

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Владимирский государственный университет
имени Александра Григорьевича и Николая Григорьевича Столетовых»

Р. И. МАКАРОВ Е. Р. ХОРОШЕВА

МЕТОДЫ АНАЛИЗА ДАННЫХ

Учебное пособие



Владимир 2021

УДК 004.451.5
ББК 32.972.134
М15

Рецензенты:

Доктор технических наук, профессор
зав. кафедрой вычислительной техники и систем управления
Владимирского государственного университета
имени Александра Григорьевича и Николая Григорьевича Столетовых
В. Н. Ланцов

Кандидат физико-математических наук, доцент
доцент Департамента математики
Финансового университета при Правительстве Российской Федерации
М. Б. Хрипунова

Издается по решению редакционно-издательского совета ВлГУ

М15 **Макаров, Р. И.** Методы анализа данных : учеб. пособие /
Р. И. Макаров, Е. Р. Хорошева ; Владим. гос. ун-т им. А. Г. и
Н. Г. Столетовых. – Владимир : Изд-во ВлГУ, 2021. – 216 с.
ISBN 978-5-9984-1399-5

Приведены общие сведения о статистических методах анализа данных, основы теории вероятностей и математической статистики, современные методы и приемы, иллюстрированные на примерах из производственной и исследовательской практики. Теоретический материал подкреплён примерами.

Предназначено для бакалавров, обучающихся по направлениям 09.03.02 – Информационные системы и технологии, 09.03.04 – Программная инженерия.

Рекомендовано для формирования общепрофессиональных компетенций в соответствии с ФГОС ВО.

Табл. 48. Ил. 50. Библиогр.: 26 назв.

УДК 004.451.5
ББК 32.972.134

ISBN 978-5-9984-1399-5

© ВлГУ, 2021

ВВЕДЕНИЕ

Курс направлен на формирование у бакалавров представления о статистических методах анализа данных как метода научного познания и на умение использовать компьютер как средство познания и обработки больших массивов данных.

Математическая статистика широко использует понятия и методы теории вероятностей, поэтому материал дисциплины содержит основные положения теории вероятностей: испытание, поле событий, операции над событиями, условные вероятности.

Анализируемые данные, как правило, рассматриваются как выборка случайных величин из генеральной совокупности данных. Закономерности, которым подчиняется исследуемая переменная, определяются комплексом условий ее наблюдения. Математически эти закономерности задаются соответствующим законом распределения вероятностей. Изучаются распределения дискретных и непрерывных случайных величин, оцениваются параметры распределения по малым выборкам.

Следующими после статистической оценки параметров распределения являются проверка гипотез о положении центра группирования, равенстве двух центров распределения, равенстве дисперсий и проверка гипотез о законе распределения.

В книге рассматриваются простейшие положения теории случайных процессов. Изучаются закономерности изменений случайных величин в зависимости от изменения неслучайного параметра – вре-

мени. Приводится классификация случайных процессов. Примером случайного процесса можно назвать флуктуационный шум, наиболее характерный для большинства каналов электросвязи. Рассматривается энергетический спектр случайного процесса.

Студенты знакомятся с методами системного анализа для изучения взаимосвязей между отдельной зависимой переменной и группой влияющих на нее показателей. Это осуществляется при помощи множественного корреляционного анализа.

Вопрос о существенности влияния того или иного фактора или комбинации факторов на рассматриваемый признак изучается с помощью дисперсионного анализа. Рассматриваются простейшие приемы дисперсионного анализа – однофакторный и двухфакторный анализы.

После того как с помощью корреляционного анализа выявлено наличие статистически значимых связей между переменными и оценена степень их тесноты, переходят к математическому описанию конкретного вида зависимости с использованием регрессионного анализа. С этой целью подбирают класс функций, отбирают наиболее информативные аргументы, вычисляют оценки неизвестных значений параметров уравнения связи и анализируют точность полученного уравнения регрессии. Рассматривается построение множественного линейного уравнения регрессии. Методом наименьших квадратов оцениваются неизвестные параметры уравнения и значимость уравнения регрессии.

В тех случаях, когда неизвестен вид уравнения, но располагают некоторой информацией о возможной зависимости, пользуются нелинейной регрессией. Анализируемые показатели часто оказываются взаимозависимыми. Структура связей между такими показателями (переменными) описывается с помощью системы одновременных

(структурных) уравнений. При оценивании коэффициентов структурной модели используется ряд методов. В книге рассматривается наиболее простой – косвенный метод наименьших квадратов.

Наличие множества исходных признаков, характеризующих процесс функционирования объектов, заставляет отбирать наиболее существенные из них и изучать меньший набор показателей. Для этого исходные признаки подвергают некоторому преобразованию, обеспечивающему минимальную потерю информации. Такое решение обеспечивается методами снижения размерности, к которым относятся факторный и компонентный анализы. Эти методы дают возможность наиболее просто объяснить многомерные структуры. Они позволяют вскрывать объективно существующие, непосредственно не наблюдаемые закономерности при помощи полученных факторов и главных компонент.

Модели, построенные по данным, характеризующим функционирование системы или процесса за ряд последовательных равноотстоящих моментов времени, называются моделями временных рядов. В пособии рассматривается простейшая модель аддитивного случайного процесса, исследуется и моделируется тренд сезонных, сезонных и периодических колебаний во временном ряду. Рассматриваются критерии точности и адекватности математических моделей. Изучаются причинно-следственные зависимости переменных, представленных в виде временных рядов.

В результате освоения курса студенты должны решать стандартные профессиональные задачи с применением естественнонаучных и общеинженерных знаний, методов математического анализа и моделирования.

1. АНАЛИЗ ДАННЫХ

В результате развития информационных технологий количество данных, накопленных в электронном виде, растет быстрыми темпами. Эти данные существуют в различных форматах: тексты, изображения, аудио, видео, гипертекстовые документы, реляционные базы данных и т. д. Однако подавляющая часть доступной информации не несет для конкретного человека какой-либо пользы, так как он не в состоянии переработать такое количество сведений. Возникает проблема извлечения полезной для пользователя информации из большого объема данных.

1.1. Понятие интеллектуального анализа данных

Понятие интеллектуального анализа данных соответствует широко распространенному термину Data Mining, который часто переводится как добыча данных, глубинный анализ данных, извлечение знаний, раскопка знаний в базах данных. Понятие Data Mining, появившееся в 1978 году, приобрело большую популярность в современной трактовке примерно с первой половины 1990-х годов. До этого времени обработка и анализ данных осуществлялись в рамках прикладной статистики, при этом в основном решались задачи по обработке небольших баз данных. Мультидисциплинарная область возникла и развивалась на базе таких наук, как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных и др.

Data Mining можно охарактеризовать как технологию, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей:

– неочевидных, так как найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем;

- объективных, так как обнаруженные закономерности будут полностью соответствовать действительности в отличие от экспертного мнения, которое всегда субъективно;
- практически полезных, так как выводы имеют конкретное значение, которому можно найти практическое применение.

1.2. Data Mining как часть рынка информационных технологий

Цель Business Intelligence – преобразование объемов данных в ценность бизнеса. Состав рынка систем Business Intelligence определяется как набор программных продуктов следующих классов:

- средства построения хранилищ данных (Data Warehousing, ХД);
- системы оперативной аналитической обработки (OLAP);
- информационно-аналитические системы (Enterprise Information Systems, EIS);
- средства интеллектуального анализа данных (Data Mining);
- инструменты для выполнения запросов и построения отчетов (query and reporting tools).

Данные обеспечивают получение информации, которая поддерживает решения. Эти понятия являются составной частью так называемой информационной пирамиды, в основании которой находятся данные, следующий уровень – это информация, затем идет решение, завершает пирамиду уровень знания (рис. 1.1).

По мере продвижения вверх по информационной пирамиде объемы данных переходят в ценность решений, т. е. ценность для бизнеса. Целью Business Intelligence является преобразование объемов данных в ценность бизнеса.

Верхний уровень приложений – это уровень бизнеса, на нем менеджеры принимают решения (перекрестные продажи, контроль ка-



Рис. 1.1. Информационная пирамида

чества, удерживание клиентов). Средний уровень действий выступает уровнем информации, именно на нем выполняются действия Data Mining (прогностическое моделирование, анализ связей, сегментация данных и др.). Нижний уровень определяет задачи Data Mining, которые необходимо решить применительно к данным, имеющимся в наличии.

Имеется ряд существенных отличий Data Mining от других методов анализа данных. Традиционные методы анализа данных (статистические методы) и OLAP в основном ориентированы на проверку заранее сформулированных гипотез (verification-driven data mining) и на «грубый» разведочный анализ, составляющий основу оперативной аналитической обработки данных (OnLine Analytical Processing, OLAP), в то время как одно из основных положений Data Mining – выявление закономерностей.

Большинство статистических методов для выявления взаимосвязей в данных используют концепцию усреднения по выборке, приводящую к операциям над несуществующими величинами, тогда как Data Mining оперирует реальными значениями. OLAP больше подходит для понимания ретроспективных данных, Data Mining опирается на ретроспективные данные для получения ответов на вопросы о будущем.

1.3. Набор данных и их атрибутов

В широком понимании данные представляют собой факты, текст, графики, картинки, звуки, аналоговые или цифровые видеосегменты. Данные могут быть получены в результате измерений, экспериментов, арифметических и логических операций и должны быть представлены в форме, пригодной для хранения, передачи и обработки. Иными словами, данные – это необработанный материал, предоставляемый поставщиками данных и используемый потребителями для формирования информации на основе данных. Набор данных может быть представлен двухмерной табл. 1.1.

Таблица 1.1

Пример набора данных

Объем реализации Y	Реклама X_1	Цена X_2	Цена конкурента X_3	Индекс потребительских расходов X_4
126	4	15	17	100
137	4,8	14,8	17,3	98,4
148	3,8	15,2	16,8	101,2
191	8,7	15,5	16,2	103,5
274	8,2	15,5	16	104,1
370	9,7	16	18	107
432	14,7	18,1	20,2	107,4
445	18,7	13	15,8	108,5
367	19,8	15,8	18,2	108,3
367	10,6	16,9	16,8	109,2
321	8,6	16,3	17	110,1
307	6,5	16,1	18,3	110,7
331	12,6	15,4	16,4	110,3
345	6,5	15,7	16,2	111,8
364	5,8	16	17,7	112,3
384	5,7	15,1	16,2	112,9

Атрибут – свойство, характеризующее объект: цвет глаз человека, температура воды и т. д. Атрибут также называют полем таблицы, измерением.

При анализе данных, как правило, нет возможности рассмотреть всю интересующую нас совокупность объектов. Изучение очень больших объемов данных – дорогостоящий процесс, требующий больших временных затрат, к тому же неизбежно приводящий к ошибкам, связанным с человеческим фактором. Вполне достаточно рассмотреть некоторую часть всей совокупности, т. е. выборку, и на ее основе получить интересующую нас информацию. Однако размер выборки зависит от разнообразия объектов, представленных в генеральной совокупности. В выборке должны быть представлены различные комбинации и элементы генеральной совокупности.

Измерение – процесс присвоения чисел характеристикам изучаемых объектов согласно определенному правилу. В процессе подготовки данных измеряется не сам объект, а его характеристики. Шкала – правило, в соответствии с которым объектам присваиваются числа.

Переменные могут быть числовыми данными либо символьными, числовые данные, в свою очередь, – дискретными и непрерывными.

Дискретные данные служат значениями признака, общее число которых конечно либо бесконечно, но может быть подсчитано при помощи натуральных чисел от одного до бесконечности. Пример дискретных данных: продолжительность маршрута троллейбуса (количество вариантов продолжительности конечно): 10, 15, 25 мин.

Непрерывные данные – данные, значения которых могут принимать какое угодно значение в некотором интервале. Измерение непрерывных данных предполагает большую точность. Пример непрерывных данных: температура, высота, вес, длина и т. д.

Наиболее часто встречаются данные, состоящие из записей, например: табличные, матричные и документальные данные, транзакционные, или операционные.

Табличные данные – данные, состоящие из записей, каждая из которых состоит из фиксированного набора атрибутов.

Транзакционные данные представляют собой особый тип данных, где каждая запись, являющаяся транзакцией, включает набор значений.

Большинство Business Intelligence (BI)-инструментов, представленных на рынке, использует слой метаданных, или репозиторий.

Метаданные (Metadata) – это данные о данных. В состав метаданных могут входить каталоги, справочники, реестры. Метаданные содержат сведения о составе данных, содержании, статусе, происхождении, местонахождении, качестве, форматах и формах представления, условиях доступа, приобретения и использования, авторских, имущественных и смежных с ними правах на данные и др. Метаданные, применяемые при управлении хранилищем, содержат информацию, необходимую для его настройки и использования.

Метаданные хранилища обычно размещаются в репозитории. Это обеспечивает возможность их использования в различных прикладных программах.

1.4. Задачи Data Mining

Задачи подразделяются по типам производимой информации, это наиболее общая классификация задач Data Mining:

- классификация;
- прогнозирование.

В результате решения задачи классификации обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных – классы; по этим признакам новый объект можно отнести к тому или иному классу. Для решения задачи классификации могут использоваться методы:

- ближайшего соседа (Nearest Neighbor);
- k-ближайшего соседа (k-Nearest Neighbor);
- байесовские сети (Bayesian Networks);
- индукция деревьев решений;
- нейронные сети (neural networks).

Кластеризация является логическим продолжением идеи классификации. Это задача более сложная, особенность кластеризации заключается в том, что классы объектов изначально не predetermined. Результатом кластеризации считается разбиение объектов на группы. В результате решения задачи прогнозирования на основе особенностей исторических данных оцениваются пропущенные или будущие значения целевых численных показателей. Для решения таких задач широко применяются методы математической статистики, нейронные сети и др.

В ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных. Отличие ассоциации от предыдущих задач в том, что поиск закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно.

В результате визуализации создается графический образ анализируемых данных. Для решения задачи визуализации используются графические методы, показывающие наличие закономерностей в данных. Пример методов визуализации – представление данных в 2D- и 3D-измерениях.

Согласно классификации по стратегиям задачи Data Mining подразделяются на следующие группы:

- обучение с учителем;
- обучение без учителя.

Категория «обучение с учителем» представлена следующими задачами: классификация, оценка, прогнозирование. Категория «обучение без учителя» представлена задачей кластеризации.

Задачи Data Mining в зависимости от используемых моделей могут быть дескриптивными и прогнозирующими. В соответствии с этой классификацией задачи Data Mining представлены группами описательных и прогнозирующих задач. В результате решения описательных задач аналитик получает шаблоны, описывающие данные, которые поддаются интерпретации. Эти задачи описывают общую концепцию анализируемых данных, определяют информативные, итоговые, отличительные особенности данных. Концепция описательных задач подразумевает характеристику и сравнение наборов данных. Характеристика набора данных обеспечивает краткое и сжатое описание некоторого набора данных. Сравнение обеспечивает сравнительное описание двух или более наборов данных.

Прогнозирующие задачи основываются на анализе данных, создании модели, предсказании тенденций или свойств новых или неизвестных данных.

1.5. Основы анализа данных

Описательная статистика, включающая технологии сбора и суммирования количественных данных, используется для превращения массы цифровых данных в форму, удобную для восприятия и обсуждения. Цель описательной статистики – обобщить первичные результаты, полученные после наблюдений и экспериментов. В состав описательной статистики входят такие характеристики, как среднее, стандартная ошибка, медиана, мода, стандартное отклонение, дисперсия выборки, эксцесс, асимметричность, интервал, минимум, максимум, сумма, счет.

Корреляционный анализ применяется для количественной оценки взаимосвязи двух наборов данных, представленных в безразмерном виде, и дает возможность установить, ассоциированы ли наборы данных по величине. Коэффициент корреляции r используется для

определения наличия взаимосвязи между двумя свойствами. Связь между признаками (по шкале Чеддока) может быть сильной, средней и слабой; тесноту связи определяют по величине коэффициента корреляции (табл. 1.2).

Таблица 1.2

Связь между признаками

Величина коэффициента корреляции r	0,1 – 0,3	0,3 – 0,5	0,5 – 0,7	0,7 – 0,9	0,9 – 1
Характеристика силы связи	Слабая	Умеренная	Заметная	Высокая	Весьма высокая

Любая зависимость между переменными обладает двумя важными свойствами: величиной и надежностью. Чем сильнее зависимость между двумя переменными, тем больше величина зависимости и тем легче предсказать значение одной переменной по значению другой переменной. Величину зависимости легче измерить, чем надежность. Надежность зависимости не менее важна, чем ее величина. Это свойство связано с представительностью исследуемой выборки.

Надежность зависимости характеризует вероятность, что эта зависимость будет снова найдена на других данных. С ростом величины зависимости переменных ее надежность обычно возрастает.

Основная особенность регрессионного анализа: при его помощи можно получить конкретные сведения о том, какую форму и характер имеет зависимость между исследуемыми переменными.

Последовательность этапов регрессионного анализа:

1. Формулировка задачи. На этом этапе формируются предварительные гипотезы о зависимости исследуемых явлений.
2. Определение зависимых и независимых (объясняющих) переменных.
3. Сбор статистических данных. Данные должны быть собраны для каждой из переменных, включенных в регрессионную модель.
4. Формулировка гипотезы о форме связи (простая или множественная, линейная или нелинейная).
5. Определение функции регрессии (заключается в расчете численных значений параметров уравнения регрессии).
6. Оценка точности регрессионного анализа.

7. Интерпретация полученных результатов. Полученные результаты регрессионного анализа сравниваются с предварительными гипотезами. Оцениваются корректность и правдоподобие полученных результатов.

8. Предсказание неизвестных значений зависимой переменной.

При помощи регрессионного анализа возможно решение задачи прогнозирования и классификации. Прогнозные значения вычисляются путем подстановки в уравнение регрессии параметров значений объясняющих переменных. Задачи классификации решаются таким образом: линия регрессии делит все множество объектов на два класса, и та часть множества, где значение функции больше нуля, принадлежит к одному классу, а та часть, где оно меньше нуля, – к другому.

Основные задачи регрессионного анализа: установление формы зависимости, определение функции регрессии, оценка неизвестных значений зависимой переменной.

Характер и форма зависимости между переменными могут образовывать следующие разновидности регрессии: положительная линейная регрессия (выражается в равномерном росте функции); положительная равноускоренно возрастающая регрессия; положительная равнозамедленно возрастающая регрессия; отрицательная линейная регрессия (выражается в равномерном падении функции); отрицательная равноускоренно убывающая регрессия; отрицательная равнозамедленно убывающая регрессия.

Задача определения функции регрессии сводится к выяснению действия на зависимую переменную главных факторов или причин при неизменных прочих равных условиях и при условии исключения воздействия на зависимую переменную случайных элементов. Функция регрессии определяется в виде математического уравнения того или иного типа.

Решение задачи оценки неизвестных значений зависимой переменной сводится к решению задачи одного из типов:

– оценка значений зависимой переменной внутри рассматриваемого интервала исходных данных, т. е. пропущенных значений; при этом решается задача интерполяции;

– оценка будущих значений зависимой переменной, т. е. нахождение значений вне заданного интервала исходных данных; при этом решается задача экстраполяции.

Обе задачи решаются подстановкой в уравнение регрессии найденных оценок параметров значений независимых переменных. Результат решения уравнения представляет собой оценку значения целевой (зависимой) переменной.

Процесс классификации состоит из двух этапов: конструирования модели и ее использования. Использование модели заключается в классификации новых или неизвестных значений. Известные значения из тестового примера сравниваются с результатами использования полученной модели. Уровень точности – процент правильно классифицированных примеров в тестовом множестве. Если точность модели допустима, возможно использование модели для классификационных примеров, класс которых неизвестен. Оценка точности классификации может проводиться при помощи кросс-проверки. Кросс-проверка (Cross-validation) – это процедура оценки точности классификации на данных из тестового множества, которое также называют кросс-проверочным множеством. Точность классификации тестового множества сравнивается с точностью классификации обучающего множества. Если классификация тестового множества дает приблизительно такие же результаты по точности, как и классификация обучающего множества, считается, что данная модель прошла кросс-проверку.

Если же выборка имеет малые объемы, рекомендуется применять специальные методы, при использовании которых обучающая и тестовая выборки могут частично пересекаться. Для классификации используются различные методы:

- классификация с помощью деревьев решений;
- байесовская (наивная) классификация;
- классификация методом опорных векторов;
- классификация при помощи метода ближайшего соседа;
- статистические методы, в частности линейная регрессия;
- классификация при помощи искусственных нейронных сетей;
- классификация при помощи генетических алгоритмов.

Прогнозирование – установление функциональной зависимости между зависимыми и независимыми переменными. Целью прогнозирования является предсказание будущих событий. Решение задачи прогнозирования требует некоторой обучающей выборки данных. Задачи прогнозирования решаются в самых разнообразных областях че-

ловеческой деятельности, таких как наука, экономика, производство и множество других сфер. Развитие методов прогнозирования непосредственно связано с развитием информационных технологий, в частности, с ростом объемов хранимых данных и усложнением методов и алгоритмов прогнозирования, реализованных в инструментах Data Mining.

