3). Прежде всего база содержит около 80 тыс. глубоководных батометрических станций, выполненных в акватории Балтийского моря с 1915 по 1980 год. В массив глубоководных наблюдений включены комплексные материалы наблюдений за гидрологическими, метеорологическими и гидрохимическими характеристикам. Логическая организация этих данных близко соответствует их представлению в таблицах ТГМ-ЗМ.

На прибрежных метеостанциях и постах в прибрежной зоне Балтийского моря наблюдения ведутся за температурой воды, соленостью, уровнем, элементами волн, ледовым покровом. В основном продолжительность подобных регулярных (1—3 сут.) наблюдений составляет 10—20 лет. Однако в массив включены данные 10 пунктов с рядами наблюдений более чем за 50 лет.

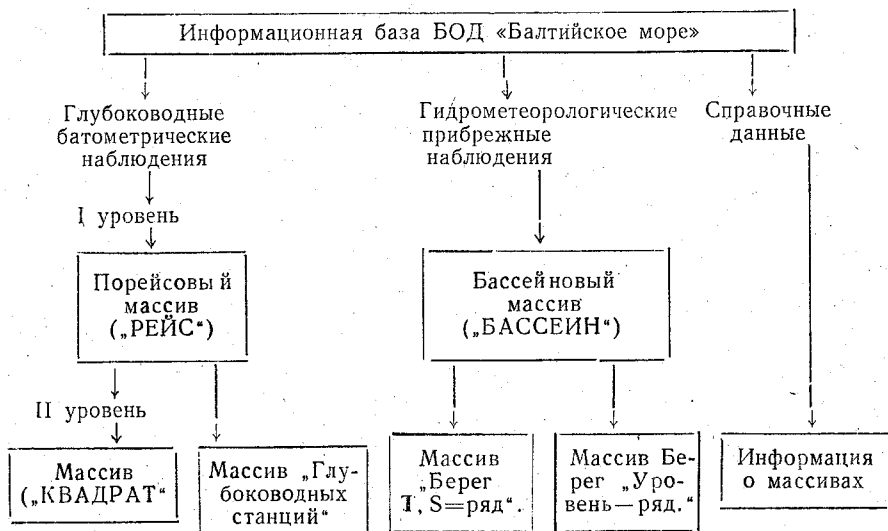


Рис. 1.3. Структура БОД «Балтийское море»

Каждый из представленных исходных массивов РЕИС и БАССЕЙН прост по своей организации, но не совсем удобен при анализе полей и рядов. Поэтому эти наблюдения в базе данных сохранены на I-м уровне. Одновременно с помощью программ, имеющихся в БОД, организованы специализированные массивы II-го уровня. Массивы II-го уровня КВАДРАТ и «ГЛУБСТАНЦИЯ — РЯД» содержат упорядоченные наблюдения из порейсового массива, которые рассортированы по пространственным трапециям $1^\circ \times 1^\circ$ или сведены в последовательные во времени ряды наблюдений в конкретных точках акватории.

Аналогично и массив прибрежных наблюдений «БАССЕЙН» на II-м уровне организации представлен двумя массивами. Массив «БЕРЕГ. Т, S-РЯД» содержит ряды Т, S-наблюдений на станциях, а массив «БЕРЕГ. УРОВЕНЬ — РЯД» — временные ряды уровней наблюдений.

Очевидно, что организация подобных массивов является весьма трудоемким технологическим процессом. Его обеспечение производится применением специализированного комплекса программ Автоматизированной Системы Обработки Океанологических Данных (АСООД). Программы АСООД обеспечивают расшифровку вновь поступающей информации, проверку ее качества, упорядочивание по пространству и времени, сортировку по значениям наблюдений, запись в соответствующие массивы I-го или II-го уровня. Кроме того, эти же программы позволяют выбрать необходимые материалы из массивов, организовать их в отдельную выборку и провести статистический анализ.

Эксплуатация БОД «Балтийское море» рассчитана на ЭВМ типа ЕС, которые имеют оперативную память более 512 кбт, необходимые накопители на магнитных дисках и лентах, печатающее устройство, устройство ввода с перфокарт и т. п.

ГЛАВА 2. ХАРАКТЕРИСТИКИ СТАТИСТИЧЕСКИХ РЯДОВ И ИХ СВОЙСТВА

2.1. Основные понятия случайных величин

Случайной величиной называется такая переменная величина, которая в результате испытания (измерений) в одинаковых условиях может принимать то или иное заранее неизвестное значение.

Различают два типа случайных величин: дискретные (прерывные) и непрерывные.

Случайная величина называется *непрерывной*, если она может принять любое значение из некоторого определенного диапазона числовой оси. Заметим, что этот диапазон может быть и бесконечным.

Случайная величина называется *дискретной*, если она принимает не любые, а только некоторые определенные значения (напри-

мер, степень волнения в баллах, количество облаков в баллах и т. п.). На практике все непрерывные случайные величины за счет ограниченной точности измерений и округлений заменяются дискретными.

Генеральная совокупность — весь возможный набор случайных величин. Генеральная совокупность может быть конечного и бесконечного объема. Хотя понятие генеральной совокупности представляет собой математическую абстракцию, оно широко используется в выводах теории вероятностей и математической статистики.

Любой ограниченный статистический ряд наблюдений $x_1, x_2, \dots, x_n$ случайной величины X можно рассматривать как выборку из генеральной совокупности. Поэтому параметры статистических рядов являются выборочными параметрами. Представительная выборка — это выборка, хорошо отражающая закономерности, присущие генеральной совокупности.

Чтобы полностью охарактеризовать (описать) случайную величину, необходимо не только указать, какие значения она может принимать, но и как часто, т. е. с какой вероятностью она принимает эти значения. Другими словами, нужно задать закон распределения случайной величины.

Законом распределения случайной величины называется 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 фигура (рис. 2.1) называется многоугольником распределения. Сумма ординат многоугольника распределения, представляющая собой сумму вероятностей всех возможных значений случайной величины, всегда равна единице.

Табличная форма закона распределения аналогичная графической форме, только в этом случае закон распределения задается в виде таблицы.

Аналитическая форма закона распределения описывается функцией распределения $F(x)$, которая определяет вероятность p того, что случайная величина X принимает значение меньше некоторого числа x : $F(x) = p(X < x)$.

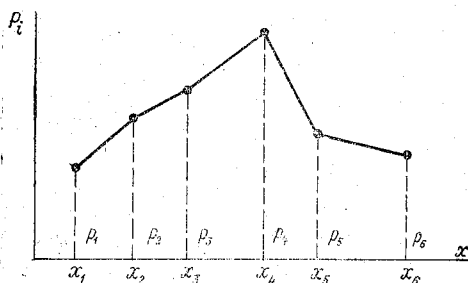
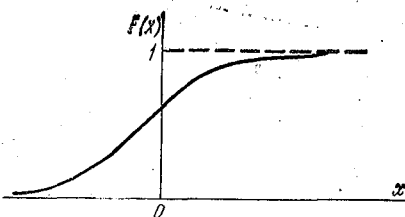


Рис. 2.1. Многоугольник распределения вероятностей случайной величины

Рис. 2.2. Функция распределения непрерывной случайной величины



Часто функцию распределения называют интегральной функцией распределения. Графиком ее для непрерывной случайной величины является непрерывная кривая, монотонно возрастающая от нуля до единицы (рис. 2.2).

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 значения. Этого недостатка лишена плотность распределения вероятностей, которая представляет собой первую производную от функции распределения, т. е.

$$f(x) = F'(x) = \lim_{\Delta x} \frac{p(x \leq X < x + \Delta x)}{\Delta x}.$$

Следовательно, плотность распределения есть предел отношения вероятности того, что случайная величина X при испытании примет значение, заключенное в интервале $[x, x + \Delta x]$, к величине Δx . Это означает, что плотность распределения есть вероятность, отнесенная к единичному интервалу изменения случайной величины.

В отличие от функции распределения плотность вероятности не является универсальной формой задания закона распределения и существует лишь для непрерывных случайных величин. График плотности распределения (рис. 2.3) называется *кривой распределения*.

2.2. Характеристики положения случайной величины

Характеристиками, определяющими положение случайной величины на числовой оси, являются математическое ожидание (среднее арифметическое), мода и медиана. Наиболее важной из них является математическое ожидание.

Математическим ожиданием дискретной случайной величины называется сумма произведений всех ее возможных значений на их вероятности, т. е.

$$M_x = x_1 p_1 + x_2 p_2 + \dots + x_n p_n. \quad (2.1)$$

Заметим, что понятие математического ожидания относится обычно к генеральной

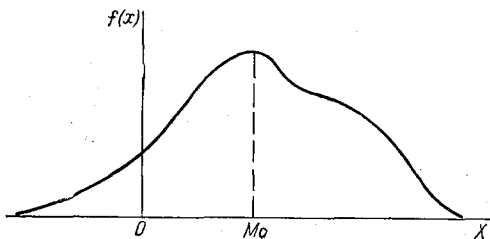


Рис. 2.3. Плотность распределения случайной величины

совокупности. При переходе от генеральной совокупности к выборке вместо математического ожидания говорят о среднем арифметическом значении или выборочном среднем.

Если вероятности случайной величины одинаковы ($p_1=p_2=\dots=p_n$), что соответствует, например, проведению измерений в одинаковых условиях, то в этом случае вместо формулы (2.1) имеем

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad (2.2)$$

где n — длина выборки (ряда). Среднее арифметическое значение характеризует центр тяжести (распределения) числового ряда.

Свойства средней арифметической.

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

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

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

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.

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

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.

Заметим, что 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.

Медианой называется значение ряда, занимающего среднее положение ранжированного ряда, т. е. расположенного в порядке убывания или возрастания значений x_i .

Если всем единицам ряда придать порядковые номера, то при нечетном числе членов ряда ($n=2m+1$) медиана определится как $Me=x_{m+1}$. При четном числе членов ряда ($n=2m$) за медиану условно принимается среднее значение между центральными значениями величин ранжированного ряда, т. е.

$$Me = \frac{1}{2} (x_m + x_{m+1}). \quad (2.3)$$

Геометрически медиана — это абсцисса точки, в которой площадь, ограниченная кривой плотности вероятности, делится пополам (рис. 2.4). Сказанное означает, что справедливо следующее равенство $p(X < Me) = p(X > Me) = 0,5$.

Модой называется наиболее часто встречающаяся в данном статистическом ряду величина. Геометрически мода представляет собой наибольшую ординату кривой плотности вероятности в случае одновершинного распределения (рис. 2.3). Поэтому одновершинное распределение называют *одномодальным*. В тех случаях, когда распределение имеет несколько вершин, его называют *многомодальным*.

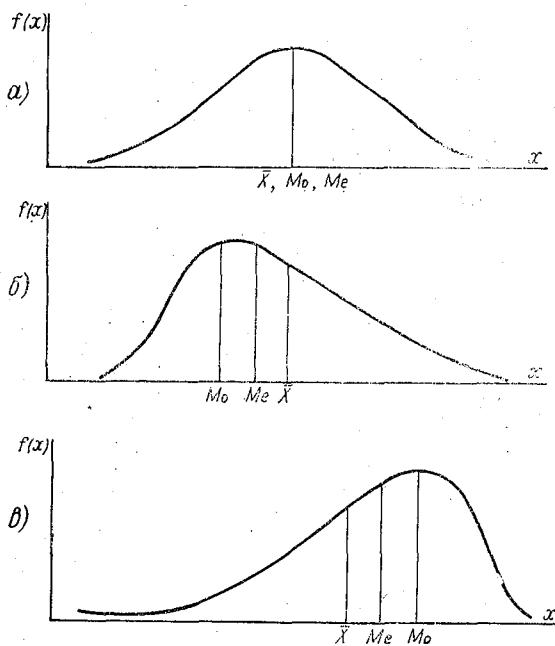


Рис. 2.4. Соотношение между средним арифметическим ($\bar{x}$), модой (Mo) и медианой (Me) для: а — симметричного распределения, б — положительно асимметричного, в — отрицательно асимметричного распределения

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

$$M_0 = \bar{x} + 3(Me - \bar{x}). \quad (2.4)$$

2.3. Характеристики рассеяния случайной величины

Простейшей мерой рассеяния (изменчивости) статистического ряда является амплитуда (размах) колебаний, определяемая как $A = x_{\max} - x_{\min}$, где $x_{\max}$, $x_{\min}$ — соответственно максимальный и минимальный члены ряда. Амплитуда 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.

Для достаточно длинных статистических рядов ($n > 30$) дисперсия определяется по следующей формуле:

$$D = \sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \quad (2.5)$$

и имеет размерность квадрата случайной величины. Среднее квадратическое отклонение представляет собой корень квадратный из дисперсии ($\sigma = \sqrt{D}$), и поэтому сохраняет размерность исходного ряда.

Коэффициент вариации $C = \sigma/\bar{x}$ является безразмерной величиной, поэтому он очень удобен для анализа рядов с различной размерностью.

Свойства дисперсии.

1. Дисперсия постоянной величины равна нулю $\sigma^2(a) = 0$.
2. Дисперсия остается постоянной, если все члены ряда увеличить или уменьшить на одно и то же число $\sigma^2(a+x) = \sigma^2$.
3. Постоянную величину можно выносить за знак дисперсии, возведя ее в квадрат $\sigma^2(ax) = a^2\sigma^2$.
4. Дисперсия алгебраической суммы независимых случайных величин равна сумме их дисперсий:

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

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

$$\sigma^2(x+y) = \sigma_x^2 + \sigma_y^2 + 2\sigma_x\sigma_y r_{xy}, \quad (2.6)$$

где r_{xy} — коэффициент корреляции между x и y , определяемый по формуле (5.1).

6. Дисперсия относительно среднего арифметического значения меньше, чем средний квадрат отклонений от любого значения в ряду x_0 , на величину $(x_0 - \bar{x})^2$.

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

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

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 обуславливает необходимость введения для описания случайных величин характеристик, позволяющих оценить степень асимметрии и крутости распределения.

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

$$As = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n\sigma^3}. \quad (2.7)$$

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. 2.4, 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. 2.4, 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. 2.4, в).

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

$$Ex = \frac{1}{n\sigma^4} \sum_{i=1}^n (x_i - \bar{x})^4 - 3. \quad (2.8)$$

$$M_P = 25,055$$

$$M_E = 26,84$$

$$M_1 = 23,04$$

$$A_5 = 0,08$$

$$E_x = -1,15$$

$$b = 1,58$$

$$\bar{x} = 25,08$$

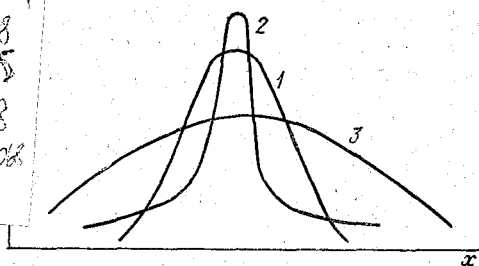


Рис. 2.5. Виды распределений с различной крутостью: 1 — нормальное распределение; 2 — распределение с положительным эксцессом; 3 — распределение с отрицательным эксцессом

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

Если эмпирическая кривая распределения является более островершинной по сравнению с нормальной кривой, то $E_x > 0$ (рис. 2.5). В результате мода эмпирического распределения должна быть больше моды нормального распределения ($M_o > M_o_N$).

Если эмпирическая кривая распределения является более плосковершинной по сравнению с нормальной кривой, то $E_x < 0$ (рис. 2.5). В этом случае $M_o < M_o_N$.

2.5. Понятие о статистических моментах

Все рассмотренные выше числовые характеристики случайной величины X являются взаимосвязанными и образуют единую систему, опирающуюся на понятие о моментах статистических совокупностей.

При описании свойств статистических рядов наиболее часто используются моменты двух видов: начальные и центральные.

Начальным моментом α_k порядка k дискретной случайной величины X называется выражение вида

$$\alpha_k = \frac{1}{n} \sum_{i=1}^n x_i^k. \quad (2.9)$$

Например, при $k=1$ имеем первый начальный момент, который эквивалентен среднему арифметическому значению.

Центральным моментом μ_k порядка k называется отклонение случайной величины X от ее среднего значения, т. е.

$$\mu_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k. \quad (2.10)$$

Соотношения между рассмотренными выше основными характеристиками статистических рядов и статистическими моментами приводятся в табл. 2.1. При этом указываются моменты до четвертого порядка включительно, так как их вполне достаточно для изучения случайных величин и решения многих практических задач.

$X_p = 93$ $X_{np} = 14$
 2,4 14

Таблица 2.1

Соотношения между числовыми характеристиками случайных величин и статистическими моментами

Характеристика статистического ряда	Выражение характеристики через моменты
$\bar{x}$	μ_1
σ^2	μ_2
σ	$\sqrt{\mu_2}$
C	$\sqrt{\mu_2/\mu_1}$
As	$\mu_3/\sqrt{\mu_2^3}$
Ex	$\mu_4/\mu_2^2 - 3$

Между центральными и начальными моментами существует определенная взаимосвязь. Так, например, первые четыре центральных момента выражаются через начальные моменты следующим образом.

Из всех моментов в качестве числовых характеристик случайных величин наиболее часто используются начальный момент первого порядка (среднее арифметическое значение) и центральный момент второго порядка (дисперсия). Кроме того, применение находят центральные моменты третьего и четвертого порядка, которые позволяют получить представление о виде кривой распределения.

2.6. Оценивание различных характеристик случайной величины по эмпирическим данным

Изменения уровня Каспийского моря представляют большой интерес для многих отраслей народного хозяйства. В межгодовом ходе уровня моря есть длительные периоды, когда он значительно понижался или повышался. Рассмотрим статистические оценки межгодовых изменений уровня по данным ст. Махачкала (табл. 2.2).

Таблица 2.2

Средние годовые значения уровня моря у ст. Махачкала, см

	0	1	2	3	4	5	6	7	8	9
1900	214	242	231	245	240	232	237	230	228	232
1910	190	192	190	179	190	210	219	222	206	204
1920	197	174	167	162	161	147	158	177	193	212
1930	24	184	191	188	173	148	128	108	75	45
1940	0	16	26	29	27	7	13	25	25	19
1950	-23	-13	-12	-26	-26	-35	-40	-31	-20	-19
1960	-37	-45	-53	-45	-38	-45	-36	-36	-46	-52
1970	-37	-46	-54	-66	-62	-75	-95	-101	-94	-51
1980		-12	-6	-2	3	12	9			

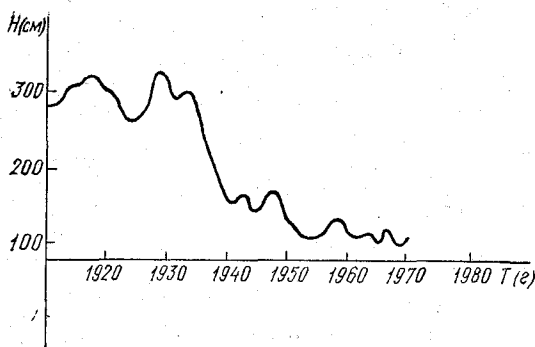


Рис. 2.6. Межгодовые колебания уровня Каспийского моря

Заметим, что приведенные в табл. 2.2 значения уровня даны в отклонениях от абсолютной отметки -28 м по отношению к Кронштадтскому футштоку. График межгодовых изменений уровня представлен на рис. 2.6.

Нетрудно увидеть, что в 1932—1940 гг. произошло резкое падение уровня (на 167 см), причины которого не вполне ясны даже в настоящее время. Затем, вплоть до 1970 г., наблюдалась определенная стабилизация. С 1972 г. началось очередное падение уровня, которое в 1977 г. достигло -29 м. Если бы уровень в течение еще нескольких лет продержался на отметках ниже -29 м, то это неминуемо привело бы к невосполнимым потерям в запасах осетровых и других ценных пород рыб. Однако уровень моря начал повышаться и впервые в истории отмечен непрерывный подъем уровня за десятилетний период.

Оценка среднеарифметического значения уровня за 86 лет составила 71,9 см или $-27,28$ м. Медиана за этот же период равняется $Me=25$ см. Величину моды для данного статистического ряда оценить практически нельзя, так как распределение величин уровня оказывается двухмодальным.

Размах колебаний уровня за 86 лет составил 345 см, т. е. более 3 м. Дисперсия и стандартное отклонение за этот же период равны 12770 см^2 и 113 см. Отсюда нетрудно определить коэффициент вариации, который составляет $C=1,57$. Коэффициент вариации очень велик, что на первый взгляд свидетельствует об исключительно высокой изменчивости уровня моря.

Однако в действительности данный коэффициент вариации является неинформативным. Если изменить отчетную шкалу для уровня, то средняя арифметическая величина также изменится, в то время как дисперсия и стандартное отклонение останутся постоянными. Например, если пересчитать данные табл. 2.2 в отклонениях от абсолютной отметки -30 м, то средняя арифметическая величина уровня окажется равной 272 см. Тогда коэффициент вариации уже составит $C=0,42$.

В статистических расчетах довольно часто используются центрированные данные $(x_i - \bar{x})$. Естественно, что среднее арифметическое центрированных данных равно нулю. В этом случае определение коэффициента вариации не имеет смысла.

Коэффициенты асимметрии и эксцесса межгодовых колебаний уровня составляют соответственно 0,2 и —1,6 см, т. е. очень малы. Следовательно, для описания межгодовых колебаний уровня моря достаточно ограничиться первыми двумя статистическими моментами.

ГЛАВА 3. ЗАКОНЫ РАСПРЕДЕЛЕНИЯ ОКЕАНОЛОГИЧЕСКИХ ЭЛЕМЕНТОВ

Закон распределения — есть способ обобщения и анализа статистических рядов, позволяющий в аналитическом виде представить их основные закономерности и структуру. В математической статистике случайная величина считается заданной, если известна ее функция распределения. Этим обстоятельством определяется фундаментальное значение законов распределения.

В настоящее время известно большое число самых разнообразных законов распределения. Рассмотрим те из них, которые получили наибольшее распространение в гидрометеорологии, а также сформулируем принципы построения эмпирической функции распределения и требования к статистическим оценкам ее характеристик.

3.1. Нормальный закон распределения. Его свойства и значение

Случайная величина X распределяется по нормальному закону (закону Гаусса), если ее плотность вероятности $f(x)$ определяется следующей формулой:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \bar{x})^2}{2\sigma^2}\right). \quad (3.1)$$

Как видно из этой формулы, достаточно определить параметры $\bar{x}$ и σ , чтобы нормальный закон случайной величины X считался заданным.

Из формулы (3.1) следует, что нормальная кривая $f(x)$ располагается симметрично относительно максимальной ординаты, равной $y = 1/\sigma\sqrt{2\pi}$ и проходящей через $\bar{x}$ (рис. 3.1). При $\bar{x} = 0$ нормальная кривая будет симметрична относительно начала координат. Заметим, что для нормального закона распределения должно выполняться условие $\bar{x} = M_0 = M_e$.

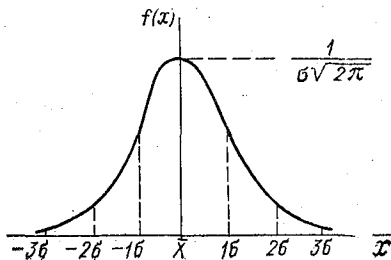


Рис. 3.1. Функция плотности вероятности нормального закона распределения

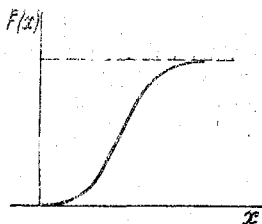


Рис. 3.2. Функция Лапласа

При изменении значения только параметра $\bar{x}$ ($\sigma = \text{const}$), кривая нормального распределения смещается параллельно оси абсцисс, не меняя своей формы. Изменение же значения параметра σ ($\bar{x} = \text{const}$) приводит к изменению формы кривой нормального закона. С возрастанием σ нормальная кривая становится все более плоской, растягиваясь вдоль оси абсцисс. При уменьшении σ , наоборот, кривая распределения сжимается с боков и вытягивается вдоль оси ординат (рис. 3.2).

Преобразуем выражение (3.1) к интегральному виду

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

Интеграл, входящий в эту формулу, может быть вычислен путем замены переменной. Обозначим

$$t = \frac{x - \bar{x}}{\sigma}, \quad (3.3)$$

где t — стандартизованная случайная величина, обладающая тем важным свойством, что при любом распределении случайной величины $\bar{t} = 0$ и $\sigma_t = 1$.

С учетом (3.3) имеем

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_0^t \exp(-t^2/2) dt. \quad (3.4)$$

Переход от величины x к t по существу означает перенос начала координат в центр распределения и выражение абсциссы в долях от стандартного отклонения. Интеграл (3.4) называется функцией Лапласа. Его численные значения в форме специальных таблиц (см. прил. 1) приводятся во многих курсах математической статистики и широко используются при решении различных задач.

При $\bar{x} \neq 0$ и $\sigma \neq 1$ между $F(x)$ и $f(x)$ существует следующее соотношение:

$$f(x) = F(x) \left(\frac{x - \bar{x}}{\sigma} \right).$$

Из прил. 1 видно, что удвоенная функция интеграла Лапласа равна единице, т.е. $2F(t=\infty) = 1$. Это означает, что площадь, ограниченная интегральной кривой распределения, также равна единице. Поэтому с помощью табулированных значений функции

Лапласа нетрудно вычислить площадь в пределах любого заданного интервала и, таким образом, определить вероятность того, что случайная величина, подчиняющаяся нормальному закону, находится в заданных границах.

Например, при $t=1, 2, 3, 4$, получаем следующие соотношения между числом наблюдений и пределами $\pm t\sigma$ на оси абсцисс нормальной кривой (рис. 3.1, 3.2): между -1 и $+1$ попадает 68,27 % всех наблюдений, между -2 и $+2$ содержится 95,40 % наблюдений, в интервале от -3 до $+3$ — 99,73 % наблюдений, и, наконец, в интервале от -4 до $+4$ содержится 99,99 % всех наблюдений.

Основная часть наблюдений попадает уже в интервал $\pm 2\sigma$. И лишь три из 1000 наблюдений имеют числовое значение, которое отклоняется от $\bar{x}$ на 3σ .

Основные свойства нормального закона. 1. Плотность вероятности всегда положительна и область существования функции $-\infty < x < \infty$.

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

3. Нормальная кривая не пересекается с осью Ox .

4. Если изменять $\bar{x}$, то кривая плотности вероятности перемещается вдоль оси x , сохраняя свою форму (с увеличением $\bar{x}$ — вправо, с уменьшением $\bar{x}$ — влево).

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

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

Значение нормального закона распределения. Главная особенность, выделяющая нормальный закон распределения среди многих других, состоит в том, что он является предельным, т. е. законом, к которому могут приближаться многие другие законы распределения при некоторых условиях.

В частности, это вытекает из центральной предельной теоремы, одна из форм которой была предложена советским математиком Ляпуновым. Теорема Ляпунова гласит: если случайные величины $X_1, X_2, \dots, X_n$ взаимно независимы, имеют конечные центральные моменты третьего порядка и если при неограниченном увеличении n выполняется условие

$$\lim_{n \rightarrow \infty} \left\{ \frac{\sum_{i=1}^n M[|X_i - \bar{X}|^3]}{D_x^{3/2}} \right\} = 0, \quad (3.5)$$

то распределение случайной величины $X = \sum_{i=1}^n X_i$ неограниченно (асимптотически) приближается к нормальному с параметрами $\bar{X} = \bar{x}_1 + \bar{x}_2 + \dots + \bar{x}_n$ и $\sigma_x = \sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2}$.

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

Поэтому в тех случаях, 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 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. Однако, как показывают результаты практических расчетов, во многих случаях достаточно четырех-пяти равноценных факторов, чтобы распределение случайной величины стало близким к нормальному закону.

Рассмотрим конкретный пример. В соответствии с уравнением теплового баланса изменения температуры поверхностного слоя воды (рис. 3.3) определяются следующими основными факторами.

1. Коротковолновым притоком солнечной радиации (R);
2. Длинноволновым эффективным излучением (J);
3. Турбулентным потоком тепла между океаном и атмосферой (H);

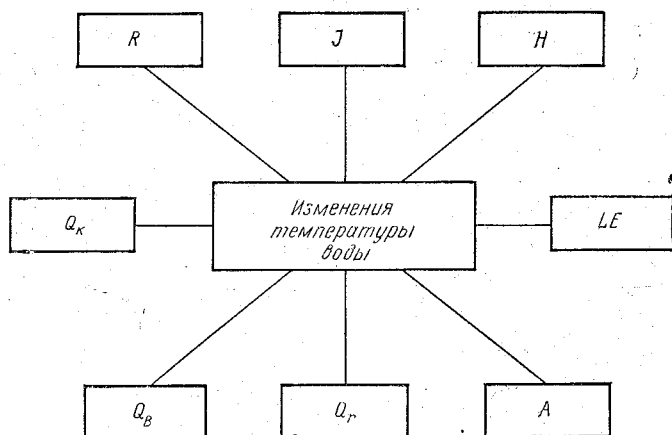


Рис. 3.3. Основные факторы изменения температуры поверхностного слоя воды

4. Затратами тепла на испарение (LE);
5. Адвекцией тепла течениями (A);
6. Горизонтальным турбулентным обменом (Q_T);
7. Вертикальным турбулентным обменом (Q_B);
8. Конвективным потоком тепла (Q_K).

Если пренебречь рядом других факторов (например, тепловыми эффектами от таяния или намерзания морских льдов), то имеем 8 факторов, влияющих на изменения температуры воды в поверхностном слое. Естественно, что значимость указанных факторов в значительной степени зависит как от масштабов временного осреднения процесса формирования теплового баланса, так и от географического района океана.

При изучении крупномасштабных процессов обычно принимают период осреднения, равный одному месяцу. Тогда наиболее важными факторами формирования годового цикла температуры воды в большинстве районов океана являются радиационные процессы. Именно вследствие преобладания радиационных факторов распределение среднемесячных величин температуры воды, как правило, не подчиняется нормальному закону.

Если же перейти к аномалиям температуры воды (раздел 5.4), исключив таким образом годовой цикл, обусловленный годовым ходом солнечной радиации, то 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.

Помимо центральной предельной теоремы, важное значение нормального закона состоит также в том, что он хорошо разработан, доступен и широко используется при решении многочисленных задач. Так, например, в математической статистике нормальное распределение играет роль некоторого стандарта, с которым сравниваются другие распределения. Можно упомянуть также, что нормальный закон лежит в основе теории ошибок. Наконец, нормальный закон распределения используется в методе наименьших квадратов, который, в свою очередь, является основой многих методов статистического анализа информации (например, корреляционный анализ).

В связи с этим проверка гипотезы нормальности распределения, т. е. степени соответствия эмпирического распределения нормальному, представляет один из наиболее важных этапов предварительной обработки наблюдений и будет рассмотрен в следующих разделах.

3.2. Законы распределения, используемые в статистических расчетах

Распределение Пирсона χ^2 . В статистических расчетах (теория ошибок, проверка гипотез, дисперсионный анализ и др.) часто

используются специальные распределения. К их числу, в частности, относится **распределение Пирсона χ^2** .

Пусть имеется n независимых случайных величин $X_1, X_2, \dots, X_n$, каждая из которых распределена по нормальному закону с нулевым средним значением и единичной дисперсией. Тогда новая случайная величина $\chi^2 = X_1^2 + X_2^2 + \dots + X_n^2$ имеет распределение, которое называется χ^2 с ν степенями свободы.

✓ Число степеней свободы — количество значений, функционально не связанных друг с другом или, другими словами, число независимых (свободных) 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 дает нам требуемую сумму. Таким образом, для статистического ряда число степеней свободы всегда равно $\nu = n - 1$.

Плотность распределения величины χ^2 имеет вид

$$f(\chi^2) = \left\{ 2^{\nu/2} \Gamma\left(\frac{\nu}{2}\right) \right\}^{-1} \chi^{2(\nu/2-1)} \exp(-0,5\chi^2), \quad (3.6)$$

где $\Gamma(\nu/2)$ — гамма-функция, $\Gamma(\nu/2) = \int_0^\infty t^{(\nu/2-1)} \exp(-t) dt$. Число степеней свободы определяется как $\nu = k - 1 - l$, где l — число параметров распределения. Поскольку $l = 1$, то $\nu = k - 2$.

Функция распределения χ^2 определяется по следующей формуле:

$$F(\chi^2) = \frac{\Gamma(\nu/2, \chi^2/2)}{\Gamma(\nu/2)} \quad \text{при } \chi^2 > 0.$$

где $\Gamma(\nu/2, \chi^2/2)$ — неполная гамма-функция.

Значения распределения χ^2 затабулированы для различных степеней свободы и уровней значимости (прил. 2).

На графике плотности вероятности распределения χ^2 для различных степеней свободы (рис. 3.4) видно, что оно резко несимметрично при малом числе ν . Однако с возрастанием ν плотность вероятности $f(\chi^2)$ становится все более симметричной и похожей

на кривую нормального распределения, что вытекает из центральной предельной теоремы. Практически при $\nu = 13$ —15 случайная величина χ^2 уже подчиняется нормальному закону.

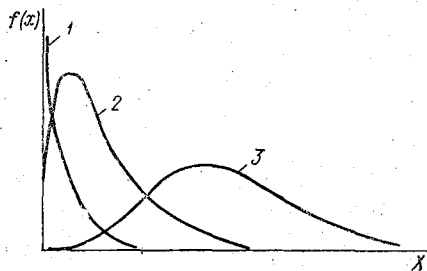


Рис. 3.4. Функция плотности вероятности χ^2 -распределения Пирсона для различных степеней свободы (ν):

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 $\bar{Z}=0$, $\sigma_Z^2=1$, а V — распределенной по закону χ^2 с ν степенями свободы. Тогда случайная величина

$$t = \frac{Z}{\sqrt{V/\nu}}$$

имеет распределение, которое называется распределением Стьюдента с ν степенями свободы. Плотность вероятности величины t выражается следующей формулой:

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi} \Gamma(\nu/2)} [1 + t^2/\nu]^{-(\nu+1)/2}. \quad (3.7)$$

Из графика плотности вероятности $f(t)$ видно, что она симметрична относительно начала координат (рис. 3.5). По мере увеличения числа степеней свободы t -распределение приближается к нормальному закону, причем скорость этого приближения выше, чем у распределения χ^2 .

Функция распределения величины t имеет довольно сложное выражение:

$$F(t) = \frac{\Gamma(\nu+1/2)}{\sqrt{\pi} \Gamma(\nu/2)} \int_{-\infty}^t \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} dt.$$

Значения t -распределения затабулированы для различных степеней свободы и уровней значимости (прил. 3).

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. Плотность вероятности этого распределения выражается следующей формулой:

$$f(F) = \frac{\nu_1^{\nu_1/2} \nu_2^{\nu_2/2} \Gamma\left(\frac{\nu_1 + \nu_2}{2}\right)}{\Gamma(\nu_1/2) \Gamma(\nu_2/2)}, \quad (3.8)$$

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 (3.8), распределение Фишера (F -распределение) не зависит от дисперсий исходных выборок, а зависит лишь

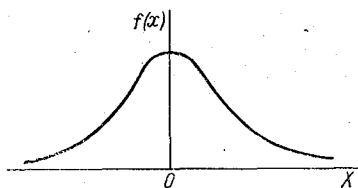


Рис. 3.5. Функция плотности вероятности t -распределения Стьюдента для степеней свободы $\nu=9$

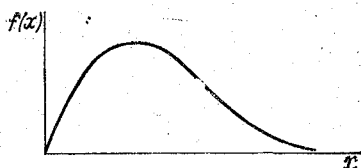


Рис. 3.6. Функция плотности вероятности F -распределения Фишера для степеней свободы $\nu_1=10$; $\nu_2=25$

от числа степеней свободы. График плотности вероятности $f(F)$ приведен на рис. 3.6.

Для F -распределения составлены таблицы значений для различных чисел степеней свободы и уровней значимости (прил. 4).

Заметим, что эти таблицы даны для одностороннего критерия значимости, т. е. когда проверяется, например, условие $D_1 > D_2$ или $D_1 < D_2$. В том случае, если необходимо проверить равенство дисперсий по двум выборкам, т. е. $D_1 = D_2$, то вероятность таблиц удваивается.

3.3. Законы распределения, используемые в гидрометеорологии

Логарифмически нормальное распределение. При решении практических задач нередко встречаются такие ситуации, когда исследуемая случайная величина X сама не является нормально распределенной, однако путем несложного ее функционального преобразования удастся получить случайную величину $Y = \varphi(X)$, распределенную по нормальному закону.

Наибольшее распространение получило логарифмическое преобразование случайной величины вида

$$Y = \log_a X,$$

которое допустимо лишь при $X > 0$. Тогда плотность распределения случайной величины Y выражается формулой

$$f(y) = \frac{1}{\sigma_y \sqrt{2\pi}} \exp \left[-\frac{(y - \bar{y})^2}{2\sigma_y^2} \right]. \quad (3.9)$$

Выполнив несложные преобразования, можно от формулы (3.9) перейти к плотности распределения исходной случайной величины X , которая будет иметь следующий вид:

$$f(x) = f(y) \frac{dy}{dx} = \frac{\log_a e}{x \sigma_y \sqrt{2\pi}} \exp \left[-\frac{(y - \bar{y})^2}{2\sigma_y^2} \right]. \quad (3.10)$$

Обычно в практических расчетах используется натуральный логарифм, т. е. $Y = \ln X$. В этом случае

$$f(x) = \frac{1}{x \sigma_y \sqrt{2\pi}} \exp \left[-\frac{(y - \bar{y})^2}{2\sigma_y^2} \right], \quad (3.11)$$

где $x > 0$, а $\bar{y}$ и σ_y — параметры логарифмически нормального закона распределения.

Рис. 3.7. Функция плотности вероятности логарифмически нормального распределения с параметрами: $C=0,4$; $As=1,3$; $Ex=2,9$

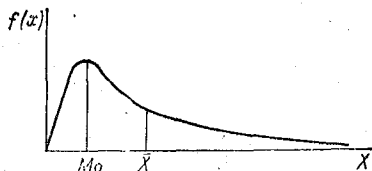


График плотности логарифмически нормального распределения приведен на рис. 3.7. Нетрудно увидеть, что данное распределение характеризуется положительной асимметрией, возрастающей с увеличением σ_y . Оно свойственно таким случайным величинам, формирование которых происходит в результате умножения большого числа влияющих на них независимых равноценных факторов.

Между параметрами нормального и логарифмически нормального распределений существуют следующие соотношения:

$$\bar{x} = \exp(\bar{y} + 0,5\sigma_y^2),$$

$$\sigma_x = \bar{x} \sqrt{\exp(\sigma_y^2) - 1},$$

или

$$\bar{y} = \ln \frac{\bar{x}}{\sqrt{1 + C_x^2}},$$

$$\sigma_y = \sqrt{\ln(1 + C_x^2)},$$

где C_x — коэффициент вариации случайной величины X .

Распределение Вейбулла. Плотность вероятности случайной величины X , распределенной по закону Вейбулла, определяется формулой

$$f(x) = \begin{cases} 0 & \text{при } X < 0 \\ m \cdot c \cdot x^{m-1} \exp(-c \cdot x^m) & \text{при } X \geq 0 \end{cases}, \quad (3.12)$$

где m и c — параметры распределения Вейбулла, которые могут иметь 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.

Функция распределения закона Вейбулла выражается формулой

$$F(x) = \begin{cases} 0 & \text{при } X < 0 \\ 1 - \exp(-c \cdot x^m) & \text{при } X \geq 0 \end{cases}. \quad (3.13)$$

Заметим, что некоторые виды распределений являются частными случаями распределения Вейбулла. Действительно, при

$m=1$ имеем *показательное* распределение, плотность вероятности которого определяется как

$$f(x) = \begin{cases} 0 & \text{при } x < 0 \\ c \cdot \exp(-c \cdot x) & \text{при } x \geq 0 \end{cases}, \quad (3.14)$$

а функция распределения показательного (экспоненциального) закона имеет вид

$$F(x) = \begin{cases} 0 & \text{при } x < 0 \\ 1 - \exp(-c \cdot x) & \text{при } x \geq 0 \end{cases}.$$

График плотности вероятности распределения показательного закона приводится на рис. 3.8. Важным свойством показательного закона является то, что его математическое ожидание и среднее квадратическое отклонение равны и функционально связаны с параметром c , т. е.

$$\bar{x} = \sigma_x = 1/c.$$

Кроме того, для показательного распределения характерно и то, что его коэффициенты вариации, асимметрии и эксцесса не зависят от параметра и имеют следующие постоянные значения:

$$c = 1, \quad As = 2, \quad Ex = 9.$$

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

$$f(x) = \begin{cases} 0 & \text{при } x < 0 \\ (x/\sigma^2) \exp(-x^2/2\sigma^2) & \text{при } x \geq 0 \end{cases}, \quad (3.15)$$

где σ — единственный определяемый параметр.

График плотности вероятности распределения Релея при различных значениях σ представлен на рис. 3.9. Нетрудно заметить, что кривая плотности вероятности, особенно при малых значениях параметра σ , является резко асимметричной.

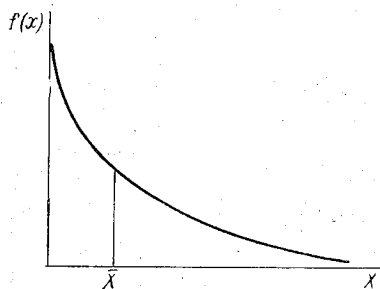


Рис. 3.8. Функция плотности вероятности показательного распределения

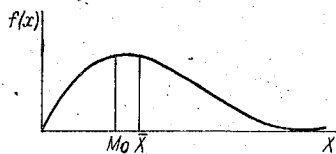


Рис. 3.9. Функция плотности вероятности распределения Релея с параметрами: $As=0,59$, $Ex=-0,17$

Функция распределения закона Релея выражается формулой

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

Статистические моменты данного распределения определяются через параметр σ следующим образом:

$$\bar{x} = \sqrt{\frac{\pi}{2}} \sigma, \quad D = \left(2 - \frac{\pi}{2}\right) \sigma^2, \\ \mu_3 = (\pi - 3) \sqrt{\frac{\pi}{2}} \sigma^3, \quad \mu_4 = \left(8 - \frac{3\pi^2}{4}\right) \sigma^4.$$

Или с учетом соотношений, приведенных в табл. 2.1, получаем, что $As \approx 0,63$, $Ex \approx -0,3$. Таким образом, кривая распределения Релея имеет большую положительную асимметрию и является более плосковершинной по сравнению с кривой нормального закона.

3.4. Требования к статистическим оценкам характеристик распределения

Так как элементы любой выборки содержат некоторые ошибки, то определенные по этой выборке параметры закона распределения не могут быть абсолютно точными. Поэтому речь может идти только об их приближенной оценке.

Допустим, что в нашем распоряжении имеется некоторая весьма обширная совокупность наблюдений. Разобьем данную совокупность на m выборок объемом в n членов каждую согласно табл. 3.1. Заметим, что для генеральной совокупности $m \rightarrow \infty$. Во всех остальных случаях m является конечной величиной.

Для каждой из m выборок может быть рассчитан тот или иной параметр a_j^* , характеризующий какие-либо свойства статистического ряда ($\bar{x}$, σ и др.). Такая статистическая оценка параметра a_j^* называется *выборочной*. Поскольку значения a_j^* можно рассматривать как случайные, то они имеют свой закон распре-

Таблица 3.1

Схема разбиения совокупности наблюдений на m выборок
объемом в n членов

$j \backslash i$	1	2	...	n	Выборочная оценка, a^*
1	x_{11}	x_{12}	...	x_{1n}	a_1^*
2	x_{21}	x_{22}	...	x_{2n}	a_2^*
...	...	...	...	...	...
m	x_{m1}	x_{m2}	...	x_{mn}	a_m^*

ления. При $m \rightarrow \infty$ выборочные оценки параметра a_j^* стремятся к истинным оценкам, т. е. $a_j^* \rightarrow a_j$.

Для того, чтобы какая-либо статистическая оценка имела бы практическую ценность, она должна удовлетворять ряду требований: быть несмещенной, состоятельной и эффективной.

Оценка называется *несмещенной*, если ее математическое ожидание равно оцениваемому параметру, т. е. $M(a^*) = a$. Это означает, что при любой продолжительности исходного ряда (объема выборки) центр распределения статистической характеристики должен совпадать с ее истинным (действительным) значением.

Оценка a^* называется *состоятельной*, если она сходится по вероятности к оцениваемому параметру при неограниченном возрастании объема выборки n , т. е.

$$\lim_{n \rightarrow \infty} p[|a^* - a| < \varepsilon] = 1,$$

где ε — сколь угодно малое наперед заданное положительное число. Для выполнения этого требования достаточно, чтобы дисперсия оценки при увеличении n стремилась к нулю, т. е.

$$\lim_{n \rightarrow \infty} D[a^*] = 0.$$

Требование состоятельности оценки означает, что с увеличением объема выборки рассеивание оценок относительно математического ожидания, совпадающего с действительным значением числовой характеристики, будет уменьшаться и при достаточно большом n отклонение a^* от a с $p \rightarrow 1$ должно быть меньше любого наперед заданного числа.

Оценка называется *эффективной*, если она является несмещенной и при этом обладает следующим свойством: $D[a^*] = D_{\min}$, где $D_{\min}$ — минимальная дисперсия среди всех несмещенных оценок параметра a^* . Это означает, что эффективная оценка имеет меньшую вероятность появления грубой ошибки при определении параметров распределения.

Для эффективности оценки самым важным является задание закона распределения. Эффективная оценка параметра генеральной совокупности для одного закона распределения не совпадает с эффективной оценкой параметра другого распределения.

Оценка среднего арифметического обладает всеми тремя вышеперечисленными свойствами: она является несмещенной, состоятельной и эффективной. Оценка дисперсии состоятельна и эффективна, но имеет отрицательное смещение. Поэтому для не очень больших выборок ($n < 25-30$) вместо n используют $n-1$, что позволяет устранить смещение.

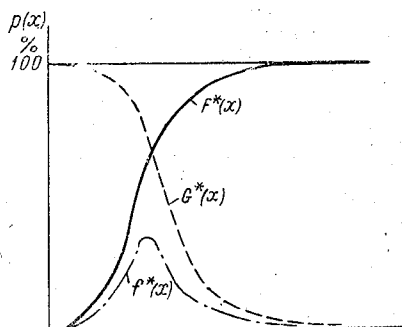
3.5. Принципы построения эмпирической функции распределения

Эмпирической (статистической) *функцией распределения* $F^*(x)$ случайной величины X называется закон изменения частоты событий $X < x$ в данной статистической выборке, т. е.

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

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

где $p^* = m/n$ — частота события $X < x$; m — число событий в данном интервале (классе) k , т. е. эмпирическая повторяемость; n — длина выборки.



При $n \rightarrow \infty$ $p^* \rightarrow p$, где p — вероятность события $X < x$ и $F^*(x) \rightarrow F(x)$, где $F(x)$ — функция распределения. В гидрометеорологических расчетах в некоторых случаях используется соотношение, имеющее следующий вид: $G(x) = p^*(X \geq x)$, которое называется *эмпирической функцией обеспеченности*. Графическое изображение эмпирической функции обеспеченности называется *эмпирической кривой обеспеченности* (рис. 3.10).

Если объем выборки n весьма велик, то построение надежных в статистическом смысле эмпирических функций распределения и обеспеченности не представляет затруднений. Однако, если в каждом интервале $m < 7-8$, то для повышения надежности результатов обычно используется ряд приемов.

Прежде всего исходный ряд наблюдений ранжируется, т. е. располагается в убывающем порядке от максимального значения до минимального $x_{\min}$. Важным представляется вопрос оптимального определения числа интервалов (классов).

Здесь 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 мало наблюдений, в результате чего групповые частоты p^* начинают обнаруживать существенные случайные колебания.

С другой стороны, при уменьшении числа интервалов случайные колебания величин p^* сглаживаются, но вместе с тем сглаживаются и характерные черты распределения.

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.

После определения количества градаций находится в первом приближении их ширина 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 зачастую принимается соответствие в числе значащих цифр Δc и x .

После этого определяются границы классов (c_1, c_2) , (c_2, c_3) , ..., (c_k, c_{k+1}) , причем для крайних границ c_1 и c_{k+1} и крайних членов выборки $x_{\max}$ и $x_{\min}$ должны выполняться условия $c_1 \leq x_{\min}$, $c_{k+1} \geq x_{\max}$.

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.

Вследствие неравноточности 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 $n \rightarrow \infty$. В средних частях кривой распределения данные формулы дают практически одинаковые результаты и различаются лишь для нижней и верхней частей кривой распределения.

3.6. Приближенные методы оценки соответствия эмпирических и теоретических функций распределения

В общем случае оценка соответствия эмпирических функций распределения теоретическим функциям может быть выполнена с помощью как упрощенных, так и более строгих методов. При-

ближенные методы позволяют производить быструю проверку нормальности распределения с помощью различных простых критериев. Более строго это можно сделать на основе законов распределения χ^2 и Колмогорова, которые получили название критериев согласия. ✓

Рассмотрим некоторые из приближенных методов:

а) если с достаточно высокой степенью точности выполняются следующие условия:

$$\bar{x} = Mo = Me,$$

$$Ex = As = 0,$$

то рассматриваемый статистический ряд распределен по нормальному закону. В качестве меры точности для первого соотношения может быть использована относительная ошибка, величина которой задается априори и зависит от цели исследования, надежности исходных данных, их изменчивости и других факторов. Оценить степень близости величин As и Ex к нулю можно путем их сравнения с вспомогательными коэффициентами U_3 и U_4 , которые вычисляются как

$$U_3 = \sqrt{\frac{6(n-1)}{(n+1)(n+3)}},$$

$$U_4 = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+5)}},$$

где n — объем выборки.

При $As/U_3 < 2-3$ и $Ex/U_4 < 2-3$ считается, что ряд распределен по нормальному закону;

б) определяется безразмерный размах варьирования ряда A/σ и составляется соотношение

$$(A/\sigma)_{\text{табл}}^{\text{ниж}} < A/\sigma < (A/\sigma)_{\text{табл}}^{\text{верх}},$$

где табличные значения нижней и верхней границ этого соотношения для различных уровней значимости приводятся в табл. 3.2. Если величина A/σ выходит за пределы доверительных границ, то исходная выборка не подчиняется нормальному закону;

в) определяется доля (часть) исходной выборки, содержащаяся в интервалах $\bar{x} \pm B\sigma$, где B — число Вестергарда (табл. 3.3). Если для всех этих интервалов доля исходной выборки (n_i/n) не превышает некоторых критических (пороговых) величин $R_{кр}$, то гипотеза о нормальности статистического ряда выполняется.

г) для не очень больших выборок ($n < 120$) следует вычислить среднее абсолютное отклонение (CAO) по формуле

$$CAO = \sum_{i=1}^n |x_i - \bar{x}|/n.$$

Таблица 3.2

Критические границы отношения A/σ

Объем выборки	Нижние границы				Верхние границы			
	Вероятность ошибки							
	0,01	0,025	0,05	0,10	0,10	0,05	0,025	0,01
10	2,51	2,59	2,67	2,76	3,57	3,686	3,777	3,875
12	2,64	2,72	2,80	2,90	3,78	3,91	4,02	4,134
14	2,75	2,83	2,92	3,02	3,95	4,09	4,21	4,34
16	2,84	2,93	3,01	3,12	4,09	4,24	4,37	4,52
18	2,92	3,01	3,10	3,21	4,21	4,37	4,51	4,67
20	2,99	3,09	3,18	3,29	4,32	4,49	4,63	4,80
25	3,15	3,24	3,34	3,45	4,53	4,71	4,87	5,06
30	3,27	3,37	3,47	3,59	4,70	4,89	5,06	5,26
35	3,38	3,48	3,58	3,70	4,84	5,04	5,21	5,42
40	3,47	3,57	3,67	3,79	4,96	5,16	5,34	5,6
50	3,62	3,73	3,83	3,95	5,14	5,35	5,54	5,77
60	3,75	3,86	3,96	4,08	5,29	5,51	5,70	5,94
70	3,85	3,96	4,06	4,19	5,41	5,63	5,83	6,07
80	3,94	4,05	4,16	4,28	5,51	5,73	5,93	6,18
90	4,02	4,13	4,24	4,36	5,60	5,82	6,03	6,27
100	4,10	4,21	4,31	4,44	5,68	5,90	6,11	6,36
150	4,38	4,48	4,59	4,72	5,96	6,18	6,39	6,64
200	4,59	4,68	4,78	4,90	6,15	6,39	6,60	6,84
500	5,13	5,25	5,37	5,49	6,72	6,94	7,15	7,42

Таблица 3.3

Условия нормальности распределения по Вестергарду

Число Вестергарда	И н т е р в а л ы	$R_{кр}$
0,3	от $\bar{x} - 0,3\sigma$ до $\bar{x} + 0,3\sigma$	0,25
0,7	от $\bar{x} - 0,7\sigma$ до $\bar{x} + 0,7\sigma$	0,50
1,1	от $\bar{x} - 1,1\sigma$ до $\bar{x} + 1,1\sigma$	0,75
3,0	от $\bar{x} - 3,0\sigma$ до $\bar{x} + 3,0\sigma$	0,99

Для выборки, имеющей приближенно нормальный закон распределения, должно быть справедливо выражение

$$|CAO/\sigma - 0,7979| < \frac{0,4}{\sqrt{n}}.$$

В противном случае считается, что выборка не подчиняется нормальному закону.

3.7. Определение доверительных границ для среднего значения и дисперсии выборки

В статистических расчетах часто возникает необходимость в так называемых интервальных оценках выборочных параметров или, другими словами, в определении с заданной степенью достоверности интервала, в который будет заключено значение оцениваемого параметра.

Аналитически интервальная оценка произвольного выборочного параметра θ^* , имеющего некоторое теоретическое распределение, может быть записана в виде $p(\theta_n < \theta^* < \theta_v) = 1 - \alpha$, где θ_n , θ_v — соответственно нижняя и верхняя доверительные границы, т. е. такие значения случайной величины, выход за пределы которых имеет наперед заданную вероятность, соответствующую уровню значимости α .

Доверительный интервал — это область значений случайной величины внутри доверительных границ. При симметричности доверительного интервала относительно оценки θ его нижняя и верхняя границы определяются как

$$\theta_n = \theta^* - \varepsilon,$$

$$\theta_v = \theta^* + \varepsilon,$$

где ε — половина длины доверительного интервала при заданном уровне значимости, показывающем процент ошибочных случаев. Так, если $\alpha = 0,05$, то это означает, что в результате случайности ошибка в выводах будет наблюдаться в 5 % наблюдений.

Для установления доверительных интервалов требуется знать закон распределения случайной величины. В тех случаях, когда в выборке $n > 30$, для всех параметров условно принимается нормальный закон распределения.

Если же статистическая совокупность мала, то при определении доверительных интервалов возникают дополнительные трудности. Из статистики известно, что в случае, если случайная величина распределена нормально, то выборка из таких средних арифметических величин распределяется по закону Стьюдента, а показатели изменчивости σ , S — по распределению χ^2 .