Основой для прогнозирования служит историческая информация, хранящаяся в базе данных в виде временных рядов. Временной ряд – последовательность наблюдаемых значений какого-либо признака, упорядоченных в неслучайные моменты времени.

В процессе определения структуры и закономерностей временного ряда предполагается обнаружение шумов и выбросов, тренда, сезонной и циклической компоненты. Тренд является систематической компонентой временного ряда, которая может изменяться во времени. Трендом называют неслучайную функцию, которая формируется под действием общих или долговременных тенденций, влияющих на временной ряд. Примером тенденции может выступать, например, фактор роста исследуемого рынка.

Сезонная компонента временного ряда является периодически повторяющейся составляющей временного ряда. Свойство сезонности означает, что через примерно равные промежутки времени форма кривой, которая описывает поведение зависимой переменной, повторяет свои характерные очертания. Определение наличия компоненты сезонности необходимо для того, чтобы входная информация обладала свойством репрезентативности.

Рекомендации по выбору параметров прогнозирования: при выборе параметров следует учитывать, что горизонт прогнозирования должен быть не меньше, чем время, необходимое для реализации решения, принятого на основе этого прогноза. Точность прогноза, требуемая для решения конкретной задачи, оказывает большое влияние на прогнозирующую систему. Ошибка прогноза зависит от используемой системы прогноза.

Наиболее распространенные виды ошибок:

1. Средняя ошибка (СО) вычисляется простым усреднением ошибок на каждом шаге. Недостаток этого вида ошибки – положительные и отрицательные ошибки аннулируют друг друга.

2. Средняя абсолютная ошибка (САО) рассчитывается как среднее абсолютных ошибок. Если она равна нулю, то мы имеем совершенный прогноз. В сравнении со средней квадратической ошибкой эта мера «не придает слишком большого значения» выбросам.

3. Сумма квадратов ошибок (SSE), среднеквадратическая ошибка. Она вычисляется как сумма (или среднее) квадратов ошибок. Это наиболее часто используемая оценка точности прогноза.

4. Относительная ошибка (ОО). Предыдущие меры использовали действительные значения ошибок. Относительная ошибка выражает качество подгонки в терминах относительных ошибок.

Прогноз может быть краткосрочным, среднесрочным и долгосрочным. Краткосрочный прогноз представляет собой прогноз на несколько шагов вперед, т. е. осуществляется построение прогноза не более чем на 3 % от объема наблюдений или на 1 – 3 шага вперед. Среднесрочный прогноз – это прогноз на 3 – 5 % от объема наблюдений, но не более 7 – 12 шагов вперед; также под этим типом прогноза понимают прогноз на один или половину сезонного цикла.

Для построения краткосрочных и среднесрочных прогнозов вполне подходят статистические методы. Долгосрочный прогноз – это прогноз более чем на 5 % от объема наблюдений. При построении данного типа прогнозов статистические методы практически не используются. Доступность данных, на основе которых будет осуществляться прогнозирование, – важный фактор построения прогнозной модели. Для выполнения качественного прогноза данные должны быть представительными, точными и достоверными. Среди распространенных методов Data Mining, используемых для прогнозирования, отметим нейронные сети, дерево решений и линейную регрессию.

Дерево решений (Decision Trees) – это метод, позволяющий предсказывать принадлежность наблюдений или объектов к тому или иному классу категориальной зависимой переменной в зависимости от соответствующих значений одной или нескольких предикторных переменных.

Цель построения дерева классификации заключается в предсказании (или объяснении) значений категориальной зависимой переменной, и поэтому используемые методы тесно связаны с более традиционными методами дискриминантного анализа, кластерного анализа, непараметрической статистики. Широкая сфера применимости дерева

классификации делает его весьма привлекательным инструментом анализа данных, но не следует поэтому полагать, что его рекомендуется использовать вместо традиционных методов статистики.

Если зависимая, т. е. целевая переменная принимает дискретные значения, то при помощи метода дерева решений выполняется задача классификации. Если же зависимая переменная принимает непрерывные значения, то дерево решений устанавливает зависимость этой переменной от независимых переменных, т. е. решает задачу численного прогнозирования. На сегодняшний день существует значительное число алгоритмов, реализующих метод дерева решений CART, C4.5, NewId, ITrule, CHAID, CN2 и т. д. Он способен решать такие задачи Data Mining, в которых отсутствует априорная информация о виде зависимости между исследуемыми данными.

Задача кластеризации сходна с задачей классификации, является ее логическим продолжением, но ее отличие в том, что классы изучаемого набора данных заранее не predetermined. Синонимами термина «кластеризация» можно считать термины «автоматическая классификация», «обучение без учителя» и «таксономия». Кластеризация предназначена для разбиения совокупности объектов на однородные группы (кластеры, или классы). Если данные выборки представить как точки в признаковом пространстве, то задача кластеризации сводится к определению «сгущений точек».

Цель кластеризации – поиск существующих структур. Кластеризация является описательной процедурой, она не делает никаких статистических выводов, но дает возможность провести разведочный анализ и изучить структуру данных.

Кластер можно охарактеризовать как группу объектов, имеющих общие свойства. Характеристиками кластера можно назвать два признака:

- внутреннюю однородность;
- внешнюю изолированность.

Существует большое количество подходов к кластеризации. Следует отметить, что в результате применения различных методов кластерного анализа могут быть получены кластеры различной формы, неодинаковые результаты, что является особенностью работы того или иного алгоритма. Однако создание схожих кластеров различными методами указывает на правильность кластеризации.

Оценка качества кластеризации может быть проведена на основе следующих процедур:

- ручной проверки;
- установления контрольных точек и проверки на полученных кластерах;
- определения стабильности кластеризации путем добавления в модель новых переменных;
- создания и сравнения кластеров с использованием различных методов.

В отличие от задач классификации кластерный анализ не требует априорных предположений о наборе данных, не накладывает ограничения на представление исследуемых объектов, позволяет анализировать показатели различных типов данных (интервальных данных, частот, бинарных данных). При этом необходимо помнить, что переменные должны измеряться в сравнимых шкалах. Кластерный анализ позволяет сокращать размерность данных, делать ее наглядной и применяется к совокупностям временных рядов, здесь могут выделяться периоды схожести некоторых показателей и определяться группы временных рядов со схожей динамикой. Задачи кластерного анализа можно объединить в следующие группы:

1. Разработка типологии или классификации.
2. Исследование полезных концептуальных схем группирования объектов.
3. Представление гипотез на основе исследования данных.
4. Проверка гипотез или исследований для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Кластер имеет следующие математические характеристики: центр, радиус, среднеквадратическое отклонение, размер кластера. Размер кластера может быть определен либо по радиусу кластера, либо по среднеквадратичному отклонению объектов для этого кластера.

Наряду со стандартизацией переменных существует вариант придания каждой из них определенного коэффициента важности, или веса, который бы отражал значимость соответствующей переменной. В качестве весов могут выступать экспертные оценки, полученные в ходе опроса экспертов – специалистов предметной области.

Методы кластерного анализа можно разделить на две группы: иерархические и неиерархические.

Иерархические алгоритмы связаны с построением дендрограмм (от греческого dendron – «дерево»), которые являются результатом иерархического кластерного анализа. Дендрограмма описывает близость отдельных точек и кластеров друг к другу, в графическом виде представляет последовательность объединения (разделения) кластеров (рис. 1.2).

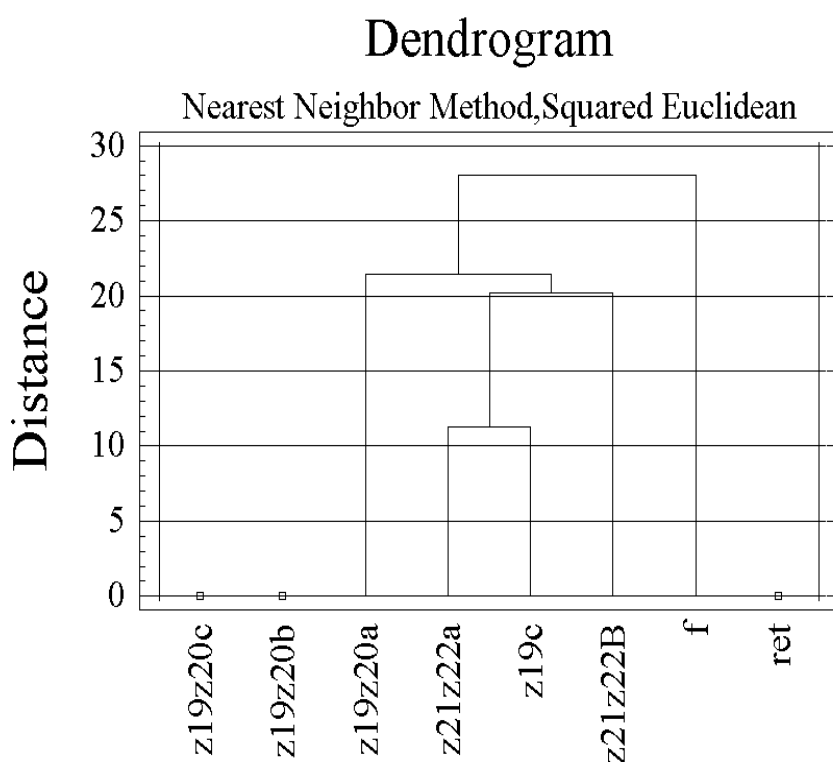


Рис. 1.2. Дендрограмма температур в камере отжига

Существует проблема определения числа кластеров. Иногда можно априорно задать это число. Однако в большинстве случаев число кластеров определяется в процессе агломерации/разделения множества объектов.

Факторный анализ – это метод, применяемый для изучения взаимосвязей между значениями переменных. Вообще факторный анализ преследует две цели: сокращение числа переменных и классификацию переменных – определение структуры взаимосвязей между переменными. Соответственно факторный анализ может использоваться

для решения задач сокращения размерности данных или решения задач классификации. Критерии, или главные факторы, выделенные в результате факторного анализа, содержат в сжатом виде информацию о существующих связях между переменными. Эта информация позволяет получить лучшие результаты кластеризации и лучше объяснить семантику кластеров. Самим факторам может быть сообщен определенный смысл. При помощи факторного анализа большое число переменных сводится к меньшему числу независимых влияющих величин, которые называются факторами. Фактор в «сжатом» виде содержит информацию о нескольких переменных. В один фактор объединяются переменные, которые сильно коррелируют между собой. В результате факторного анализа отыскиваются такие комплексные факторы, которые как можно более полно объясняют связи между рассматриваемыми переменными.

1.6. Задача визуализации

Визуализация – это инструментарий, который позволяет увидеть конечный результат вычислений, организовать управление вычислительным процессом и даже вернуться назад к исходным данным, чтобы определить наиболее рациональное направление дальнейшего движения. В результате использования визуализации создается графический образ данных. Применение визуализации помогает в процессе анализа данных увидеть аномалии, структуры, тренды. При рассмотрении задачи прогнозирования мы использовали графическое представление временного ряда и увидели, что в нем присутствует сезонная компонента. Главное преимущество визуализации – практически полное отсутствие необходимости в специальной подготовке пользователя. С возрастанием количества накапливаемых данных становится все сложнее интерпретировать полученные результаты.

К способам визуального, или графического, представления данных относят графики, диаграммы, таблицы, отчеты, списки, структурные схемы, карты и т. д. Визуализация традиционно рассматривалась как вспомогательное средство при анализе данных, однако сейчас все больше исследований говорят о ее самостоятельной роли.

Существует такой распространенный и наиболее простой способ представления модели, как «черный ящик». В этом случае пользова-

тель не понимает поведения той модели, которой пользуется. Однако несмотря на непонимание, он получает результат – выявленные закономерности. Классическим примером такой модели служит модель нейронной сети.

Примерами средств визуализации, при помощи которых можно оценить качество модели, являются диаграмма рассеивания, таблица сопряженности, график изменения величины ошибки.

Диаграмма рассеивания представляет собой график отклонения значений, прогнозируемых при помощи модели, от реальных. Эти диаграммы используют для непрерывных величин. Визуальная оценка качества построенной модели возможна только по окончании процесса построения модели.

Таблица сопряженности используется для оценки результатов классификации. Такие таблицы применяются для различных методов классификации. Оценка качества построенной модели возможна только по окончании процесса построения модели.

График изменения величины ошибки демонстрирует изменение величины ошибки в процессе работы модели. Например, в процессе работы нейронных сетей пользователь может наблюдать за изменением ошибки на обучающем и тестовом множествах и остановить обучение для недопущения «переобучения» сети. Здесь оценка качества модели и его изменения может определяться непосредственно в процессе построения модели.

Методы визуализации в зависимости от количества используемых измерений принято делить на две группы:

- представление данных в одном, двух и трех измерениях;
- представление данных в четырех и более измерениях.

При использовании двух- и трехмерного представления информации пользователь имеет возможность увидеть закономерности набора данных:

- его кластерную структуру и распределение объектов на классы (например, на диаграмме рассеивания);
- топологические особенности;
- наличие трендов;
- информацию о взаимном расположении данных;
- существование других зависимостей, присущих исследуемому набору данных.

Если набор данных имеет более трех измерений, то возможны такие варианты:

- использование многомерных методов представления информации;
- снижение размерности до одно-, двух- или трехмерного представления.

1.7. Основные этапы интеллектуального анализа

В общем случае процесс интеллектуального анализа и обработки данных состоит из следующих шести этапов: отбор данных, очистка, обогащение, кодирование, извлечение знаний и сообщение.

Отбор данных. Как правило, для решения конкретной задачи нужны не все данные из хранилища данных. Сначала необходимо выбрать то их подмножество, которое будет подвергнуто анализу. При этом, возможно, потребуется объединить несколько таблиц, а полученные записи отфильтровать.

Очистка. Существует несколько типов очистки данных (удаление дублирующих записей, исправление типографских ошибок, добавление отсутствующей информации и т. д.). Некоторые из них могут выполняться заранее, в то время как другие вызываются только после обнаружения загрязнения на этапах кодирования или обнаружения. Очень важным элементом очистки следует назвать устранение дублирования записей.

Извлечение знаний. Данный этап является ядром процесса интеллектуального анализа и обработки знаний. Фактически в технологии обнаружения знаний необходимо различать четыре различных типа знания, которые могут быть извлечены из данных:

1. Поверхностное знание. Это информация, которая может быть легко найдена из баз данных, использующих инструментальное средство запроса типа структурированного языка запросов (SQL).

2. Многомерное знание. Это информация, которая может быть проанализирована при использовании интерактивных аналитических инструментальных средств обработки OLAP с помощью инструментальных средств.

3. Скрытое знание. Это информация, которая может быть найдена относительно легко с помощью алгоритмов распознавания образ-

цов или машинного обучения. Для нахождения этих образцов также можно было бы использовать средства SQL, но это потребует много времени.

4. Глубокое знание. Это информация, которая хранится в базе данных, но может быть обнаружена только в том случае, если имеется ключ, который сообщит нам, где смотреть. Почти невозможно декодировать сообщение, которое зашифровано, т. е. если нет ключа, который указывает, что искать.

Сообщение. Сообщение о результатах процесса обнаружения знаний может принимать много форм. В общем случае можно использовать любой редактор сообщений или графическое инструментальное средство, чтобы сделать доступными результаты процесса.

1.8. Инструментальные средства анализа данных

Инструменты Data Mining во многих случаях рассматриваются как составная часть BI-платформ, в состав которых также входят средства построения хранилищ и витрин данных, средства обработки неожиданных запросов (ad-hoc query), средства отчетности (reporting), а также инструменты OLAP.

К категории наборов инструментов относятся универсальные средства, включающие методы классификации, кластеризации и предварительной подготовки данных. К этой группе относятся следующие известные коммерческие инструменты:

- IBM Intelligent Miner for Data; инструмент предлагает последние Data Mining-методы, поддерживает полный Data Mining-процесс – от подготовки данных до презентации результатов; поддерживает языки XML и PMML;

- Oracle Data Mining; инструмент обеспечивает GUI, PL/SQL-интерфейсы, Java-интерфейс; используемые методы: байесовская классификация, алгоритмы поиска ассоциативных правил, кластерные методы, SVM и др.;

- SAS Enterprise Miner; интегрированный набор, который обеспечивает дружественный GUI; поддерживается методология SEMMA;

- SPSS – один из наиболее популярных инструментов, поддерживается множество методов Data Mining;

– Statistica Data Miner; инструмент обеспечивает всесторонний, интегрированный статистический анализ данных, имеет мощные графические возможности, управление базами данных, а также приложение разработки систем;

– Polyanalyst – набор, обеспечивающий всесторонний Data Mining.

Вторая группа задач представлена инструментами, реализующими следующие решения:

– инструментарий для поиска ассоциативных правил;

– агенты;

– оценивание, регрессии и прогнозирование;

– анализ связей;

– последовательные шаблоны и временные ряды;

– инструменты BI (Business Intelligence), Database and OLAP software;

– инструменты преобразования и очистки данных;

– библиотеки, компоненты и инструментальные наборы для разработчиков создания встроенных приложений Data Mining;

– Web Mining: анализ поведения сайтов, XML mining;

– поиск на Web;

– Audio and Video Mining.

Контрольные вопросы

1. Сформулируйте понятие интеллектуального анализа данных.
2. В чем отличие Data Mining от других методов анализа данных?
3. Какие бывают виды данных и их атрибуты?
4. Раскройте задачи Data Mining.
5. Назовите цель описательной статистики и ее состав.
6. В чем заключаются особенности корреляционного анализа?
7. В чем заключаются особенности регрессионного анализа?
8. Как выполняется оценка неизвестных значений зависимой переменной по уравнению регрессии?
9. Охарактеризуйте этапы процесса классификации объектов.
10. Назовите методы классификации.
11. В чем сущность прогнозирования?
12. Какие виды ошибок прогнозирования наиболее распространены?

13. Какие методы используются для прогнозирования?
14. В чем суть метода дерева решений для решения задач классификации?
15. В чем цель кластеризации?
16. В чем заключаются цели факторного анализа?
17. Что понимают под задачей визуализации?
18. Приведите пример модели «черный ящик».
19. Охарактеризуйте этапы процесса интеллектуального анализа и обработки данных.
20. Назовите инструментальные средства анализа данных.

2. СЛУЧАЙНЫЕ СОБЫТИЯ

2.1. Испытание. Поле событий. Операции над событиями

Испытанием называется осуществление какого-нибудь определенного комплекса условий, который может быть воспроизведен сколь угодно большое число раз [2]. Результат испытания называется событием. Некоторые события происходят неизбежно в результате каждого испытания. Они называются достоверными, а другие не могут происходить и называются невозможными. В результате испытаний в зависимости от случайных обстоятельств может произойти то или иное событие из множества событий, возможных при данных испытаниях. Такое множество называется *полем событий*, связанным с испытанием, а события этого поля называются *случайными*.

Поле может содержать равновозможные события: E_1, E_2, \dots, E_n . Эти события называются элементарными исходами испытаний. Каждому возможному событию A_i поля событий отвечает некоторая часть или подмножество элементарных исходов, из которых как бы составлено A_i .

События могут быть *взаимобусловленными*. В этом случае говорят, что событие A влечет за собой событие B , если при наступлении A неизбежно наступает B : $A \subset B$.

Если $A \subset B$ и одновременно $B \subset A$, то события A и B называют *эквивалентными*, что обозначается знаком равенства $A = B$.

Можно сказать, что каждое событие поля представляет логическую сумму некоторых событий из множества (E_1, E_2, \dots, E_n) . Так, со-

бытие A (7, 10, 12) можно записать так: $A = E_7 + E_{10} + E_{12}$, здесь знак плюс заменяет союз «или».

Сумма $S = A_1 + A_2 + \dots + A_k$ представляет событие, заключающееся в появлении A_1 или A_2 или...или A_k , или некоторых из них вместе.

Пример. События (1, 2, 3) и (3, 4, 5) совместимы: они наступают вместе в тех испытаниях, в которых исход имеет номер 3.

Сумме событий отвечает подмножество элементов, полученных объединением исходов. Так, сумма событий (1, 2, 3) + (1, 2) будет равна (1, 2, 3). Каждый элемент входит в сумму один раз.

Два события поля A и A^c называются *противоположными* (взаимно дополнительными), если они несовместимы и в сумме составляют достоверное событие. Так, два события «появление отказа в ЭВМ» и «отсутствие отказа в ЭВМ» в течение рабочего времени противоположны. По определению

$$A + A^c = U.$$

Таким образом, достоверно, что наступит A или не A^c .

Невозможное событие V противоположно достоверному, т. е. $U + V = U$.

Под произведением событий A_1, A_2, \dots, A_k нашего поля понимают одновременное или совместное наступление их всех.

Произведение несовместимых событий (противоположных) есть невозможное событие.

Пример. Произведение событий $A(2, 3, 4, 5, 6)$ и $B(1, 2, 4, 7, 8)$ есть событие $C = AB = C(2, 4)$, так как A и B наступают вместе тогда, когда наступит событие (2) или (4).