Доверительный интервал для параметра x определяется по формуле:

$$\bar{x} - t_{2\alpha} \frac{\sigma}{\sqrt{n}} < M < \bar{x} + t_{2\alpha} \frac{\sigma}{\sqrt{n}}. \quad (3.16)$$

Данная формула означает, что истинная оценка среднего арифметического, т. е. математическое ожидание, лежит в пределах

$$\left(\varepsilon = \bar{x} \pm t_{2\alpha} \frac{\sigma}{\sqrt{n}} \right).$$

Для определения доверительного интервала дисперсии нормально распределенной случайной величины X используют следующее выражение:

$$p\left(\frac{nD^*}{\chi_2^2} < D < \frac{nD^*}{\chi_1^2}\right) = 1 - \alpha,$$

где D^* — выборочное значение дисперсии, D — истинная оценка дисперсии; значения χ_1^2 и χ_2^2 выбираются так, чтобы выполнялось следующее условие: $p(\chi^2 < \chi_1^2) = p(\chi^2 > \chi_2^2) = \alpha/2$.

С учетом этого условия доверительный интервал для дисперсии приближенно определяется формулой

$$n \frac{D^*}{\chi_2^2} < D < n \frac{D^*}{\chi_1^2}. \quad (3.17)$$

Аналогичным образом находится истинная оценка для стандартного отклонения нормально распределенной случайной величины

$$\sqrt{n} \frac{\sigma^*}{\chi_2} < \sigma < \sqrt{n} \frac{\sigma^*}{\chi_1}.$$

Числа χ_1^2 и χ_2^2 находятся по значениям $v = n - 1$ и α из таблицы χ^2 -распределения как квантили $\chi_{\alpha/2, v}^2$ и $\chi_{(1-\alpha/2), v}^2$.

✓ 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. 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.

✓ 4.1. Понятие о нулевой гипотезе и уровне значимости

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

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

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

Если имеется нулевая гипотеза, то обязательно должны существовать альтернативные (противоположные) гипотезы. Вообще говоря, их может быть бесчисленное множество, однако в некоторых простых случаях их удастся свести к единственной альтернативе. Например, в рассматриваемом примере альтернативная гипотеза имеет вид

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

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 параметрические и непараметрические критерии.

В результате 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. Сказанное можно проиллюстрировать табл. 4.1.

Только принятие правильной или отклонение неправильной гипотезы можно считать верным решением. Если нулевая гипотеза отвергается, в то время как на самом деле она верна, то возникает

Таблица 4.1

Возможные комбинации принятия (отвержения) нулевой гипотезы

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

ошибка, называемая ошибкой первого рода. Наоборот, если ошибочная гипотеза принимается, то совершается ошибка второго рода.

Вероятность появления ошибки первого рода называется уровнем значимости α . Если величина α всегда задается заранее, то, вообще говоря, вероятность появления 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 .

Таким образом, 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 ($\bar{x}_1 - \bar{x}_2$) можно ожидать не менее чем пять раз на 100 испытаний. Если частота появления исследуемой статистики окажется меньше указанной разности, то гипотеза опровергается.

Значения статистики, при которых гипотеза опровергается, т. е. вероятность которых меньше заданного уровня значимости α , образуют критическую область проверяемой гипотезы. В связи с этим задача проверки гипотезы сводится к построению критической области для выбранного уровня значимости. Если статистика попадет в критическую область (рис. 4.1), то это указывает на несоответствие гипотезы наблюдаемым данным и нулевая гипотеза опровергается.

Вообще говоря, выбор уровня значимости является произвольным. Обычно он устанавливается на основе опыта как уровень, дающий практическую уверенность, что ошибочные заключения будут сделаны только в очень редких случаях. Наиболее часто используются уровни значимости 1, 5 и 10 %.

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

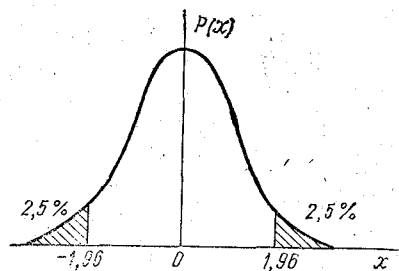


Рис. 4.1. Кривая нормального распределения с двумя критическими областями (заштрихованы), охватывающими 5 % площади под кривой

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.

При проверке гипотез следует различать двусторонний и односторонний уровни значимости. Если для нас одинаково представляют интерес как положительные, так и отрицательные разности между изучаемыми объектами, то, например, при 5%-ном уровне значимости берется по 2,5 % вероятности отклонений соответственно в положительной и отрицательной области.

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.

4.2 Оценка расхождения между значениями двух выборок

Пусть мы имеем две выборки объемом n_x и n_y , причем не обязательно $n_x = n_y$, взятые из генеральных совокупностей X и Y . Требуется оценить степень расхождения между выборочными средними $\bar{x}$ и $\bar{y}$.

Задаемся нулевой гипотезой, что математические ожидания случайных величин X и Y равны, т. е. $H_0: M_x = M_y$.

Тогда альтернативная гипотеза имеет вид $H_0: M_x \neq M_y$.

При больших объемах исходных выборок для проверки H_0 используется нормированная разность средних выборочных значений

$$z^* = \frac{|\bar{y} - \bar{x}|}{\sigma_{(\bar{y} - \bar{x})}}, \quad (4.1)$$

где

$$\sigma_{(\bar{y} - \bar{x})} = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}.$$

Этот критерий употребляется приблизительно при $n > 30$, когда величина z распределяется нормально. Поэтому проверка нулевой гипотезы осуществляется в следующем порядке.

1. Задается уровень значимости α .
2. По уровню значимости α из таблицы значений функции Лапласа (прил. 1) находится критическое значение $z_{кр}$ как квантиль $z_{(1-\alpha/2)}$ при двусторонней критической области.
3. По формуле (4.1) вычисляется выборочное значение z^* .
4. Проверяется соответствие выдвинутой нулевой гипотезы результатам наблюдений с использованием следующего решающего правила: если $|z^*| \geq z_{кр}$, то нулевая гипотеза отвергается как

несостоятельная; т. е. математические ожидания случайных величин X и Y не равны друг другу.

Если в нашем распоряжении имеются две выборки малого объема с неизвестным распределением, то при условии равенства дисперсий $\sigma_x^2 = \sigma_y^2$ для проверки H_0 используется критерий

$$t^* = \frac{|\bar{y} - \bar{x}|}{\sqrt{n_x \sigma_x^2 + n_y \sigma_y^2}} \sqrt{\frac{n_x n_y (n_x + n_y - 2)}{n_x + n_y}}. \quad (4.2)$$

Критерий t^* подчиняется распределению Стьюдента с числом степеней свободы $\nu = n_x + n_y - 2$. Последовательность проверки нулевой гипотезы такая же, как в предыдущем случае, только величина $t_{кр}$ выбирается из таблиц Стьюдента (прил. 3). Если $t^* > t_{кр}$, то нулевая гипотеза о равенстве математических ожиданий величин X и Y отвергается.

В более общем случае при $\sigma_x^2 \neq \sigma_y^2$ эта задача, называемая иногда проблемой Беренса — Фишера, становится в теоретическом отношении более сложной. Обычно методы ее решения основаны на использовании критерия

$$w^* = \frac{\bar{y} - \bar{x}}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}},$$

для которого известен ряд приближенных распределений. Например, $w = w(\alpha, n_x, n_y, c)$, где

$$c = \frac{(n_x - 1) \sigma_x^2}{(n_x - 1) \sigma_x^2 + (n_y - 1) \sigma_y^2}.$$

Если вычисленное выборочное значение критерия w^* превышает $w_{кр}$ ($w^* > w_{кр}$), то нулевая гипотеза о равенстве математических ожиданий опровергается.

Заметим также, что в приближенных расчетах используются более простые критерии, например, «правило трех сигм». Суть его состоит в том, что оценка различий средних значений двух выборок производится с помощью соотношения

$$t > 3\sigma,$$

где

$$t = \frac{|\bar{y} - \bar{x}|}{\sigma(\bar{y} - \bar{x})}, \quad \sigma = \sqrt{\frac{n_x + n_y - 2}{n_x + n_y - 4}}.$$

Если $t > 3\sigma$, то расхождение между средними значениями двух выборок считается неслучайным.

Пример. Оценим степень расхождения между выборочными средними межгодовых колебаний уровня Каспийского моря по данным табл. 2.2. С этой целью разделим весь ряд ($n=86$) на два равных отрезка ($n_x = n_y = 43$). В соответствии с критерием (4.1)

были рассчитаны статистические характеристики (среднее арифметическое, дисперсия) обоих отрезков исходного ряда, которые оказались равны: $\bar{x}=172,5$ см, $\sigma_x^2=4106$ см², $\bar{y}=-28,5$ см, $\sigma_y^2=1047$ см². После этого вычисляем критерий z^* ($z^*=201/10,9=18,4$). Принимая в качестве уровня значимости, например $\alpha=0,05$, находим из прил. 1 величину $z_{кр}$ ($z_{кр}=1,8$). Поскольку $z^*>z_{кр}$, то нулевая гипотеза о равенстве средних отвергается. Данный вывод, вообще говоря, представляется очевидным, так как из рис. 4.2 наглядно видно систематическое отличие временного хода отметок уровня в первую и вторую половины рассматриваемого ряда.

4.3 Оценка расхождения между дисперсиями двух выборок

Пусть мы имеем две выборки объемом n_x и n_y , взятые из независимых генеральных совокупностей X и Y , распределенных по нормальному закону.

Задаемся нулевой гипотезой, что дисперсии случайных величин X и Y равны, т. е. $H_0: D_x = D_y$. В качестве критерия для проверки такой гипотезы наиболее часто используется критерий Фишера, который определяется по формуле

$$F^* = \frac{D_x^*}{D_y^*}, \quad (4.3)$$

где $D_x^* > D_y^*$. Величина F^* подчиняется распределению Фишера, зависящему только от числа степеней свободы $\nu_x = n_x - 1$ и $\nu_y = n_y - 1$. Это позволяет по заданному уровню значимости α из таблиц Фишера (прил. 4) определить величину $F_{кр}$.

Проверка соответствия выдвинутой нулевой гипотезы результатам наблюдений осуществляется по следующему правилу: если $F^* > F_{кр}$, то расхождение существенно и гипотеза о равенстве дисперсий должна быть опровергнута.

Следует помнить, что если нулевая гипотеза формулируется относительно равенства двух дисперсий ($D_x^* = D_y^*$), то необходимо брать двусторонний критерий 2α (вероятность таблиц Фишера удваивается). Если же H_0 формулируется для $D_x > D_y$ или $D_x < D_y$, то берется односторонний критерий α .

В приближенных расчетах для оценки расхождений между дисперсиями двух выборок может быть использован критерий Романовского, который удобен тем, что не требует специальных таблиц.

Вначале рассчитывается вспомогательная величина θ :

$$\theta = \frac{\nu_y - 2}{\nu_y} F^*, \quad (4.4)$$

где F^* — дисперсионное отношение ($D_x^* > D_y^*$); ν_y — число степеней свободы, большее четырех.

Математическое ожидание θ для выборок из нормальной совокупности с одинаковой дисперсией равно единице, а стандартное отклонение определяется по следующей формуле:

$$\sigma(\theta) = \sqrt{\frac{2(v_x + v_y - 2)}{v_x(v_y - 4)}}.$$

С доверительной вероятностью $p=0,89$ можно ожидать, что абсолютная величина отклонения θ от единицы по модулю не превышает 3 σ . Отсюда следует, что если две выборки относятся к одной генеральной совокупности, то

$$R = \frac{|\theta - 1|}{\sigma(\theta)} < 3, \quad (4.5)$$

т.е. расхождение выборочных оценок дисперсий не существенно. Если $R > 3$, то расхождение существенно.

Пример. Используя приведенные в предыдущем разделе значения дисперсий межгодовых колебаний уровня моря, оценим степень их расхождения. С этой целью было вычислено дисперсионное отношение, которое оказалось равным $F^* = 3,92$. Определив число степеней свободы ($v_x = v_y = 42$) и задаваясь уровнем значимости $\alpha = 0,05$, находим из прил. 4 величину $F_{кр}$ ($F_{кр} = 2,07$). Поскольку выполняется условие $F^* > F_{кр}$, то гипотеза о равенстве дисперсий должна быть опровергнута.

4.4. Оценка соответствия эмпирических и теоретических функций распределения на основе критериев согласия

Как уже указывалось в предыдущей главе, наиболее строгая оценка соответствия эмпирической и теоретической функций распределения может быть получена с помощью критериев согласия. При этом важным является то обстоятельство, что в качестве теоретического может использоваться не только нормальный, но и любой другой закон распределения.

Наиболее широкое распространение получил критерий Пирсона на χ^2 . Данный критерий применяется для рядов, имеющих достаточно большой объем выборки, причем число градаций должно быть не меньше пяти.

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

$$H_0: F^*(\bar{x}^*, \sigma_x^*) = F(\bar{x}, \sigma_x).$$

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

$$\chi_p^2 = n \sum_{i=1}^k \frac{(p_i^* - p_i)^2}{p_i} = \sum_{i=1}^k \frac{(m_i - np_i)^2}{np_i}, \quad (4.6)$$

где p_i^* — эмпирическая вероятность в i -й градации, p_i — теоретическая вероятность; k — число градаций в выборке объемом n ; m_i — абсолютная частота (число событий) в i -й градации.

После того как по эмпирическим данным по формуле (4.6) вычисляется величина χ_p^2 осуществляется проверка условия $\chi_p^2 > \chi_{\text{табл.}}^2$, при числе степеней свободы $\nu = k - 1 - l$ (l — число параметров теоретического распределения) и уровне значимости α . Если это условие выполняется, то гипотеза о соответствии эмпирического и теоретического распределений отвергается; расхождение между ними носит случайный характер. Заметим, что в практических расчетах наиболее часто табличные значения $\chi_{\text{табл.}}^2$ принимают для следующих уровней значимости: 0,01; 0,05; 0,10.

Критерий согласия Колмогорова обычно используется применительно к интегральным функциям распределения и при достаточно больших объемах выборок, причем в этом случае он не зависит от вида распределения.

Расчетная схема использования данного критерия сводится к определению статистики λ , представляющей собой произведение критерия D на корень квадратный из объема совокупности, т. е.

$$\lambda = D \sqrt{n}, \quad (4.7)$$

где $D = \max |F^*(x) - F(x)|$ — максимальное значение модуля разности между статистической $F^*(x)$ и теоретической $F(x)$ функциями распределения.

Распределение случайной величины D установлено Колмогоровым, и при неограниченном возрастании объема выборки оно зависит лишь от числа λ

$$\lim_{n \rightarrow \infty} P(D \sqrt{n} \geq \lambda) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 \lambda^2} = P(\lambda_0).$$

Главное условие к исходной информации — непрерывность, т. е. эмпирическая функция распределения должна быть построена по несгруппированным в градации величинам или по сгруппированным, когда интервалы эмпирического распределения будут достаточно малыми (различия менее 1%).

Значения функции распределения Колмогорова затабулированы и приводятся в табл. 4.2 в зависимости от уровня значимости α .

Общая схема проверки нулевой гипотезы состоит в следующем. Вначале по формуле (4.7) вычисляется выборочное значение λ^* .

Далее, задаваясь уровнем значимости (обычно принимают $\alpha = 5\%$), по таблице распределения Колмогорова (табл. 4.2) опре-

Критерий Колмогорова D при различных уровнях значимости

%	99,9	95	90	50	10	5	1	0,1
D	0,374	0,520	0,571	0,828	1,224	1,358	1,627	1,950

деляют критическое значение $\lambda_{\text{табл.}}$. Если $\lambda^* > \lambda_{\text{табл.}}$, то нулевая гипотеза отвергается и предполагается гипотеза несоответствия наблюдаемых данных принятой теоретической функции распределения.

Следует иметь в виду, что при использовании данного критерия учитывается лишь наибольшее отклонение 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 аналитической кривой, но не настолько, чтобы повысить максимальное отклонение, т. е. величину D . В этих случаях критерий Колмогорова будет показывать на хорошее согласие теоретической и эмпирической функций распределения, в то время как в действительности будет наблюдаться их очевидное несоответствие.

4.5. Проверка гипотезы о принадлежности двух выборок одной генеральной совокупности

Предположим, что имеются две независимые выборки случайных величин X и Y , для которых предполагается непрерывность неизвестных функций распределения $F_1(x)$ и $F_2(y)$.

Запишем

$$x_1, x_2, \dots, x_m,$$

$$y_1, y_2, \dots, y_n,$$

т. е. длина выборок может быть различной.

Если статистические параметры этих рядов (среднее арифметическое, стандартное отклонение и др.) отличаются друг от друга, то возникает вопрос, являются ли наблюдаемые расхождения следствием объективного различия законов распределения $F_1(x)$ и $F_2(y)$ или они могут быть объяснены случайностью выборки.

Наиболее простым способом проверки нулевой гипотезы $H_0: F_1(x) = F_2(y)$ является использование непараметрического критерия Уилкоксона. Суть этого критерия состоит в следующем. Расположим все наблюдения, полученные в двух выборках, в общую последовательность в порядке возрастания их значений. Например, $x_1 y_2 x_3 x_2 y_3 y_1 x_4 y_4 x_5 \dots$ и будем считать, что если какому-

либо элементу этого ряда x_i и предшествует элемент y_i ($y_i < x_i$), то данная пара элементов образует инверсию U .

В рассматриваемом примере x_3 и x_2 образуют по одной инверсии, x_4 образует три инверсии, x_5 — четыре. Таким образом, всего имеем девять инверсий.

Критерием, лежащим в основе проверки гипотезы однородности сравниваемых выборок, может служить общее число инверсий. Критерий Уилкоксона распределяется асимптотически нормально с математическим ожиданием $M_U = nl/2$ и дисперсией $D_U = nl(n+l+1)/12$, причем нормальное приближение оказывается достаточно точным уже при $(n+l) > 20$.

Для проверки нулевой гипотезы необходимо подсчитать фактическое число инверсий и, вычислив

$$t_U^* = \frac{\sum U - M_U}{\sqrt{D_U}}, \quad (4.8)$$

сравнить с t -критерием Стьюдента при заданном уровне значимости и числе степеней свободы $n+l-2$.

Если $|t^*| > t_{кр}$, то нулевая гипотеза опровергается и делается вывод, что выборки принадлежат разным генеральным совокупностям. В противном случае можно полагать, что рассматриваемые выборки однородны.

РАЗДЕЛ 2

ПОСТРОЕНИЕ ЭМПИРИЧЕСКИХ ЗАВИСИМОСТЕЙ

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

Функциональная связь. Если каждому допустимому значению одной переменной соответствует одно и только одно значение другой переменной, то такая зависимость носит название функциональной (рис. 5.1, а). Функциональные зависимости обычно могут быть доказаны теоретическим путем (с помощью логических рассуждений) и не нуждаются в опытной проверке.

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. 5.1, б). В общем случае условие независимости выражается следующим образом: $F(x, y) = F_1(x)F_2(y)$, т. е. функция распределения системы независимых случайных величин равна произведению функций распределения отдельных случайных величин.

Стохастическая связь. Если каждому значению одной переменной с определенной вероятностью соответствует значение другой переменной, то такая зависимость называется стохастической. Естественно, что она носит промежуточный характер между функциональным и случайным типами связи (рис. 5.1, в). Стохастическая зависимость характеризуется **теснотой** связи. Чем выше

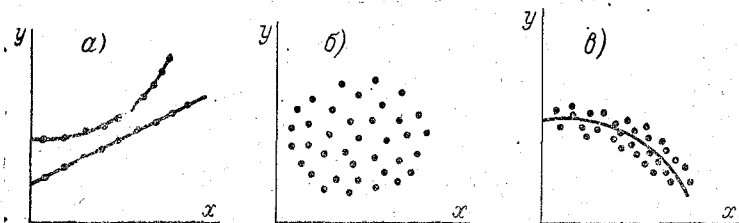


Рис. 5.1. Виды связей океанологических элементов: а — функциональная, б — случайная, в — стохастическая

теснота связи, тем ближе стохастическая зависимость приближается к функциональной. Наоборот, по мере уменьшения тесноты связи, сопоставляемые переменные становятся все более независимыми.

Заметим, что связь между исследуемыми процессами может носить как линейный, так и нелинейный характер. Для описания линейных зависимостей обычно используется аппарат корреляционного анализа, который будет рассмотрен в гл. 5.

Что касается построения нелинейных зависимостей, то это более сложная задача. Одним из простейших способов ее решения является подбор эмпирической формулы, аппроксимирующей связь между переменными по их известным значениям. Однако 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 наблюдаемыми 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 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 и в настоящее время и широко используется в численных расчетах.

Способы построения нелинейных зависимостей обсуждаются в гл. 6.

Наконец, учитывая, что при обработке экспериментальных данных, мы всегда сталкиваемся с погрешностями измерений и расчетов, то в гл. 7 рассматриваются методы их анализа.

ГЛАВА 5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

5.1. Коэффициент корреляции, его свойства и оценка достоверности

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 Спирмена, коэффициенты сопряженности Пирсона и Чупрова и ряд других показателей.

К параметрическим критериям относится линейный коэффициент корреляции, коэффициенты регрессии и др.

Наибольшее распространение из указанных показателей получил коэффициент корреляции, который является характеристикой меры линейной взаимосвязи двух случайных величин X и Y вне зависимости от их размерности, т. е.

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}. \quad (5.1)$$

Формула (5.1) справедлива для длинных статистических рядов. Вследствие того что выборочный коэффициент корреляции имеет отрицательное смещение, для коротких рядов в знаменатель формулы (5.1) вводится множитель $(n-1)$ вместо n .

Формула (5.1) удобна для физического анализа, однако в численных расчетах на ЭВМ обычно используется формула

$$r = \frac{\sum xy - (n^{-1})(\sum x)(\sum y)}{\sqrt{[\sum x^2 - (n^{-1})(\sum x)^2][\sum y^2 - (n^{-1})(\sum y)^2]}}. \quad (5.2)$$

Отметим некоторые важные свойства коэффициента корреляции.

Его величина 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.

Из (5.1) вытекает, что коэффициент корреляции изменяется в пределах $-1 \leq r \leq 1$. При $r = \pm 1$ связь между X и Y становится функциональной. В этом случае облако точек на графике связи вырождается в прямую линию, наклоненную под некоторым углом к оси абсцисс (табл. 5.1).

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.

В том случае, когда величины X и Y являются независимыми в статистическом смысле, то $r = 0$ и в результате облако точек превращается в круг или квадрат (рис. 5.1).

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 (рис. 5.1).

Распределение выборочных коэффициентов корреляции r^* очень сильно зависит от длины сравниваемых рядов и от величин r^* . При малых величинах r^* и достаточно больших объемах

выборки распределение r^* 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 r^* 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 $|r^*| < 0,5$ и $n > 30-40$.

Исходя из нормального закона распределения, для оценки среднеквадратических погрешностей коэффициентов корреляции используется формула Романовского

$$\sigma_r = \frac{1 - r^{*2}}{\sqrt{n-1}} \sqrt{1 + \frac{11r^{*2}}{2n} + \frac{75r^{*2} - 13}{2n^2}} \quad (5.3)$$

или

$$\sigma_r \approx \frac{1 - r^{*2}}{\sqrt{n-1}}.$$

Доверительные интервалы при заданном уровне значимости α определяются с использованием t -статистики Стьюдента

$$r^* - t_\alpha \sigma_r < r < r^* + t_\alpha \sigma_r.$$

✓ Оценка значимости r^* обычно осуществляется на основе нулевой гипотезы, которая в этом случае выбирается относительно проверки r^* на равенство 0, т.е. $H_0: r=0$. Строится критическая область вида $|r^*| > t_\alpha \sigma_r$ с уровнем значимости α . Если полученное по данным выборки значение r^* окажется внутри критической области, то H_0 отвергается. Таким образом, если $|r^*| < t_\alpha \sigma_r$, то r^* — значим.

б) оценка коэффициентов корреляции при $|r^*| > 0,5$ и $n < 30-40$.

В этом случае, как уже говорилось выше, распределение выборочных коэффициентов корреляции не подчиняется нормальному закону. Поэтому точность r^* обычно оценивается с помощью преобразования Фишера, основанного на использовании специальной переменной z , функционально связанной с r^* следующим выражением:

$$z = 0,5 \ln \frac{1 + r^*}{1 - r^*} = \operatorname{arcth} r^*, \quad (5.4)$$

где th — гиперболический тангенс.

Распределение величины z (прил. 5) почти не зависит от n и r , а с возрастанием n оно быстро приближается к нормальному со средним значением

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

и дисперсией

$$\sigma_z^2 = \frac{1}{n-3}. \quad (5.5)$$

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

$$\sigma_z = 1/\sqrt{n-3}.$$

Доверительные границы для z записываются следующим образом:

$$z^* - t_\alpha \sigma_z \leq z \leq z^* + t_\alpha \sigma_z.$$

Построив доверительные границы для z , нетрудно от них перейти к доверительным границам для r^* , используя для этого обратное преобразование $r^* = f(z)$. Так как $r^* = \text{th } z$, то

$$\text{th} \left[\frac{1}{2} \ln \frac{1+r^*}{1-r^*} - \frac{t_\alpha}{\sqrt{n-3}} \right] \leq r^* < \text{th} \left[\frac{1}{2} \ln \frac{1+r^*}{1-r^*} + \frac{t_\alpha}{\sqrt{n-3}} \right]. \quad (5.6)$$

в) оценка коэффициентов корреляции при $|r^*| < 0,5$, $n < 30-40$ и при $|r^*| > 0,5$, $n > 30-40$.

В этих случаях приближенно можно считать, что распределение выборочных коэффициентов корреляции не очень заметно отличается от нормального, поэтому для оценки точности величин r^* можно пользоваться вариантом а).

Коэффициенты корреляции как мера линейной связи между процессами получила широкое распространение в статистических расчетах. Так, многие методы многомерной статистики основаны на анализе корреляционных матриц.

Если мы имеем значения какого-либо элемента (например, температуры воды), измеренного в N точках за промежуток времени M ($M > N$), то нетрудно составить матрицу исходных значений температуры размером $N \times M$ или

$$T = \begin{vmatrix} T_{11} & T_{12} & \dots & T_{1N} \\ T_{21} & T_{22} & \dots & T_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ T_{M1} & T_{M2} & \dots & T_{MN} \end{vmatrix}.$$

В результате вычисления коэффициентов корреляции данная матрица может быть преобразована в квадратную симметричную матрицу следующего вида:

$$R = \begin{vmatrix} r_{11} & r_{12} & \dots & r_{1N} \\ r_{21} & r_{22} & \dots & r_{2N} \\ \dots & \dots & \dots & \dots \\ r_{N1} & r_{N2} & \dots & r_{NN} \end{vmatrix}$$

Диагональные элементы этой матрицы всегда равны $r_{11}=r_{22}=\dots=r_{NN}=1$, 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. На полигоне в Норвежском море в рамках программы «Полэкс — Север» четыре научно-исследовательских судна выполнили одновременные многосуточные станции в течение 30 суток. В результате расчетов коэффициентов корреляции для временных рядов температуры воды на поверхности моря была получена следующая матрица:

$$R = \begin{vmatrix} 1 & 0,2 & 0,9 & 0,1 \\ 0,2 & 1 & 0,2 & -0,7 \\ 0,9 & 0,2 & 1 & -0,8 \\ 0,1 & -0,7 & -0,8 & 1 \end{vmatrix}$$

Элемент r_{ij} , находящийся в этой матрице на пересечении i -й строки и j -го столбца, показывает корреляционную связь температуры воды между i -м и j -м судами. Прежде всего проверим значимость полученных коэффициентов корреляции. Вычисленные значения средних квадратических погрешностей коэффициентов корреляции σ_r можно объединить в матрицу вида

$$\sigma_R = \begin{vmatrix} 0 & \underline{0,16} & 0,04 & \underline{0,18} \\ \underline{0,16} & 0 & \underline{0,16} & 0,09 \\ 0,04 & \underline{0,16} & 0 & 0,07 \\ \underline{0,18} & 0,09 & 0,07 & 0 \end{vmatrix}$$

Величина t_α при уровне значимости $\alpha=10\%$ составляет $t_\alpha=1,7$. Очевидно, что в критическую область вида $t_\alpha\sigma_r$ попадают значения r_{12} , r_{21} , r_{23} , r_{32} , r_{14} , r_{41} (подчеркнуты). Следовательно, можно сделать вывод об отсутствии статистической связи в колебаниях температуры воды, наблюдаемых на первом и втором, втором и третьем, первом и четвертом судах. Остальные коэффициенты корреляции значимы. Положительный знак коэффициента r_{13} указывает на синфазные колебания температуры воды, наблюдаемые

на первом и третьем судах. Наоборот, отрицательные значения r_{24} , r_{34} демонстрируют противофазность колебаний температуры, наблюдаемых на втором и четвертом, третьем и четвертом судах.