2.2. Операции над событиями

События A и B называются независимыми, если вероятность осуществления одного из них не зависит от вероятности осуществления другого

$$P(A) = P(A/B) = P(A/B^c).$$

Вероятность произведений независимых событий равна произведению их вероятностей

$$P(AB) = P(A)P(B).$$

Произведение вероятностей зависимых событий

$$P(AB) = P(A)P(B/A).$$

Сумма несовместимых событий

$$P(A + B) = P(A) + P(B).$$

Вероятность противоположного события

$$P(A^*) = 1 - P(A).$$

Для совместимых событий формула сложения имеет вид

$$P(A + B) = P(A) + P(B) - P(AB).$$

Операции над событиями можно графически интерпретировать (рис. 2.1).

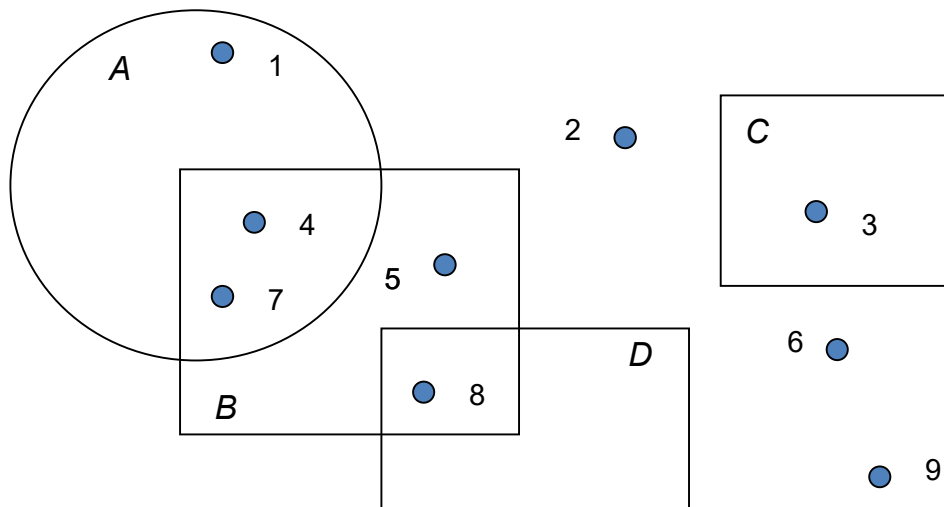


Рис. 2.1. Графическая интерпретация поля события

На рисунке отражены четыре события $A(1, 4, 7)$, $B(4, 5, 7, 8)$, $C(3)$, $D(8)$.

Несовместимыми являются события A , C , D , совместимыми – события A и B .

Операции над событиями:

$$A \cdot B = (4, 7); A + B = (1, 4, 5, 7, 8); B + D = B(4, 5, 7, 8); B \cdot D = D(8).$$

2.3. Частость и вероятность

Рассмотрим серию из N испытаний, произведенных в одних и тех же условиях. Допустим, что нас интересует определенное событие A поля испытаний. Если в нашей серии испытаний событие A произошло $k_N(A)$ раз, то отношение $k_N(A) / N = W_N(A)$ называется частотой: $0 \leq W_N(A) \leq 1$.

Если событие A невозможно $W_N(A) = 0$, если событие достоверно, то $W_N(A) = 1$. Если событие A невозможно, $A = V$, то в любой серии произведенных испытаний будем иметь $k_N(A) = 0$ и $W_N(A) = 0$. Если событие A достоверно, то $k_N(A) = 1$ и $W_N(A) = 1$.

Отношение $W_N(A)$ представляет случайную величину, его значение зависит от случайных обстоятельств, сопутствующих испытанию. Если N невелико, то частота может сильно изменяться при повторении серии из N испытаний. В обширном классе случаев частота события A обнаруживает устойчивость. Это число меньше единицы, представляет как бы количественную меру возможности наступления события A в испытании, называется его вероятностью.

Пример. Если выборка небольшая, то среди пяти отобранных изделий из партии могут оказаться бракованными 1, 2, 3, 4 и даже 5 изделий с признаком A – брака. Повторяя выборку много раз, будем получать сильно отличающиеся друг от друга частоты $W_N(A)$. С ростом объема N разброс частоты уменьшается, приближаясь к доле признака A во всей обследуемой партии.

Задача теории вероятностей заключается в том, чтобы, зная вероятности некоторых простейших событий, получить путем анализа или вычислений вероятности интересующих нас сложных событий, т. е. иметь возможность предсказывать частоты этих событий при массовом производстве испытаний.

2.4. Основные аксиомы теории вероятностей

Аксиома 1. С каждым событием A данного поля испытаний связывается число $P(A)$, называемое вероятностью и удовлетворяющее условию $0 \leq P(A) \leq 1$.

Аксиома 2. Вероятность достоверного события U поля равна единице, так что $P(U) = 1$.

Аксиома 3. Правило сложения вероятностей несовместимых событий.

Если событие S поля подразделяется на несовместимые события A_1, A_2, \dots, A_m того же поля, т. е. представляет собой сумму этих событий, так что

$$S = A_1 + A_2 + \dots + A_m \text{ и } A_i \cdot A_j = V \text{ при любых } i \text{ и } j (i, j = 1, 2, \dots, m), \text{ то} \\ P(S = A_1 + A_2 + \dots + A_m) = P(A_1) + P(A_2) + \dots + P(A_m) \text{ и } A_i A_j = V.$$

При любых i, j ($i, j = 1, 2, \dots, m$), то вероятность суммы несовместимых событий поля равна сумме их вероятностей.

Наступление события S согласно аксиоме 3 может осуществляться или в виде A_1 , или в виде A_2 , ... в виде A_m .

Аксиома 4 называется правилом сложения вероятностей несовместимых событий.

Если некоторые два события A_1 и A_2 не являются несовместимыми, то

$$P(A_1 + A_2) \leq P(A_1) + P(A_2).$$

Предположим, что рассматриваем такое испытание, при котором элементарные исходы E_1, E_2, \dots, E_n несовместимы и образуют полную группу

$$E_1 + E_2 + \dots + E_n = U.$$

Если событие A составлено из элементарных событий $E_1 + E_2 + \dots + E_n$ так, что $A = E_1 + E_2 + \dots + E_n$, то согласно правилу сложения получим

$$P(A) = P(E_1) + P(E_2) + \dots + P(E_n).$$

Таким образом, вероятность каждого события поля есть сумма вероятностей элементарных событий, составляющих данное событие. В частности,

$$P(U) = P(E_1) + P(E_2) + \dots + P(E_n) = 1.$$

Если событие A влечет за собой событие B , т. е. все элементы события A входят в состав события B , кроме того, в состав B могут входить и другие элементы, тогда вероятность события A не больше вероятности события B $P(A) \leq P(B)$.

Рассмотрим такое испытание, в котором элементарные исходы E_1, E_2, \dots, E_n по самой постановке опыта совершенно равновероятны, например выборка карточки из картотеки, то любое событие можно представить как сумму некоторых событий из этой полной группы. При этих условиях

$$P(E_1) = P(E_2) = P(E_3) = \dots = P(E_n) = 1/n = p.$$

и сумма $\sum_{i=1}^n P(E_i) = np = 1$.

Если событие A можно представить как сумму m составляющих его элементарных событий E_i

$$A = E_{i1} + E_{i2} + \dots + E_{im}, \text{ то}$$

$$P(A) = P(E_{i1}) + P(E_{i2}) + \dots + P(E_{im}) = m \cdot 1/n = m/n.$$

Мы пришли к классическому определению вероятности: вероятность события A есть отношения числа m благоприятствующих этому событию исходов к общему числу n всех возможных элементарных несовместимых и равно возможных исходов испытаний.

Применительно к выборке большого объема S (с возвращением) частота каждого элементарного исхода E_k равна примерно $1/n$, а частота события A будет приближенно равна m/n . Для частоты события A будем иметь

$$W(A) \approx mN / nN = \frac{m}{n} = P(A).$$

Пример. Пусть допуск валика принтера подразделен на четыре равных групповых допуска. В сборочном цехе имеется партия валиков численностью 100 шт., из которых 15 шт. с размерами в пределах B_1 группы, 40 шт. в пределах второй группы B_2 , 30 шт. – B_3 , 15 шт. – B_4 .

Сборщик наугад вынимает из группы один валик, т. е. будет вынут один из 100 валиков. 100 элементарных исходов равновероятны, поэтому вероятность событий будет:

$$P(B_1) = 15/100 = 0,15; \quad P(B_2) = 40/100 = 0,4; \quad P(B_3) = 30/100 = 0,3; \\ P(B_4) = 15/100 = 0,15.$$

Невозможное событие V заключается в том, что вынутая из партии наугад деталь окажется не относящейся по размеру ни к одной группе:

$$\bar{B}_1 \bar{B}_2 \bar{B}_3 \bar{B}_4 = V, \text{ тогда как } B_1 + B_2 + B_3 + B_4 = U.$$

Событие \bar{B}_1 противоположное событию B_1 , заключается в том, что вынутая из партии деталь окажется не принадлежащей по своим размерам к первой группе, т. е. она будет принадлежать к группам B_2 , или B_3 , или B_4 :

$$\bar{B}_1 = B_2 + B_3 + B_4, \quad P(\bar{B}_1) = (40 + 30 + 15)/100 = 85/100 = 0,85 \text{ и т. д.}$$

События $\bar{B}_1 \bar{B}_2$ совместимы, так как вынутая деталь окажется принадлежащей к третьей и четвертой группам, оба они наступят одновременно. Отсюда следует, что событие $B_3 + B_4$ влечет за собой как событие \bar{B}_1 , так и событие \bar{B}_2

$$B_3 + B_4 \subset \bar{B}_1, \quad B_3 + B_4 \subset \bar{B}_2.$$

$$\text{Мы видим, что } P(B_3 + B_4) = 0,3 + 0,15 = 0,45 < \overline{P(\bar{B}_1)} = 0,85.$$

2.5. Случайные события

Теория вероятностей – наука о закономерностях массовых случайных событий, т. е. событий, которые при определенных условиях могут произойти, а могут и не произойти [2]. Случайным событием может стать выход из строя мобильного телефона во время гарантийного периода. Степень возможности осуществления события характеризуется вероятностью события.

Случайность событий можно рассматривать как результат некоторого эксперимента со случайным исходом, поставленного специально. Предположим, что эксперимент можно повторить в одних и тех же условиях неоднократно. Если из серии опытов N событие A произошло M раз, то отношение $W(A) = M/N$ можно назвать относительной частотой события A .

При небольших значениях N частота одного и того же события может колебаться в широких пределах. Однако при большом числе опытов она стабилизируется, приближается к некоторому пределу, называемому вероятностью осуществления рассматриваемого события, $P(A) = W(A) = M/N$.

При $M = 0$ имеем невозможное событие, которое при определенных условиях никогда не произойдет. В реальных условиях имеет место событие, вероятность которого близка к нулю. Такое событие называют практически невозможным.

При $M = N$ имеем достоверное событие, которое обязательно произойдет при заданных условиях. Вероятность его равна единице.

Для любого события A вероятность его лежит в пределах $0 \leq P(A) \leq 1$.

Событие \bar{A} , состоящее в том, что событие A не произойдет, называется противоположным событию A .

Пример. Охрана окружающей среды при производстве стекла связана прежде всего с уменьшением выбросов и сбросов загрязняющих веществ в атмосферу и водоемы. Анализ материалов заболеваемости осуществлялся с использованием статистических показателей: частоты заболеваемости в числе случаев на 100 работающих [3]. Для изучения причинно-следственных зависимостей заболеваемости работников производства от содержания вредных веществ в выбросах и сбросах были использованы данные за 33 месяца работы производства. Построена модель «черного ящика» исследуемой системы (рис. 2.2)

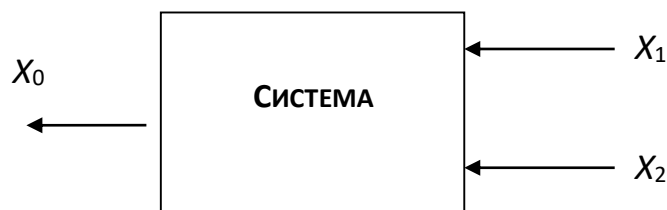


Рис. 2.2. Модель «черного ящика»:
 X_0 – целевой показатель, заболеваемость Z_{ab} ;
 X_1 – факторная переменная, содержание
 железа в сточных водах после отстойников Fe;
 X_2 – содержание взвешенных веществ
 в сточных водах W_{zw}

Один из приемов приведения числовых переменных к дискретной форме состоит в разбиении интервала вариации переменной на квантили Q -интервалы, обладающие тем свойством, что вероятности попадания значения переменной в каждый из них равны. На практике часто выделяют квантили приближенно, пользуясь непосредственно эмпирическими данными (табл. 2.1 для $Q = 2$).

Таблица 2.1

Разбиение интервала вариации переменных на квантили

Интервал	X_0, Z_{ab}	X_1, Fe	X_2, W_{zw}
Квантиль 1	2 – 4	0,175 – 0,49	0,1 – 7,4
Квантиль 2	4,1 – 9	0,491 – 0,95	7,41 – 19,5

После разбивки интервала вариации переменных на квантили каждое значение переменной заменяется номером квантили, которой оно соответствует. В результате получаем отображение непрерывного множества значений переменной на конечное дискретное множество значений. В рассматриваемом примере выбираем два квантиля $Q = 2$ с вероятностью $p = 1/Q = 0,5$ того, что значения переменных принадлежат требуемому интервалу. Отображение непрерывного множества значений переменных на конечное дискретное множество приведено в табл. 2.2.

Таблица 2.2

Отображение непрерывного множества значений переменных на конечное дискретное множество

Переменная		Z_{ab}	
		1	2
Fe X_1	1	8	8
	2	8	9
W_{zw} X_2	1	8	9
	2	8	8

Рассчитаем условные вероятности значений входных переменных X_1 и X_2 для различных значений выходной переменной X_0 (табл. 2.3).

Таблица 2.3

Условные вероятности отображения непрерывного множества на конечное дискретное множество

Параметр		Z_{ab}	
		1	2
Fe X_1	1	0,5	0,47
	2	0,5	0,53
W_{zw} X_2	1	0,5	0,53
	2	0,5	0,47

Вероятности значений выходной переменной определяются путем последовательного использования формулы Байеса для учета информации о состоянии каждой входной переменной. Формула Байеса имеет вид

$$p(A_i / B_{gh}) = \frac{p(A_i)p(B_{gh} / A_i)}{\sum_{j=1}^n p(A_j)p(B_{gh} / A_j)}, \quad (2.1)$$

где $p(A_i / B_{gh})$ – вероятность i -го значения выходной переменной при условии, что имеет место h -е значение входной переменной X_g ; n – число возможных значений выходной переменной; $p(B_{gh} / A_i)$, $p(B_{gh} / A_j)$ – вероятность h -го значения входной переменной X_g при усло-

вии i -го (j -го) значения выходной переменной; $p(A_i) = 16/33 = 0,485$, $p(A_j) = 17/33 = 0,515$ – вероятность i -го (j -го) значения выходной переменной.

Заданы условные вероятности значений входных переменных X_1 и X_2 при условии различных значений выходной переменной X_0 (см. табл. 2.3):

$$\begin{aligned} p(X_1 = 1|X_0 = 1) &= 0,5; \\ p(X_1 = 2|X_0 = 1) &= 0,5; \\ p(X_1 = 1|X_0 = 2) &= 0,47; \\ p(X_1 = 2|X_0 = 2) &= 0,53; \\ p(X_2 = 1|X_0 = 1) &= 0,5; \\ p(X_2 = 2|X_0 = 1) &= 0,5; \\ p(X_2 = 1|X_0 = 2) &= 0,53; \\ p(X_2 = 2|X_0 = 2) &= 0,47. \end{aligned}$$

Положим, что поступила информация о значении второй входной переменной $X_2 = 1$. Согласно формуле (2.1) вероятность события $X_0 = 1|X_2 = 1$ составит

$$p(X_0 = 1 / X_2 = 1) = \frac{0,485 \cdot 0,5}{0,485 \cdot 0,5 + 0,515 \cdot 0,53} = \frac{0,2425}{0,51545} = 0,47.$$

Аналогично можно рассчитать вероятность события $X_0 = 2|X_2 = 1$, которая составит

$$p(X_0 = 2 / X_2 = 1) = \frac{0,515 \cdot 0,53}{0,485 \cdot 0,5 + 0,515 \cdot 0,53} = \frac{0,27295}{0,51545} = 0,53.$$

Если входные переменные независимы, можно вычислить вероятность i -го значения выходной переменной при условии, что известны значения некоторых или всех входных переменных. Например, предположим, что требуется определить вероятность $p(A_i / (B_{gh} \cup B_{qw}))$ события A_i при условии, что $X_g = h$ и $X_q = w$. Для этого можно использовать формулу Байеса в следующей форме:

$$p(A_i / (B_{gh} \cup B_{qw})) = \frac{p(A_i / B_{gh})p(B_{qw} / A_i)}{\sum_{j=1}^n p(A_j / B_{gh})p(B_{qw} / A_j)}, \quad (2.2)$$

где значение $p(A_i / B_{gh})$ ранее определено по формуле (2.1).

В дополнение к имеющейся информации о второй переменной $X_2 = 1$ поступила информация еще о первой: $X_1 = 2$. Согласно формуле (2.2) вероятность события $X_0 = 1 | (X_2 = 1 \cup X_1 = 2)$ равна

$$\begin{aligned} p(X_0 = 1 | (X_2 = 1 \cup X_1 = 2)) &= \\ &= \frac{p(X_0 = 1 | X_2 = 1)p(X_1 = 2 | X_0 = 1)}{p(X_0 = 1 | X_2 = 1)p(X_0 = 1 | X_1 = 2) + p(X_0 = 2 | X_2 = 1)} \times \\ &\times \frac{1}{p(X_1 = 2 | X_0 = 2)} = \frac{0,47 \cdot 0,5}{0,47 \cdot 0,5 + 0,53 \cdot 0,53} = 0,456. \end{aligned}$$

Энтропия переменной X_0 до получения информации о входных переменных составляла $-0,485 \cdot \log_2 0,485 - 0,515 \cdot \log_2 0,515 = 0,999$ бит, после получения первого сигнала $X_2 = 1$ она стала равной $-0,47 \cdot \log_2 0,47 - 0,53 \log_2 \cdot 0,53 = 0,997$ бит, а после второго сигнала $X_1 = 2$ сократилась до $-0,456 \cdot \log_2 0,456 - 0,544 \cdot \log_2 0,544 = 0,994$ бит.

На практике поступление новой информации может не только снижать, но и увеличивать энтропию.

В общем случае для определения вероятности i -го значения выходной переменной формулу Байеса применяют ровно столько раз, сколько имеется известных значений входных переменных.

Контрольные вопросы

1. Что называется испытанием и событием в испытании?
2. Какие бывают события в испытаниях?
3. Что представляет собой сумма событий, в чем она заключается?
4. Что понимается под произведением событий?
5. Приведите пример графической интерпретации операций над событиями.
6. Охарактеризуйте свойства поля событий.
7. Что называется частотой и вероятностью событий?
8. Как влияет объем выборки на частоту?
9. Назовите основные аксиомы теории вероятностей.
10. Раскройте свойства суммы вероятностей.
11. В чем сущность классического определения вероятности?

3. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ. ЗАКОНЫ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН

3.1. Определение случайной величины

Под случайной величиной понимают числовой результат эксперимента со случайным исходом. При оценке надежности ЭВМ случайная величина – это количество отказов за некоторый промежуток времени или время наработки на отказ. В некоторых случаях множество значений случайной величины может быть конечным или счетным (количество отказов, количество дефектных изделий). Такая случайная величина называется дискретной. В других ситуациях случайная величина принимает любое значение из некоторого промежутка – это непрерывная случайная величина.

Случайная величина считается заданной, если известен закон распределения – соотношение, устанавливающее связь между множеством значений случайной величины и их вероятностями. Дискретную случайную величину можно задать в виде табл. 3.1.

Таблица 3.1

Дискретная случайная величина

x_i	x_1	x_2	...	x_n
p_i	p_1	p_2		p_n

Другой способ задания случайной величины – с помощью функции распределения – вероятности того, что случайная величина X окажется меньше некоторого x , $F(x) = P(X < x)$.

Этот способ задания случайной величины универсален и может использоваться и для дискретной, и непрерывной случайной величины.

Для дискретных событий соотношение между возможными значениями случайной величины x_i и их вероятностями p_i называют законом распределения и либо записывают их в виде ряда (таблицы), либо представляют в виде зависимостей $F\{x\}$ (рис. 3.1, а) или $p(x)$ (рис. 3.1, в).

Для непрерывных случайных величин (процессов) закон распределения представляют либо в виде функции распределения (интегральный закон распределения – рис. 3.1, б), либо в виде плотности вероятностей (дифференциальный закон распределения – рис. 3.1, г).

Отметим основные свойства функции распределения: ее значение лежит в промежутке от нуля до единицы $0 \leq F(x) \leq 1$.

При этом $F(-\infty) = 0$. $F(+\infty) = 1$. Функция $F(x)$ монотонно неубывающая: при $x_2 > x_1$ $F(x_2) \geq F(x_1)$.

Вероятность попадания случайной величины в полуинтервал $[x_1, x_2)$ определяется по формуле $p(x_1 \leq X < x_2) = F(x_2) - F(x_1)$.

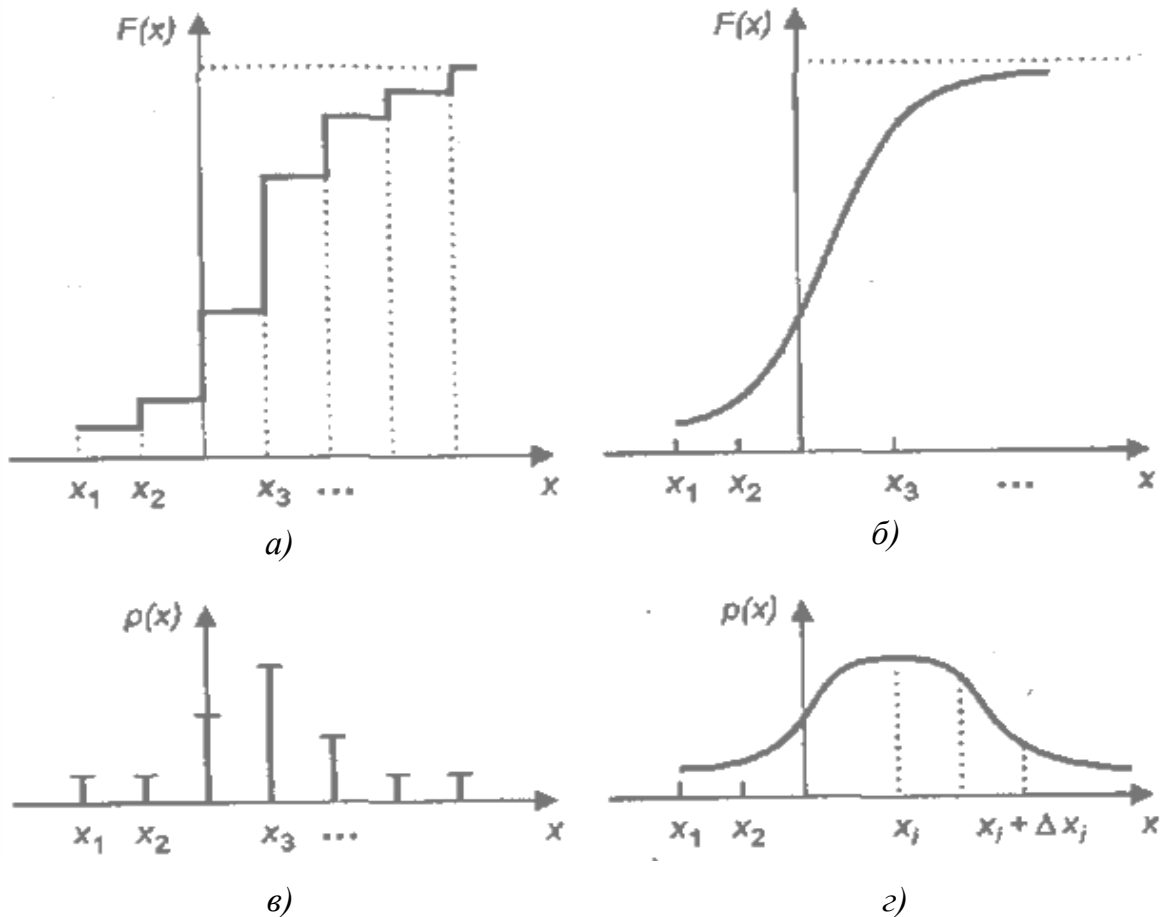


Рис. 3.1. Способы задания случайных величин

Функция распределения непрерывной случайной величины обладает этими же свойствами, кроме того, она непрерывна и дифференцируема.

Часто непрерывная случайная величина задается иначе, с помощью производной от функции распределения, которая называется плотностью распределения вероятности, $f(x) = F'(x)$.

Закон распределения служит удобной формой статистического отображения системы. Однако получение закона (даже одномерного)

или определение изменений этого закона при прохождении через какие-либо устройства или среды представляет собой трудную, часто невыполнимую задачу. Поэтому в ряде случаев пользуются не распределением, а его характеристиками – начальными и центральными моментами.

Наибольшее применение получили:

Первый начальный момент – математическое ожидание, или среднее значение случайной величины:

– для дискретных величин $m_x = \sum x_i p_i$;

– для непрерывных величин $m_x = \int_{-\infty}^{+\infty} x f(x) dx$.

Второй центральный момент – дисперсия случайной величины:

– для дискретных величин $D_x = \sum (x_i - m_x)^2 p_i$;

– для непрерывных величин $D_x = \int_{-\infty}^{+\infty} (x - m_x)^2 f(x) dx$.

На практике иногда используется не оценка дисперсии s^2 , а среднее квадратическое отклонение s .

Свойства математического ожидания:

– математическое ожидание постоянной равно постоянной;

– постоянный множитель можно вынести из-под знака математического ожидания;

– математическое ожидание суммы случайных величин равно сумме математического ожидания слагаемых $m(C) = C$; $m(CX) = C m(X)$.

Свойства дисперсии:

– дисперсия постоянной равна нулю;

– постоянный множитель можно вынести из-под знака дисперсии, возведя его в квадрат;

– для независимых случайных величин дисперсия суммы равна сумме дисперсий слагаемых $D(C) = 0$; $D(kX) = k^2 D(X)$; $D(X + Y) = D(X) + D(Y)$.

Обобщение понятия дисперсии – центральный момент k -го порядка

$$\mu_k[x] = M[x - m_x]^k.$$

Для симметричного распределения коэффициенты асимметрии и эксцесса равны нулю $a_x = 0$, $e_x = 0$.

Центральные моменты используются для расчета характеристик формы кривой распределения. Коэффициент асимметрии характеризует несимметричность кривой распределения $a_x = \frac{\mu_3[X]}{\sigma_x^3}$.

Коэффициент эксцесса характеризует крутость кривой распределения. Положительный эксцесс имеют распределения более островершинные $e_x = \frac{\mu_4[X]}{\sigma_x^4} - 3$.

3.2. Законы распределения дискретных случайных величин

Из множества распределений случайной величины, имеющих важное значение в задачах управления качеством, рассмотрим из дискретных биномиальное распределение и распределение Пуассона. Из непрерывных – нормальное, экспоненциальное, равномерное, а также распределения, используемые в статистических расчетах: χ -квадрат, Фишера, Стьюдента.

Практическое применение получили в основном одномерные распределения, что связано со сложностью получения статистических закономерностей и доказательства адекватности их применения для конкретных приложений, которое базируется на понятии выборки.

Под выборкой понимается часть изучаемой совокупности явлений, на основе исследования которой получают статистические закономерности, присущие всей совокупности и распространяемые на нее с какой-то вероятностью. Для того чтобы полученные при исследовании выборки закономерности можно было распространить на всю совокупность, выборка должна быть представительной (репрезентативной), т. е. обладать определенными качественными и количественными характеристиками.

Биномиальное распределение

Пусть производится эксперимент, нас интересует, произошло событие A или нет. Случай, в котором событие A произошло, назовем успехом. Вероятность успеха $P(A) = p$. Если событие A не произошло, то его вероятность равна $P(A) = 1 - p$.

Серия испытаний такого типа произошла n раз. Нас интересует вероятность события, состоящего в том, что оно произошло m раз. Решение этой задачи имеет вид (биномиальный закон распределения (рис. 3.2))

$$P(X = m) = C_n^m p^m q^{n-m},$$

где $C_n^m = \frac{n!}{m!(n-m)!}$ – число сочетаний из n элементов по m .

Можно доказать, что математическое ожидание случайной величины X равно $m_X = np$, дисперсии $D_X = npq$.

Коэффициент асимметрии биномиального распределения

$$a_x = \frac{(q-p)}{\sqrt{npq}}.$$

Коэффициент эксцесса

$$e_x = \frac{(1-6p+6p^2)}{\sqrt{npq}}.$$

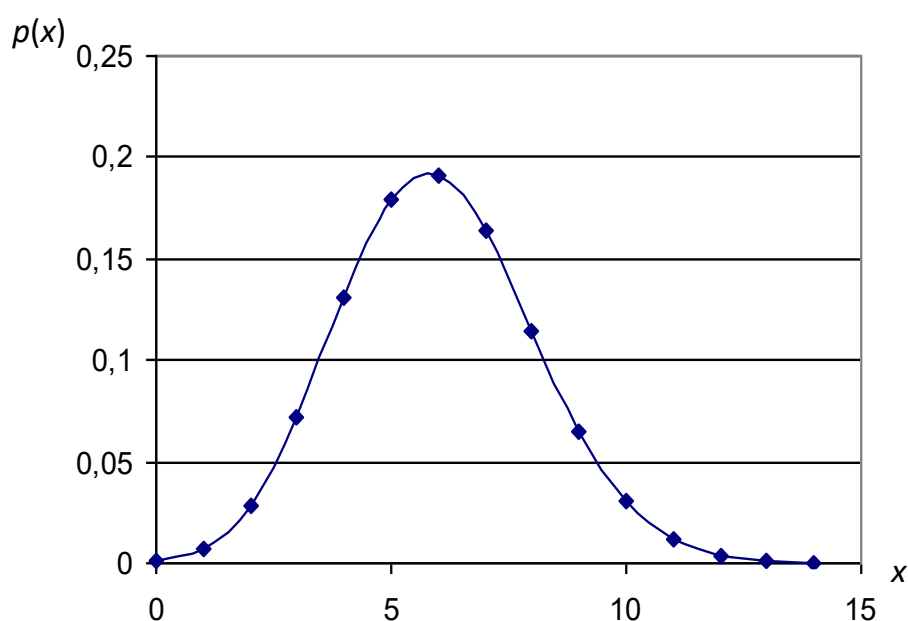


Рис. 3.2. Вероятности биномиального распределения при значении $p = 0,3$, $n = 20$

С ростом n $a_x \rightarrow 0$, $e_x \rightarrow 0$ биномиальный закон приближается к нормальному закону распределения.

Пример. Рассмотрим выборку с возвращением объемом $n = 30$ из большой партии изделий. При соблюдении случайного отбора оно

соответствует схеме Бернулли. Роль вероятности p здесь играет доля дефектных изделий во всей партии. Допустим, что $p = 0,05$. Мы можем при этом предположении рассчитать вероятность обнаружения в выборке того или иного числа x дефектных изделий. Расчетная вероятность обнаружения в выборке m дефектных изделий приведена в табл. 3.2.

Таблица 3.2

Расчетная вероятность обнаружения в выборке m дефектных изделий

m	$P_{30}(x = m)$
0	0,2146
1	0,3389
2	0,2586
4	0,0451
6	0,0027
9	0,000001

Закон распределения Пуассона

Если число испытаний n неограниченно возрастает, а математическое ожидание числа появлений события остается постоянным и равным Λ , то вероятность $p_n(x)$ биномиального распределения при каждом $x = 0, 1, 2, \dots$ стремится к пределу $p_n(x) = \frac{e^{-\lambda} \lambda^x}{x!}$.

Предельное значение образует распределение Пуассона. Для практических целей приближение биномиального распределения к пуассоновскому получается при $n \geq 60$.

Математическое ожидание случайной величины, следующей закону Пуассона, равно параметру этого закона Λ $m(X) = \Lambda$.

Дисперсия этой случайной величины равна тому же параметру Λ : $D(X) = \Lambda$.

Коэффициент асимметрии распределения Пуассона всегда положительная величина $a_x = \frac{1}{\sqrt{\lambda}}$.

Экссесс распределения Пуассона всегда положителен $e_x = \frac{1}{\lambda}$.

Если случайная величина представляет сумму двух независимых случайных величин, следующих каждая закону Пуассона, то сумма также следует закону Пуассона.

3.3. Законы распределения непрерывных случайных величин

Экспоненциальное распределение

Экспоненциальным (показательным) законом называется распределение непрерывной случайной величины, плотность вероятностей которой равна $f(x) = \lambda e^{-\lambda x}$ при $x > 0$ (при $x = 0$ $f(x) = \lambda$).

Функция распределения получается интегрированием плотности распределения $F(x) = 1 - e^{-\lambda x}$.

График плотности функции экспоненциального распределения приведен на рис. 3.3.

Математическое ожидание и дисперсия случайной величины X , имеющей экспоненциальное распределение, равны $m_x = \frac{1}{\lambda}$, $D_x = \frac{1}{\lambda^2}$.

Плотностью распределения может служить любая интегрируемая функция $p(x)$, удовлетворяющая двум условиям:

$$p(x) \geq 0,$$

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

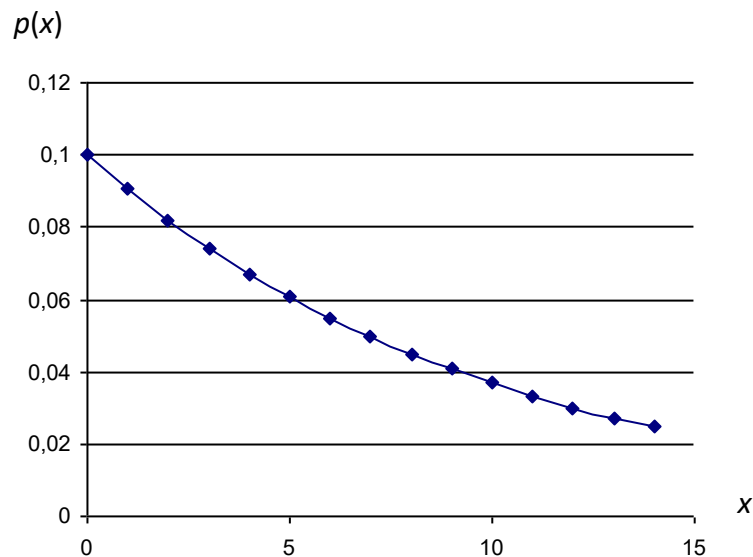


Рис. 3.3. График плотности функции экспоненциального распределения при $\Lambda = 0,1$

Теория вероятностей показывает, что вероятность наступления случайного явления в промежутке времени, равном t секунд, определяется соотношением $P(t) = 1 - \exp(-\Lambda t)$, где Λ – постоянное положи-

тельное число. Оказывается, что схема, лежащая в основе этого вывода, с бóльшим или мéньшим приближением может быть применена к поступлению телефонного вызова на автоматической телефонной станции, к отказу ЭВМ из-за неполадок, к распаду атома радиоактивного вещества и др.

Равномерное распределение

Равномерное распределение имеет ошибки округления при измерениях или вычислениях.

Например, взвешиваем образец товара из проверочной закупки. Его масса может принимать любое значение, но мы, глядя на шкалу весов, определяем ближайшее деление. Ошибка снятия показаний по шкале имеет равномерное распределение.

Непрерывная случайная величина X равномерно распределена в интервале $[a; b]$ на рис. 3.4, если ее плотность вероятности в этом интервале постоянна, т. е. если все значения в этом интервале равновероятны

$$f(x) = \begin{cases} c, & a \leq x \leq b \\ 0, & x < a, \quad x > b \end{cases}$$

Значение постоянной c определяется из условия нормировки

$$1 = \int_{-\infty}^{+\infty} f(x) dx = 0 + \int_a^b c dx + 0 = c(b - a) \Rightarrow c = \frac{1}{b - a}.$$

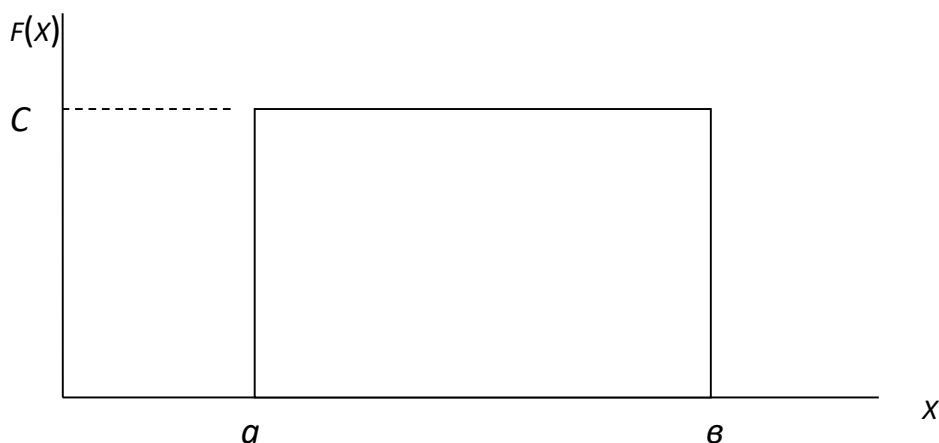


Рис. 3.4. График плотности вероятности равномерно и непрерывно распределенной на отрезке (a, b) случайной величины

Функция распределения

$$F(x) = \begin{cases} 0, & x < a, \\ (x-a)/(b-a), & a \leq x \leq b, \\ 1, & x > b. \end{cases}$$

Числовые характеристики равномерно распределенной случайной величины определяются так:

– математическое ожидание

$$\begin{aligned} M[X] &= \int_{-\infty}^{+\infty} xf(x)dx = \int_{-\infty}^a xf(x)dx + \int_a^b xf(x)dx + \int_b^{+\infty} xf(x)dx = \\ &= \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a} \frac{x^2}{2} \Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}; \end{aligned}$$

– дисперсия

$$\begin{aligned} D[X] &= \int_{-\infty}^{+\infty} (x-a)^2 f(x)dx = \\ &= \int_{-\infty}^a (x-a)^2 f(x)dx + \int_a^b (x-a)^2 f(x)dx + \int_b^{+\infty} (x-a)^2 f(x)dx = \\ &= \int_a^b \frac{(x-a)^2}{b-a} dx = \frac{1}{b-a} \frac{(x-a)^3}{3} \Big|_a^b = \frac{\left(x - \frac{a+b}{2}\right)^3}{3(b-a)} \Big|_a^b = \\ &= \frac{\left(b - \frac{a+b}{2}\right)^3 - \left(a - \frac{a+b}{2}\right)^3}{3(b-a)} = \frac{(b-a)^2}{12}. \end{aligned}$$

Среднее квадратичное отклонение равномерного распределения равно

$$\sigma_x = \frac{b-a}{2\sqrt{3}}.$$

Стандартное равномерное распределение имеет нулевое математическое ожидание и единичную дисперсию

$$x \approx U(0,1), \quad m_x = 0, \quad D_x = 1;$$

- коэффициент асимметрии $a_x = 0$,
- коэффициент эксцесса $e_x = -6/5$.

Пример. Законом равномерного распределения описываются погрешности расчетов в ЭВМ, определяемые числом разрядов вычислителя. Зная цену младшего разряда вычислителя Δ , можно рассчитать дисперсию погрешности вычислительной операции $\sigma^2 = \Delta^2/12$.

При необходимости можно определить параметры a и b равномерного распределения по известным значениям математического ожидания m_x и дисперсии D_x случайной величины. Для этого составляется система уравнений следующего вида:

$$\begin{cases} \frac{a+b}{2} = m_x, \\ \frac{b-a}{2\sqrt{3}} = \sigma_x. \end{cases}$$

Вероятность попадания равномерно распределенной случайной величины в интервал $[\alpha, \beta)$ определяется так:

$$P(\alpha < X \leq \beta) = \int_{\alpha}^{\beta} f(x)dx = \int_{\alpha}^{\beta} \frac{1}{b-a} dx = \frac{\beta - \alpha}{b - a},$$

где $[\alpha, \beta] \in [a, b]$.

Контрольные вопросы

1. Дайте определение случайной величины и назовите способы задания случайной величины.
2. Назовите основные свойства функции распределения.
3. Какие из законов распределения дискретных случайных величин получили практическое применение в задачах управления качеством?
4. Охарактеризуйте биномиальное распределение дискретных случайных величин.
5. Сформулируйте закон распределения Пуассона дискретных случайных величин.
6. В чем состоит суть экспоненциального распределения непрерывных случайных величин?
7. В чем состоит суть равномерного распределения непрерывных случайных величин?

4. НЕПРЕРЫВНЫЕ СЛУЧАЙНЫЕ ВЕЛИЧИНЫ. ЗАКОНЫ РАСПРЕДЕЛЕНИЯ

4.1. Нормальное распределение

Нормальным распределением (законом Гаусса*) называется распределение непрерывной случайной величины, плотность которого определяется по формуле

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

где m и σ – параметры распределения.