5.2. Понятие о методе наименьших квадратов

Если мы имеем совокупность наблюдений $x_1, \dots, x_n$ и $y_1, \dots, y_n$ и между ними существует некоторое приближенное соотношение $y = f(x; a_1, a_2, \dots, a_m)$, то для отыскания неизвестных коэффициентов $a_1, \dots, a_m$ обычно используется метод наименьших квадратов. Его сущность заключается в том, что сумма квадратов отклонений точек от линии связи должна быть наименьшей, т. е.

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n [y_i - f(x; a_1, \dots, a_m)_i]^2 = \min, \quad (5.7)$$

причем $n \geq m$. Метод наименьших квадратов впервые был сформулирован в 1805 г. Лежандром, поэтому (5.7) иногда называется принципом Лежандра. Видно, что формула (5.7) минимизирует отклонения наблюденных значений переменной y от вычисленных по зависимости $f(x_i)$.

Так как в уравнении (5.7) x_i и y_i известны (результаты наблюдений), то при заданном общем виде сглаживающей функции ве-

личину $\sum_{i=1}^n \varepsilon_i^2 = S$ можно рассматривать лишь как функцию от неизвестных коэффициентов $a_1, a_2, \dots, a_m$. Для их нахождения в соответствии с общим правилом определения экстремума дифференцируемой функции необходимо найти частные производные от нее по этим коэффициентам и приравнять их к нулю:

$$\frac{\partial S}{\partial a_1} = -2 \sum_{i=1}^n [y_i - f(x; a_1, a_2, \dots, a_m)_i] \left(\frac{\partial f}{\partial a_1} \right) = 0,$$

$$\frac{\partial S}{\partial a_2} = -2 \sum_{i=1}^n [y_i - f(x; a_1, a_2, \dots, a_m)_i] \left(\frac{\partial f}{\partial a_2} \right) = 0,$$

.....

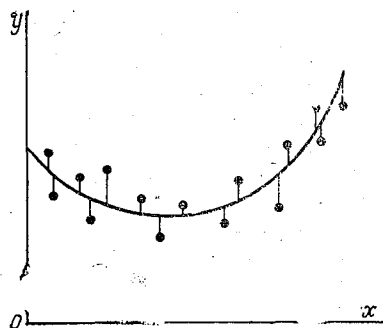
$$\frac{\partial S}{\partial a_m} = -2 \sum_{i=1}^n [y_i - f(x; a_1, a_2, \dots, a_m)_i] \left(\frac{\partial f}{\partial a_m} \right) = 0.$$

Данная система уравнений называется системой нормальных уравнений. Число уравнений в системе всегда равно числу неизвестных, поэтому задача определения коэффициентов $a_1, \dots, a_m$ является определенной и имеет единственное решение.

К недостаткам метода наименьших квадратов относится то, что он предполагает нормальное распределение опытных данных, весьма чувствителен к выбросам и должен быть линеен относительно параметров $a_1, \dots, a_m$.

$$S = \sum_{i=1}^n (y_i - y_{\text{фак}})^2 \rightarrow \min$$

Рис. 5.2. Отклонения наблюдаемых значений y_i от аппроксимирующей кривой y



Пример. Предположим, две выборки экспериментальных данных x_i и y_i образуют в координатной плоскости xOy некоторую криволинейную статистическую связь (рис. 5.2). Наша задача состоит в наилучшем подборе вида аппроксимирующей функции и определении ее коэффициентов. В соответствии с методом наименьших квадратов необходимо минимизировать отклонения ε_i (рис. 5.2), т.е. подобрать такую кривую, которая дает наилучшее приближение к исходным точкам в смысле (5.7).

Очевидно, зависимость на рис. 5.2 может быть представлена в виде полинома второй степени $y = a_0 + a_1x + a_2x^2$. Тогда имеем систему из 3-х линейных уравнений:

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

В результате несложных преобразований этой системы получим

$$\begin{cases} a_0n + a_1 \sum_{i=1}^n x_i + a_2 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i, \\ a_0 \sum_{i=1}^n x_i + a_1 \sum_{i=1}^n x_i^2 + a_2 \sum_{i=1}^n x_i^3 = \sum_{i=1}^n x_i y_i, \\ a_0 \sum_{i=1}^n x_i^2 + a_1 \sum_{i=1}^n x_i^3 + a_2 \sum_{i=1}^n x_i^4 = \sum_{i=1}^n x_i^2 y_i. \end{cases} \quad (5.8)$$

Решение данной системы линейных нормальных уравнений с использованием современных ЭВМ не представляет каких-либо затруднений. Однако следует иметь в виду, что в систему (5.8) входят высокие степени переменной x , причем самая высокая степень равна удвоенной степени полинома. Это обстоятельство является главным источником вычислительных ошибок.

5.3. Линейная регрессия двух переменных

Для исследования статистической взаимосвязи между двумя природными процессами или явлениями чаще всего используется математическая модель в виде уравнения регрессии. Простейшей моделью регрессии является линейная регрессия двух переменных.

При построении уравнения регрессии предполагается наличие линейной зависимости между известной (зависимой) переменной x_i и неизвестной (независимой) переменной y_i в виде

$$y_i = a_1 x_i + a_0, \quad i = 1, \dots, n, \quad (5.9)$$

где a_1 — коэффициент регрессии; a_0 — свободный член; n — длина рядов.

Если известно, что все наблюдения x_i и y_i в статистическом смысле независимы и подчиняются нормальному закону распределения, то в этом случае для определения параметров a_0 и a_1 , может быть использован метод наименьших квадратов.

В соответствии с методом наименьших квадратов требуется минимизировать разность квадратов фактических и рассчитанных значений y_i , т. е.

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

Отсюда нетрудно получить систему из двух нормальных линейных уравнений

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

или, после несложных преобразований,

$$\begin{cases} a_0 n + a_1 \sum x_i = \sum y_i, \\ a_0 \sum x_i + a_1 \sum x_i^2 = \sum x_i y_i. \end{cases} \quad (5.10)$$

Решая (5.10) относительно параметров a_0 и a_1 , имеем

$$a_1 = \frac{\sum_{i=1}^n (x_i y_i - n \bar{x} \bar{y})}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}, \quad (5.11)$$

$$a_0 = \bar{y} - a_1 \bar{x}. \quad (5.12)$$

Заметим, что параметры a_0 и a_1 могут быть представлены в несколько ином виде. Так, из (5.11) можно получить

$$a_1 = r \frac{\sigma_y}{\sigma_x}. \quad (5.13)$$

Отсюда видно, что коэффициент регрессии прямо пропорционален коэффициенту корреляции. Естественно, при отсутствии корреляции $a_1=0$.

Подставляя (5.13) в (5.12), имеем

$$a_0 = \bar{y} - r \frac{\sigma_y}{\sigma_x} \bar{x}. \quad (5.14)$$

С учетом выражений (5.13) и (5.14) классическое уравнение регрессии может быть переписано в следующем виде:

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x}). \quad (5.15)$$

Аналогичным образом может быть представлено и уравнение регрессии x по y

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y}). \quad (5.16)$$

Следует иметь в виду, что уравнения (5.15) и (5.16) являются различными самостоятельными зависимостями, взаимно не получающимися одно из другого. Модель (5.15) условно называют прямой регрессией, а модель (5.16) — обратной регрессией.

Физический смысл параметров a_0 и a_1 становится очевидным, если обратиться к рис. 5.3. Коэффициент регрессии a_1 представляет собой тангенс угла наклона линии регрессии к оси абсцисс ($a_1 = \operatorname{tg} \alpha$), а свободный член — расстояние от начала координат до точки пересечения оси ординат с линией регрессии.

В зависимости от знака при параметрах a_0 и a_1 линия регрессии занимает различное положение в декартовой системе координат (рис. 5.4). Если $a_0=0$, то линия регрессии проходит через начало координат.

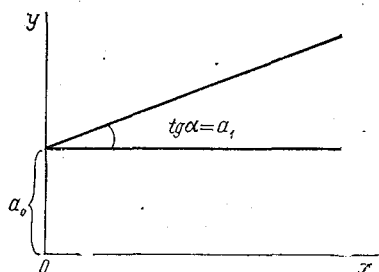


Рис. 5.3. Интерпретация коэффициентов в уравнении линейной регрессии

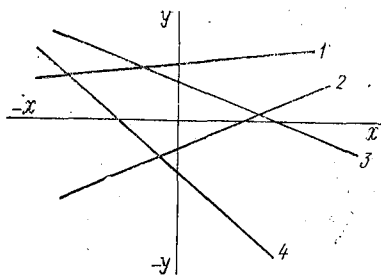


Рис. 5.4. Вид графика линейной регрессии в зависимости от значений параметров: 1 — $a_0 > 0$, $a_1 > 0$; 2 — $a_0 < 0$, $a_1 > 0$; 3 — $a_0 > 0$, $a_1 < 0$; 4 — $a_0 < 0$, $a_1 < 0$

В качестве критерия точности уравнения регрессии используется среднее квадратическое отклонение между наблюдаемыми величинами и рассчитанными по уравнению регрессии

$$\sigma_{y(x)} = \sqrt{\frac{\sum_{i=1}^n (y_{i \text{ рас}} - y_{i \text{ наб}})^2}{n}}. \quad (5.17)$$

В силу отрицательной смещенности выборочных оценок $\sigma_{y(x)}^*$ вместо n используют $(n-1)$, а для очень коротких рядов ($n < 10$) рекомендуется даже брать $n-2$. Выражение (5.17) может быть преобразовано к следующему виду:

$$\sigma_{y(x)} = \sigma_y \sqrt{1 - r^{*2}}.$$

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

Соответствующие стандартные погрешности параметров a_0 и a_1 определяются как

$$\sigma_{a_1} = \frac{\sigma_y}{\sigma_x} \frac{1 - r^{*2}}{\sqrt{n}}, \quad (5.17a)$$

$$\sigma_{a_0} = \frac{\sigma_{y(x)}}{\sqrt{n}} = \frac{\sigma_y \sqrt{1 - r^{*2}}}{\sqrt{n}}.$$

Рассмотрим оценку адекватности соответствия регрессионной модели исходным данным, основываясь на положениях дисперсионного анализа. Прежде всего вспомним важное свойство дисперсий двух рядов

$$D(x + y) = \sigma_x^2 + \sigma_y^2 + 2r\sigma_x\sigma_y,$$

т. е. дисперсия суммы двух переменных равна сумме дисперсий плюс удвоенное произведение средних квадратических отклонений переменных x и y на коэффициент корреляции между ними. При $r=0$ переменные являются некоррелированными.

Представим независимую переменную y_i в виде $y_i = \tilde{y}_i + \varepsilon$, где $\tilde{y}_i$ — вычисленные по уравнению регрессии значения y_i ; ε — ошибка регрессии, для которой предполагается нормальное распределение с нулевым средним ($\bar{\varepsilon}=0$) и отсутствие корреляции с переменной x_i . В этом случае дисперсия переменной y_i будет равна

$$Dy = D\tilde{y} + D\varepsilon.$$

Но так как

$$D\tilde{y} = D(a_0 + a_1 x) = a_1^2 Dx = r \frac{\sigma_y^2}{\sigma_x^2} \sigma_x^2 = r^2 Dy,$$

то получим

$$\frac{\widetilde{Dy}}{Dy} = r^2, \quad \frac{D\varepsilon}{Dy} = 1 - r^2. \quad (5.18)$$

Проанализируем эти соотношения. Мы имеем три характеристики:

1. Дисперсию переменной y_i , характеризующую ее общую изменчивость

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

2. Дисперсию $\widetilde{y}$, характеризующую изменчивость линии регрессии относительно среднего значения

$$D\widetilde{y} = \frac{1}{n-1} \sum_{i=1}^n (\widetilde{y}_i - \bar{y})^2.$$

3. Дисперсию ε , характеризующую отклонение прямой, построенной по методу наименьших квадратов, от результатов наблюдений (переменной y_i)

$$D\varepsilon = \frac{1}{n-1} \sum_{i=1}^n (\widetilde{y}_i - y_i)^2.$$

Отношение $\widetilde{Dy}/Dy$ показывает качество (степень) приближения вычисленных значений $\widetilde{y}_i$ по уравнению регрессии к фактическим значениям y_i . Квадрат коэффициента корреляции обычно называют *коэффициентом детерминации*.

Отношение $D\varepsilon/Dy$ показывает степень неопределенности уравнения регрессии, т. е. долю изменчивости переменной y_i , которая не может быть объяснена переменной x_i . Фактически отношение $D\varepsilon/Dy$ — это некоторая шумовая составляющая уравнения регрессии.

Естественно, чем выше коэффициент детерминации, тем выше качество приближения вычисленных значений $\widetilde{y}_i$ к фактическим значениям y_i и тем меньше уровень шума (зашумленность) уравнения регрессии.

На практике для оценки адекватности регрессионной модели обычно составляется таблица по следующей форме (см. табл. 5.1).

По результатам расчетов проверяется нулевая гипотеза о равенстве дисперсий $H_0: \widetilde{Dy} = D\varepsilon$. Вычисленное значение F^* -критерия ($F^* = \widetilde{Dy}/D\varepsilon$) сравнивается с $F_{\text{табл}}$ при степенях свободы $v_1 = 1$, $v_2 = n - 2$ и заданном уровне значимости α .

Если $F^* > F_{\text{табл}}$, то нулевая гипотеза о равенстве дисперсий опровергается и делается предположение, что в этом случае имеет место альтернативная гипотеза $H_1: \widetilde{Dy} \neq D\varepsilon$. Это означает, что дисперсия, описываемая линейной регрессией, неслучайным образом отличается от дисперсии шума и, следовательно, можно предположить, что регрессионная модель является адекватной.

Проверка адекватности линейной регрессионной модели

$$y = a_1x + a_0$$

Источник изменчивости	Дисперсия	Число степеней свободы	Критерий Фишера
Линейная регрессия	$D\tilde{y} = \frac{1}{n-1} \sum_{i=1}^n (\tilde{y}_i - \bar{y})^2$	1	
Отклонение от регрессии	$D\varepsilon = \frac{1}{n-1} \sum_{i=1}^n (\tilde{y}_i - y)^2$	$n-2$	
Общая изменчивость	$Dy = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$	$n-1$	$F^* = \frac{D\tilde{y}}{D\varepsilon}$

Значения $F_{\text{табл}}$, составленные Фишером для нормально распределенных выборок, приводятся в табл. 5.2.

Таблица 5.2

Значения критерия Фишера при $v_1=1$ и $\alpha=0,05$

v_2	5	10	15	20	30	40	100
$F_{\text{табл}}$	6,61	4,96	4,54	4,32	4,17	4,08	3,96

Таким образом, при построении регрессионной модели необходима проверка исходных выборок на нормальность.

Пример. По результатам натурных наблюдений в Норвежском море рассмотрим возможность построения линейной регрессионной модели взаимосвязи температуры и солености воды вида: $S = a_0 + a_1T$. Коэффициент парной корреляции исследуемых рядов наблюдений длиной по 40 значений составил $r=0,72$. Оценим значимость коэффициента корреляции по формуле (5.3): $\sigma_r = (1 - 0,72^2)/(40 - 1) = 0,08$, $r/\sigma_r = 9,4$. Полученное значение r/σ_r превышает табличное значение t -критерия Стьюдента $t(\alpha=5\%, v=39) = 1,68$, что убеждает в значимости корреляционной связи температуры и солености. Вычислим по формулам (5.13), (5.14) значения коэффициентов регрессии $a_0=34,671$, $a_1=0,069$ и составим уравнение регрессии: $S=0,069T+34,671$.

Далее необходимо определить значимость полученных коэффициентов регрессии. Для этого по формулам (5.17а) вычислим их стандартные погрешности

$$\sigma_{a_1} = \frac{\sigma_s}{\sigma_T} \frac{1 - r^2}{\sqrt{n}} = \frac{0,19}{1,1} \frac{1 - 0,72^2}{\sqrt{40}} \approx 0,012,$$

$$\sigma_{a_0} = \frac{\sigma_s \sqrt{1 - r^2}}{\sqrt{n}} = \frac{0,19 \sqrt{1 - 0,72^2}}{\sqrt{40}} \approx 0,014.$$

Сравнивая величины $a_1/\sigma_{a_1} = 5,8$ и $a_0/\sigma_{a_0} = 2476,4$ с табличным t -критерием ($t_{кр} = 1,68$), можно сделать вывод о значимости коэффициентов регрессии.

Для проверки общей адекватности модели регрессии исходным наблюдениям заполним таблицу вида 5.1.

Источник дисперсии	Дисперсия	Число степеней свободы	Значение
Линейная регрессия	$Dy = 0,032$	$\nu_1 = 1$	$F^* = 6,4$
Отклонения от регрессии	$D\varepsilon = 0,005$	$\nu_2 = 39$	
Общая изменчивость	$Dy = 0,036$	38	

Табличное значение F -распределения Фишера при степенях свободы $\nu_1 = 1$, $\nu_2 = 39$ составит $F_{табл} = 4,1$, что существенно меньше расчетного. Это позволяет сделать вывод об адекватности полученного уравнения регрессии исходным данным.

5.4. Понятие о ложной корреляции

Если два ряда переменных 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 приблизительном равенстве ($C_1 \simeq C_2 \simeq C_3$), коэффициент корреляции между Z_1 и Z_2 будет составлять $r_{z_1, z_2} = 0,5$. Если $C_3 \simeq 2C_1 \simeq 2C_2$, то $r_{z_1, z_2} = 0,8$ и если $C_3 = 3C_1 = 3C_2$, то уже $r_{z_1, z_2} = 0,9$. Таким образом, из некоррелированных рядов мы получили почти функ-

циональную 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 элементов. Годовой ход, как известно, обусловлен солнечной радиацией, а суточный ход — вращением Земли вокруг собственной оси. Поэтому без исключения годового и суточного хода гидрометеорологических элементов, корреляция между ними заведомо завышается.

Для исключения годового и суточного хода используют обычно ряд приемов, простейший из которых заключается в вычислении аномалий данного элемента. Корреляция между аномалиями гидрометеорологических элементов в значительной степени уменьшает ложную корреляцию. Поэтому в общем случае, когда третья переменная неизвестна, эффект ложной корреляции приближенно может быть определен как $r_{\text{ложн}} = |r_{\text{набл}}| - |r_{\text{ист}}|$, где $r_{\text{ист}}$ — корреляция между аномалиями рассматриваемых рядов.

Разумеется, приведенным выше примером не исчерпываются возможности появления эффекта ложной корреляции при изучении гидрометеорологических процессов или явлений. 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.

Однако необходимо отметить, что существует преобразование, которое не приводит к появлению ложной корреляции и, тем не менее, увеличивает коэффициент корреляции. Если мы приведем переменные x_1 и x_2 к стандартизованному виду

$$z_1 = \frac{x_1 - \bar{x}_1}{\sigma x_1} \quad \text{и} \quad z_2 = \frac{x_2 - \bar{x}_2}{\sigma x_2},$$

то в этом случае, как показано Г. А. Алексеевым, $r_{z_1, z_2} \geq r_{x_1, x_2}$.

Пример. Для оценки эффекта ложной корреляции воспользуемся ежемесячными картами температуры поверхности океана, составляемых Гидрометцентром с 1977 г. Для двух пятиградусных квадратов с центрами (первый $\varphi = 60^\circ$ с. ш., $\lambda = 10^\circ$ з. д.; второй $\varphi = 65^\circ$ с. ш., $\lambda = 20^\circ$ 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 месяца были рассчитаны аномалии температуры воды как $\Delta x = x_i - \bar{x}$, где $\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i$. В резуль-

тате получены две новые выборки, корреляция между которыми оказалась равной $r_{\Delta T} = 0,15$. Следовательно, ложная корреляция, обусловленная эффектом годового хода солнечной радиации, составляет $r_{\text{лож}} = 0,79$.

ГЛАВА 6. АНАЛИЗ НЕЛИНЕЙНЫХ ЗАВИСИМОСТЕЙ

6.1. Общая схема построения нелинейных зависимостей

Существуют различные способы установления связи между гидрометеорологическими процессами или явлениями, которая в общем случае может носить любой заранее неизвестный характер. Одним из простейших способов является подбор эмпирической формулы, аппроксимирующей связь между переменными по их известным значениям.

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.

Решение задачи построения эмпирической формулы можно разделить на 3 этапа: I этап — выбор типа формулы; II этап — определение параметров выбранной формулы; III этап — оценка достоверности полученной формулы.

Рассмотрим вкратце каждый из этапов.

Выбор типа формулы. Если нет каких-либо 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 $y = a_0 + a_1x$, 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 $x' = \varphi(x, y)$, $y' = \psi(x, y)$, которые связаны между собой линейной зависимостью. Вычисляя для заданных значений x_i и y_i соответствующие им новые значения x'_i и y'_i и нанося 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. 6.1. 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.

Значительно более сложным является вариант, когда необходимо подобрать три параметра эмпирической формулы, т. е. 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 остав-

Таблица 6.1

Функции и их линеаризация

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

шихся двух параметров используется метод выравнивания и таким образом оценивается возможность применения этой формулы для описания исходных данных.

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 разделе 5.2, 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, находятся a_0 и a_1 .

Например, если выбранная формула имеет вид $y = a_0 x^{a_1}$, то уравнение в линейной форме записывается как $\lg y = \lg a_0 + a_1 \lg x$, т. е. $y' = \lg y$, $a_0' = \lg a_0$, $a_1' = a_1$, $x' = \lg x$. Далее составляется система нормальных уравнений

$$\begin{aligned} a_0' n + a_1' \sum x_i' &= \sum y_i', \\ a_0' \sum x_i' + a_1' \sum x_i'^2 &= \sum x_i' y_i' \end{aligned}$$

и вычисляются параметры a_1' и a_0'

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

Отсюда можно найти параметры исходной формулы a_0 и a_1' .

Заметим, что в приближенных расчетах, не требующих высокой точности, могут быть использованы метод *выбранных точек* и метод *средних*.

Суть метода выбранных точек заключается в том, что на координатной плоскости xOy с возможной аккуратностью проводим плавную кривую Γ , наиболее близко примыкающую к исходным точкам x_i, y_i . На кривой Γ выбираем систему m (по числу параметров выбранной формулы) точек, не обязательно совпадающих с исходными точками. Желательно, чтобы эти точки были по возможности равномерно распределены по всей рабочей части кривой Γ и в то же время не лежали бы близко к не очень надежным крайним значениям x_i и y_i . После этого тщательно измеряют координаты $\tilde{x}_j, \tilde{y}_j$.

Если исходные данные x_i, y_i описываются линейной зависимостью, то в этом случае имеем следующую систему линейных уравнений:

$$\begin{aligned} \tilde{y}_1 &= a_1 \tilde{x}_1 + a_0, \\ \tilde{y}_2 &= a_1 \tilde{x}_2 + a_0, \end{aligned}$$

которая решается относительно параметров a_0 и a_1 .

В рассмотренном выше примере, когда выбранная формула имеет вид $y = a_0 x^{a_1}$, составляется следующая система уравнений:

$$\begin{cases} \lg \tilde{y}_1 = a_1 \lg \tilde{x}_1 + \lg a_0, \\ \lg \tilde{y}_2 = a_1 \lg \tilde{x}_2 + \lg a_0. \end{cases}$$

По известным значениям $\lg y$ и $\lg x$ нетрудно определить a_1 , $\lg a_0$, а затем a_0 .

Суть метода средних состоит в том, что параметры эмпирической формулы определяются практически только из одного условия, а именно, условия равенства нулю суммы всех отклонений наблюдаемой величины от среднего значения

$$\sum_{i=1}^n \varepsilon_i = 0, \quad (6.1)$$

где $\varepsilon_i = y_i - \bar{y}$. Метод средних позволяет вычислить до трех параметров на основе лишь одного условия (6.1).

Например, при построении линейной зависимости $y = a_0 + a_1 x$ на основе метода средних прежде всего выборка переменной x_i и соответствующие им значения y_i делятся пополам. Для каждой из половин составляем уравнение

$$\sum_{i=1}^m (a_0 + a_1 x_i - y_i) = 0, \quad \sum_{i=m+1}^n (a_0 + a_1 x_i - y_i) = 0,$$

где $m = n/2$ при четном n и $m = (n+1)/2$ при n нечетном. Отсюда нетрудно получить систему двух уравнений

$$\begin{cases} a_1 \sum_{i=1}^m x_i + m a_0 = \sum_{i=1}^m y_i, \\ a_1 \sum_{i=m+1}^n x_i + (n-m) a_0 = \sum_{i=m+1}^n y_i, \end{cases}$$

из которой определяются параметры a_0 и a_1 .

Оценка достоверности полученной формулы. Первичная проверка заключается в вычислении по найденной формуле значений y_i и их сравнении с наблюдаемыми значениями y_i . При этом основное внимание уделяется анализу остатков, определяемых как $\varepsilon_i = y_i - \tilde{y}_i$.

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 считается, 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. Естественно, что при малых n ограничиваются лишь проверкой данного условия.

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:

$$\eta = \sqrt{\frac{\widetilde{D}_y}{D_y}} = \sqrt{\frac{\sum (\widetilde{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}}. \quad (6.2)$$

Корреляционное отношение изменяется в пределах от 0 до 1. Оно оказывается равным нулю только в том случае, когда корреляционная связь между x_i и y_i вообще отсутствует, т. е. когда линия регрессии является горизонтальной прямой. При $\eta=1$, наоборот, связь между переменными x_i и y_i становится функциональной.

Заметим также, что при нелинейной регрессии во всех случаях $|r| < \eta$. При $|r| = \eta$ связь между x_i и y_i становится линейной. Таким образом, различие между η и $|r|$ может служить показателем степени нелинейности связи рассматриваемых рядов. Средняя квадратическая ошибка эмпирической формулы определяется как $\sigma_y(x) = \sigma_y \sqrt{1 - \eta^2}$.

Однако следует помнить, что наиболее действенным способом оценки точности эмпирической формулы является расчет значений $y_{i \text{ рас}}$ по независимым данным (данным, не вошедшим в исходную выборку для оценки параметров формулы) и последующее сравнение с наблюдаемыми значениями $y_{i \text{ наб}}$.

Естественно, что не всегда в результате проверки эмпирической формулы точность ее может нас удовлетворить. В этом случае можно попытаться произвести уточнение полученной эмпирической формулы. Один из наиболее простых приемов заключается в следующем. Требуется уменьшить сумму квадратов отклонений исходных точек y_i от полученной эмпирической кривой $f(x_i)$, т. е.

$$S = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n [y_i - f(x_i)]^2. \quad (6.3)$$

Рассмотрим функцию $\hat{y} = f(x) + c$, где c — некоторая постоянная величина.

Подберем число c так, чтобы сумма квадратов новых отклонений

$$\hat{S} = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n [v_i - f(x_i) - c]^2$$

была бы минимальной. Очевидно, что можно записать

$$\hat{S} = \sum_{i=1}^n (c - \varepsilon_i)^2.$$

В результате имеем следующее нормальное уравнение:

$$\frac{\partial \hat{S}}{\partial c} = 2 \sum_{i=1}^n (c - \varepsilon_i) = 0.$$

Отсюда $nc = \sum_{i=1}^n \varepsilon_i$ и, следовательно,

$$c = \frac{1}{n} \sum_{i=1}^n \varepsilon_i. \quad (6.4)$$

Итак, наилучшей постоянной c является среднее арифметическое величин ε_i . Поэтому, если $c \neq 0$, то прибавление к правой части эмпирической формулы постоянного числа c , определяемого по формуле (6.4), приводит к уточнению этой формулы в смысле (6.3). Если $c = 0$ или близко к нулю, то указанный прием не дает нужного эффекта.