Функция нормального распределения

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-m)^2}{2\sigma^2}} dt.$$

Для краткой записи нормального распределения используют выражение $N(m, \sigma)$. В частном случае параметры $m = 0$, $\sigma = 1$, распределение называют стандартным нормальным распределением. В этом случае плотность распределения записывается формулой

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

Кривая распределения имеет колоколообразный вид (рис. 4.1), вертикальная ось является осью симметрии, горизонтальная – асимптотой. Максимальное значение ординаты равно $\frac{1}{\sqrt{2\pi}}$. При значении аргумента $x = \pm 3$ значение функции близко к нулю. При общей площади под кривой, равной единице, в этом диапазоне лежит 99,73 % данных. Коэффициент асимметрии и эксцесс равны нулю $a_x = 0$, $e_x = 0$.

Функция стандартного нормального распределения иногда называется функцией Лапласа. Она имеет специальное обозначение

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

* Гаусс (1777 – 1855) – знаменитый немецкий математик – заложил основы теории случайных ошибок и метода наименьших квадратов, широко используемых в науке и технике.

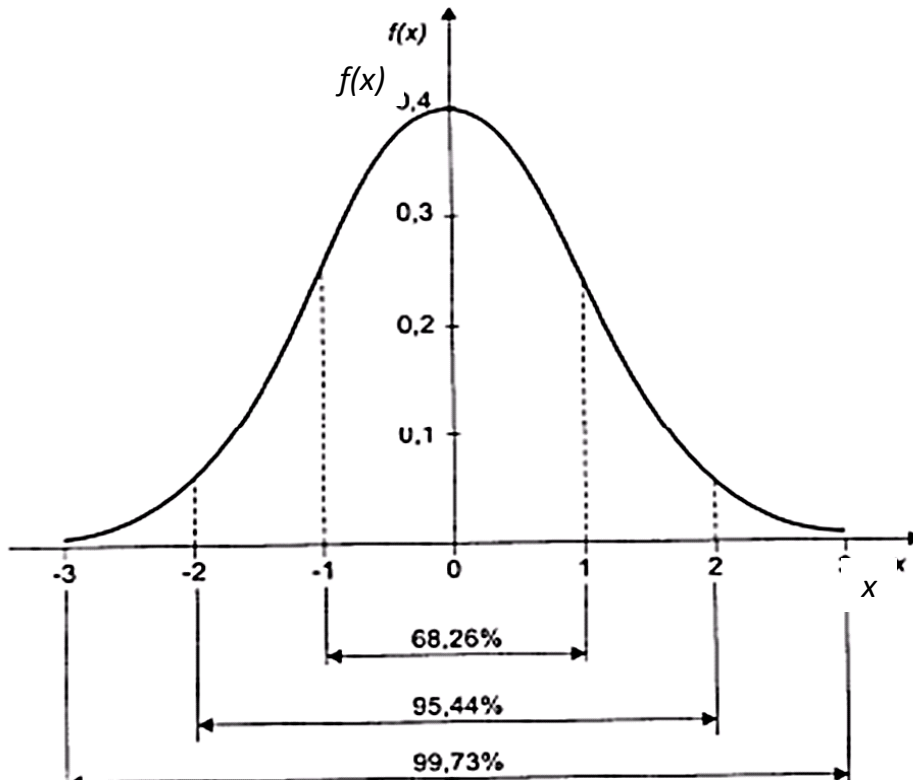


Рис. 4.1. Кривая нормального закона распределения

Эта функция табулирована. График функции показан ниже на рис. 4.2. Из симметрии графика вытекает соотношение $\Phi(-x) = 1 - \Phi(x)$.

Квантилем нормального распределения порядка p называется число u_p , для которого функция стандартного нормального распределения $\Phi(u_p) = p$. Табулированы квантили нормального распределения. Из симметрии графика функции стандартного нормального распределения вытекает, что $u_{1-p} = -u_p$.

Можно установить связь между функцией распределения $F(x)$ для распределения $N(m, \sigma)$ и функцией стандартного нормального распределения

$$F(x) = \Phi\left(\frac{x - m}{\sigma}\right).$$

Вероятность попадания нормального распределения случайной величины в интервал от x_1 до x_2 определяются по формуле

$$P(x_1 \leq X < x_2) = \Phi\left(\frac{x_2 - m}{\sigma}\right) - \Phi\left(\frac{x_1 - m}{\sigma}\right).$$

Параметр σ характеризует форму кривой распределения, это характеристика рассеяния. Наибольшая ордината кривой распределения обратно пропорциональна σ . При увеличении σ максимальная ордината уменьшается, так как площадь кривой распределения всегда должна оставаться равной единице. При увеличении σ кривая распределения становится более плоской. На рис. 4.3 показаны три нормальные кривые при $m = 0$ с различными значениями σ .

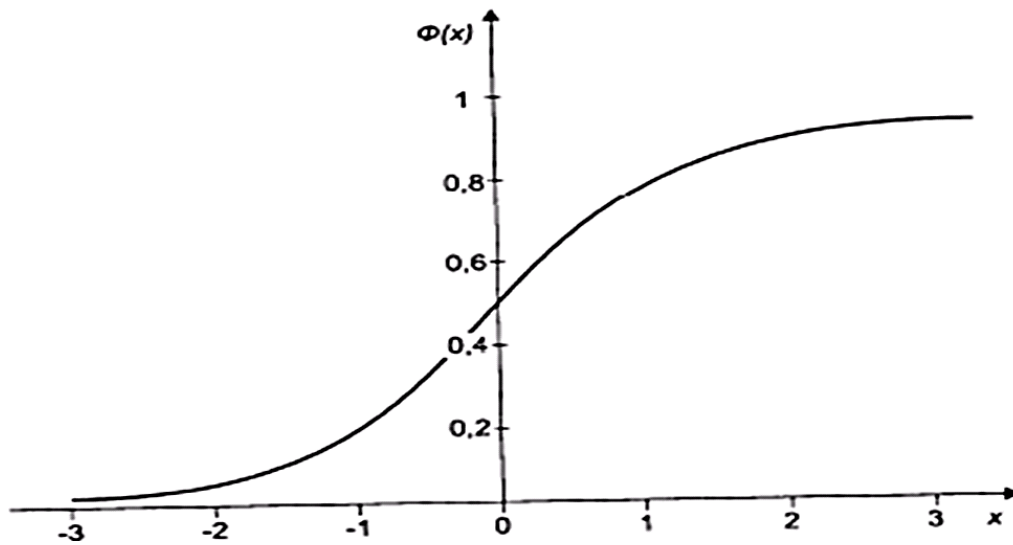


Рис. 4.2. График функции стандартного нормального распределения

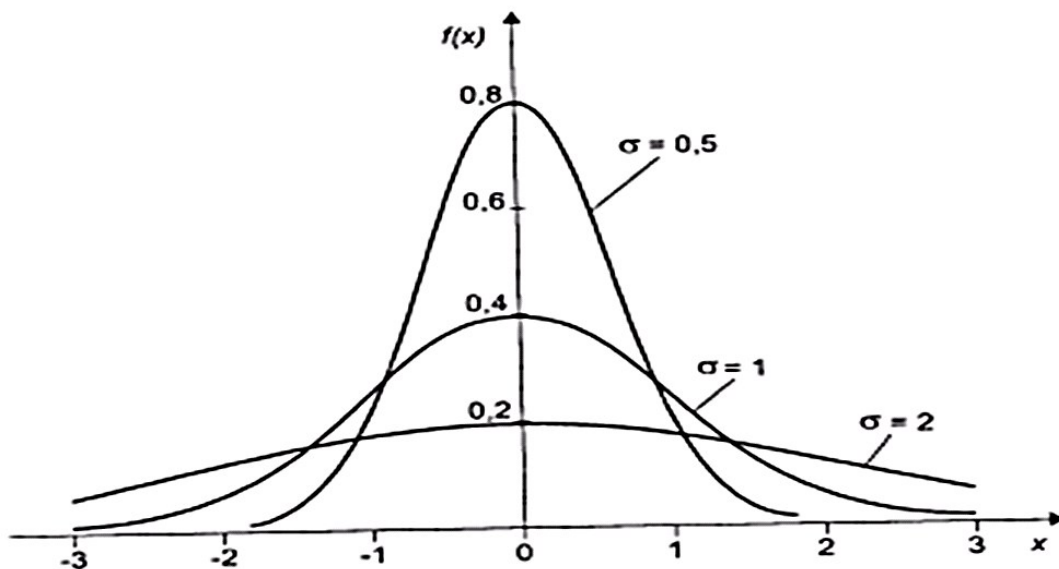


Рис. 4.3. Влияние параметра σ на вид кривой нормального распределения

Часто в расчетах надо найти вероятность того, что случайная величина X не слишком сильно отклоняется от своего математического ожидания m

$$P(|X - m| < \varepsilon) = 2\Phi(\varepsilon/\sigma) - 1.$$

Пусть, например, $\xi = 3\sigma$. Используя таблицы функции стандартного нормального распределения, находим

$$P(|X - m| < 3\sigma) = 2\Phi(3) - 1 = 2 \cdot 0,99865 - 1 = 0,9973.$$

Вероятность того, что случайная величина отклонится от математического ожидания больше чем 3σ , ничтожно мала $P(|X - m| > 3\sigma) = 0,0027$. Такое событие практически невозможно.

На практике часто используется правило «трех сигм»: отклонение нормально распределенной случайной величины от ее математического ожидания, как правило, не превышает утроенного стандартного отклонения.

Пример. На станке-автомате изготавливают валики для принтеров диаметром 10 мм. Точность изготовления валиков на станке оценивается стандартным отклонением, равным $\sigma = 0,03$ мм. Сколько в среднем валиков из 100 удовлетворяют требованиям, чтобы отклонение диаметра от номинального не превышало 0,05 мм?

$$P(|X - m| < 0,05) = 2\Phi(0,05/0,03) - 1 = 2\Phi(1,67) - 1 = 2 \cdot 0,9522 - 1 = 0,9044,$$

т. е. примерно 90 валиков из каждых 100 удовлетворяют требованиям.

Широкое распространение нормального распределения обосновывается центральной предельной теоремой, которая устанавливает условия, в которых справедливо нормальное распределение. Упрощенная формулировка теоремы такова. Пусть X_1, X_2, \dots, X_n независимые одинаково распределенные случайные величины. Тогда при увеличении n закон распределения суммы этих величин неограниченно приближается к нормальному.

Главная особенность, выделяющая нормальный закон среди других видов, состоит в том, что он является предельным законом, к которому приближаются другие законы распределения при весьма часто встречающихся типичных условиях.

4.2. Распределение χ -квадрат

Пусть X_i случайные величины, имеющие стандартное нормальное распределение $N(0, 1)$. Распределение суммы квадрат этих величин $\chi^2(k) = \sum_{i=1}^k X_i^2$ называется распределением χ -квадрат с k степенями свободы.

График кривых распределений χ -квадрат показан на рис. 4.4. Он представляет собой одномодальную асимметричную кривую распределения с максимумом в точке $x = k - 2$.

Квантили распределения χ -квадрат обозначаются $\chi^2 p(k)$, они табулированы и их значение определяется числом степеней свободы k и порядком квантиля p . Например, $\chi^2 0,975(15) = 276,5$.

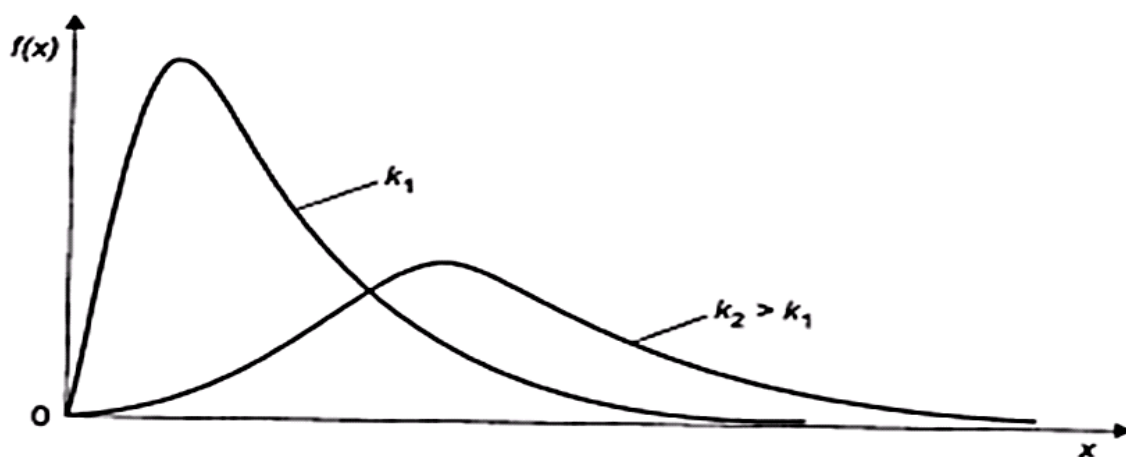


Рис. 4.4. Кривые распределения χ -квадрат

4.3. Распределение Стьюдента

Пусть X случайная величина, имеющая стандартное нормальное распределение $N(0, 1)$, а Y случайная величина, распределенная по закону χ -квадрат с k степенями свободы. Распределение величины $t(k) = \frac{X}{\sqrt{Y/k}}$ называется распределением Стьюдента с k степенями свободы. График кривой распределения показан на рис. 4.5.

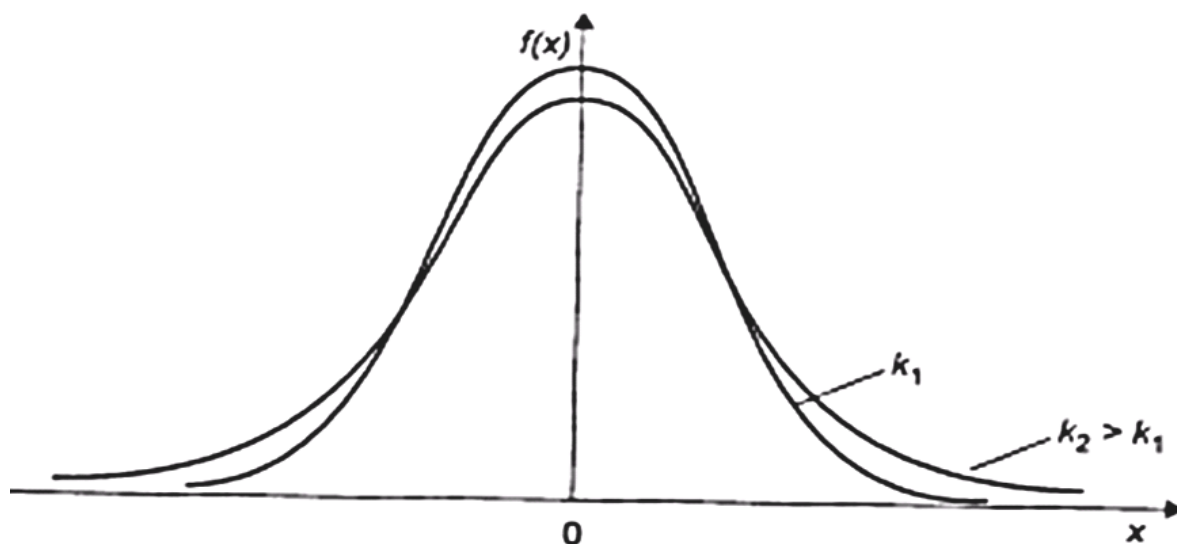


Рис. 4.5. Кривые t -распределения Стьюдента

Кривая распределения Стьюдента симметрична относительно оси ординат как при нормальном распределении. Для больших значений k она очень близка к нормальной кривой $N(0, 1)$. Квантили распределения Стьюдента $tp(k)$ табулированы. Например, $t_{0,99}(12) = 2,681$.

4.4. Распределение Фишера

Важные приложения в дисперсионном анализе имеет распределение Фишера. Пусть Y_1 случайная величина, распределенная по χ -квадрат с k_1 степенями свободы. Случайная величина Y_2 тоже распределена по закону χ -квадрат со степенями свободы k_2 . Тогда распределение величины

$$F(k_1, k_2) = \frac{Y_1 / k_1}{Y_2 / k_2}$$

называется распределением Фишера с k_1 степенями свободы в числителе и k_2 в знаменателе. График кривой распределения изображен на рис. 4.6. Квантиль распределения $F_p(k_1, k_2)$ табулирован. При данных k_1 и k_2 мы можем найти по таблице значения F_q такие, что $P(F > F_q) = \frac{q}{100}$, где F_q будет q -процентным верхним пределом распределения. Например, $F_{0,95}(5,10) = 3,33$.

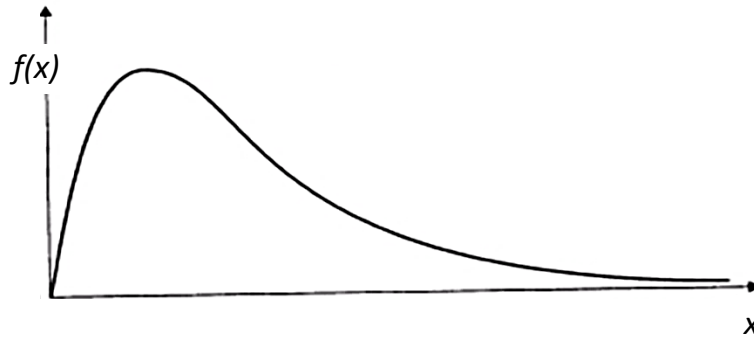


Рис. 4.6. Кривая распределения Фишера

Контрольные вопросы

1. Что понимают под нормальным распределением?
2. Каково влияние параметра σ на вид кривой нормального распределения.
3. Сформулируйте центральную предельную теорему.
4. Назовите особенность, выделяющую нормальный закон среди других видов.
5. Что называется распределением χ -квадрат?
6. Что называется распределением Стьюдента?
7. Что понимают под распределением Фишера?

5. МНОГОМЕРНОЕ РАСПРЕДЕЛЕНИЕ ДИСКРЕТНЫХ И НЕПРЕРЫВНЫХ СЛУЧАЙНЫХ ВЕЛИЧИН

5.1. Двумерное дискретное распределение

Часто приходится рассматривать одновременно системы из двух, трех и более чисел величин, например помехи в электронном усилителе со случайными амплитудой и фазой и т. д. Такие величины в зависимости от случайного исхода испытания принимают систему из двух, трех и более чисел значений и могут изображаться точкой в пространстве соответствующего числа измерений. Они носят название двумерных, трехмерных и так далее по числу компонент такой величины. Статистические данные о таких величинах принимают форму таблиц с двумя, тремя и так далее входами.

Дискретная случайная величина задается таблицей распределения (см. таблицу).

Двумерное дискретное распределение

Y	X				p(y _i)
	x ₁	x ₂	...	x _n	
y ₁	p(x ₁ , y ₁)	p(x ₂ , y ₁)		p(x _n , y ₁)	p(y ₁)
y ₂	p(x ₁ , y ₂)	p(x ₂ , y ₂)		p(x _n , y ₂)	p(y ₂)
...					
y _m	p(x ₁ , y _m)	p(x ₂ , y _m)		p(x _n , y _m)	p(y _m)
	p(x ₁)	p(x ₂)		p(x _n)	p(y)

Вероятность совпадения событий ($X = x_i$) и ($Y = y_j$) записана в клетках таблицы $p(x_i, y_j)$. Предполагается, что все комбинации X, Y составляют полную группу событий, поэтому сумма вероятностей, стоящих в таблице, равна единице

$$\sum_i \sum_j p(x_i, y_j) = 1.$$

Если просуммировать все вероятности, стоящие в столбце, то получим $\sum_j p(x_i, y_j) = p(x_i)$, а стоящих в строке $\sum_i p(x_i, y_j) = p(y_j)$.

Из таблицы распределения двумерной величины определяются одномерные законы распределения величин X и Y . Условная вероятность события $Y = y_j$, если наблюдалось событие $X = x_i$, определяется соотношением

$$p(y_j / x_i) = \frac{p(x_i, y_j)}{p(x_i)}.$$

Сумма условных вероятностей, отвечающих одному и тому же условию $X = x_i$, называется условным распределением Y при $X = x_i$

$$\sum_j p(y_j / x_i) = \frac{\sum_j p(x_i, y_j)}{p(x_i)} = \frac{p(x_i)}{p(x_i)} = 1.$$

Наиболее важной характеристикой является условное математическое ожидание $M(Y/x)$ величины Y при фиксированном значении $X = x$,

где x может равняться x_1, x_2, \dots, x_i . Это математическое ожидание определяется равенством

$$M(Y/x) = \sum_j y_j p(y_j/x).$$

Аналогично вводятся условная дисперсия и условные моменты более высоких порядков.

При переходе от одного столбца таблицы к другому изменяется и $M(Y/x)$ для значений x_1, x_2, \dots, x_i . Эта функция называется регрессией Y по X и описывает изменение центров тяжести масс вероятностей на вертикальных прямых $X = x = \text{const}$.

Аналогично можно рассматривать условные законы величины X при фиксированном значении $Y = y_j$. Эта функция описывает изменение центров тяжести масс вероятностей на горизонтальных прямых $Y = y = \text{const}$.