Рассмотрим конкретный пример построения эмпирической формулы. 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 $\alpha_e = \Delta e / e_0$, $\alpha_T = \Delta T / T_0$, e_0 — насыщающая упругость водяного пара при температуре поверхности океана T_0 .

На основе данных по влажности, температуре воздуха и воды для 9 судов погоды за 20-летний период путем пространственного осреднения была получена зависимость перепада влажности от перепада температуры (табл. 6.2).

Таблица 6.2

Сезонные изменения осредненных за многолетний период
и для девяти судов погоды величин α_T и α_e

Параметры	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
$\alpha_T \cdot 10^2$	0,85	0,81	0,66	0,42	0,23	0,12	0,07	0,16	0,31	0,46	0,64	0,78
α_e	0,31	0,31	0,28	0,24	0,19	0,15	0,13	0,17	0,21	0,25	0,28	0,30
α_e	0,31	0,305	0,28	0,24	0,19	0,16	0,14	0,17	0,21	0,24	0,28	0,30

Проверка всех эмпирических формул, приведенных в табл. 6.1, 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:

$$\alpha_e = \sqrt{a_0 + a_1 \alpha_T}. \quad (6.5)$$

Для выравнивания данного выражения достаточно α_e возвести в квадрат, т. е. $y' = \alpha_e^2$. Как следует из графика связи величин α_e^2 и α_T , все точки находятся практически на прямой. Это свидетельствует о том, что формула (6.5) может быть использована для аппроксимации зависимости $\alpha_e = f(\alpha_T)$. Численные значения коэффициентов a_0 и a_1 , найденные методом наименьших квадратов, составляют: $a_0 = 0,012$, $a_1 = 10$. Предвычисленные по формуле (6.5) значения α_e также представлены в табл. 6.2. Нетрудно видеть, что они почти полностью совпадают с заданными значениями α_e , поэтому анализ остатков не имеет смысла. Корреляционное отношение оказалось равным $\eta = 0,99$.

6.2. 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.

$$y = \sum_{i=0}^m a_i x^i = a_0 + a_1 x + a_2 x^2 + \dots + a_m x^m, \quad (6.6)$$

где m — максимальная степень полинома, называемая степенью уравнения.

Если в (6.6) 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 с одной точкой экстремума, третьей степени — кубическая кривая с двумя точками экстремума, четвертой степени — кривая четвертой степени с тремя точками экстремума (рис. 6.1). Следовательно, $q = m - 1$, где q — число экстремумов функции (6.6).

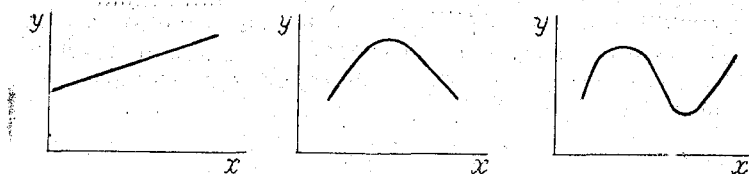


Рис. 6.1. Вид полиномиальной кривой в зависимости от степени функции

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.

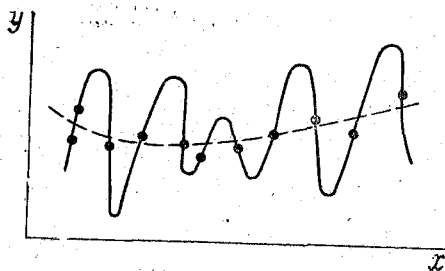
Действительно, это наглядно видно, если обратиться к рис. 6.2, на котором представлена полиномиальная аппроксимация в виде уравнения 10-й степени. Кривая 10-й степени проходит точно через все заданные точки ($n=11$), однако пользоваться этой кривой практически нельзя. В то же время достаточно точно аппроксимацию исходных точек дает парабола (уравнение 2-й степени). Поэтому на практике обычно редко используют максимальную степень полинома $m>3-4$.

Для нахождения коэффициентов $a_0, a_1, \dots, a_m$ применяется метод наименьших квадратов

$$S = \sum_{i=1}^n [y_i - (a_0 + a_1x + \dots + a_mx^m)]^2 = \min^2.$$

Приравняв частные производные от S по всем неизвестным параметрам $a_0, a_1, \dots, a_m$ к нулю

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



получим систему нормальных уравнений, решая которую можно определить параметры $a_0, a_1, \dots, a_m$.

Для оценки качества регрессии находится корреля-

Рис. 6.2. Аппроксимация опытных данных t_w полиномом 10-й степени

ционное отношение $\eta = \sqrt{\widetilde{Dy}/Dy}$ и среднеквадратическая ошибка $\sigma_y(x)$.

Обычно процедура расчетов при полиномиальной аппроксимации состоит в следующем. Вначале принимается $m=1$ и для этого случая определяются параметры регрессии и корреляционное отношение η_1 . Затем принимается $m=2$ и заново рассчитываются параметры регрессии и корреляционное отношение η_2 . Если выполняется условие $(\eta_2 - \eta_1) > \varepsilon$, где ε — некоторое заданное положительное число, то делается вывод о продолжении расчетов. Принимается $m=3$ и вся процедура повторяется. Когда добавление члена полинома более высокого порядка уже не будет давать существенного увеличения точности аппроксимации, делается вывод о прекращении расчетов. Естественно, что это зависит и от выбора числа ε , которое в зависимости от поставленной задачи может быть различным. Обычно ε принимается в пределах от 0,01 до 0,1.

Видно, что большим неудобством классического уравнения полиномиальной регрессии является необходимость заново пересчитывать все коэффициенты регрессии при увеличении степени полинома. Этого недостатка можно избежать, если воспользоваться ортогональными многочленами П. Л. Чебышева, благодаря которым удастся производить добавление новых слагаемых более высокого порядка, не изменяя при этом вычисленные ранее коэффициенты.

Сущность способа Чебышева состоит в том, что аппроксимирующий многочлен отыскивают не непосредственно в виде суммы степеней x , а в виде некоторой комбинации многочленов. В результате этого уравнение регрессии (6.6) записывается в виде

$$y = \alpha_0 \varphi_0(x) + \alpha_1 \varphi_1(x) + \dots + \alpha_m \varphi_m(x).$$

Здесь $\varphi_0(x), \varphi_1(x), \dots, \varphi_m(x)$ — многочлены Чебышева, которые 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

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

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 α_{m+1} .

Коэффициенты $\alpha_0, \alpha_1, \dots, \alpha_m$ определяются с помощью метода наименьших квадратов. Опуская достаточно громоздкие промежуточные выкладки, приведем сразу окончательные формулы

$$\alpha_0 = \frac{\sum y_i}{n}, \quad \alpha_1 = \frac{\sum y_i \varphi_1(x_i)}{\sum \varphi_1^2(x_i)},$$

$$\alpha_2 = \frac{\sum y_i \varphi_2(x_i)}{\sum \varphi_2^2(x_i)}, \dots, \alpha_m = \frac{\sum y_i \varphi_m(x_i)}{\sum \varphi_m^2(x_i)}.$$

Отсюда видно, что 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 α_j .

6.3. Тренд-анализ

Тренд-анализ — это название математического метода разделения эмпирических данных на две части: систематическую и случайную. Систематическая часть трактуется как поверхность тренда, случайная часть — отклонения поверхности тренда от системы исходных эмпирических данных. В основе метода лежит представление тренда как некоторой функции пространственных координат, построенной по эмпирическим данным таким образом, чтобы сумма квадратов отклонений их от поверхности тренда была бы минимальна.

Тренд-анализ может быть одномерным, двухмерным, трехмерным и четырехмерным. Наиболее широкое применение тренд-анализ получил при анализе пространственных полей. В этом случае одномерный тренд-анализ представляет собой полиномиальную регрессию, в которой зависимой переменной x_i является какая-либо пространственная координата.

Основная формула двухмерного тренд-анализа может быть записана следующим образом:

$$G(x, y) = \sum_{i=0}^m \sum_{j=0}^{m-1} a_{ij} x^i y^j, \quad (6.7)$$

где x, y — географические координаты; m — показатель степени.

Очевидно, 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 тренд-анализа (6.7) приобретет следующий вид:

$$G(x, y) = a_0 + a_1x + a_2y + a_3x^2 + a_4y^2 + a_5xy + a_6x^3 + a_7y^3 + a_8x^2y + a_9xy^2. \quad (6.8)$$

В данном выражении первые три слагаемых дают линейное уравнение тренда, а первые шесть слагаемых — уравнение поверхности тренда второго порядка.

Коэффициенты a_{ij} в (6.7) находятся методом наименьших квадратов. Так, например система линейных нормальных уравнений для определения линейного уравнения тренда имеет вид

$$\begin{aligned} \sum G &= a_0n + a_1 \sum x + a_2 \sum y, \\ \sum xG &= a_0 \sum x + a_1 \sum x^2 + a_2 \sum xy, \\ \sum yG &= a_0 \sum y + a_1 \sum xy + a_2 \sum y^2. \end{aligned}$$

Аналогичным образом составляются системы нормальных уравнений для поверхностей тренда более высокого порядка.

Для оценки точности аппроксимации рассчитывается дисперсия фактических значений поля

$$D_T = \frac{1}{n-1} \sum_{i=1}^n (G_{Ti} - \bar{G})^2$$

(n — число исходных точек) и дисперсия вычисленных по уравнению (6.8) значений G_R

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

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

Естественно, что прежде всего двухмерный тренд-анализ может использоваться в задаче объективного анализа гидрометеорологических полей. Кроме того, тренд-анализ можно применить для аппроксимации некоторой таблично заданной функции в зависимости от двух переменных. Так, например, 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 влажности 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. Доволь-

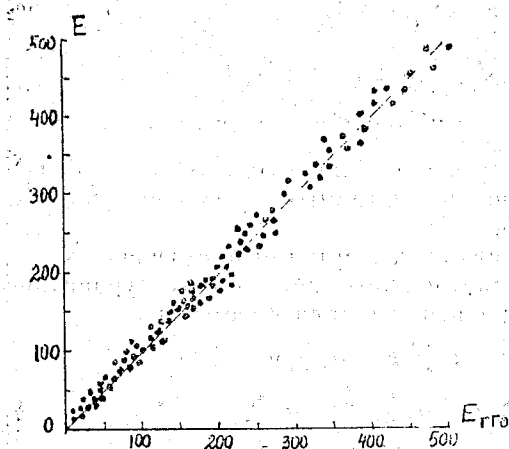


Рис. 6.3. Сравнение среднемесячных величин испарения

но 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 двумерного тренд-анализа. В результате расчетов было установлено, что коэффициент влагообмена может быть аппроксими-

рован поверхностью тренда второй степени

$$c_E = 0,85 \cdot 10^{-3} + 0,762 \cdot 10^{-4} \cdot U_{10} + 0,882 \cdot 10^{-4} \Delta T_{10} - 0,591 \cdot 10^{-6} \cdot U_{10}^2 - 0,11 \cdot 10^{-5} \cdot \Delta T_{10}^2 - 0,191 \cdot 10^{-5} \cdot U_{10} \cdot \Delta T_{10}, \quad (6.9)$$

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. 7.1).

Сравнение среднемесячных величин испарения для пяти судов погоды, вычисленных с использованием (6.9), с аналогичными значениями испарения, рассчитанными по методике ГГО, показало (рис. 6.3), 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.

6.4. Понятие о рациональных функциях

При построении нелинейных зависимостей все более широкое распространение получают так называемые рациональные функции (приближения), являющиеся отношением двух многочленов (полиномов) $P(x)$ и $Q(x)$, т. е.

$$R_{k,l}(x) = P(x)/Q(x) = \sum_{i=0}^k a_i x^i / \sum_{i=0}^l b_i x^i. \quad (6.10)$$

Отличительной особенностью рациональных приближений является то, что они имеют меньшее число параметров (коэффициентов) и требуют меньшего числа операций при определении значений функций по сравнению с полиномиальными приближениями для заданной точности. В тех случаях, когда на ЭВМ скорость выполнения операции деления близка к скорости умножения, то эффективность использования рациональной аппроксимации возрастает.

Кроме того, рациональные приближения позволяют аппроксимировать более широкий класс функций. Так, с помощью многочленных приближений нельзя достаточно эффективно аппроксимировать функции, которые обращаются в бесконечность на заданном интервале. При использовании рациональных приближений допустима аппроксимация таких функций и, кроме того, можно аппроксимировать прямые параллельные оси Ox при больших значениях x .

Наконец, рациональные функции обладают свойством оставаться рациональными при линейных преобразованиях их аргументов.

Рациональная функция $R_{k,l}(x)$ называется наилучшим равномерным приближением с весом $\omega(x)$ к функции $f(x)$ на множестве X , если оно обеспечивает минимальное значение следующему функционалу:

$$S_{k,l}(f) = \max_{x \in X} \left| \frac{f(x) - R_{k,l}(x)}{\omega(x)} \right|. \quad (6.11)$$

В соответствии с теоремой Чебышева наилучшее равномерное рациональное приближение $R_{k,l}(x)$ с весом $\omega(x)$ к непрерывной функции $f(x)$ существует и единственно.

Если вид рационального приближения не диктуется физическими соображениями, то стремятся так подобрать степени числителя и знаменателя (6.10), чтобы достигалось минимальное значение максимального взвешенного отклонения (6.11). Для большого класса функций рациональные приближения дают хорошую аппроксимацию уже при малых значениях k и l ($k \neq 0, l \neq 0$). Естественно, что при $l=0$, $\omega(x)=1$, имеем обычную полиномиальную регрессию.

Неизвестные коэффициенты a_i, b_i определяются методом наименьших квадратов

$$S = \sum_{i=1}^n \omega_i \left[f_i - \frac{P(x_i)}{Q(x_i)} \right]^2 = \min. \quad (6.12)$$

В общем случае из условия (6.12) следует, что необходимо решать систему нелинейных уравнений, что весьма сложно. Поэтому числитель и знаменатель рациональной аппроксимации проще строить с помощью последовательных итераций.

Пусть $P_0(x)$ и $Q_0(x)$ — начальные приближения. Тогда

$$S = \sum_{i=0}^n \omega_i \left[f_i - \frac{P_0(x_i) + \Delta P_0 x_i}{Q_0(x_i) + \Delta Q_0(x_i)} \right]^2 = \min. \quad (6.13)$$

Рациональную функцию $P(x)/Q(x)$ можно линеаризовать относительно ΔP_0 и ΔQ_0 следующим образом:

$$\frac{P_0 + \Delta P_0}{Q_0 + \Delta Q_0} \approx \frac{P_0}{Q_0} \left(1 + \frac{\Delta P_0}{P_0} - \frac{\Delta Q_0}{Q_0} \right).$$

Подставляя это выражение в (6.12), получим

$$S = \sum_{i=1}^n \frac{\omega_i}{Q_0^2(x_i)} \left[f_i Q_0(x_i) - P_0(x_i) - \left(\Delta P_0(x_i) - \frac{P_0(x_i)}{Q_0(x_i)} \Delta Q_0(x_i) \right) \right]^2. \quad (6.14)$$

Для построения итерационного процесса введем

$$f_{0i} = f_i Q_0(x_i) - P_0(x_i),$$

$$\omega_{0i} = \omega_i / Q_0^2(x_i).$$

Приращения $\Delta Q_0(x_i)$ и $\Delta P(x_i)$ можно представить в виде многочленов

$$\Delta P_0(x) = \sum_{j=1}^k c_j x^j, \quad \Delta Q_0(x) = \sum_{j=k+l}^{k+l} c_j x^{j-(k+1)}, \quad (6.15)$$

где c_j — неизвестные коэффициенты. Свободный член для приращения $\Delta Q_0(x)$ можно положить равным нулю. Такое представление вполне оправдано, поскольку числитель и знаменатель рационального приближения определяются с точностью до общего множителя. Поэтому в качестве Q_0 можно взять постоянную величину $Q_0=1$.

Тогда соотношение (6.14) может быть записано в виде, аналогичном (6.13), т. е.

$$S_0 = \sum_{i=1}^n \omega_{0i} [f_{0i} - g_{0i}]^2,$$

где

$$g_{0i} = \sum_{j=1}^n c_j x^j + \frac{p_0(x_i)}{Q_0(x_i)} \sum_{j=k+l+1}^{k+l+1} c_j x^{j-(k+1)}.$$

Коэффициенты c_j вычисляются с учетом того, что они должны минимизировать функционал S_0 . Это требование уже приводит к построению системы нормальных уравнений, решение которой и дает численные значения коэффициентов c_j . После этого первый шаг приближения к числителю и знаменателю рациональной аппроксимации можно построить как

$$P_1 = P_0 + \Delta P_0; \quad Q_1 = Q_0 + \Delta Q_0.$$

Второй шаг приближения строится аналогично первому и т. д. На $m+1$ шаге строится $m+1$ приближение

$$P_{m+1} = P_m + \Delta P_m; \quad Q_{m+1} = Q_m + \Delta Q_m.$$

Условием сходимости итерационного процесса является выполнение неравенства

$$\left| \frac{F_m - F_{m+1}}{F_{m+1}} \right| \leq \varepsilon.$$

ГЛАВА 7. АНАЛИЗ ПОГРЕШНОСТЕЙ ОКЕАНОЛОГИЧЕСКИХ ИЗМЕРЕНИЙ И РАСЧЕТОВ

7.1. Общая схема анализа погрешностей

Вообще говоря, практически уравнение любого типа (балансовое, гидродинамическое и т. д.) может быть представлено в виде

$$\sum_{i=1}^k x_{i, \text{ист}} = 0, \quad (7.1)$$

где k — число членов исходного уравнения; $x_{i, \text{ист}}$ — «истинная» величина каждой компоненты исходного уравнения. Данное уравнение предполагает отсутствие погрешностей измерений и расчетов, что в действительности практически никогда не выполняется. Вследствие этого алгебраическая сумма всех членов уравнения (7.1) обычно не равна нулю. С учетом сказанного, «истинное» уравнение (7.1) приобретает вид

$$\sum_{i=1}^k x_i = \eta, \quad (7.2)$$

где x_i — наблюдаемая компонента исходного уравнения; η — суммарная погрешность определения всех компонент уравнения (7.2), называемая невязкой.

В самом общем случае невязка может быть представлена следующим образом:

$$\eta = \sum_{i=1}^k \xi_i + \sum_{i=1}^k \delta_i + \sum_{j=1}^l \mu_j, \quad (7.3)$$

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 $\oint (P - E) d\Gamma + Q = 0$, где P — осадки; E — испа-

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

В настоящее время для оценки погрешностей существующих методов определения составляющих уравнения (7.2) используют четыре метода:

1. Сравнение результатов различных независимых методов расчета отдельных составляющих уравнения (7.2).
2. Сравнение расчетов составляющих уравнения (7.2) 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 (7.2) при независимом определении всех его членов.

Если первые три способа позволяют оценить лишь ошибки отдельных составляющих уравнения (7.2), то последний дает непосредственно величину невязки. Как показывает опыт, даже в сравнительно простых случаях 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 (7.2).

Проиллюстрируем анализ погрешностей применительно к Каспийскому морю. С этой целью представим уравнение водного баланса моря для годовых интервалов времени

$$Q_6 + P_m - E_m - \Delta V = \eta, \quad (7.4)$$

где Q_6 — приток речных вод к морю; P_m — осадки, выпадающие на зеркало моря; E_m — испарение с акватории моря, ΔV — изменения полезного объема моря.

В уравнении (7.4) не учитываются прежде всего приток подземных вод Q_u и плотностные (стерические) изменения объема за счет колебаний во времени температуры и солености деятельного слоя моря ΔV_p , т. е.

$$\sum_{j=1}^2 \mu_j = Q_u + \Delta V_p.$$

Если рассматривать межгодовые изменения составляющих уравнения баланса (7.4), то по различным оценкам величина Q_u изменяется от нескольких км³/год до нескольких десятков км³/год. На наш взгляд, первая из этих оценок представляется более обоснованной. Что касается величины ΔV_p , то, по нашим оценкам, ее межгодовые колебания, во-первых, носят случайный характер, а во-вторых, ее абсолютные значения также не превышают нескольких км³/год.

Несколько по-иному обстоит дело, если рассматривать внутригодовые изменения составляющих водного баланса. В этом случае, вследствие отчетливо выраженного сезонного хода температуры воды, амплитуда ΔV_p уже будет составлять десятки км³.

Сравнение величин Q_u и ΔV_p со средними за многолетний (1930—1986 гг.) период составляющими водного баланса (табл. 7.1) показывает, что их роль в формировании баланса незначительна. Наиболее важный вклад в формировании баланса, а следовательно, в колебаниях уровня моря, играют приток речных вод и испарение. Так как средняя величина ΔV оказалась за рассматриваемый период отрицательной, то это свидетельствует, что произошло понижение уровня моря.

Межгодовые изменения величины невязки η приводятся на рис. 7.1. 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.

Как видно из рис. 7.1, распределение величин η носит преимущественно случайный характер. Среднее многолетнее значение невязки составляет лишь 3 км³/год, что свидетельствует в целом о достаточно высокой степени достоверности (надежности) годовых величин балансов Каспийского моря.

Обратимся теперь к табл. 7.2, в которой представлены коэффициенты корреляции между отдельными составляющими водного баланса за указанный выше период.

Таблица 7.1

Статистические характеристики составляющих водного баланса
Каспийского моря за 1930—1986 гг., км³/год

Характеристики	Q_6	P_m	E_m	ΔV	η
$\bar{X}$	290	76,8	373,7	—9,9	3,0
σ	45,5	13,4	19,3	54,1	11,2
$X_{\max}$	385,0	109,7	410	102,6	
$X_{\min}$	213,0	49,1	303	—114,5	

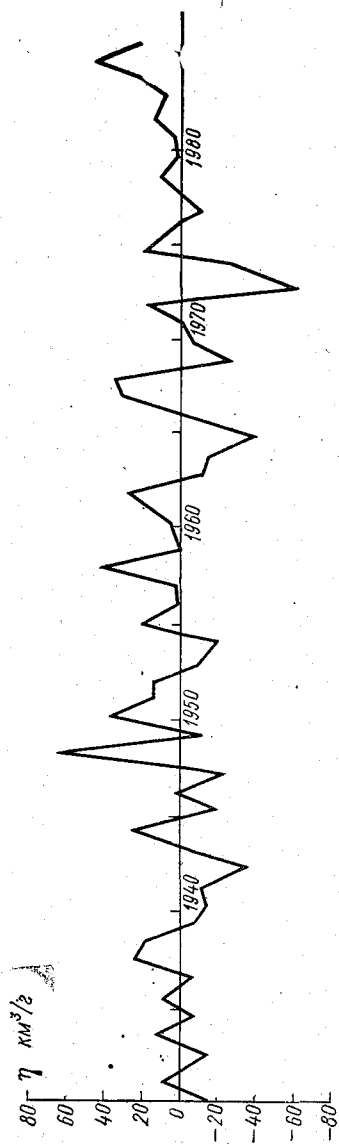


Рис. 7.1. Межгодовой ход величин невязок уравнений водного баланса Каспийского моря

Таблица 7.2

Корреляционная матрица составляющих водного баланса Каспийского моря

Составляющие	η	P	Q	E	ΔV
η	1,00				
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

Как следует из табл. 7.2, наиболее высокая корреляция величин невязки отмечается с испарением и изменениями полезного объема моря. Это означает, что данные составляющие водного баланса определяются, очевидно, с наибольшей случайной погрешностью.

На первый взгляд, сомнительным выглядит результат корреляции между η и P . Однако следует иметь в виду, что на оценку величин осадков в большей степени влияет величина систематической погрешности, которая в коэффициенте корреляции по существу не отражается.

Что касается статистической связи между отдельными составляющими баланса, то наиболее высока она между ΔV и Q . Это означает, что наибольший вклад во внутригодовые изменения уровня Каспийского моря вносит приток речных вод.

7.2. Случайные и систематические погрешности

Случайной погрешностью величины x_i называется погрешность, изменяющаяся 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 не представляется возможным.

Мерой случайной погрешности единичного значения любого члена x_i может служить его среднее квадратическое отклонение, определяемое, как известно, по формуле

$$\sigma_x = \sqrt{\frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2}.$$

В качестве случайной погрешности осредненной величины $\bar{x}$ принимается выражение

$$\sigma_{\bar{x}} = \sigma_x / \sqrt{n}. \quad (7.5)$$

Из (7.5) следует важное следствие. С увеличением n случайная погрешность уменьшается и при достаточно больших значениях

и $\sigma_x \rightarrow 0$, в результате чего ею можно пренебречь. Это позволяет, увеличивая длину исходной выборки, исключать из уравнения (7.2) случайные погрешности и выделить систематические.

Кроме того, существуют и другие меры случайной погрешности. Абсолютная случайная погрешность — погрешность, выражаемая в единицах измеряемой (рассчитываемой) величины, т. е.

$$|x - x_{\text{ист}}| = \delta.$$

Относительная случайная погрешность выражается в долях единицы или в %: $\varepsilon = \delta/x_{\text{ист}}$. Иногда абсолютную погрешность соотносят со стандартным отклонением случайной величины X следующим образом: $\delta \approx 0,8\sigma$. Наконец, в некоторых случаях используется вероятная случайная погрешность, определяемая как $\delta_v \approx \frac{2}{3}\sigma$.

При условии, что исходные данные распределены по нормальному закону, а внутрирядная связь отсутствует, случайные погрешности различных характеристик статистических рядов могут быть рассчитаны следующим образом:

ошибка выборочного среднего

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n-1}},$$

ошибка среднего квадратического отклонения

$$\sigma_{\sigma} \approx \frac{\sigma}{\sqrt{2n-1}},$$

ошибка коэффициента асимметрии

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

ошибка коэффициента эксцесса

$$\sigma_{\varepsilon} = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+3)}} \approx \sqrt{\frac{24}{n}}.$$

Оценим влияние связности ряда на величину случайных погрешностей. Допустим, что статистический ряд может быть представлен в виде простой цепи Маркова, т. е. имеет место корреляция только между смежными значениями ряда. Сказанное означает, что $r(\tau=1) \neq 0$, $r(\tau>1) = 0$, где τ — сдвиг ряда.

В этом случае ошибка выборочного среднего имеет вид

$$\sigma_{\bar{x}} \approx \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{1+r(\tau=1)}{1-r(\tau=1)}}, \quad (7.6)$$

а ошибка стандартного отклонения

$$\sigma_o \approx \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{1+r^2(\tau=1)}{1-r^2(\tau=1)}}. \quad (7.7)$$

Из приведенных формул видно, что с увеличением степени связности ряда величина случайной погрешности возрастает. При сильной инерционности процесса (например, $r(\tau=1)=0,80$) ошибка выборочного среднего возрастает в 3 раза, а ошибка стандартного отклонения более чем в 6 раз.

Систематической погрешностью величины x_i называется погрешность, изменяющаяся по определенному закону. В общем случае систематические погрешности можно разделить на четыре вида: инструментальные ошибки, возникающие из-за дефектов приборов измерений; личные ошибки, связанные с ограниченностью возможностей 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.

Обычно в составляющих уравнения (7.2) присутствуют все виды систематических ошибок, поэтому отделить их друг от друга, как правило, не представляется возможным. Очевидно, имеет смысл определить лишь суммарную систематическую погрешность

каждой компоненты $x_i(\xi_i)$ или даже их общую сумму $\sum_k \xi_i$.

Для выявления систематических погрешностей можно, например, использовать сравнение измеренных (рассчитанных) компонент уравнения (7.2) 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.

7.3. Выявление и устранение грубых погрешностей

Грубые ошибки (промахи) по характеру своего происхождения относятся к случайным погрешностям и являются следствием неисправностей в аппаратуре, неправильных вычислений, описок и т. п. В тех случаях, когда грубые ошибки отличаются от основной массы данных на порядок и более, они могут быть выявлены и устранены визуально. Однако если грубые ошибки не слишком сильно отличаются от других результатов, то для их обнаружения приходится прибегать к специальным приемам.

Рассмотрим наиболее простые методы, основанные на использовании критерия Стьюдента. Для выборок небольшого объема ($n \leq 25$) вычисляется максимальное относительное отклонение

$$\frac{|x_m - \bar{x}|}{\sigma} = t^*, \quad (7.9)$$

где x_m — крайний (наибольший или наименьший) элемент выборки. Затем максимальное относительное отклонение, соответствующее t -статистике Стьюдента, сравнивается с табличным значением t при заданном уровне значимости α

$$t^* \leq t_\alpha.$$

Если это неравенство выполняется, то крайнее наблюдение не отсеивается, если не выполняется, то наблюдение исключается из выборки. 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 вместо (7.9) вычисляется отношение

$$\frac{|x_m - \bar{x}|}{\sigma \sqrt{(n-1)/n}} = t^*. \quad (7.10)$$

Далее процедура обнаружения грубых погрешностей аналогична изложенной выше. Введение множителя $\sqrt{(n-1)/n}$ учитывает смещенность оценок при малых объемах выборки.

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

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

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

$$|x_m - \bar{x}| > 3\sigma, \quad (7.10a)$$

где x_m — 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

$$|x_m - \bar{x}| > 2\sigma, \quad (7.10б)$$

которому соответствует вероятность появления случайной ошибки в исследуемом ряду, равная $\approx 5\%$.

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 критериями (7.9) — (7.10) крайние значения статистического ряда, реальность существования которых сомнений не вызывает.

Например, в январе 1989 г. температура воздуха в г. Ленинграде составила 0°C . Средняя многолетняя величина и стандартное отклонение температуры воздуха за 50 лет соответственно равны -8°C и $2,5^\circ\text{C}$. По формуле (7.9) получаем $t^* = 3,5$, а критерий Стьюдента при уровне значимости $\alpha = 10\%$ составляет $t = 1,96$. 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.

Поскольку аномальные условия формирования гидрометеорологических процессов играют 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. Поэтому многие исследования направлены на изыскание пригодных для практического использования приближенных формул и на развитие методов изучения характеристик выбросов путем статистического моделирования на ЭВМ.

7.4. Понятие о косвенных погрешностях

Под косвенными погрешностями понимаются погрешности величины Y , которая непосредственно не определяется, но может быть вычислена через измеряемые параметры $x_1, x_2, \dots, x_m$, т. е. $y = f(x_1, x_2, \dots, x_m)$.

Предположим, что погрешности подчиняются нормальному закону распределения, по абсолютной величине значительно уступают $x_1, x_2, \dots, x_m$ ($\sigma_1 \ll x_1$ и т. д.) и некоррелированы друг с другом. В этом случае для оценки случайных погрешностей может быть использован так называемый метод частных погрешностей

$$\sigma_y = \sqrt{\left(\frac{\partial f}{\partial x_1}\right)^2 \sigma_1^2 + \left(\frac{\partial f}{\partial x_2}\right)^2 \sigma_2^2 + \dots + \left(\frac{\partial f}{\partial x_n}\right)^2 \sigma_m^2} \quad (7.11)$$

Придавая функции f конкретный вид, можно получить аналитические формулы для оценки случайных погрешностей. Формулы для некоторых простейших зависимостей приводятся в табл. 7.3.

Формула для оценки случайных погрешностей различных зависимостей

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

Если ошибки коррелированы друг с другом, что, вообще говоря, нельзя упускать из виду, то формулы для оценки случайных погрешностей усложняются. Так, например, если $y = x_1 + x_2$, то $\sigma_y = \sqrt{\sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2 r_{1,2}}$, где $r_{1,2}$ — коэффициент корреляции между погрешностями x_1 и x_2 . Чем больше величина $r_{1,2}$, тем существеннее роль последнего слагаемого в формировании суммарной погрешности σ_y . Однако в зависимости от знака $r_{1,2}$ суммарная погрешность σ_y может либо увеличиться, либо уменьшиться. При $r_{1,2} > 0$ σ_y увеличивается, при $r_{1,2} < 0$ σ_y уменьшается.

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

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

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

Пример. Соленость в автоматизированных системах зондирования морских вод, как правило, непосредственно не измеряется. К непосредственно измеряемым с необходимой точностью параметрам относятся температура t , удельная электропроводность κ , гидростатическое давление P и скорость распространения звука c . Для расчета солености достаточно измерить три из указанных че-

тырех параметров. Если, например, принять $S=f(t, x, P)$, то в этом случае средняя квадратическая погрешность вычисления солёности будет определяться выражением

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

Средние квадратические погрешности первичных параметров, а также значение частных производных, приводятся ниже.

$$\begin{aligned} \sigma_t &= 0,05^\circ\text{C}; \quad \sigma_x = 0,05 \text{ мСм} \cdot \text{см}^{-1}; \quad \sigma_P = 1 \cdot 10^{-2} \text{ МПа}; \quad (\partial S / \partial t)_{x, P} = \\ &= -1,145^\circ /_{00} / ^\circ\text{C}; \quad (\partial S / \partial x)_{t, P} = 1,34\% (\text{мСм} \cdot \text{см}^{-1}); \quad (\partial S / \partial P)_{t, x} = \\ &= -0,06^\circ /_{00} / \text{МПа}. \end{aligned}$$

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

Если

$$S = f(t, x, c), \quad \sigma_S = 0,0096^\circ /_{00},$$

$$S = f(t, c, p), \quad \sigma_S = 0,075^\circ /_{00},$$

$$S = f(x, c, p), \quad \sigma_S = 0,038^\circ /_{00}.$$

Итак, наиболее точные результаты при расчете солёности могут быть получены, если в качестве первичных параметров служат температура, удельная электропроводность и гидростатическое давление.

РАЗДЕЛ 3

АНАЛИЗ ОКЕАНОЛОГИЧЕСКИХ ВРЕМЕННЫХ РЯДОВ

Под временным рядом обычно понимают некоторую конечную реализацию случайной величины, взятой через равные промежутки времени. Такие промежутки времени называют *интервалом* (шагом) дискретизации Δt . Естественно, что возможен широкий спектр значений Δt от самых малых (например, секунда, минута) до самых больших (например, сто лет, тысяча лет). При этом каждому из интервалов дискретизации соответствует свой временной ряд, описывающий процессы того или иного масштаба (периода).

Колебания во времени океанологических элементов 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; другие явления.

6. *Внутривековая изменчивость* (периоды в десятки лет), взаимосвязанная с внутривековыми колебаниями климата. Пример — потепление Арктики в тридцатых годах нашего 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.

Поэтому в данном разделе рассматриваются основные положения теории случайных функций и стандартные методы анализа временных рядов, причем разбиение глав 9 и 10 является условным (спектральный анализ и фильтрация рядов также представляют собой методы статистического анализа) и продиктовано целесообразностью, а также объемом изложения материала.

Что касается вопросов моделирования и прогноза временных рядов, на наш взгляд, они выходят за рамки данного курса и, следовательно, должны излагаться в других учебных дисциплинах.

ГЛАВА 8. ОСНОВНЫЕ ПОЛОЖЕНИЯ ТЕОРИИ СЛУЧАЙНЫХ ФУНКЦИЙ

8.1. Понятие о случайной функции

Понятие случайной функции представляет собой обобщение понятия случайной величины. *Случайной функцией* называется такая функция, значения которой при каждом значении аргумента меняются произвольным образом.

Конкретный вид, который случайная функция принимает в результате испытаний (наблюдений), называется *реализацией случайной функции*. Фактически всякая функция, встречающаяся на практике, представляет собой случайную функцию, так как 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 временные ряды, в то время как случайные поля можно отождествить с пространственными картами.

В общем случае случайную функцию $X(t)$ можно рассматривать как совокупность всех ее возможных реализаций. Действительно, в результате n независимых испытаний (наблюдений) мы

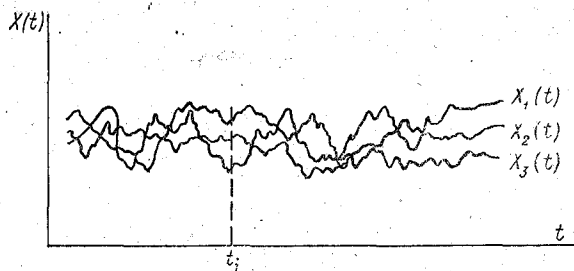


Рис. 8.1. Представление случайной функции $X(t)$ семейством ее реализаций

получим n реализаций этой случайной функции $x_1(t), x_2(t), \dots, x_n(t)$, каждая из которых носит неслучайный характер (рис. 8.1).

Зафиксируем некоторое значение аргумента (например, $t=t_j$). Тогда в момент t_j получим n значений случайной величины, т. е. $X_1(t_j), X_2(t_j), \dots, X_n(t_j)$. Их набор представляет собой сечение случайной величины $X(t_j)$. Число сечений соответствует длине реализации случайной функции.

Учитывая, что исчерпывающей характеристикой случайной величины является закон распределения, то аналогичным образом исчерпывающей характеристикой случайной функции также является закон распределения.

Однако, в отличие от случайной величины, имеющей одномерный закон распределения, исчерпывающее описание случайной функции $X(t)$ дает лишь закон совместного распределения системы бесчисленного множества ее произвольных сечений $X(t_1), X(t_2), \dots, X(t_n)$. На практике получить и использовать такую вероятностную характеристику случайной функции абсолютно невозможно. Поэтому в практических приложениях обычно используются не более чем двумерные законы распределения.

Основными числовыми характеристиками случайной функции являются математическое ожидание, дисперсия, автокорреляционная функция и др. Например, под математическим ожиданием случайной функции $X(t)$ понимается неслучайная функция $M_x(t)$, которая при каждом значении аргумента t равна математическому ожиданию соответствующего сечения случайной функции, т. е.

$$M_x(t) = \int_{-\infty}^{\infty} xf(x, t)dx$$
, где $f(x, t)$ — дифференциальная функция распределения каждого сечения случайной функции $X(t)$. Итак, геометрически, математическое ожидание можно рассматривать как некую среднюю кривую всех возможных реализаций данной случайной функции.

Дисперсией случайной функции $X(t)$ называется неслучайная функция $D_x(t)$, которая при каждом значении аргумента t равна дисперсии соответствующего сечения случайной функции, т. е.

$$D_x(t) = \int_{-\infty}^{\infty} [x(t - M_x(t))]^2 f(x, t) dx.$$

Соответственно среднее квадратическое отклонение случайной функции имеет вид $\sigma_x(t) = \sqrt{D_x(t)}$.

Что касается понятий о других характеристиках случайной функции, то они будут рассмотрены в следующих разделах.

8.2. Стационарность случайных функций

Случайные функции подразделяются в зависимости от устойчивости их вероятностных характеристик на стационарные и нестационарные.

Стационарные случайные функции описывают процессы, которые протекают во времени почти однородно, их вероятностные характеристики с изменением аргумента практически не изменяются и поэтому не зависят от начала отсчета.

Нестационарные случайные функции описывают процессы, которые протекают во времени так, что их вероятностные характеристики с изменением аргумента существенно изменяют свои значения и, следовательно, зависят от начала отсчета.

Различают стационарность в широком и узком смысле.

Случайная функция $X(t)$ называется стационарной в широком смысле, если она обладает следующими свойствами: ее математическое ожидание и дисперсия постоянны $M_x(t) = \text{const}$; $D_x(t) = \text{const}$, а ее корреляционная 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. Величину τ обычно называют сдвигом.

Случайная функция $X(t)$ называется стационарной в узком смысле, если ее многомерные распределения при одновременном прибавлении ко всем аргументам $t_1, t_2, \dots, t_n$ одного и того же числа τ остаются неизменными, т. е.

$$\begin{aligned} f(x_1; x_2, \dots, x_n; t_1, t_2, \dots, t_n) = \\ = f(x_1, x_2, \dots, x_n; t_1 + \tau, \dots, t_n + \tau). \end{aligned}$$

Стационарность случайной функции в широком смысле является простейшим частным случаем ее стационарности в узком смысле. В гидрометеорологии обычно используется понятие стационарности в широком смысле.

Следует отметить, что многие нестационарные процессы на некоторых участках, где нет заметных изменений математического ожидания и дисперсии, могут считаться стационарными. Кроме того, большое число нестационарных процессов может быть приведено к стационарному виду путем некоторых преобразований. Если нестационарность функции вызвана изменяющимся во вре-

мени математическим ожиданием, то для приведения ее к стационарному виду можно, например, использовать первые разности, т. е.

$$\Delta' x(t) = X(t_{i+1}) - X(t_i), \quad i = \overline{1, n}.$$

В том случае, если первые разности случайного процесса еще носят нестационарный характер, то 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 и нестационарной функций приводится на рис. 8.2.

Проверка стационарности требует наличия, по крайней мере, нескольких реализаций процесса, которые далеко не всегда имеются в нашем распоряжении. Чаще всего располагают лишь одной реализацией достаточной длины. Предполагая эргодичность процесса, эту реализацию разбивают на отдельные, одинаковые по продолжительности интервалы. Затем для каждого из них вычисляются оценки среднего арифметического, дисперсии и автокорреляционной функции.

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.

Что касается сравнения оценок автокорреляционной функции, то его рекомендуется производить для тех ординат, абсолютные нормированные значения которых лежат в пределах $0,4 \leq r(\tau) \leq 0,6$. Если при этом расхождения между оценками превышают 2—3 среднеквадратические ошибки их определения, то процесс следует считать нестационарным.

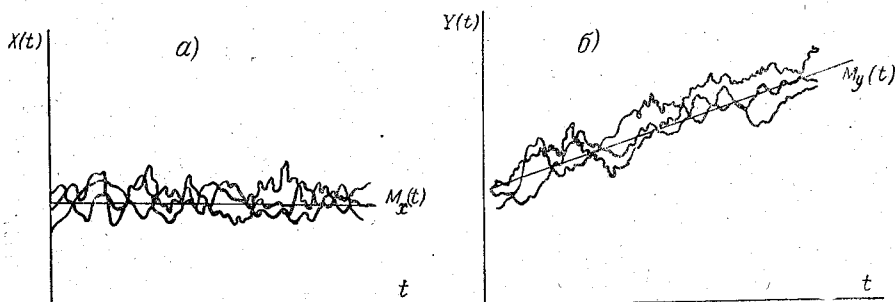


Рис. 8.2. Реализации стационарных (а) и нестационарных (б) случайных функций

8.3. Эргодичность стационарных случайных функций

Стационарные случайные функции делятся на эргодические и неэргодические. Если одна стационарная случайная функция достаточно большой длины содержит в себе фактически всю информацию об основных свойствах этого процесса, т. е. может заменить при обработке множество реализаций той же продолжительности, то такая стационарная функция является *эргодической*. Случайные функции, не обладающие таким свойством, называются *неэргодическими* стационарными случайными функциями. На рис. 8.3 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.

Для случайной функции $X(t)$ характерно то, что каждая из ее реализаций обладает почти одним и тем же средним значением, близким к $M_x(t)$, и почти одним и тем же размахом колебаний, примерно равным размаху колебаний случайной функции. Каждая же реализация случайной функции $Y(t)$ имеет свое среднее значение и свой средний размах колебаний, которые существенно отличаются от математического ожидания и дисперсии самой функции.

Отсюда следует, что оценки соответствующих характеристик случайной функции $X(t)$, вычисленные по одной достаточно длинной реализации и по множеству реализаций той же общей длины, будут между собой равнозначны. Совершенно очевидно, что для случайной функции $Y(t)$ этого не наблюдается. Поэтому стационарная функция $X(t)$ является эргодической, а функция $Y(t)$ свойством эргодичности не обладает.

При численных расчетах свойство эргодичности играет большое значение, так как позволяет существенно сократить объемы вычислений, а с другой стороны, принимая во внимание, что продолжительность многих гидрометеорологических рядов невелика, эргодичность есть единственная возможность использования аппарата теории случайных функций применительно к статистическим рядам.

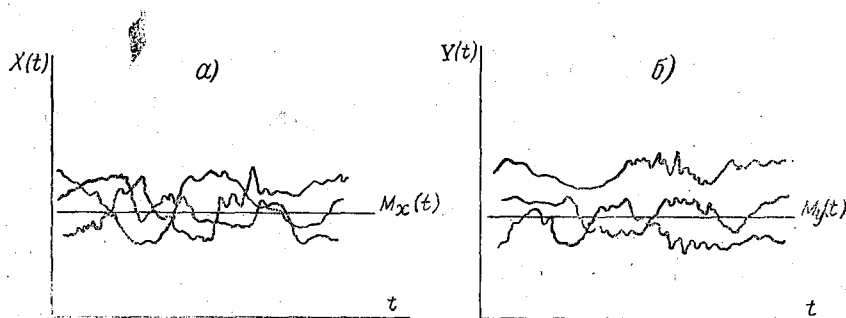


Рис. 8.3. Реализации эргодических (а) и неэргодических (б) стационарных случайных функций

Однако на практике доказать свойство эргодичности очень сложно, поэтому обычно о нем 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. $r(\tau) \rightarrow 0$ при $\tau \rightarrow \infty$.

ГЛАВА 9. МЕТОДЫ АНАЛИЗА ВРЕМЕННЫХ РЯДОВ

9.1. Общая схема исследования временной изменчивости

Если какая-либо случайная величина X измерена в N точках пространства, причем длина каждой реализации составляет M , то получим матрицу размером $N \times M$. 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, то 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 $\tau = 1$ мес.), т. е. фильтруются синоптические колебания. В этом случае временной ряд может быть представлен как

$$X(t) = \Phi(t) + Q(t), \quad (9.1)$$

где $\Phi(t)$ — межгодовая (низкочастотная) изменчивость; $Q(t)$ — внутригодовая (высокочастотная) изменчивость.

Задачей разложения (9.1) является разделение разномасштабных составляющих. Это достигается многими приемами и, в частности, методами фильтрации. Если же необходимо выделить лишь межгодовую изменчивость, то для этого достаточно осреднить среднемесячные данные за годовые интервалы времени.

Составляющую $Q(t)$ можно выразить как

$$Q(t) = \sum_{i=1}^k A_i \cos(\omega_i t + \varphi_i) + \varepsilon(t), \quad (9.2)$$

где A_i , ω_i , φ_i — амплитуда, частота, фаза i -й гармоники полигармонического ряда сезонного хода; k — число гармоник; $\varepsilon(t)$ — стационарный случайный процесс с нулевым средним $M_\varepsilon = 0$ и известной корреляционной функцией $r(\tau)$. Заметим, что представление ряда в виде (9.2) в определенной степени гипотетично, так как не всегда соблюдается требование к стационарности $\varepsilon(t)$ и независимости слагаемых в правой части (9.2).

Анализ ряда $Q(t)$ заключается в том, что вначале методами гармонического анализа или периодограммного анализа производится оценка годового хода. После исключения годового хода исследуется его остаточная часть $\varepsilon(t)$.

Одним из важных вопросов оказывается нахождение критерия разделения составляющей $Q(t)$ на гармоническую и случайную части. В предположении их независимости (некоррелированности) можно записать

$$D(Q) = D(G) + D(\varepsilon), \quad (9.3)$$

где дисперсия гармонической части представляется как $D(G) = \sum_{i=1}^k D(G)_i$, т. е. суммой дисперсий отдельных гармоник.

Рассчитывая последовательно гармоники, будем проверять выполнение условия

$$D(G)_i \leq D(\varepsilon), \quad (9.4)$$

где $D(\varepsilon)$ определяется из выражения (9.3). Если условие (9.4) 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 ошибки измерений и расчетов.

Рассмотрим теперь структуру межгодовой изменчивости $\Phi(t)$. По аналогии с (9.2) ее можно представить в виде следующего разложения:

$$\Phi(t) = T(t) + \sum_{j=1}^l A_j \cos(\omega_j t + \varphi_j) + P(t), \quad (9.5)$$

где $T(t)$ — трендовая составляющая; A_j , ω_j , φ_j — параметры полигармонического ряда, характеризующего регулярные межгодовые колебания; $P(t)$ — остаточная часть, характеризующая нерегулярные межгодовые колебания.

При анализе межгодовой изменчивости прежде всего выделяется трендовая составляющая, под которой обычно понимают

некоторое плавное изменение процесса с периодом, существенно превышающим длину исходной реализации. Приемы выделения тренда будут рассмотрены в разделе 9.3.

После обнаружения и исключения тренда, получим некоторый ряд, который следует 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 полусный прилив). Однако соотношение $C(t)/P(t)$ (где $C(t)$ — регулярная часть) 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 принимая дисперсию ряда при использовании нормированной спектральной плотности $D_\Phi = 1$, получим

$$\frac{C(t)}{P(t)} = \frac{D_c}{1 - D_c}.$$

Таким образом, основные этапы исследования временной изменчивости состоят в следующем (рис. 9.1):

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.

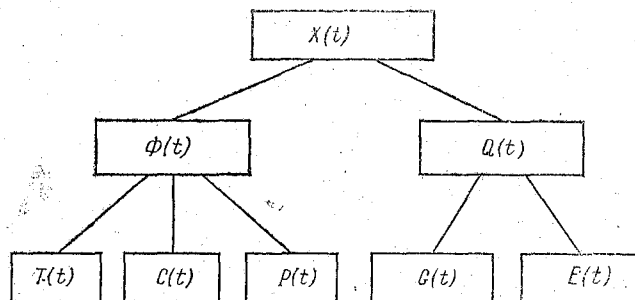


Рис. 9.1. Общая схема анализа временных рядов

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.

9.2. Гармонический анализ

Для описания периодических и квазипериодических процессов обычно используется гармонический анализ. В общем случае любую кривую, даже достаточно сложной формы, можно описать с помощью разложения Фурье. Если мы имеем функцию $f(t)$, то в интервале $[T, T+2\pi]$ ее можно представить в виде ряда Фурье:

$$f(t) = \bar{f} + \sum_{n=1}^{\infty} (a_n \cos(n\omega t) + b_n \sin(n\omega t)), \quad (9.6)$$

где $\bar{f}$ — математическое ожидание функции $f(t)$; ω — частота; a_n и b_n — коэффициенты, определяемые как

$$a_n = \frac{2}{T} \int_0^T f(t) \cos(n\omega t) dt,$$

$$b_n = \frac{2}{T} \int_0^T f(t) \sin(n\omega t) dt.$$

Разложение в ряд Фурье не только точно представляет функцию $f(t)$ при $N \rightarrow \infty$, но и обеспечивает при фиксированном 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 $U_n = a_n \cos(n\omega t) + b_n \sin(n\omega t)$, $n=1, 2, 3, \dots$ называются *гармониками*. Если ввести угол φ_n так, что

$$\cos \varphi_n = \frac{a_n}{\sqrt{a_n^2 + b_n^2}}, \quad \sin \varphi_n = \frac{b_n}{\sqrt{a_n^2 + b_n^2}},$$

то гармоники можно представить в виде

$$U_n = A_n \cos(n\omega t - \varphi_n), \quad A_n = \sqrt{a_n^2 + b_n^2}.$$

Величины A_n и φ_n называются соответственно *амплитудой* и *фазой* гармоники.

Заметим, что разложение (9.6) 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.

Итак, 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 .

Ряд Фурье тем точнее представляет значения функции, чем многочисленнее имеющиеся исходные данные. В этом случае возрастает число гармоник. Так, если число наблюдений составляет N , то число гармоник равно $N/2$.

Пример. При исследовании годового цикла ($N=12$ мес.) первая гармоника имеет период, равный длине всего основного периода (12 месяцев), вторая — период равный половине основного (6 месяцев), третья — период, равный $1/3$ основного (4 месяца). Наконец, последняя шестая гармоника имеет период, равный 2 месяцам (рис. 9.2).

В природе, как известно, существуют явления со строгой периодичностью (годовой и суточный ход основных гидрометеорологических элементов). Поэтому при исследовании рядов за годовой и суточный промежутки выделенные гармоники будут иметь четкий физический смысл.

Пример. Рассмотрим особенности годового хода изменений теплосодержания деятельного слоя океана в десятиградусных широтных зонах северного полушария. В табл. 9.1 приводятся амплитуды A (Вт/м^2), фазы Φ (месяцы) первого максимума и изменчивость (%), обусловленная годовой и полугодовой гармониками.

Из табл. 9.1 видно, что максимальная амплитуда 1 гармоники годового хода изменений теплосодержания океана наблюдается в зоне $40-30^\circ\text{N}$ и быстро уменьшается по направлению к экватору и полюсу от этой зоны.

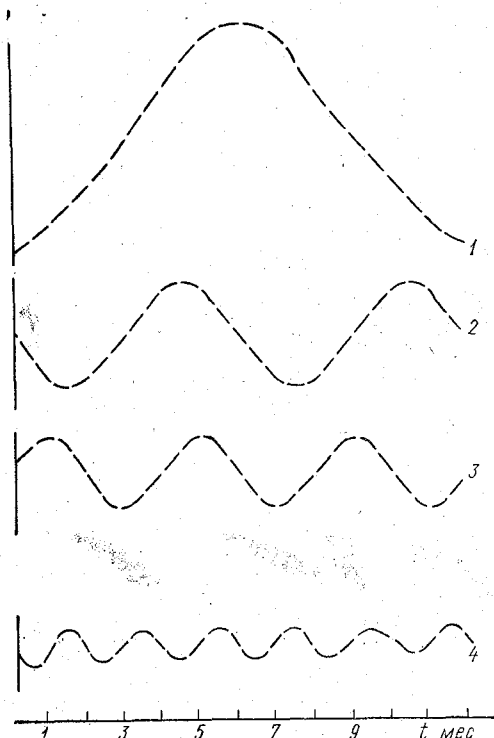


Рис. 9.2. Гармоники годового хода: 1, 2, 3, 4 — первая, вторая, третья, четвертая гармоники соответственно

Таблица 9.1

Гармонический анализ изменений теплосодержания деятельного слоя океана
в десятиградусных широтных зонах северного полушария

	90—80	80—70	70—60	60—50	50—40	40—30	30—20	20—10	10—0
A_1	31	37	27	58	86	109	66	50	16
A_2	6	16	6	2	10	0	19	25	11
Φ_1	4,9	6,5	5,8	5,7	6,0	5,8	5,7	2,9	4,2
Φ_2	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

Время наступления максимума изменений теплосодержания океана очень слабо меняется с широтой и почти везде (исключая зону 20—10°N) отмечается в летний период, когда происходит интенсивное прогревание верхних слоев океана за счет солнечной радиации.

Полугодовая гармоника играет существенную роль в сезонном ходе изменений теплосодержания только в низких широтах, причем в экваториальной зоне (10—0°N) ее амплитуда сравнима с амплитудой 1 гармоник. Сумма 1 и 2 гармоник с высокой точностью описывают годовой ход изменений теплосодержания океана в пределах 10—70°N, т. е. в большей части акватории океана северного полушария. В экваториальной зоне и в Северном Ледовитом океане уже заметна роль неперiodических (случайных) колебаний в формировании годового хода изменений теплосодержания океана.

9.3. Выделение тренда

В настоящее время нет общепринятой точки зрения, что следует понимать под трендом во временных рядах. По нашему мнению, как уже указывалось выше, тренд представляет собой некоторую медленно изменяющуюся составляющую процесса, период которой превышает длину исходной реализации.

Отсюда следует, что в качестве тренда временных рядов могут служить циклы более высокого порядка. Если, например, шаг дискретизации временного ряда составляет 1 сутки, то в соответствии с классификацией океанологических процессов трендом может являться годовой период. Однако это будет наблюдаться не во всех случаях. Чтобы убедиться в этом, достаточно обратиться к рис. 9.3, на котором представлены суточные значения температуры воды за годовой промежуток ($N=365$).