5.2. Двумерное непрерывное нормальное распределение

Плотность распределения задается выражением

$$p_{xy}(x, y) = \{1/(2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2})\} \exp\{-1/2 \cdot Q(x, y)\}, \quad (5.1)$$

где $Q(x, y) = 1/(1 - \rho_{xy}^2) \{ (x - m_x)^2/\sigma_x^2 + (y - m_y)^2/\sigma_y^2 - 2\rho_{xy}(x - m_x)/\sigma_x \times (y - m_y)/\sigma_y \}$; m_x и m_y – центры распределения случайных величин X и Y ; σ_x и σ_y – стандартные отклонения случайных величин X и Y .

Выражение (5.1) является плотностью двумерного распределения двух линейно коррелированных величин X и Y , каждая из которых в отдельности нормально распределена с соответствующими значениями центра и дисперсии.

Если величины X и Y независимы и нормально распределены с плотностями соответственно $N(x, m_x, \sigma_x)$ и $N(y, m_y, \sigma_y)$, то плотность их совместного распределения получается из (5.1) при $\rho_{XY} = 0$ как произведение плотностей $N(x, m_x, \sigma_x)$ и $N(y, m_y, \sigma_y)$ их одномерных распределений

$$\Psi_{XY}(x, y) = \{1/(2\pi\sigma_x\sigma_y)\} \exp\{-1/2[(x - m_x)^2/\sigma_x^2 + (y - m_y)^2/\sigma_y^2]\}.$$

Из этого следует, если нормально распределенные величины некоррелированы, то они вместе с тем и независимы. Этот вывод не подходит для произвольного закона распределения, а только для нормального.

Рассмотрим условное нормальное распределение, его плотность равна

$$p(y/x) = \{1/(2\pi\sigma_y\sqrt{(1-\rho_{xy}^2)})\} \exp\{-1/2[((y - m_y) - \rho_{xy}(\sigma_y/\sigma_x)(x - m_x)) / (\sigma_y\sqrt{(1-\rho_{xy}^2)})]^2\}.$$

Плотность условного распределения Y при данном значении x является нормальным распределением с центром

$$M(Y/x) = m_{Y/x} = m_Y + \rho_{XY}(\sigma_Y/\sigma_X)(x - m_X),$$

который представляется математическим ожиданием Y при данном x

$$M(Y/x) = m_{Y/x}.$$

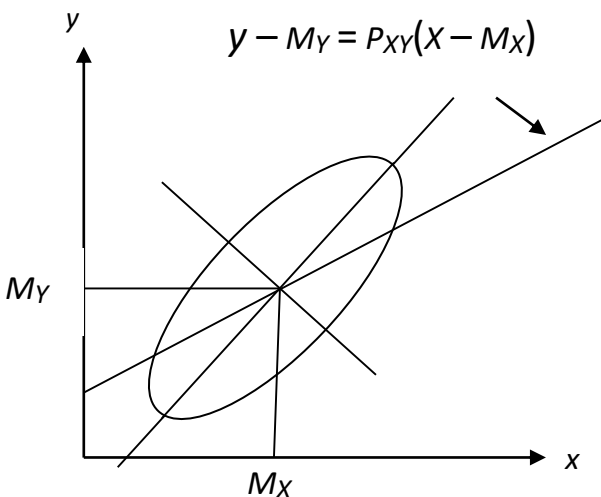
Также условное стандартное отклонение будет

$$\sigma_{Y/x} = \sigma_Y \sqrt{(1-\rho_{XY}^2)}. \quad (5.2)$$

Уравнение (5.2) представляет вместе с тем уравнение линии нормальной регрессии Y по X , которая является прямой линией. Аналогично регрессия X по Y будет также линией, а условная дисперсия равна

$$\sigma_{X/y}^2 = \sigma_X^2 ((1 - \rho_{XY}^2)).$$

Величина (5.2) представляет теоретическое среднее квадратическое отклонение погрешностей оценки ожидаемого значения Y по x .



Сечение плотности двумерного нормального распределения

Отсюда следует, что оценка Y по x с помощью линии регрессии одинакова при всех значениях x . Функция плотности вероятности двумерного распределения может быть наглядно отображена в трехмерной плоскости (см. рисунок).

Рассекая поверхность нормального распределения плоскостью, параллельной поверхности $x - y$, в сечении получаем эллипс, за исключением вырожденного случая $\rho_{xy} = \pm 1$. Сечения в разных плоскостях будут да-

вать эллипсы различных размеров с одинаковой ориентацией их глав-

ных осей, составляющих некоторый угол с осями координат. Главные оси не могут быть параллельными линии регрессии.

Если значения X , Y некоррелированы, $\rho_{xy} = 0$, то в сечении будем иметь эллипс с центром m_x , m_y и главными осями, параллельными осям координат x и y .

Таким образом, с увеличением силы корреляционной связи между величинами X и Y происходит все больший поворот главных осей эллипсов относительно координатных осей.

5.3. Многомерное распределение

В многомерном статистическом анализе рассматривается множество признаков, которые обозначаются вектором x , имеющим k компонент, каждый из которых характеризует соответствующий признак (x_1, x_2, \dots, x_k) .

Функция распределения случайного вектора (детерминированная неотрицательная величина) определяется формулой

$$F(x) = P(X_1 < x_1, X_2 < x_2, \dots, X_k < x_k) = P(X < x).$$

Детерминированная неотрицательная величина обладает следующими свойствами:

$F(x) = 0$, если среди x_j имеется хотя бы одна компонента, равная $-\infty$;

$F(x) = 1$, если все компоненты вектора x равны $+\infty$.

Различают непрерывные k -мерные, дискретные k -мерные и смешанные k -мерные случайные величины.

Непрерывная k -мерная случайная величина имеет плотность распределения вероятностей $p(x) = p(x_1, x_2, \dots, x_k) \geq 0$, удовлетворяющую условию

$$F(x) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_k} p(x_1, x_2, \dots, x_k) d_{x_1} d_{x_2} \dots d_{x_k}.$$

Плотность $p(x)$ обладает свойствами

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} p(x_1, x_2, \dots, x_k) d_{x_1} d_{x_2} \dots d_{x_k} = 1.$$

Вероятность попадания точки (x_1, x_2, \dots, x_k) в какую-нибудь область G равна

$$\int_G \dots \int p(x_1, x_2, \dots, x_k) d_{x_1} d_{x_2} \dots d_{x_k}.$$

Из определения плотности следует

$$\int_{-\infty}^{+\infty} p(x_1, x_2) d_{x_1} = p(x_2);$$

$$: \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x_1, x_2, x_3) d_{x_2} d_{x_3} = p(x_1);$$

$$\int_{-\infty}^{+\infty} p(x_1, x_2, x_3) d_{x_1} = p(x_2, x_3).$$

Плотности вероятности функции распределения подсистем $L (1 \leq L < k)$ случайных величин системы k случайных величин называют частными, или маргинальными, распределениями.

Условными распределениями случайного вектора x называются распределения подсистемы $L (1 \leq L < k)$ его компонент при условии, что остальные $k - L$ компоненты являются фиксированными (отделяются косой чертой).

Рассмотрим условное распределение двумерной случайной величины (x_1, x_2) , являющейся подсистемой системы $(x_1, x_2, x_3, x_4, x_5)$ при условии, когда в ней фиксированы три последние компоненты.

Для дискретной случайной величины

$$P_{i_1, i_2 / i_3, i_4, i_5} = \frac{P_{i_1, i_2, i_3, i_4, i_5}}{P_{i_3, i_4, i_5}}.$$

Для непрерывной случайной величины

$$P_{x_1, x_2 / x_3, x_4, x_5} = \frac{P_{x_1, x_2, x_3, x_4, x_5}}{P_{x_3, x_4, x_5}};$$

где $P_{x_3, x_4, x_5} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x_1, x_2, x_3, x_4, x_5) d_{x_1} d_{x_2}.$

Компоненты X называются независимыми, если

$$F(x_1, x_2, \dots, x_k) = F(x_1)F(x_2) \dots F(x_k).$$

Рассмотрим случай дискретной двумерной величины, заданной в таблице вероятностей. Каково бы ни было значение $X = x$ для условной вероятности Y , в случае независимости будем иметь

$$p(y/x) = P(Y = y/X = x) = p_y(y),$$

т. е. условная вероятность совпадает с безусловной, т. е. распределение величины Y не реагирует на изменение величины X .

Точно так же $p(x/y) = p(x)$. Поэтому все условные математические ожидания $M(Y/x)$ величины Y будут независимы от x и равны безусловному математическому ожиданию Y

$$\bar{y}(x) = \sum_j y_j p(y_j/x) = MY.$$

В случае непрерывного распределения величины X и Y называются независимыми, если выполняется равенство $P(X < x, Y < y) = P(X < x)P(Y < y)$, справедливое для любых x и y .

Пользуясь определением функции распределения, можно записать $F_{XY}(x, y) = F_X(x)F_Y(y)$.

Аналогичное равенство имеет место для плотностей:

$$p_{XY}(x, y) = p_X(x)p_Y(y).$$

Совместная плотность n независимых непрерывных величин X_1, X_2, \dots, X_n

$p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$ выразится через плотности компонент

$$p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = p_{X_1}(x_1)p_{X_2}(x_2) \dots p_{X_n}(x_n).$$

5.4. Закон больших чисел

Неравенство Чебышева. Мы видели, что характеристика $\sigma_X = \sqrt{D_X}$ представляет некоторую среднюю меру, стандарт, отклонения от центра распределения. Следует ожидать, что отклонения, значительно превышающие по абсолютной величине σ_X , должны быть маловероятны. В случае нормального распределения эта вероятность равна

$$Q(t) = P(|X - m_X| > t \sigma_X), \quad (5.3)$$

где $t > 0$ изображается площадью под нормальной кривой вне интервала $(-t, +t)$. Для $t = 3$ эта вероятность составляет 0,0027; при $t = 4$ вероятность уменьшается до 0,000063, при $t = 6$ вероятность уменьшается до $2 \cdot 10^{-9}$ и т. д.

Заслуга Чебышева состоит в доказательстве неравенства, показывающего, что убывание вероятности $Q(t)$ при возрастании t хотя и не всегда совершается столь быстро, как в нормальном случае, но оно происходит всегда не медленнее, чем по закону $1/t^2$.

При любом законе распределения, обладающем моментом двух первых порядков (математическое ожидание и дисперсия), верхняя граница вероятности

$$Q(t) = P(|X - m_X| > t \sigma_X) \leq 1/t^2.$$

Простота и универсальность позволяют использовать неравенство Чебышева для важных теоретических заключений, хотя для практических расчетов оно оказывается слишком грубым [4].

5.5. Основные предельные законы теории вероятностей

Рассмотрим две фундаментальные теоремы теории вероятностей, имеющие обширный круг приложений. Эти теоремы представляют обобщение теорем Я. Бернулли и Лапласа, относящиеся к закону распределения частот (или числа появлений) случайного события в данной серии независимых испытаний.

Число появлений событий в n независимых испытаниях можно рассматривать как сумму n независимых величин. После каждого испытания наблюдатель записывает результат, ставя 1 или 0 в зависимости от того, появилось или не появилось событие в этом испытании. С испытанием связана случайная двузначная величина X_s , $s = 1, 2, \dots$. Все величины независимы между собой и одинаково распределены согласно таблице распределений

$X = 0$	$X = 1$
$q = 1 - p$	p

Мы можем представить величины X_s как разные экземпляры одной и той же величины X (без номера). Сумма $S_n = X_1 + X_2 + \dots + X_n$ равна числу m появлению событий в серии испытаний.

Частота событий m/n представляется средним арифметическим величины X_s

$$S_n/n = (X_1 + X_2 + \dots + X_n)/n.$$

Для этого представления частоты можно рассчитать основные характеристики ее распределения, которые совпадают с биномиальным распределением:

- математическое ожидание $M(S_n/n) = p$;
- дисперсия $D(S_n/n) = pq/n$. (5.4)

Дисперсия частоты согласно (5.4) стремится к нулю при неограниченном возрастании n . Опираясь на неравенство Чебышева (5.3), получаем теорему Якова Бернулли

$$P\{|S_n/n - p| \geq \xi\} = P\{|(X_1 + X_2 + \dots + X_n)/n - M(X_1 + X_2 + \dots + X_n)/n| > \xi\} \rightarrow 0 \quad (5.5)$$

при $n \rightarrow \infty$.

Теорема Бернулли (5.5) утверждает, что среднее арифметическое большого числа независимых величин (частного вида – двухзначных) почти наверное будет как угодно близко к своему математическому ожиданию – постоянной величине p .

То обстоятельство, что дисперсия величины S_n/n стремится к нулю при $n \rightarrow \infty$, имеет следствием устойчивость среднего арифметического. Распределение среднего (частоты) концентрируется в сколь угодно малом интервале $(p - \xi, p + \xi)$, а вероятность, приходящаяся на значения вне этого интервала, как угодно мала при достаточно большом n .

В этом случае говорят, что последовательность средних арифметических при $n \rightarrow \infty$ «сходится по вероятности» к постоянной величине

$$P = M(S_n/n).$$

Факт устойчивости средних арифметических большого числа одинаково распределенных независимых величин имеет место при произвольном распределении каждого слагаемого, если только при этом распределении величины обладают конечной дисперсией. В этом случае

$$X_{\text{нсп}} = (X_1 + X_2 + \dots + X_n)/n \text{ имеет место } D(X_{\text{нсп}}) = D(X)/n$$

и поэтому $D(X_{\text{нсп}}) \rightarrow 0$ при $n \rightarrow \infty$.

На основании неравенства Чебышева получим при сколь угодно малом (но постоянном) $\xi > 0$:

$$P(|X_{\text{нсп}} - m_X| > \xi) < D(X_n)/\xi^2 \rightarrow 0,$$

$$P(|X_{\text{нсп}} - m_X| \leq \xi) \xi^2 \rightarrow 1,$$

что доказывает сходимость последовательности $X_{\text{нсп}}$ по вероятности к пределу m_X при $n \rightarrow \infty$.

Осредняя достаточно большое число независимых и одинаково распределенных случайных величин, получаем с вероятностью, как угодно близкой к единице, значение, сколь угодно мало отличающееся от общего математического ожидания величин. Это положение, называемое «законом больших чисел», было установлено П. Л. Че-

бышевым (1821 – 1894). Закон выражает основную и общую закономерность, имеющую первостепенное значение как для обоснования статистических методов, так и для теоретического объяснения большого круга явлений.

Контрольные вопросы

1. Охарактеризуйте двумерное дискретное распределение.
2. Охарактеризуйте двумерное непрерывное нормальное распределение.
3. Чем характеризуется многомерное распределение?
4. Что показывает неравенство Чебышева?
5. Назовите фундаментальные теоремы теории вероятностей, имеющие обширный круг приложений.
6. Что утверждает теорема Якова Бернулли?
7. Что выражает закон больших чисел?

6. СТАТИСТИЧЕСКИЕ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ ПО МАЛЫМ ВЫБОРКАМ

Установление закономерностей, которым подчинены массовые случайные явления, основано на изучении методами теории вероятностей статистических данных – результатов наблюдений. Первая задача математической статистики – указать способы сбора и группировки статистических сведений, полученных в результате наблюдений или специально поставленных экспериментов.

Вторая задача математической статистики – разработать методы анализа статистических данных в зависимости от целей исследования. Сюда относятся:

а) оценка неизвестной вероятности события, неизвестной функции распределения и оценка параметров распределения, вид которого известен;

б) проверка статистических гипотез о виде неизвестного распределения или о величине параметров распределения, вид которого известен.

Итак, задача математической статистики состоит в создании методов сбора и обработки статистических данных для получения научных и практических выводов.

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

Генеральной совокупностью называют совокупность объектов, из которых производится выборка.

Объемом совокупности (выборочной или генеральной) называют число объектов этой совокупности. Например, если из 1000 деталей отобрано для обследования 100 деталей, то объем генеральной совокупности $N = 1000$, а объем выборки $n = 100$.

Статистическое распределение выборки

Пусть из генеральной совокупности извлечена выборка объемом n : x_1, x_2, \dots, x_n , причем x_1 наблюдалось m_1 раз, x_2 — m_2 раз и т. д.

Объем выборки равен $\sum_{i=1}^k m_i = n$, где m_i — частоты. Их отношения

к объему выборки представляют относительные частоты $\frac{m_i}{n} = p_i^*$.

Статистическим рядом распределения называют перечень всех значений x_i из выборки и соответствующих им частот или относительных частот. Статистическое распределение можно задать также в виде интервального статистического ряда, т. е. последовательности интервалов и соответствующих им частот (в качестве частоты, соответствующей интервалу, принимают сумму частот, попавших в этот интервал).

Пример. Пусть объем выборки $n = 20$ и

x_i	2	6	12
m_i	3	10	7

Найдем относительные частоты:

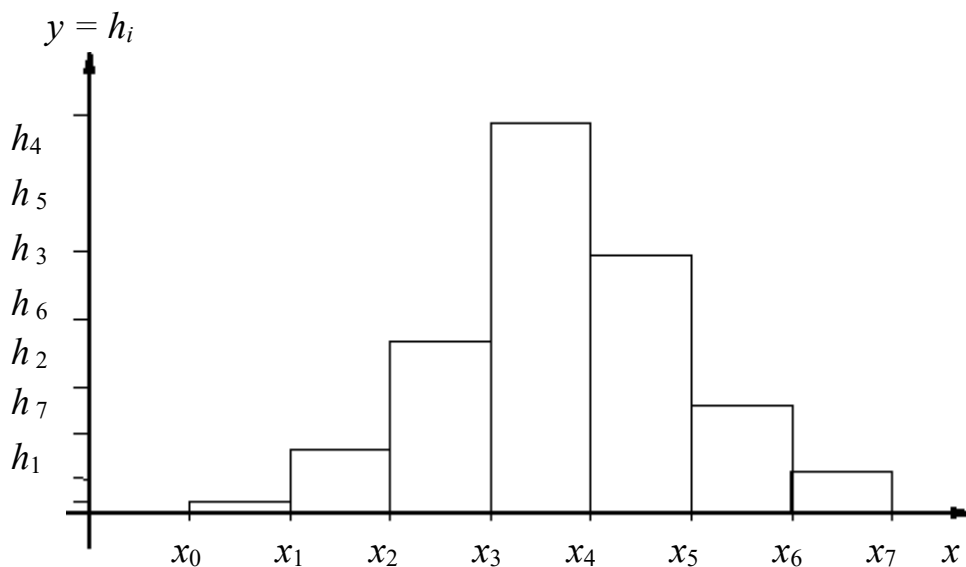
$$p_1^* = \frac{3}{20} = 0,15, \quad p_2^* = \frac{10}{20} = 0,50, \quad p_3^* = \frac{7}{20} = 0,35.$$

Тогда распределение относительных частот

x_i	2	6	12
p_i^*	0,15	0,50	0,35

Контроль: $0,15 + 0,50 + 0,35 = 1$.

Гистограммой частот называют ступенчатую фигуру (см. рисунок), состоящую из прямоугольников, основаниями которой служат частичные интервалы длиной Δx , а высоты которых равны отношению $h_i = p_i^* / \Delta x$, ($i = 1, 2, \dots, k$).



Гистограмма частот

Площадь i -го частичного прямоугольника равна p_i^* – относительной частоте значений выборки, попавших в i -й интервал. Следовательно, вся площадь равна сумме всех относительных частот, т. е. единице.

Для построения гистограммы частот на оси абсцисс откладывают частичные интервалы, а над ними проводят отрезки, параллельные оси абсцисс на расстоянии $h_i = p_i^* / \Delta x$, ($i = 1, 2, \dots, k$).

Ступенчатая кусочно-постоянная линия $y = h_i$ при $x \in [x_{i-1}; x_i)$, ($i = 1, 2, \dots, k$) при стремлении числа интервалов $k \rightarrow \infty$, а соответственно $\Delta x \rightarrow 0$ превращается в график функции $y = f^*(x)$, который является эмпирическим аналогом дифференциальной функции распределения $y = f(x)$.

6.1. Получение точечных оценок параметров генеральной совокупности по выборке

Для общности обозначим статистическую оценку неизвестного параметра S распределения (т. е. $M(X)$, $D(X)$ и т. д.) через S^* (т. е. \bar{X} , \bar{D} и т. д.).

Точечной статистической оценкой S^* некоторого параметра S генеральной совокупности (или теоретического распределения случайной величины X) называется приближенное значение этого параметра, рассчитанное по данным выборки. Иными словами, точечная статистическая оценка S^* некоторого параметра S генеральной совокупности (теоретического распределения случайной величины X) есть некоторая функция выборки (результатов наблюдений за случайной величиной) $S^* = S^*(x_1, x_2, \dots, x_n)$, где x_1, x_2, \dots, x_n — n отобранных элементов генеральной совокупности (реализации случайной величины X в первом, втором, ..., n -м опытах).

Во многих случаях использование числовых характеристик выборки требует ответа на вопрос: насколько точно выборочные оценки соответствуют статистическим характеристикам генеральной совокупности, т. е. можно ли утверждать, что генеральная совокупность описывается числовыми характеристиками конечной выборки? Сформулируем некоторые критерии, которые следует предъявить к выборочным оценкам для положительного ответа на поставленный вопрос.