Действительно, если мы рассматриваем весь ряд, то мы имеем отчетливо выраженный сезонный ход температуры воды, на который накладываются короткопериодные синоптические колебания. Но поскольку длина исходной реализации позволяет выделить годовой период, то тренд в ряду отсутствует.

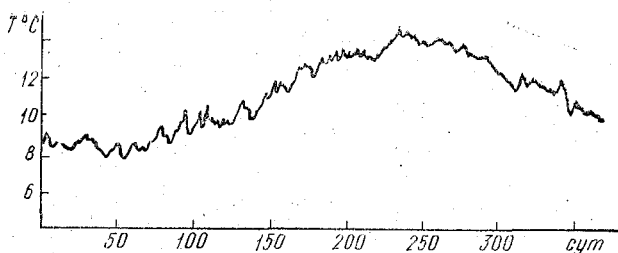


Рис. 9.3. Годовая серия наблюдений за температурой поверхности океана.

Если же уменьшить длину временного ряда (например, до 11 месяцев), то годовой период выделить уже не удастся, поэтому можно сделать вывод о наличии в ряду трендовой составляющей. Из рис. 9.3 видно, что тренд в этом случае носит нелинейный характер и, очевидно, может быть аппроксимирован полиномом второй степени:

$$T(t) = a_0 + a_1 t + a_2 t^2. \quad (9.7)$$

По мере дальнейшего уменьшения длины исходной реализации тренд будет претерпевать определенные изменения и для некоторых сравнительно небольших частей этого ряда вполне достаточным уже оказывается представление тренда линейным уравнением

$$T(t) = a_0 + a_1 t. \quad (9.8)$$

Таким образом, вид тренда и численные значения коэффициентов, определяемые методом наименьших квадратов, существенно зависят от длины временного ряда.

Разумеется, помимо выражений (9.7), (9.8), возможны и другие способы аппроксимации тренда. Например, в некоторых случаях нелинейный тренд описывается экспоненциальной зависимостью вида $T(t) = a_0 + a_1 \cdot \exp(a_2 t)$.

Следует иметь в виду, что все выше приведенные формулы характеризуют тренд по отношению к средней арифметической временного ряда, т. е. ее сдвиг во времени. Это наиболее часто встречающийся тип тренда на практике.

Однако в некоторых случаях может встречаться тренд по дисперсии временного ряда. Характерный пример такого тренда — полумесечное 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 (рис. 9.4) амплитуда его колебаний плавно изменяется от наибольших значений в сизигию до наименьших в квад-

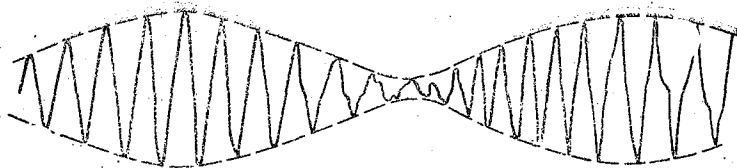


Рис. 9.4. График временного хода уровня моря

ратуру. Огибающая амплитуды свидетельствует о наличии тренда по дисперсии, в то время как математическое ожидание остается постоянным.

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 (9.8), то, обозначая последовательные моменты времени через $t_1, t_2, t_3, \dots, t_n$, исходный временной ряд $\Phi(t)$ можно записать следующим образом

$$\Phi_1 = a_0 + a_1 + d_1 \quad (t = 1),$$

$$\Phi_2 = a_0 + 2a_1 + d_2 \quad (t = 2),$$

$$\Phi_3 = a_0 + 3a_1 + d_3 \quad (t = 3),$$

$$\dots \dots \dots$$

$$\Phi_n = a_0 + na_1 + d_n \quad (t = n),$$

где $d_1, d_2, d_3, \dots, d_n$ — отклонения величин временного ряда от тренда, т. е. $d(t) = \Phi(t) - T(t)$.

Найдем первые разности

$$\Delta_1' = \Phi_2 - \Phi_1 = a_1 + (d_2 - d_1),$$

$$\Delta_2' = \Phi_3 - \Phi_2 = a_1 + (d_3 - d_2),$$

$$\dots \dots \dots$$

$$\Delta_{n-1}' = \Phi_n - \Phi_{n-1} = a_1 + (d_n - d_{n-1}).$$

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. Оценим тренд в межгодовых колебаниях уровня моря (табл. 2.2). Для этого методом наименьших квадратов рассчитаны полиномы первой и второй степени, которые соответственно имеют вид

$$T(t) = 253,4 - 4,17t,$$

$$T(t) = 297,9 - 7,2t + 0,035t^2.$$

Отношение дисперсии линейного тренда к дисперсии исходного ряда составляет $\eta = 0,85$. Аналогичное отношение для нелинейного тренда равно $\eta^2 = 0,88$. Следовательно, тренд описывает почти всю информацию об исходном процессе, причем существенных различий в типе (линейный или нелинейный) тренда не наблюдается. Оказалось также, что остаточные ряды, полученные в результате исключения тренда, $\Phi'(t) = \Phi(t) - T(t)$ очень мало отличаются друг от друга и содержат лишь некоторые нерегулярные (случайные) колебания. Однако, несмотря на это, остаточные ряды не подчиняются нормальному закону распределения, но в то же время их можно считать приближенно стационарными.

9.4. 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 стационарного эргодического 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 АФ имеет вид

$$r(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T X(t) X(t + \tau) d\tau. \quad (9.9)$$

Переходя от (9.9) к дискретному стационарному процессу конечной продолжительности, получим

$$r(\tau) = \frac{\sum_{i=1}^{N-\tau} (x_i - \bar{x})(x_{i+\tau} - \bar{x})}{(N - \tau) \sigma^2}, \quad (9.10)$$

где N — длина реализации, которая предполагается весьма значительной.

Помимо автокорреляционной функции, в статистике часто используют понятие автоковариационной функции, также характеризующей внутреннюю структуру процесса и определяемой по формуле

$$R(\tau) = \frac{\sum_{i=1}^{N-\tau} (x_i - \bar{x})(x_{i+\tau} - \bar{x})}{N - \tau}. \quad (9.11)$$

Из формул (9.10) и (9.11) следует, что

$$\frac{R(\tau)}{r(\tau)} = \sigma^2, \quad (9.12)$$

т. е. автокорреляционная и автоковариационная функции отличаются друг от друга на величину дисперсии.

Перечислим свойства автокорреляционной функции для типичных гидрометеорологических процессов:

1. АФ — четная функция, т. е. $r(-\tau) = r(\tau)$, что позволяет сократить процесс вычислений, рассчитывая $r(\tau)$ только для положительных сдвигов.

2. АФ — убывающая функция, т. е. $r(\tau=0) \geq r(\tau)$, причем при $\tau \rightarrow \infty$ $r(\tau) \rightarrow 0$.

3. Так как при $\tau=0$ $r=1$, то дисперсия АФ $\sigma^2=1$.

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:

$$\tau_{\text{кор}}^{(1)} = \int_0^{\infty} |r(\tau)| d\tau. \quad (9.13)$$

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. В качестве приближенной оценки максимального интервала корреляции в некоторых случаях принимают сдвиг τ_i , начиная с которого $|r(\tau)| \leq 0,05$.

Максимальный интервал корреляции может быть отождествлен с характерным временным масштабом физического процесса. В частности, в теории турбулентности максимальный интервал корреляции принимается за интегральный масштаб турбулентности.

Существует и другое определение интервала корреляции

$$\tau_{\text{кор}}^{(2)} = \int_0^{\infty} |r(\tau)|^2 d\tau, \quad (9.14)$$

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. Приблизненно 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

$$r(\tau) = e^{-\alpha|\tau|}, \quad (9.15)$$

$$r(\tau) = e^{-\alpha|\tau|} \cos \beta\tau, \quad (9.16)$$

где α — коэффициент затухания; β — частота колебания АФ. Подставив (9.15) в (9.13) и (9.14), имеем

$$\tau_{\text{кор}}^{(1)} = \frac{1}{\alpha}, \quad \tau_{\text{кор}}^{(2)} = \frac{1}{2\alpha}.$$

Таким образом, если АФ выражается экспоненциальной зависимостью, то $\tau_{\text{кор}}^{(1)} = 2\tau_{\text{кор}}^{(2)}$.

Подставив (9.16) в (9.13) и (9.14), получим

$$\begin{aligned} \tau_{\text{кор}}^{(1)} = \frac{1}{\alpha(1+\mu^2)} & \left[1 + \mu \exp\left(\frac{2\pi}{\mu}\right) + \frac{\mu}{\exp\left(\frac{2\pi}{\mu} - 1\right)} \left(\exp\left(-\frac{\pi}{2\mu}\right) + \right. \right. \\ & \left. \left. + 2 \exp\left(\frac{\pi}{2\mu}\right) + \exp\left(\frac{3\pi}{2\mu}\right) \right) \right], \\ \tau_{\text{кор}}^{(2)} = \frac{2\alpha^2 + \beta^2}{4\alpha(\alpha^2 + \beta^2)}, \end{aligned}$$

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 (9.16), имеем

$$\alpha = \ln \cos \frac{\beta \tau_2}{r(\tau_A)}, \quad \beta = \frac{\pi}{2\tau_1},$$

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 необходимости продолжительности наблюдений для надежной оценки ординат АФ, если в исследуемом ряду имеется цикличность. С этой целью представим временной ряд в виде разложения Фурье:

$$X(t) = \sum_{k=1} A_k \sin(\omega_k t + \varphi_k).$$

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 среднеквадратической ошибкой σ :

$$N \geq \tau_m \left(1 + \frac{1}{2\pi\sigma} \right). \quad (9.17)$$

Из (9.17) следует, например, что, для вычисления ординат оценки АФ с погрешностью 3 % от общей дисперсии процесса, необходимо, чтобы

$$N \approx (5 - 6) \tau_m,$$

а с погрешностью 2 %

$$N \approx 9\tau_m.$$

Таким образом, надежное определение ординат оценки 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 исключают из реализации.

При расчетах АФ для эргодического стационарного процесса с не очень большой длиной ряда статистические характеристики $(\bar{x}, \sigma)$, рассчитанные при различных величинах сдвига τ по формуле (9.10), могут заметно отличаться друг от друга. В связи с этим для коротких рядов значения АФ целесообразно рассчитывать по формуле

$$r^*(\tau) = \frac{\sum_{i=1}^{N-\tau} (x_i - \bar{x}_{1, N-\tau}) (x_{i+\tau} - \bar{x}_{\tau+1, N})}{(N - \tau - 1) \sigma_{1, N-\tau} \sigma_{\tau+1, N}}. \quad (9.18)$$

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 корреляции окажется значимым, если будет выполняться следующее условие:

$$|r(\tau)| > t_{\alpha} \sigma_r(\tau),$$

где t — критерий Стьюдента с $N - \tau - 1$ степенями свободы при заданном уровне значимости α ; $\sigma_r(\tau)$ — стандартное отклонение ординат АФ, рассчитанное в предположении $r(\tau)=0$, т. е.

$$\sigma_r(\tau) = \frac{1}{\sqrt{N - \tau - 1}}.$$

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 нормальному закону. Тогда доверительные границы для оценки ординат АФ при уровне значимости $\alpha=0,05$ могут быть найдены как

$$r_n(\tau_i) = r^*(\tau_i) - \frac{2}{\sqrt{N - \tau_i}} [1 - r^{*2}(\tau_i)],$$

$$r_v(\tau_i) = r^*(\tau_i) + \frac{2}{\sqrt{N - \tau_i}} [1 - r^{*2}(\tau_i)],$$

где $r_n(\tau_i)$, $r_v(\tau_i)$ — соответственно нижняя и верхняя доверительные границы.

9.5. 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 является взаимнокорреляционная функция (ВКФ), которая применительно к непрерывным процессам бесконечной продолжительности определяется как

$$r_{xy}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T X(t) Y(t + \tau) dt. \quad (9.19)$$

Функция r_{xy} 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 .

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 $r_{xy}(\tau)$:

$$\left. \begin{aligned} r_{xy}(\tau) &= r_{yx}(-\tau) \\ r_{yx}(\tau) &= r_{xy}(-\tau) \end{aligned} \right\}.$$

Функция $r_{xy}(\tau)$ характеризует степень зависимости при опережении $X(t)$ относительно $Y(t)$; $r_{yx}(\tau)$ — при запаздывании $X(t)$ по отношению к $Y(t)$.

При переходе от (9.19) к дискретным процессам большой продолжительности формулы для вычисления ВКФ приобретают вид

$$r_{xy}(\tau) = \frac{\sum_{i=0}^{N-\tau} (x_i - \bar{x})(y_{i+\tau} - \bar{y})}{(N - \tau) \sigma_x \sigma_y}, \quad (9.20)$$

$$r_{yx}(\tau) = \frac{\sum_{i=0}^{N-\tau} (y_i - \bar{y})(x_{i-\tau} - \bar{x})}{(N - \tau) \sigma_x \sigma_y}. \quad (9.21)$$

Естественно, что при небольших значениях N на точность вычисления ВКФ начинают уже сказываться изменения статистических характеристик: средней арифметической и стандартного отклонения. Учет их изменений производится аналогично (9.18).

Взаимная корреляционная функция связана со взаимной автокорреляционной функцией следующим образом:

$$r_{xy}(\tau) = \frac{R_{xy}(\tau)}{\sigma_x(0) \sigma_y(0)},$$

$$r_{yx}(\tau) = \frac{R_{yx}(\tau)}{\sigma_x(0) \sigma_y(0)},$$

где произведение $\sigma_x(0) \sigma_y(0)$ при нулевом сдвиге представляет взаимную дисперсию переменных X и Y .

ВКФ является более информативной по сравнению с АФ в том смысле, что она даст 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 . 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. Проанализируем колебания вертикального градиента температуры в термоклизе по годовому ряду среднесуточных наблюдений в пункте, находящемся в районе станции погоды «Танго» ($T(\varphi=29N, \lambda=135E)$). Исходный ряд центрировался относительно среднегодового значения. Затем вычислялись значения АФ до $\tau=60$ суткам (рис. 9.5). Одновременно были рассчитаны также нижняя и верхняя доверительные границы.

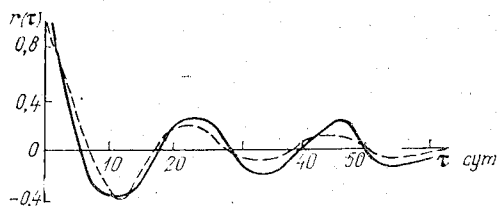


Рис. 9.5. Автокорреляционная функция колебаний вертикального градиента температуры в термоклизе (по Григоркиной, Губеру, Фуксу).

— АФ после исключения низкочастотной составляющей; — — — аппроксимация АФ зависимостью (9.16)

Как видно из рис. 9.5, график АФ представляет собой затухающую косинусоиду. Это свидетельствует о том, что в исходном процессе присутствует квазипериодическая составляющая, период которой, определенный по коррелограмме, близок к 20—22 суткам. Для аппроксимации АФ может быть использована формула вида (9.16). После нахождения численных значений параметров α ($\alpha \approx 0,07$ 1/сут) и β ($\approx 0,28$ 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. Используя найденные параметры α и β , нетрудно получить

$$\tau_{\text{кор}}^{(1)} \approx 20 \text{ суток}, \quad \tau_{\text{кор}}^{(2)} \approx 4 \text{ суток}.$$

Заметим, что интервал $\tau_{\text{кор}}^{(1)}$, определенный по числу пересечений реализацией процесса нулевого уровня, оказался равным 22 суткам.

Таким образом, линейная зависимость вертикального градиента температуры от предшествующей ситуации в данном пункте достаточно велика на интервале, не превышающем 4 суток. Через 20—22 суток значения процесса практически не коррелированы.

ГЛАВА 10. СПЕКТРАЛЬНЫЙ АНАЛИЗ

10.1. Спектральная плотность

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.

Некоторая функция частоты $S(\omega)$, связанная с автоковариационной функцией $R(\tau)$ обратным преобразованием Фурье, получила название спектральной плотности (энергетического спектра, плотности спектра мощности и т. п.), т. е.

$$S(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} R(\tau) e^{-i\omega\tau} d\tau, \quad (10.1)$$

где $e^{-i\omega\tau}$ — показательная функция мнимого аргумента. Обратным преобразованием для (10.1) является следующее выражение:

$$R(\tau) = \int_{-\infty}^{\infty} S(\omega) e^{i\omega\tau} d\omega.$$

Физический смысл $S(\omega)$ состоит в том, что она означает плотность дисперсии случайной стационарной 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 , то дисперсия такого процесса запишется выражением

$$\sigma^2 = \frac{1}{2} \sum_{i=1}^n A_i^2, \quad (10.2)$$

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.

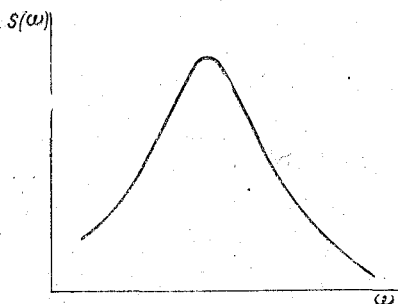
Если по оси абсцисс отложить частоты или соответствующие им периоды, а по оси ординат — плотность дисперсии, приходящуюся на различные частоты, то полученная кривая будет называться спектральной плотностью (рис. 10.1). В интегральном виде (10.2) имеет вид

$$\sigma^2 = \int_0^{\infty} S(\omega) d\omega,$$

т. е. дисперсия случайного процесса равна площади под кривой спектральной плотности.

В гидрометеорологии часто используют нормированную спектральную плотность, отражающую распределение плотности дисперсии по частотам в долях единицы, т. е.

$$s(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} r(\tau) e^{-i\omega\tau} d\tau. \quad (10.3)$$



В данном случае площадь под кривой спектральной плотностью равна единице.

К основным свойствам спектральной плотности относятся четность $S(\omega) = S(-\omega)$ и неотрицательность $S(\omega) \geq 0$.

Для вещественных случай-

Рис. 10.1. График спектральной плотности

ных процессов в силу четности функции $s(\omega)$ выражение (10.3) можно преобразовать к виду

$$s(\omega) = \frac{1}{\pi} \int_0^{\infty} r(\tau) \cos(\omega\tau) d\tau. \quad (10.4)$$

Аналитическое вычисление спектральной плотности возможно, если известно соответствующее аналитическое выражение корреляционной функции. Так, например, если АФ задается в виде (9.15) и (9.16), то получим следующие выражения для спектральной плотности:

$$s(\omega) = \frac{1}{\pi} \frac{\alpha}{\omega^2 + \alpha^2}, \quad (10.5)$$

$$s(\omega) = \frac{1}{\pi} \frac{\alpha}{(\omega - \beta)^2 + \alpha^2}. \quad (10.6)$$

Спектр вида (10.5) характерен для случайных процессов, интенсивность которых монотонно убывает с увеличением частоты. Максимум спектральной плотности (10.5) находится на частоте $\omega=0$. При $\alpha \gg \omega$ кривая спектральной плотности выравнивается и переходит в прямую линию, параллельную оси частот и находящуюся от нее на расстоянии $1/\pi\alpha$.

Спектр вида (10.6) содержит доминирующее по интенсивности колебание на частоте $\omega_0 > 0$. Данная частота связана с параметрами α и β зависимостью

$$\omega_0 = \sqrt{-\alpha^2 - \beta^2 + 2\beta \sqrt{\alpha^2 + \beta^2}}, \quad (10.7)$$

а ширина основания максимума $\Delta\omega_{эф}$ определяется выражением

$$\Delta\omega_{эф} = \frac{\pi\alpha}{1 + \frac{\alpha^2}{\alpha^2 + 4\beta^2}}. \quad (10.8)$$

В том случае, если $\alpha \ll \beta$, то зависимости (10.7) и (10.8) упрощаются:

$$\omega_0 \approx \beta, \quad \Delta\omega_{эф} \approx \pi\alpha. \quad (10.9)$$

Как видно из (10.9), чем меньше корреляция между ординатами процесса, тем больше $\Delta\omega_{эф}$. При $\alpha \rightarrow \infty$ $\Delta\omega_{эф} \rightarrow \infty$, в результате чего $s(\omega)$ становится одинакова для всех частот. При $\alpha \rightarrow 0$ 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.

В большинстве случаев $r(\tau)$ реальных океанологических процессов очень трудно поддается точной аппроксимации. В связи с этим в практических расчетах интеграл (10.3) обычно заменяют сум-

мированием с некоторым верхним пределом m , называемым длинной реализации спектра (максимальным сдвигом):

$$s(\omega) = \frac{1}{\pi} \sum_{\tau=1}^m r(\tau) \cos(\omega\tau) \Delta\tau. \quad (10.10)$$

Существует целый ряд численных методов аппроксимации выражения (10.10). Например, официальное признание получил метод, рекомендованный Всемирной метеорологической организацией, согласно которому

$$s(i) = r_0/m + \frac{2}{m} \sum_{\tau=1}^{m-1} \left[r_{\tau} \cos\left(\frac{360}{2m} i\tau\right) \right] + (-1)^m \frac{r_m}{m},$$

где i — номер гармоники ($i=1, 2, \dots, n$); 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 $T=2m/i$. 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, а с другой — неопределенностью выбора величины максимального сдвига.

Спектральные функции, рассчитанные по рядам наблюдений, обычно имеют много пиков, из которых часть является чисто случайной и должна быть исключена из анализа.

В качестве первого приближения для оценки надежности значений спектральной плотности используется метод Тьюки. 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

$$\nu = \frac{2N - m/2}{m}.$$

Для нахождения χ^2/ν составлены специальные таблицы, входными параметрами для которых являются уровень значимости и число степеней свободы (табл. 10.1).

Доверительные интервалы составляются относительно среднего спектра

$$S_0(\omega) = \sum_{\omega=0,01}^{2\pi} S(\omega)/N$$

по следующей формуле:

$$S_0(\omega) - \frac{\chi^2}{\nu} < S(\omega) < S_0(\omega) + \frac{\chi^2}{\nu}.$$

Доверительные границы χ^2/ν

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	1,83	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

В качестве уровня значимости используют обычно $\alpha=5\%$ и $\alpha=10\%$. Как показывают результаты расчетов, только самые высокие пики многих океанологических процессов превышают заданный уровень значимости.

10.2. Понятие о частотной весовой функции

При оценке спектра по формуле (10.10) возникает противоречие, обусловленное выбором предела m при неизменной длине выборки. С уменьшением m возрастает достоверность (эффективность) оценки спектра, но увеличивается его сглаженность и, следовательно, уменьшается подробность описания исходного процесса (рис. 10.2). Наоборот, с увеличением m на графике спектральной плотности (рис. 10.2) увеличивается количество пиков, т. е. возрастает степень подробности описания исходного процесса, однако при этом уменьшается эффективность

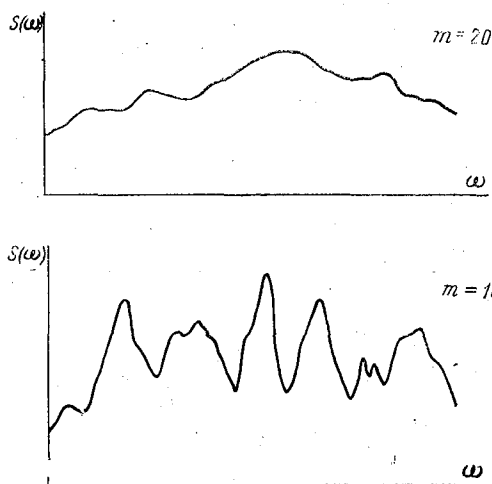


Рис. 10.2. Вид спектральной функции при различных значениях максимального сдвига (m)

оценок спектра. Таким образом, выигрывая в одном, мы всегда проигрываем в другом.

В принципе, до настоящего времени не разработано метода, позволяющего устранить отмеченное выше противоречие. Однако его можно ослабить путем подстановки в (10.10) так называемой частотной весовой функции или временного окна, т. е.

$$\tilde{s}(\omega) = \frac{1}{\pi} \sum_{i=1}^m \lambda(\tau) r(\tau) \cos(\omega\tau) \Delta\tau. \quad (10.11)$$

При этом идеальная частотная функция должна обладать следующим свойством:

$$\lambda(\tau) = \begin{cases} 1 & \text{при } \tau \leq m \\ 0 & \text{при } \tau > m \end{cases}. \quad (10.12)$$

Однако в действительности добиться выполнения данного условия чрезвычайно сложно.

Существует значительное число различных видов весовых функций, получивших применение в практических расчетах. К ним относятся:

функция Бартлетта

$$\lambda(\tau) = \begin{cases} 1 - |\tau|/\tau_m & \text{при } |\tau| \leq \tau_m \\ 0 & \text{при } |\tau| > \tau_m \end{cases},$$

функция Хемминга

$$\lambda(\tau) = \begin{cases} 0,54 + 0,46 \cos \pi\tau/\tau_m & \text{при } 0 \leq \tau \leq \tau_m \\ 0 & \text{при } \tau > \tau_m \end{cases},$$

функция Парзена

$$\lambda(\tau) = \begin{cases} 1 - \frac{6\tau^2}{\tau_m} \left(1 - \frac{\tau}{\tau_m}\right) & \text{при } 0 \leq \tau < \frac{\tau_m}{2} \\ 2 \left(1 - \frac{\tau}{\tau_m}\right)^3 & \text{при } \frac{\tau_m}{2} \leq \tau \leq \tau_m \\ 0 & \tau > \tau_m \end{cases}$$

а также целый ряд других весовых функций, сведения о которых можно найти в специальной литературе.

Характерной особенностью практически всех частотных весовых функций является то, что они производят сглаживание автокорреляционных оценок (весовая функция Хемминга может приводить даже к появлению отрицательных ординат спектра) и, следовательно, искажают оценки спектральной плотности. Но поскольку истинные оценки спектра, как правило, неизвестны, то очень трудно определить и возможные погрешности за счет введения частотной весовой функции.

Как видно из приведенных выше весовых функций, все они зависят лишь от одного аргумента — значения τ_m , представляю-

щего собой точку отсечения корреляционной функции. Естественно, что должна существовать взаимосвязь между m и τ_m , причем $m \leq \tau_m$. Для простоты положим, что $m = \tau_m$. В этом случае, задавая лишь один параметр (более удобным представляется τ_m), нетрудно рассчитать оценки спектральной плотности с использованием любой частотной функции.

Из результатов численных расчетов следует, что оценки спектральной плотности значительно больше зависят от величины τ_m , чем от вида временного 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 $\tau_m = 0,1 - 0,3N$. 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 анализом полученных спектров друг с другом. Это позволяет определить, какие особенности спектра являются закономерными, т. е. присущими исследуемому процессу.

10.3. Виды спектральных функций океанологических процессов

Большое разнообразие временной изменчивости океанологических процессов заключено между двумя эталонами: изменчивостью детерминированных процессов и изменчивостью чисто случайного процесса типа «белого шума». Очень редко океанологические характеристики полностью соответствуют какому-либо классическому виду процесса. Как правило, они представляют собой смесь нескольких составляющих изменчивости. Поэтому практический анализ вычисленных корреляционных и спектральных оценок, состоит в том, чтобы провести сравнение полученных эмпирических и теоретических функций. Рассмотрим корреляционные и спектральные функции некоторых типовых процессов (рис. 10.3).

Характеристика, которая не меняется во времени и остается постоянной величиной в течение всего периода наблюдений, описывается автокорреляционной функцией в виде прямой линии. Соответствующая ей спектральная функция имеет ненулевые значения только на нулевой частоте (рис. 10.3, а).

Процесс, характеристики которого строго периодически меняются во времени, носит название гармонического. Корреляционная функция такого процесса также имеет вид гармонического косинусоидального незатухающего колебания с периодом τ_0 . Соответствующий график спектральной функции имеет единственный максимум на частоте $f_0 = 1/\tau_0$ (или $\omega = 2\pi/\tau$), пропорциональный дисперсии периодических колебаний (рис. 10.3, б). Подобными характеристиками обладают временные ряды, описывающие колебания суточного и годового хода океанологических процессов — температуры воды, уровня моря, а также параметры морского волнения.

форм.

Б.Ш.

тренз.

Б.Ш.
+
форм.

Б.Ш.
форм
+
тренз.

полнз.

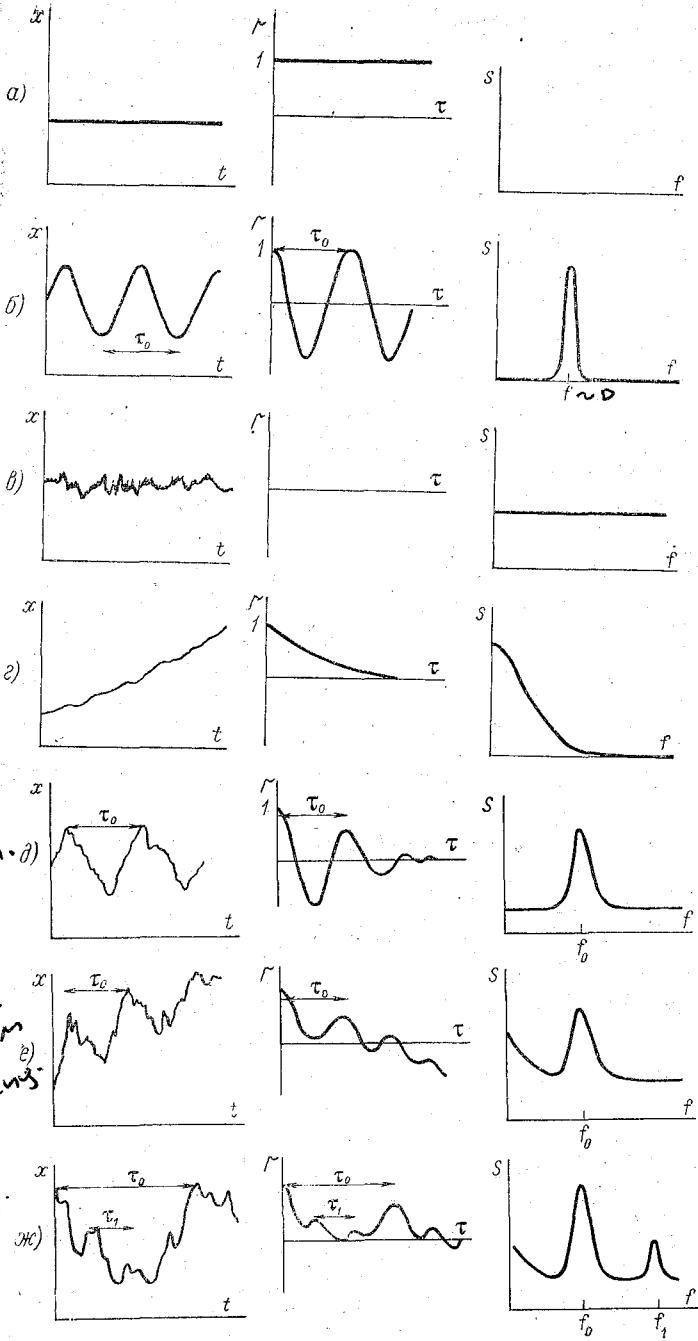


Рис. 10.3. Вид спектральных $S(f)$ и автокорреляционных функций $r(\tau)$ типовых океанологических процессов $X(t)$.

Процесс типа «белого шума» характеризуется дисперсией σ^2 , которая возникает под влиянием неперiodических стохастических колебаний, не содержащих в себе скрытой закономерной изменчивости. Автокорреляционная функция подобного процесса не отличается от нуля для всех τ_0 , кроме $\tau=0$, где $r(\tau)=1$. Спектральная функция процесса белого шума имеет вид горизонтальной линии, что соответствует постоянству дисперсии стохастических колебаний на всех частотах (рис. 10, в).

Многие океанологические характеристики содержат в своей изменчивости трендовые составляющие, которые представляют собой неперiodические колебания с периодом, превышающим длину исследуемой выборки. Автокорреляционная функция трендового процесса представляется экспонентой, а спектральная функция — плавной кривой, затухающей по мере увеличения частоты колебаний (рис. 10.3, г). Подобная инерционность часто сопоставляется со спектрами «красного шума». Такие спектральные функции типичны для большинства климатических параметров океана.

В изменчивости реальных океанологических характеристик чаще всего регистрируется наложение нескольких элементарных процессов. В большинстве регистрируемых параметров периодические изменения, как правило, проходят на фоне случайных колебаний «белого шума». Наложение гармонических колебаний на шум регистрируется на графике спектральной функции частичным снижением пика, соответствующего гармонике, и присутствием фонового ненулевого уровня спектра на всех частотах (рис. 10.3, д). Если дополнительно присутствуют трендовые колебания в исследуемых рядах, то спектральная функция имеет характерный подъем в интервале низких частот (рис. 10.3, е). Нередко гидрометеорологические ряды содержат не одно, а несколько характерных гармонических колебаний (полигармонический процесс). В этом случае на графике спектральной функции прослеживается несколько характерных пиков (рис. 10.3, ж). Полученный в итоге график описывает все возможные колебания в гидрологических рядах. В процессе анализа представляется возможным выделить и описать параметры всех элементарных процессов, которые рассмотрены выше.

10.4. Понятие о взаимной спектральной плотности

Взаимная спектральная плотность (ВСП) связана с взаимной корреляционной функцией двух стационарных случайных процессов $X(t)$ и $Y(t)$ обратным преобразованием Фурье, т. е.

$$s_{xy}(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} r_{xy}(\tau) e^{-i\omega\tau} d\tau. \quad (10.13)$$

В общем случае ВСП представляет собой комплексную величину, состоящую из вещественной части, называемой коспектром $C_{0xy}(\omega)$ и мнимой части — квадратурного спектра $Q_{xy}(\omega)$, т. е.

$$s_{xy}(\omega) = C_{0xy}(\omega) + iQ_{xy}(\omega).$$

Вещественная часть ВСП (косинус-спектр) находится как косинус преобразования Фурье четной части взаимной корреляции

$$C_{0xy}(\omega) = \frac{1}{\pi} \int_0^{\infty} r_+(\tau) \cos \omega\tau d\tau, \quad (10.14)$$

где

$$r_+(\tau) = \frac{r_{xy}(-\tau) + r_{xy}(\tau)}{2}. \quad (10.15)$$

Так как $r_+(\tau)$ — четная функция, то и коспектр является четной функцией и его называют четной (симметричной) частью ВСП.

Операцией (10.15) осуществляется сглаживание эффекта асимметрии функции взаимной корреляции $r_{xy}(\tau)$. Симметрия $r_+(\tau)$ 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:

$$Q_{xy}(\omega) = \frac{1}{\pi} \int_0^{\infty} r_-(\tau) \sin \omega\tau d\tau, \quad (10.16)$$

где

$$r_-(\tau) = \frac{r_{xy}(-\tau) - r_{xy}(\tau)}{2}. \quad (10.17)$$

Преобразованием (10.17) осуществляется усиление эффекта асимметрии функции $r_{xy}(\tau)$. Поскольку $r_-(\tau)$ является нечетной функцией ($r_-(\tau) = -r_-(-\tau)$), то асимметрия функции $r_-(\tau)$ проявляется в смещении максимума $r_-(\tau)$ на некоторый сдвиг $\tau \neq 0$.

При отсутствии асимметрии $r_-(\tau)$, $\tau \neq 0$. Неравенство нулю $r_-(\tau)$ 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 энергии гармоник различных частот в общую корреляцию ряда при условии, что гармоники, содержащиеся во временном ряде $X(t)$, сдвинуты на четвертьпериода назад по отношению к временному ряду $Y(t)$.

Взаимная спектральная плотность связана с взаимной дисперсией следующим образом:

$$r_{xy}(0) = \int_{-\infty}^{\infty} s_{xy}(\omega) d\omega$$

или

$$r_{xy}(0) = \int_{-\infty}^{\infty} C_{0xy}(\omega) d\omega.$$

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

$$F^2(\omega) = \frac{|s_{xy}^2(\omega)|}{s_x(\omega)s_y(\omega)} = \frac{C_{0xy}^2(\omega) + Q_{xy}^2(\omega)}{s_x(\omega)s_y(\omega)}. \quad (10.18)$$

Функция $F(\omega)$, называемая *когерентностью*, характеризует линейную статистическую связь спектральных компонент одинаковой частоты и по смыслу аналогична коэффициенту детерминации, но в отличие от него зависит от частоты.

Вследствие того, что всегда справедливо неравенство

$$C_{0xy}^2(\omega) + Q_{xy}^2(\omega) \leq s_x(\omega)s_y(\omega),$$

называемое неравенством когерентности, величина $F^2(\omega)$ изменяется в пределах от 0 до 1.

Из (10.18) следует, что при $Q_{xy}(\omega_i) = 0$, $C_{0xy}(\omega_i) \neq 0$ разность фаз колебаний на частоте ω_i должна быть равна нулю, так как взаимосвязь процессов будет существовать только за счет их синхронного взаимодействия. При $C_{0xy}(\omega_i) = 0$, $Q_{xy}(\omega_i) \neq 0$ 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 $C_{0xy}(\omega_i) \neq 0$, $Q_{xy}(\omega_i) \neq 0$, разность фаз спектральных компонент фиксированной частоты вычисляется по формуле

$$\varphi_y(\omega) - \varphi_x(\omega) = \theta_{xy}(\omega) = \arctg \frac{Q_{xy}(\omega)}{C_{0xy}(\omega)}.$$

Величина $\theta_{xy}(\omega)$, называемая при геометрической интерпретации ВСП фазовым спектром, 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 $\sqrt{C_{0xy}^2(\omega) + Q_{xy}^2(\omega)}$, можно назвать амплитудным спектром.

Одновременно когерентность служит мерой устойчивости разности фаз. Если разность фаз двух процессов постоянна, то $F(\omega) = 1$, если разность фаз неустойчива, то должно выполняться условие $F^2(\omega) \rightarrow 0$.

10.5. Фильтрация 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 преобразования вида

$$\tilde{X}(t) = \int_{-\infty}^{\infty} h(\tau) X(t + \tau) d\tau, \quad (10.19)$$

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

$$c(\omega) = \int_{-\infty}^{\infty} h(\tau) \cos 2\pi\omega\tau d\tau. \quad (10.20)$$

Как следует из выражения (10.20), весовая функция $h(\tau)$ такого четного фильтра связана с частотной характеристикой обратным преобразованием Фурье

$$h(\tau) = \int_{-\infty}^{\infty} c(\omega) \cos 2\pi\omega\tau d\omega. \quad (10.21)$$

Таким образом, для фильтрации процесса $X(t)$ следует вначале задать частотную характеристику $c(\omega)$, а затем из выражения (10.21) найти весовую функцию и выполнить преобразование (10.19).

Задача фильтрации решается наилучшим образом, если гармоники с частотами $\omega_1 \leq \omega \leq \omega_2$ на выходе фильтра сохраняются неизменными, а амплитуды всех остальных гармоник обратятся в нуль. Частотная характеристика такого идеального фильтра имеет вид

$$c(\omega) = \begin{cases} 1 & \text{при } \omega_1 \leq \omega \leq \omega_2 \\ 0 & \text{при } \omega < \omega_1, \omega > \omega_2 \end{cases}. \quad (10.22)$$

Подставив характеристику (10.22) в уравнение (10.21), можно найти весовую функцию идеального фильтра:

$$h(\tau) = \frac{\cos \pi \tau (\omega_1 + \omega_2) \sin \pi \tau (\omega_2 - \omega_1)}{\pi \tau}. \quad (10.23)$$

Величина спектра после фильтрации определяется как

$$\hat{S}(\omega) = c(\omega) \tilde{S}(\omega).$$

Однако частотная характеристика (10.22) может быть реализована в том случае, если весовая функция (10.23), 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 $\omega_1 \leq \omega \leq \omega_2$, 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 (10.22) 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.

Переходя к сумме по конечному интервалу $(2M+1)$, выражение (10.19) запишем как

$$\tilde{X}_i = \sum_{j=-M}^M h_j X_{i+j}. \quad (10.24)$$

В зависимости от выделяемого диапазона частот фильтры подразделяются на низкочастотные или сглаживающиеся ($0 \leq \omega \leq \omega_2$), высокочастотные ($\omega_1 \leq \omega \leq \infty$) и полосовые ($\omega_1 \leq \omega \leq \omega_2$).

Первые применяются в случае необходимости исследования долгопериодных колебаний или тренда в исходном ряду, а также для подавления случайных колебаний, связанных с ошибками измерений.

Высокочастотные фильтры позволяют стационаризовать ряд и исследовать высокочастотные колебания до определенного заданного 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, например,

относятся скользящее среднее, треугольный фильтр (фильтр Барлетта), 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

$$h_j = \begin{cases} \frac{1}{2M+1} & \text{при } |j| \leq M, \\ 0 & |j| > M, \end{cases} \quad (10.25)$$

где $(2M+1)$ — интервал сглаживания.

Общая формула для отфильтрованного ряда при низкочастотной фильтрации может быть записана как

$$\tilde{X}_i = \frac{1}{2M+1} \sum_{j=-M}^M X_{i+j}.$$

При использовании скользящего осреднения в качестве высокочастотного фильтра общее выражение для $\tilde{X}_i$ приобретает вид

$$\tilde{X}_i = X_i - \frac{1}{2M+1} \sum_{j=-M}^M X_{i+j}.$$

И, наконец, для полосового фильтра общую формулу для ряда X_i можно представить как

$$\tilde{X}_i = \frac{1}{2M_1+1} \sum_{j=-M_1}^{M_1} X_{i+j} - \frac{1}{2M_2+1} \sum_{j=-M_2}^{M_2} X_{i+j},$$

где $M_2 > M_1$.

Частотная характеристика низкочастотного фильтра, определяемая выражением

$$c_t(\omega) = \frac{\sin \pi \omega (2M+1)}{\pi \omega (2M+1)}$$

приводится на рис. 10.4, а.

Частотную характеристику высокочастотного фильтра $c_h(\omega)$ можно получить через частотную характеристику сглаживающего фильтра (рис. 10.4, б)

$$c_h(\omega) = 1 - c_t(\omega).$$

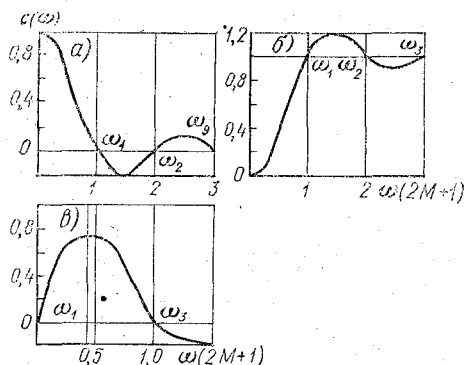
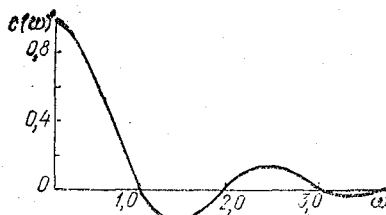


Рис. 10.4. График частотной характеристики различных фильтров: а — высокочастотный фильтр; б — низкочастотный фильтр; в — полосовой фильтр

Рис. 10.5. Частотная характеристика фильтра скользящего среднего

Наконец, частотная характеристика полосового фильтра может быть представлена как разность частотных характеристик двух сглаживающих фильтров (рис. 10.4, в)



$$c(\omega) = c_l(\omega, M_1) - c_l(\omega, M_2).$$

Рассмотрим, например, частотную характеристику сглаживающего фильтра. Как видно из рис. 10.5, величина $c_l(\omega)$ на участке главного лепестка (от значения $\omega=0$ 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 $p=1, 3, 5, \dots$, в процессе $X(t)$ на выходе фильтра меняют фазу на 180° .

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

$$\frac{1}{2(N-M)} < \omega \leq \frac{1}{2}$$

или с периодами в диапазоне $2 \leq p < 2(N-M)$, где p выражается через число интервалов дискретности;

3) средние арифметические значения рядов X_i и $\tilde{X}_i$ связаны между собой соотношением

$$\sum_{i=1}^N \tilde{X}_i / \sum_{i=1}^N X_i = \sum_{j=-M}^M h_j.$$

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.

Для фильтра «скользящее среднее» характерно также то, что дисперсия сглаженного ряда всегда меньше дисперсии исходного ряда и 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 ее. 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.

2000 / 30
180 / 66.000
200

ПРИЛОЖЕНИЯ

Приложение 1

Функция Лапласа

	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0	0,000	0,040	0,079	0,118	0,155	0,192	0,226	0,258	0,288	0,316
1	0,341	0,364	0,385	0,403	0,419	0,433	0,445	0,455	0,464	0,471
2	0,477	0,482	0,486	0,489	0,492	0,494	0,495	0,496	0,497	0,498
3	0,498	0,499	0,499	0,499	0,499	0,499	0,499	0,499	0,499	0,499
4	0,500									

Приложение 2

Критическое значение χ^2 -распределения
с ν степенями свободы при заданном уровне значимости α

ν	Уровень значимости α (%)			
	95%	90%	10%	5%
1	10^{-3}	0,01	2,7	3,8
2	0,1	0,2	4,6	5,9
4	0,7	1,0	7,7	9,4
6	1,6	2,2	10,6	12,5
8	2,7	3,4	13,3	15,5
10	3,9	4,8	15,9	18,3
15	7,2	8,5	22,3	24,9
20	10,8	12,4	28,4	31,4
25	14,6	16,4	34,3	37,6
30	18,4	20,5	40,2	43,7
40	26,5	29,0	51,8	55,7
50	34,7	37,6	63,7	67,5
60	43,1	46,4	74,4	79,0
80	60,3	64,2	96,5	101,8
100	77,9	82,3	118,5	124,3

Критические значения t -критерия при ν степенях свободы
и заданном уровне значимости

Число степеней свободы ν	Уровень значимости α , %			
	10	5	2,5	1
1	3,078	6,314	12,706	31,821
2	1,886	2,920	4,303	6,965
3	1,638	2,353	3,182	4,541
4	1,533	2,132	2,776	3,747
5	1,476	2,015	2,571	3,365
6	1,440	1,943	2,447	3,143
7	1,415	1,895	2,365	2,998
8	1,397	1,860	2,306	2,896
9	1,383	1,833	2,262	2,821
10	1,372	1,812	2,228	2,764
11	1,363	1,796	2,201	2,718
12	1,356	1,782	2,179	2,681
13	1,350	1,771	2,160	2,650
14	1,345	1,761	2,145	2,624
15	1,341	1,753	2,131	2,602
16	1,337	1,746	2,120	2,583
17	1,333	1,740	2,110	2,567
18	1,330	1,734	2,101	2,552
19	1,328	1,729	2,093	2,539
20	1,325	1,725	2,086	2,528
21	1,323	1,721	2,080	2,518
22	1,321	1,717	2,074	2,508
23	1,319	1,714	2,069	2,500
24	1,318	1,711	2,064	2,492
25	1,316	1,708	2,060	2,485
26	1,315	1,706	2,056	2,479
27	1,314	1,703	2,052	2,473
28	1,313	1,701	2,048	2,467
29	1,311	1,699	2,045	2,462
30	1,310	1,697	2,042	2,457

Приложение 4

Критические значения F -распределения с ν_1 и ν_2 степенями свободы для пятипроцентного уровня значимости ($\alpha=5\%$)

ν_2	Число степеней свободы числителя ν_1						
	1	2	3	4	5	10	20
2	18,5	19,0	19,1	19,2	19,3	19,4	19,4
5	6,6	5,7	5,4	5,1	5,0	4,7	4,5
10	4,9	4,1	3,7	3,4	3,3	2,9	2,7
15	4,5	3,6	3,2	3,0	2,9	2,5	2,3
20	4,3	3,4	3,1	2,8	2,7	2,3	2,1
25	4,2	3,3	2,9	2,7	2,6	2,2	2,0
30	4,1	3,3	2,9	2,6	2,5	2,1	1,9
40	4,0	3,2	2,8	2,6	2,4	2,0	1,8
60	4,0	3,1	2,7	2,5	2,3	1,9	1,7

Примечание. Дисперсионное соотношение $F^*=\sigma_1/\sigma_2$ при сравнении с табличным F -критерием должно быть составлено так, чтобы число степеней свободы числителя ν_1 не превосходило число степеней свободы знаменателя ν_2 .

$$G_1 = 1 \quad G_2 = n - 1$$

Приложение 5

Значение величин Z для значений r от 0,00 до 0,99

r	0	1	2	3	4	5	6	7	8	9
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	0,698	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	0,974	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

ЛИТЕРАТУРА

ОСНОВНАЯ

1. Бельшев А. П., Клеванцов Ю. П., Рожков В. А. Вероятностный анализ морских течений. — Л.: Гидрометеониздат, 1982. — 296 с.
2. Беляев В. И. Обработка и теоретический анализ океанографических наблюдений. — Киев: Наукова думка, 1973. — 293 с.
3. Григоркина Р. Г., Губер П. К., Фукс В. Р. Прикладные методы корреляционного и спектрального анализа крупномасштабных океанологических процессов. — Л., изд. ЛГУ, 1973. — 171 с.
4. Груза Г. В., Рейтенбах Р. Г. Статистика и анализ гидрометеорологических данных. — Л.: Гидрометеониздат, 1982. — 216 с.
5. Драган Я. П., Рожков В. А., Яворский И. Н. Методы вероятностного анализа ритмики океанологических процессов. — Л.: Гидрометеониздат, 1987. — 318 с.
6. Исаев А. А. Статистика в метеорологии и климатологии. — М.: изд. МГУ, 1988. — 245 с.
7. Калашиников П. А. Первичная обработка гидрологической информации: физико-химические свойства морской воды. — Л.: Гидрометеониздат, 1985. — 152 с.
8. Коняев К. В. Спектральный анализ случайных океанологических полей. — Л.: Гидрометеониздат, 1981. — 208 с.
9. Рожков В. А. Методы вероятностного анализа океанологических процессов. — Л.: Гидрометеониздат, 1980. — 280 с.
10. Шелутко В. А. Численные методы в гидрологии. — Л.: изд. ЛПИ, 1983. — 155 с.

ДОПОЛНИТЕЛЬНАЯ

1. Беляев Б. Н. Численные методы обработки и анализа океанологической информации. — Л.: изд. ЛПИ, 1978. — 114 с.
2. Бендат Дж., Пирсол А. Применения корреляционного и спектрального анализа. — М.: Мир, 1983. — 312 с.
3. Бронштейн И. Н., Семендяев К. А. Справочник по математике. — М.: Наука, 1986. — 544 с.
4. Демидович Б. П., Марон И. А., Шувалова Э. З. Численные методы анализа. — М.: Наука, 1967. — 368 с.
5. Монин А. С., Яглом А. М. Статистическая гидромеханика. — М.: Наука, 1965. 4.1—639 с., 4.2—420 с.
6. Николаев Ю. В. Численные методы обработки и анализа океанологической информации (конспект лекций). — Л.: изд. ЛПИ, 1973. — 75 с.
7. Поляк И. И. Методы анализа случайных процессов и полей в климатологии. — Л.: Гидрометеониздат, 1979. — 256 с.
8. Рождественский А. В., Чеботарев А. И. Статистические методы в гидрологии. — Л.: Гидрометеониздат, 1974. — 421 с.

ОГЛАВЛЕНИЕ

	Стр.
Введение	3
РАЗДЕЛ 1. ПЕРВИЧНАЯ ОБРАБОТКА ОКЕАНОЛОГИЧЕСКОЙ ИН- ФОРМАЦИИ	5
Глава 1. Накопление и хранение океанологической информации	6
1.1. Океанологическая информация	—
1.2. Единая система сбора, хранения и переработки гидрометеорологической информации	8
1.3. Понятие банка океанологических данных	10
1.4. Характеристика современных банков океанологических данных	13
Глава 2. Характеристики статистических рядов и их свойства	17
2.1. Основные понятия случайных величин	—
2.2. Характеристики положения случайной величины	19
2.3. Характеристики рассеяния случайной величины	22
2.4. Характеристики формы кривой распределения случайной величины	23
2.5. Понятие о статистических моментах	24
2.6. Оценивание различных характеристик случайной величины по эмпирическим данным	25
Глава 3. Законы распределения океанологических элементов	27
3.1. Нормальный закон распределения. Его свойства и значение	—
3.2. Законы распределения, используемые в статистических расчетах	31
3.3. Законы распределения, используемые в гидрометеорологии	34
3.4. Требования к статистическим оценкам характеристик распределения	37
3.5. Принципы построения эмпирической функции распределения	38
3.6. Приближенные методы оценки соответствия эмпирических и теоретических функций распределения	40
3.7. Определение доверительных границ для среднего значения и дисперсии выборки	43
Глава 4. Статистическая проверка гипотез	44
4.1. Понятие о нулевой гипотезе и уровне значимости	45
4.2. Оценка расхождения между средними значениями двух выборок	47
4.3. Оценка расхождения между дисперсиями двух выборок	49
4.4. Оценка соответствия эмпирических и теоретических функций распределения на основе критерия согласия	50
4.5. Проверка гипотезы о принадлежности двух выборок одной генеральной совокупности	52
РАЗДЕЛ 2. ПОСТРОЕНИЕ ЭМПИРИЧЕСКИХ ЗАВИСИМОСТЕЙ	54
Глава 5. Корреляционный анализ	55
5.1. Коэффициент корреляции, его свойства и оценка достоверности	—

5.2. Понятие о методе наименьших квадратов	Стр. 60
5.3. Линейная регрессия двух переменных	62
5.4. Понятие о ложной корреляции	67
Глава 6. Анализ нелинейных зависимостей	69
6.1. Общая схема построения нелинейных зависимостей	—
6.2. Полиномиальная регрессия	75
6.3. Тренд-анализ	78
6.4. Понятие о рациональных функциях	80
Глава 7. Анализ погрешностей океанологических измерений и расчетов	83
7.1. Общая схема анализа погрешностей	—
7.2. Случайные и систематические погрешности	87
7.3. Выявление и устранение грубых погрешностей	89
7.4. Понятие о косвенных погрешностях	91
РАЗДЕЛ 3. АНАЛИЗ ОКЕАНОЛОГИЧЕСКИХ ВРЕМЕННЫХ РЯДОВ	94
Глава 8. Основные положения теории случайных функций	95
8.1. Понятие случайной функции	—
8.2. Стационарность случайных функций	97
8.3. Эргодичность стационарных случайных функций	99
Глава 9. Методы анализа временных рядов	100
9.1. Общая схема исследования временной изменчивости	—
9.2. Гармонический анализ	103
9.3. Выделение тренда	105
9.4. Автокорреляционный анализ	108
9.5. Понятие о взаимокорреляционной функции	112
Глава 10. Спектральный анализ	115
10.1. Спектральная плотность	—
10.2. Понятие о частотной весовой функции	119
10.3. Виды спектральных функций океанологических процессов	121
10.4. Понятие о взаимной спектральной плотности	123
10.5. Фильтрация рядов	126

ПАВЕЛ АНТОНОВИЧ ВАЙНОВСКИЙ
ВАЛЕРИЙ НИКОЛАЕВИЧ МАЛИНИН

МЕТОДЫ ОБРАБОТКИ АНАЛИЗА ОКЕАНОЛОГИЧЕСКОЙ ИНФОРМАЦИИ. ЧАСТЬ 1. ОДНОМЕРНЫЙ АНАЛИЗ

Учебное пособие

Редактор О. С. Крайнова

Формат 60×90¹/₁₆. Бумага тип. № 2. Лит. гарн. Печать высокая.
Сдано в набор 28.11.90 г. Подписано к печати 25.02 91
Печ. л. 8,5 Уч.-изд. л. 9,0 Тираж 350 экз. Темплан 1991 г. поз. 200
ЛГМИ. 195196. Ленинград, Малоохтинский пр., 98. Цена 72 коп.

Типография ВАС