Пусть по выборке объемом n найдена оценка S_1^* . Повторим опыт — извлечем из генеральной совокупности другую выборку того же объема и найдем новую оценку S_2^* . Повторяя, можно получить числа $S_1^*, S_2^*, \dots, S_k^*$. Ясно, что выборочные оценки S^* можно рассматривать как случайную величину, а числа $S_1^*, S_2^*, \dots, S_k^*$ как ее возможные значения. Следовательно, к оценке соответствия характеристик S^* и S можно подойти с вероятностных позиций.

При получении чисел $S_1^*, S_2^*, \dots, S_k^*$ будут, естественно, случайные отклонения. Но для достаточного числа измерений можно утверждать, что $M(S^*) = S$, т. е. математическое ожидание оценки S^* равно оцениваемому параметру S при любом объеме выборки. Такая статистическая оценка S^* называется несмещенной. Если же равенство не соблюдается, т. е. $M(S^*) \neq S$, то оценку S^* называют смещенной.

Не всегда можно утверждать, что несмещенная оценка дает хорошее приближение для S . Действительно, возможные значения S^* могут быть сильно рассеяны вокруг своего среднего значения $M(S^*)$, т. е. дисперсия $D(S^*)$ может быть значительной. В этом случае, если, к

примеру, взять по одной выборке оценку S_1^* , то она может быть сильно удалена от $M(S^*)$, а значит, и от S . Приняв S_1^* , мы допустили бы большую ошибку. Но если потребовать, чтобы дисперсия для S^* была малой, то возможность допустить большую ошибку будет исключена. По этой причине к S^* предъявляется требование эффективности.

Эффективной называют статистическую оценку S^* , которая имеет наименьшую из возможных дисперсию при заданном объеме выборки n . И, наконец, при рассмотрении выборок большого объема к статистическим оценкам предъявляется требование состоятельности.

Состоятельной называют статистическую оценку S^* , которая при увеличении объема выборки стремится к оцениваемому параметру генеральной совокупности, т. е. для любого достаточно маленького $\varepsilon > 0$ выполняется следующее предельное равенство $\lim_{n \rightarrow \infty} p(|s^* - s| < \varepsilon) = 1$.

К примеру, если дисперсия несмещенной оценки $D(S^*)$ при увеличении объема n стремится к нулю, то, очевидно, S^* будет и состоятельной.

Пусть из генеральной совокупности X извлечена выборка объемом n со значениями x_1, x_2, \dots, x_n . В качестве несмещенной эффективной состоятельной оценки математического ожидания используют выборочную среднюю

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n},$$

а в качестве оценки дисперсии – выборочную дисперсию

$$\bar{D} = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}.$$

Для случая интервального статистического ряда распределения

$$\bar{X} = \frac{\sum_{i=1}^k m_i x_i}{n} \approx \frac{\sum_{i=1}^k m_i c_i}{n},$$

$$\bar{D} = \bar{\sigma}^2 = \frac{\sum_{i=1}^k m_i (x_i - \bar{X})^2}{n-1} \approx \frac{\sum_{i=1}^k m_i (c_i - \bar{X})^2}{n-1},$$

где k – число интервалов, а $c_i = \frac{x_{i-1} + x_i}{2}$ – середины соответствующих интервалов.

Для вычислений \bar{X} и $\bar{\sigma}^2$ по данным эмпирических наблюдений полезно составлять вспомогательную расчетную таблицу.

Расчет оценок по эмпирическим наблюдениям

№ п/п	Интервал	Частота m_i	Относительная частота P_i^*	Высота гистограммы h_i	Середина интервалов c_i	$m_i c_i$	$m_i (c_i - \bar{X})^2$
1	$[x_0, x_1)$						
2	$[x_1, x_2)$						
...							
k	$[x_{k-1}, x_k)$						

6.2. Интервальное оценивание параметров генеральной совокупности по выборке

Наряду с точечными оценками (приближенными численными значениями исследуемых параметров) используются также интервальные оценки. Интервальной оценкой параметра θ называют числовой интервал (определяемый его начальной \tilde{L}_n и конечной \tilde{U}_n точками – концами интервала), который с заданной вероятностью γ накрывает (охватывает) неизвестное значение исследуемого параметра θ генеральной совокупности. Интервал, содержащий оцениваемый параметр генеральной совокупности, называют доверительным интервалом, а вероятность γ – доверительной вероятностью, уровнем доверия или надежностью оценки.

Границы \tilde{L}_n и \tilde{U}_n и величина интервальной оценки вычисляются по данным выборки и поэтому являются случайными величинами.

Доверительный интервал для математического ожидания нормально распределенной случайной величины при неизвестной дисперсии с заранее заданной надежностью находят из формулы [5]

$$\bar{x} - \frac{s}{\sqrt{n-1}} t_{\alpha(n-1)} < a < \bar{x} + \frac{s}{\sqrt{n-1}} t_{\alpha(n-1)},$$

где $t_{\alpha(n-1)}$ – критерий Стьюдента, который определяется по таблицам для уровня значимости α и числа степеней свободы $(n-1)$.

Интервал $(\bar{x} - \frac{s}{\sqrt{n-1}} t_{\alpha(n-1)}, \bar{x} + \frac{s}{\sqrt{n-1}} t_{\alpha(n-1)})$ будет доверительным интервалом для оценки a , отвечающим доверительной вероятности $1 - \alpha$.

Пример 1. Требуется построить $P = 99\%$ доверительный интервал для оценки генерального среднего диаметра a валика по пробе из 10 деталей. Находим средний диаметр валика $x^{cp} = 2$ мк, стандартное отклонение $s = 2,3$ мк. Уровень значимости критерия равен $\alpha = (100 - 99) : 100 = 0,01$, $n = 10$. По таблице Стьюдента для $\alpha = 0,01$ и $n = 10$ находим $t_{0,01;9} = 3,25$.

Рассчитаем $t_{0,01;9} \frac{s}{\sqrt{n-1}} = 3,25 \frac{2,3}{3} = 2,49$. Откуда определяем доверительный интервал для диаметра валика $2 - 2,49 < a < 2 + 2,49$ или $-0,49 < a < 4,49$.

Таким образом, допустимые с надежностью 99% значения параметра a лежат в интервале $(-0,49; 4,49)$.

Формула расчета доверительного интервала дисперсии нормально распределенной генеральной совокупности при неизвестном математическом ожидании [5]

$$\frac{n_s^2}{\chi_{\alpha/2, (n-1)}^2} < \sigma^2 < \frac{n_s^2}{\chi_{1-\alpha/2, (n-1)}^2},$$

где $\chi_{\alpha/2, (n-1)}^2$ – квантиль распределения χ -квадрат с $(n-1)$ степенью свободы порядка $\alpha/2$, определяемый по таблице.

Пример 2. Найти 95% -ный доверительный интервал для математического ожидания твердости сплавов (в условных единицах), если по результатам измерений получены следующие значения: 14,2; 14,8; 14,0; 14,7; 13,9; 14,8; 15,1; 15,0; 14,5. Объем выборки 9.

Выборочное среднее $x_{cp} = (14,2 + 14,8 + \dots 14,5)/9 = 14,56$; выборочная дисперсия $\sigma^2 = (14,2^2 + 14,8^2 + \dots 14,5^2) - 14,56^2 = 0,17$; несмещенная дисперсия $s^2 = 9 \cdot 0,17/8 = 0,19$, $s = 0,43$; доверительная вероятность $p = 0,95$; уровень значимости $\alpha = 0,05$; $1 - \alpha/2 = 1 - 0,05/2 = 0,975$; квантиль распределения Стьюдента $t_{0,975(8)} = 2,306$ (по таблице). Тогда доверительный интервал математического ожидания составит $14,56 - 0,33 < m_x < 14,56 + 0,33$. С вероятностью 0,95 математическое ожидание твердости сплава лежит в пределах от 14,23 до 14,89.

Пример 3 [5]. Требуется определить при доверительной вероятности 0,96 доверительные границы дисперсии высоты штамповок колец подшипника по данным выборки объемом 20 шт. Распределение штамповок по высоте принимается нормальным. Средняя арифметическая высоты штамповок равна $x_{cp} = 32,2975$ мм, $20s^2 = 2,5282$, $\alpha = 1 - 0,96 = 0,04$. По таблице находим $\chi^2_{1-\alpha/2(n-1)} = \chi^2_{0,98(19)} = 8,6$ и $\chi^2_{\alpha/2(n-1)} = \chi^2_{0,02(19)} = 33,7$.

Доверительный интервал для дисперсии высоты штамповок колец подшипника можно записать

$$\frac{2,5282}{33,7} < s^2 < \frac{2,5282}{8,6}; \quad 0,075 < s^2 < 0,294 \text{ или } 0,27 \text{ мм} < s < 0,53 \text{ мм.}$$

Контрольные вопросы

1. Что называют гистограммой частот?
2. Поясните состоятельность, несмещенность и эффективность оценок.
3. Для чего используются точечные оценки параметров генеральной совокупности?
4. Что называют интервальной оценкой параметров генеральной совокупности?
5. Как определить доверительный интервал параметра и доверительную вероятность?
6. Как рассчитать доверительный интервал для математического ожидания нормально распределенной случайной величины?
7. Как определить доверительный интервал дисперсии нормально распределенной генеральной совокупности?

7. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ. ОСНОВНЫЕ ПОНЯТИЯ, ИСПОЛЬЗУЕМЫЕ ПРИ ПРОВЕРКЕ ГИПОТЕЗ

Статистическая гипотеза – любое предположение, касающееся неизвестного распределения случайных величин (элементов), соответствующее некоторым представлениям об изучаемом явлении. В частном случае это может быть утверждение о значениях параметров распределения генеральной совокупности [5, 6].

Различают нулевую и альтернативную гипотезы. Нулевая гипотеза – гипотеза, подлежащая проверке. Альтернативная гипотеза – каждая допустимая гипотеза, отличная от нулевой. Нулевую гипотезу обозначают H_0 , альтернативную – H_1 (от *Hypothesis* – «гипотеза» (англ.)).

Конкретная задача проверки статистической гипотезы полностью описана, если заданы нулевая и альтернативная гипотезы. При обработке реальных данных большое значение имеет правильный выбор гипотез. Принимаемые предположения, например, нормальность распределения, должны быть тщательно обоснованы, в частности, статистическими методами. Необходимо помнить, что в подавляющем большинстве конкретных прикладных задач распределение результатов наблюдений в той или иной степени отлично от нормального.

7.1. Уровень значимости и мощность критерия. Ошибки при проверке гипотез

При проверке статистической гипотезы возможны ошибки. Есть два рода ошибок.

1. Ошибка первого рода заключается в том, что отвергают нулевую гипотезу, в то время как в действительности эта гипотеза верна. Вероятность ошибки первого рода называется уровнем значимости и обозначается α .

2. Ошибка второго рода состоит в том, что принимают нулевую гипотезу, в то время как в действительности эта гипотеза неверна.

Обычно используют не вероятность ошибки второго рода, а ее дополнение до единицы. Эта величина носит название мощности критерия. Итак, мощность критерия – это вероятность того, что нулевая гипотеза будет отвергнута, когда альтернативная гипотеза верна.

Понятия уровня значимости и мощности критерия объединяются в понятие функции мощности критерия – функции, определяющей вероятность того, что нулевая гипотеза будет отвергнута.

Наглядным способом интерпретации ошибок является их графическое представление.

Предположим, что проверяется гипотеза $H_0: \mu_1 = \mu_0$ о равенстве среднего значения генеральной совокупности заданной величине μ_0 (известной, например, из предыдущих экспериментов). Для этого берут выборку объемом n , находят ее среднее арифметическое \bar{x}_B и по его величине судят о справедливости гипотезы H_0 .

Распределение среднего арифметического \bar{x}_B при условии, что верна гипотеза H_0 , будет $f(\bar{x}_B/\mu_0)$. Это распределение качественно представлено на рис. 7.1. Распределение среднего арифметического \bar{x}_B при условии, что верна альтернативная гипотеза $H_1: \mu_1 \neq \mu_0$, будет уже другим.

Будем считать, что гипотеза H_0 отвергается, если выборочное среднее арифметическое \bar{x}_B окажется больше некоторого критического значения, т. е. $\bar{x}_B > K$, как показано на рис. 7.1.

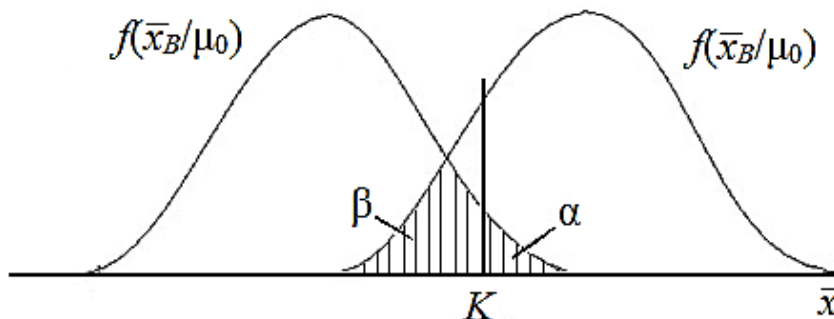


Рис. 7.1. Ошибки первого и второго рода

Область непринятия гипотезы H_0 называется критической областью критерия. Она показана на рисунке наклонной штриховкой α . Уровень значимости будет соответствовать площади критической области.

Вероятность ошибки второго рода β будет равна площади под кривой распределения $f(\bar{x}_B/\mu_1)$, показанной на рисунке вертикальной штриховкой. Величина $1 - \beta$ называется мощностью критерия.

Исследователь всегда должен формулировать гипотезу и задавать уровень значимости до получения экспериментальных данных, по которым эта гипотеза будет проверяться.

При выборе уровня значимости исследователь исходит из практических соображений, отвечая на вопрос, какую вероятность ошибки он считает допустимой для его конкретной задачи? Обычно считают достаточным уровень значимости 0,05 (5 %), иногда 1 или 10 %, редко 0,1 %.

7.2. Статистические критерии

Статистический критерий строится с помощью некоторой статистики $U(x_1, x_2, \dots, x_n)$ – функции от результатов наблюдений x_1, x_2, \dots, x_n . В пространстве значений статистики U выделяют критическую область Ψ , т. е. область со следующим свойством: если значения применяемой статистики принадлежат данной области, то нулевую гипотезу отклоняют (иногда говорят – отвергают), в противном случае – не отклоняют (т. е. принимают).

Статистику U , используемую при построении определенного статистического критерия, называют статистикой этого критерия.

Общая схема проверки гипотез

Процедура проверки гипотез обычно проводится по следующей схеме:

1. Формулируются гипотезы H_0 и H_1 .
2. Выбирается уровень значимости критерия.
3. По выборочным данным вычисляется значение некоторой случайной величины, называемой статистикой критерия, или просто статистическим критерием, который имеет известное стандартное распределение (нормальное, t -распределение Стьюдента и т. п.).
4. Вычисляются критическая область и область принятия гипотезы, т. е. находят критическое (граничное) значение критерия при выбранном уровне значимости.
5. Найденное значение критерия сравнивается с критическим и по результатам сравнения делается вывод: отвергнуть гипотезу или не отвергнуть. Если вычисленное по выборке значение критерия меньше чем критическое, то нулевую гипотезу H_0 не отвергают на заданном уровне значимости.

В этом случае наблюдаемое по экспериментальным данным различие генеральных совокупностей можно объяснить только случайностью выборки. Однако это совсем не означает доказательства равенства параметров генеральных совокупностей. Просто имеющийся в распоряжении статистический материал не дает оснований для отклонения гипотезы о том, что эти параметры одинаковы. Возможно, появится другой экспериментальный материал, на основании которого эта гипотеза будет отклонена.

Если вычисленное значение критерия больше критического, то гипотеза H_0 отклоняется в пользу гипотезы H_1 при данном уровне значимости. В этом случае наблюдаемое различие генеральных совокупностей уже нельзя объяснить только случайностями и говорят, что наблюдаемое различие значимо (статистически значимо) на выбранном уровне значимости.

Следует подчеркнуть разницу между статистической и практической значимостью. Заключение о практической значимости всегда делается человеком, изучающим данное явление. И здесь истинным критерием являются опыт и интуиция исследователя, а статистические критерии значимости – лишь формально точный инструмент, используемый в исследовании. Чем больше исследователь знает об изучаемом явлении, тем точнее будут сформулированная им гипотеза и выводы, сделанные с помощью критериев значимости.

В настоящее время при проверке гипотез, особенно с использованием специализированных программных средств, уровень значимости до эксперимента точно не устанавливается, а по экспериментальным данным вычисляется вероятность P того, что критерий (статистика критерия) выйдет за пределы значения, рассчитанного по выборке. Таким образом, P – это экспериментальный (эмпирический) уровень значимости. Точное значение P обычно не указывают, а окончательные результаты приводят, сравнивая вычисленное значение критерия со стандартными значениями. Если, например, P не превосходит 0,05, то на уровне значимости 5 % различие считается статистически незначимым.

7.3. Критерий значимости при биномиальном распределении

Наиболее распространенной задачей проверки гипотез при биномиальном распределении, когда проводятся повторные независимые испытания, следует рассматривать сравнение вероятности успеха p с заданным значением p_0 , т. е. нулевая гипотеза имеет вид $H_0: p = p_0$.

Предположим, что в серии из n испытаний успех имел место m раз. Тогда при определенных условиях для проверки рассматриваемой нулевой гипотезы можно использовать статистику, имеющую стандартное нормальное распределение:

$$u = \frac{\frac{m}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Пример. При контроле выборки из $n = 100$ деталей $m = 6$ из них оказались дефектными. Можно ли считать, что доля дефектных изделий в партии превышает 3 %? Проверяется нулевая гипотеза $H_0: p = p_0 = 0,03$ при альтернативной гипотезе $H_1: p > p_0$. Примем уровень значимости $\alpha = 0,05$.

Выборочное значение статистики равно

$$U = (6/100 - 0,03)/[0,03(1 - 0,03)/100]^{1/2} = 1,759.$$

Положение границы критической области при правостороннем критерии определяется квантилью $u_{0,95} = 1,645$. Выборочное значение статистики попало в критическую область, нулевая гипотеза отвергается, следовательно, доля дефектных изделий превышает 3 %.

Критерии значимости подразделяются на три типа:

1. Критерии значимости, которые служат для проверки гипотез о параметрах распределений генеральной совокупности (чаще всего нормального распределения). Эти критерии называются параметрическими.

2. Критерии, которые для проверки гипотез не используют предположений о распределении генеральной совокупности. Эти критерии не требуют знания параметров распределений, поэтому называются непараметрическими.

3. Особую группу критериев составляют критерии согласия, служащие для проверки гипотез о согласии распределения генеральной совокупности, из которой получена выборка, с ранее принятой теоретической моделью (чаще всего нормальным распределением).

7.4. Односторонние и двусторонние критерии

Пусть цель исследования в том, чтобы выявить различие параметров двух генеральных совокупностей, которые соответствуют различным ее естественным условиям (условия жизни, возраст испытуемых и т. п.). Зачастую неизвестно, в какой из совокупностей рассматриваемый параметр будет больше, а в какой меньше. Например, если сравнивают средние оценки учащихся в контрольной и экспериментальной группах, то заранее неизвестно, в какой группе средняя оценка будет больше. В этом случае нулевая гипотеза состоит в том, что средние равны между собой, а цель исследования – доказать обратное, т. е. выявить различие между средними. При этом допускается, что различие может быть любого знака. Такие гипотезы называются двусторонними.

Но иногда задача состоит в том, чтобы доказать увеличение или уменьшение параметра; например, средний результат в экспериментальной группе выше (ниже), чем в контрольной. При этом уже не допускается, что различие может быть другого знака. Тогда альтернативная гипотеза $H_1: \mu_2 \geq \mu_1$ при нулевой $H_0: \mu_2 \leq \mu_1$ (или $H_1: \mu_2 < \mu_1$, если нулевая $H_0: \mu_2 \geq \mu_1$). Такие гипотезы называются односторонними.

Пример. Утверждается, что шарики для подшипников, изготовленные автоматическим станком, имеют средний диаметр $d_0 = 10$ мм. Используя односторонний критерий с $\alpha = 0,05$, проверить эту гипотезу, если в выборке из $n = 16$ шариков средний диаметр оказался равным 10,3 мм, а дисперсия известна и равна $s^2 = 1$ мм².

Решение. Нулевая гипотеза: $H_0: \mu = 10$. Вычисляем наблюдаемое значение критерия $U_{\text{набл}} = \frac{(\bar{x} - \mu)}{\sigma} \sqrt{n} = \frac{10,3 - 10}{1} \sqrt{16} = 1,2$.

По таблице функции распределения Лапласа найдем критическую точку для односторонней критической области (при гипотезе $H_1: \mu > 10$) по уровню значимости $\alpha = 0,05$: $\Phi(U_{\text{кр}}) = \frac{1 - 2\alpha}{2} = 0,45$, откуда $U_{\text{кр}} = 1,96$.

Так как $U_{\text{набл}} = 1,2 < 1,96 = U_{\text{кр}}$, то нулевую гипотезу можно принять и считать, что средний диаметр действительно равен $d_0 = 10$ мм.

Критерии значимости, служащие для проверки двусторонних гипотез, называются двусторонними, а для односторонних гипотез –

односторонними. Выбор односторонней или двусторонней гипотезы находится за пределами формальных статистических методов и полностью зависит от целей исследования.

Например, необходимо доказать различие средних значений генеральных совокупностей (средних значений некоторого результата исследований) при двух различных методиках, применяемых в контрольной и экспериментальной группах. Если неизвестно, какая группа покажет в среднем лучший результат, то нужно выдвинуть нулевую гипотезу $H_0: \mu_2 = \mu_1$ против двусторонней альтернативы $H_1: \mu_2 \neq \mu_1$. Различие доказывается по разности средних арифметических результатов в контрольной и экспериментальной группах ($\bar{x}_2 - \bar{x}_1$). Распределение разности $\bar{x}_2 - \bar{x}_1$ при условии, что верна нулевая гипотеза H_0 , схематично представлено на рис. 7.2.

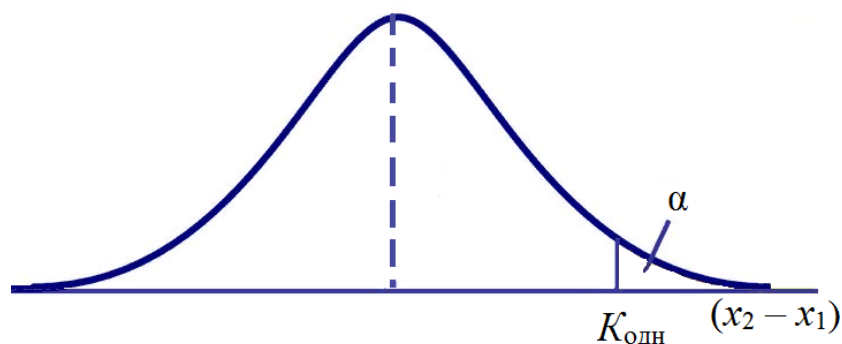


Рис. 7.2. Уровни значимости при одностороннем критерии

Решение об отклонении гипотезы H_0 принимается в том случае, если разность $\bar{x}_2 - \bar{x}_1$ выходит за пределы некоторого значения двустороннего критерия (допустимы отклонения в обе стороны от нуля). Ошибка, которая при этом допускается, равна, как известно, уровню значимости α . Но поскольку отклонения возможны в обе стороны, то при симметричном распределении вероятности отклонений, больших $K_{дв}$ и меньших $K_{дв}$, будут одинаковы и составят $\alpha/2$.

Если предположить, что в экспериментальной группе будут показаны в среднем более высокие результаты, то можно выдвинуть одностороннюю альтернативу $H_1: \mu_2 > \mu_1$. В этом случае при той же нулевой гипотезе $H_0: \mu_2 = \mu_1$ распределение разности $\bar{x}_2 - \bar{x}_1$ будет таким же, как и для двустороннего критерия (рис. 7.3). Но теперь представляют интерес только положительные значения разности $\bar{x}_2 - \bar{x}_1$. Решение об отклонении H_0 принимается, когда $\bar{x}_2 - \bar{x}_1$ окажется больше не-

которого значения одностороннего критерия. При том же уровне значимости α $K_{\text{одн}}$ будет всегда меньше $K_{\text{дв}}$, поэтому нулевая гипотеза будет при одностороннем критерии отклоняться чаще (рис. 7.3).

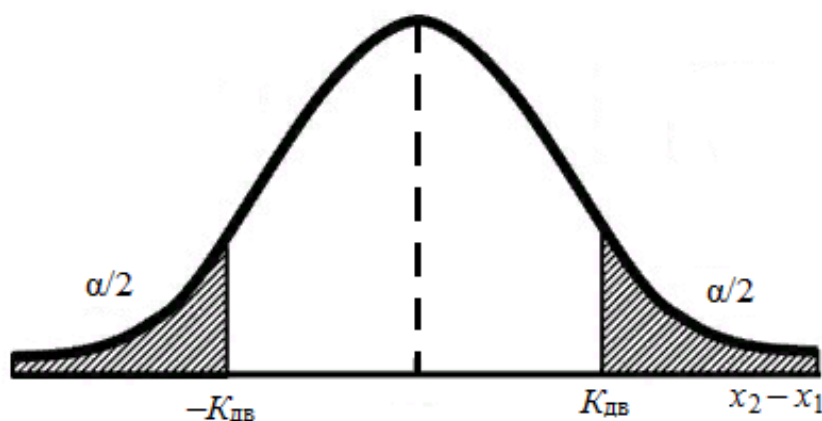


Рис. 7.3. Уровни значимости при двустороннем критерии

Таким образом, двусторонние критерии оказываются более консервативными, чем односторонние. В этом нет никакого противоречия или доказательства несостоятельности статистических методов. Просто в первом случае, используя двустороннюю гипотезу, мы допускали и отрицательный эффект в экспериментальной группе. В такой ситуации выводы должны быть более осторожными, чем в случае односторонней гипотезы, когда имеется дополнительная информация, позволяющая сделать предположение о положительном эффекте новой программы, что, естественно, дает возможность сделать более точный вывод. Правда, следует отметить, что если превышение критического значения в каком-либо исследовании незначительно, то в достоверности вывода о наличии положительного эффекта можно усомниться. В такой ситуации следует провести дополнительные исследования.

7.5. Некоторые типичные задачи проверки параметрических гипотез

Рассмотрим некоторые наиболее часто встречающиеся задачи, решаемые с помощью проверки гипотез [6]:

– задачи сравнения (сравнение выборочных характеристик с нормативными характеристиками);

– сравнение характеристик двух выборок между собой (для проверки гипотезы о принадлежности этих выборок к одной генеральной совокупности).

Типичными непараметрическими задачами считаются:

- проверка гипотез о виде выборочного распределения;
- проверка значимости расхождения выборочных характеристик.

Проверка гипотез о среднем значении

1. Сравнение среднего значения с нормативным.

Такие задачи встречаются при проверке качества продукции, характеризуемого некоторым средним показателем:

- среднее время работы устройства;
- средний размер детали и т. д.

2. Сравнение средних значений двух совокупностей.

Пусть имеются две совокупности, характеризующиеся средними значениями \bar{X} и \bar{Y} , дисперсиями δ_x^2 и δ_y^2 .

Выдвигается гипотеза, что эти средние равны, т. е. $H_0: \bar{X} = \bar{Y}$.

Для проверки основной гипотезы используют критерий

$$\theta = \frac{\bar{X}_\varepsilon - \bar{Y}_\varepsilon}{\sqrt{D(\bar{X}_\varepsilon - \bar{Y}_\varepsilon)}}.$$

Так как $M(\bar{X}_\varepsilon) = \bar{X}$, $M(\bar{Y}_\varepsilon) = \bar{Y}$ при справедливости нулевой гипотезы H_0 будем иметь $M(\theta) = 0$.

Используя свойства дисперсии и предполагая выборки независимыми, получим

$$\delta^2(\bar{X}_\varepsilon - \bar{Y}_\varepsilon) = \delta^2(\bar{X}_\varepsilon) + \delta^2(\bar{Y}_\varepsilon) = \frac{\delta_x^2}{n_1} + \frac{\delta_y^2}{n_2}.$$

Сделав дополнительное предположение, что дисперсии обеих совокупностей равны, т. е. $\delta_x^2 = \delta_y^2 = \delta^2$ получим

$$\delta^2(\bar{X}_\varepsilon - \bar{Y}_\varepsilon) = \delta^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right).$$

Предположение о равенстве дисперсий нуждается в специальной проверке, о чем речь пойдет в следующем разделе. Подставляя это выражение в формулу для критерия, получаем

$$\theta = \frac{\bar{X}_\varepsilon - \bar{Y}_\varepsilon}{\delta \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Если обе выборки достаточно большого объема, то случайные величины \bar{X}_δ и \bar{Y}_δ имеют нормальное распределение, поэтому нормально будет распределен и критерий θ .

Заменяя неизвестную дисперсию генеральной совокупности δ^2 на ее несмещенную выборочную оценку S^2

$$S^2 = \frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{n_1 + n_2 - 2},$$

придем к нормально распределенному критерию

$$Z = \frac{\bar{X}_\varepsilon - \bar{Y}_\varepsilon}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Дальнейшая проверка ведется обычным образом с использованием таблиц функций распределения Лапласа.

Если выборки малого объема и применение нормального распределения могут привести к ошибкам, для того же критерия Z используют t -распределение Стьюдента с числом степеней свободы $l = n_1 + n_2 - 2$.

Пример. Проводится сравнение роста 20-летних юношей, проживающих в Москве и Новосибирске. На основе двух случайных выборок, выполненных в двух городах, были получены следующие данные. В Москве отобрали 75 юношей, по величинам роста которых были вычислены две характеристики: средний рост юношей, который оказался равным 179 см, и стандартное отклонение, которое оказалось равным 8 см; в Новосибирске были случайно отобраны 57 юношей, их средний рост оказался равным 176 см со стандартным отклонением 10 см. На основе этих экспериментальных данных следует проверить гипотезу о примерном равенстве роста московских и новосибирских 20-летних юношей. Принять доверительную вероятность равной 90 %. Предполагается, что рост юношей подчиняется нормальному закону распределения.

Иная постановка вопроса к тем же исходным данным может звучать так: следует выяснить, значимо или незначимо отличаются

друг от друга выборочные средние значения. Если будет показано, что выборочные средние отличаются незначимо, то отсюда можно будет сделать вывод о справедливости нулевой гипотезы о примерном равенстве роста юношей, проживающих в различных городах. В противном случае будет сделан вывод о существенном различии роста юношей из этих городов.

Решение.

Постановка задачи:

$H_0: \bar{x}_r = \bar{y}_r$, здесь \bar{x}_r – средний рост 20-летних юношей Москвы,
 \bar{y}_r – средний рост 20-летних юношей Новосибирска;

$H_1: \bar{x}_r \neq \bar{y}_r$.

При такой постановке задачи следует строить двустороннюю критическую область. Вычислим границы этой области на основе табличного решения уравнения

$$\Phi_0(t_{кр}) = \frac{\gamma}{2} \rightarrow \Phi_0(t_{кр}) = \frac{0,90}{2} = 0,45 \rightarrow t_{кр} \approx 1,65.$$

Вычислим на основе экспериментальной информации наблюдаемое значение критерия

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \rightarrow t_{набл} = \frac{179 - 176}{\sqrt{\frac{8^2}{75} + \frac{10^2}{57}}} \approx 1,86.$$

Изобразим результаты графически (рис. 7.4).

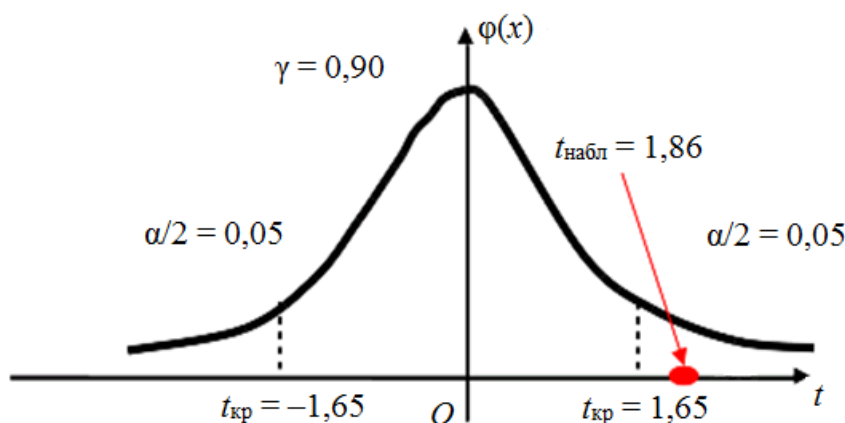


Рис. 7.4. Графические результаты уровней значимости при двустороннем критерии

Поскольку наблюдаемое значение критерия попало в критическую область значений параметра, то следует отвергнуть основную гипотезу в пользу альтернативной гипотезы и сказать, что средний рост московских и новосибирских 20-летних юношей отличается значимо.

Сравнение дисперсий двух совокупностей

Пусть имеются две нормально распределенные совокупности, дисперсии которых равны $\delta_1^2 = \delta_2^2$; нулевая гипотеза $H_0 : \delta_1^2 = \delta_2^2$.

Так как дисперсии генеральных совокупностей неизвестны, гипотеза проверяется на основе сопоставления выборочных дисперсий S_1^2 и S_2^2 . Если отношение $S_1^2 : S_2^2$ близко к 1, нет оснований отклонять нулевую гипотезу, если значительно отличается – гипотеза отклоняется. Для решения вопроса, насколько большим должно быть отличие выборочных дисперсий, чтобы отклонение нулевой гипотезы было достаточно обоснованным, используется отношение

$$F(v_1, v_2) = \frac{\max(S_1^2, S_2^2)}{\min(S_1^2, S_2^2)}.$$

Распределение этого отношения, называемое F -распределением Фишера – Снедекора, зависит от двух параметров: чисел степеней свободы числителя и знаменателя $v_1 = n_1 - 1$ и $v_2 = n_2 - 1$, где n_1 и n_2 – объемы выборок. Числа v_1 и v_2 указываются в фигурных скобках рядом с вычисленным значением F

$$F = \frac{s_2^2}{s_1^2}; \left\{ \begin{array}{l} v_1 \\ v_2 \end{array} \right\}.$$

Критическая область строится в зависимости от вида альтернативной гипотезы:

1. Нулевая гипотеза $H_0 : \delta_1^2 = \delta_2^2$. Альтернативная гипотеза $H_1 : \delta_1^2 > \delta_2^2$, если $(S_1^2 > S_2^2)$.

По заданному α и известным v_1 и v_2 по таблице распределения Фишера – Снедекора находим критическое значение $F_{\text{табл}}$. Проверка гипотезы H_0 сводится к следующему правилу: если отношение выборочных дисперсий $F_{\text{набл}} > F_{\text{табл}}$, гипотеза H_0 отклоняется; если $F_{\text{набл}} < F_{\text{табл}}$, гипотеза H_0 не отклоняется.

2. Альтернативная гипотеза $H_1: \sigma_2^2 \neq \sigma_1^2$.

В этом случае строим симметричную двустороннюю критическую область с критическими точками F_1 и F_2 , определяемыми из неравенств

$$P(F < F_1) = \alpha/2, P(F > F_2) = \alpha/2.$$

Правая критическая точка находится непосредственно по таблице критических точек распределения Фишера – Снедекора для уровня значимости $\alpha/2$ и степеней свободы ν_1 и ν_2 . Левых критических точек F_1 таблица не содержит, но при выбранном симметричном способе построения критической области достигается попадание критерия F в критическую область с вероятностью, равной уровню значимости α . Так как из определения уровня значимости $P(F < F_1) + P(F > F_2) = \alpha$, то, выбирая $P(F > F_2) = \alpha/2$, мы одновременно достигаем и $P(F < F_2) = \alpha/2$.

Гипотеза H_0 проверяется по тому же правилу, что и в случае односторонней критической области, но табличные значения критерия ищут для значения $\alpha/2$, вдвое меньшего, чем заданный уровень значимости: если отношение выборочных дисперсий $F_{\text{набл}} > F_2$, нулевая гипотеза H_0 отклоняется, если $F_{\text{набл}} < F_2$, гипотеза H_0 не отклоняется.

Пример. При обработке валиков на двух станках-автоматах были отобраны две пробы численностью $n_1 = 10$ шт. и $n_2 = 15$ шт. [5]. По данным этих проб были рассчитаны эмпирические дисперсии, оказавшиеся равными $s_1^2 = 9,6$ мк² и $s_2^2 = 5,7$ мк², откуда $F = s_1^2 / s_2^2 = 9,6/5,7 = 1,68$.

Чтобы проверить гипотезу о равенстве дисперсий $H_0: \sigma_1^2 = \sigma_2^2$ (при альтернативной гипотезе $H_1: \sigma_1^2 \neq \sigma_2^2$), необходимо построить критическую область для критерия F . Тогда можно судить о том, будет ли полученное нами значение слишком большим или слишком малым. За критическую область принимают два интервала: больших значений $F > F_2$ и малых значений $F < F_1$. Критическую область подбирают при уровне значимости $P(F > F_2) = \alpha/2$ и $P(F < F_1) = \alpha/2$. Такой выбор критической области обеспечивает большую чувствительность критерия F .

Выбирая уровень значимости $\alpha = 0,1$, находим при $\alpha/2 = 0,1/2 = 0,05$, $\nu_1 = 10 - 1 = 9$ и $\nu_2 = 15 - 1 = 14$ значение $F_{\text{таб}} = 2,65$.

Выборочное значение $F = 1,68 < F_{\text{таб}} = 2,65$ является незначимым, т. е. предположение о равенстве дисперсий не противоречит наблюдениям, т. е. нет еще оснований считать, что станки обладают различной точностью.

7.6. Непараметрические гипотезы. Критерии согласия

Ранее рассматривались методы проверки гипотезы относительно отдельных параметров генерального распределения. Особое место занимают гипотезы относительно согласованности выборочного распределения с теоретическим (генеральным) распределением.

Критерии согласия позволяют ответить на вопрос о том, являются ли различия между выборочным и теоретическим распределением столь незначительными, что они могут быть приписаны влиянию случайных факторов, или нет.

Пусть закон распределения генеральной совокупности неизвестен, но есть основания предполагать, что он имеет определенный вид, в частности:

- если выполняются условия центральной предельной теоремы, есть основание ожидать, что генеральное распределение нормальное;
- если выборочное среднее и выборочная дисперсия равны, то можно предполагать, что генеральная совокупность распределена по закону Пуассона и т. д.

Эти утверждения носят характер гипотез и должны быть подвергнуты статистической проверке.

Для проверки гипотезы H_0 закон распределения имеет данный вид (нормальный, равномерный, показательный), используется специально подобранная случайная величина, которая называется критерием согласия. Критерий согласия есть критерий проверки гипотезы о предполагаемом законе неизвестного распределения.

Имеется несколько критериев согласия:

χ^2 -квадрат Пирсона, критерий Колмогорова, Мизеса – Смирнова и др.

7.7. Критерий Пирсона

Рассмотрим случай, когда выборка представляется интервальным статистическим рядом. Для изучения случайной величины X

проведено n опытов, диапазон наблюдавшихся значений величины X разбит на q интервалов. Ряд распределения имеет вид

Интервал	(x_1, x_2)	(x_2, x_3)	$(x_q, x_q + 1)$
$W_i^* = P_i^* = m_i/n$	P_1^*	P_2^*	P_q^*

m_i – количество экспериментальных данных в i -м интервале, $\sum m_i = n$.

В соответствии с предполагаемым теоретическим законом распределения вычислим вероятности попадания случайной величины в соответствующий интервал $p_i = P(x_i < X < x_{i+1})$ и рассмотрим величину

$$\chi^2 = \sum_{i=1}^q \frac{n}{p_i} (p_i^* - p_i)^2,$$

которая характеризует степень расхождения теоретических и эмпирических данных. Учитывая, что $W_i = P_i^* = \frac{m_i}{n}$, получим

лучим
$$\chi_{\text{набл}}^2 = \sum_{i=1}^q \frac{(m_i - n p_i)^2}{n p_i}.$$

Можно показать, что при $n \rightarrow \infty$ распределение этой случайной величины независимо от того, каков закон распределения генеральной совокупности, стремится к распределению Пирсона χ^2 с числом степеней свободы $\nu = q - 1 - k$, где k – число параметров генерального распределения, оцениваемых на основании наблюдаемых данных. Если проверяется согласие выборочного распределения с распределением Пуассона, единственный параметр которого оценивается по выборочным данным, то $\nu = q - 2$, если проверяется согласие с нормальным распределением, для которого по выборочным данным оцениваются два параметра \bar{X} , σ , то $\nu = q - 3$ и т. д.

При полном совпадении теоретического и экспериментального распределений $\chi^2 = 0$, в противном случае $\chi^2 > 0$. Задавшись уровнем значимости α , находим табличное критическое значение $\chi_{\text{табл}}^2$ при $\chi_{\text{табл}}^2 > \chi_{\text{набл}}^2$ принимаем гипотезу H_0 , при $\chi_{\text{табл}}^2 \leq \chi_{\text{набл}}^2$ отклоняем гипотезу H_0 о виде распределения.

В связи с асимптотическим характером закона Пирсона χ^2 должны выполняться следующие условия [5]:

В связи с асимптотическим характером закона Пирсона χ^2 должны выполняться следующие условия [5]:

- выборка должна образовываться в результате случайного отбора;
- объем выборки n должен быть достаточно большим (практически не менее 50 единиц);
- численность каждой группы должна быть не менее 5 (если это условие не выполняется, производится объединение малочисленных интервалов).

Пример. Отдел технического контроля проверил n партий одно-типных изделий и установил, что число X нестандартных изделий в одной партии имеет эмпирическое распределение, приведенное в табл. 7.1, в одном столбце которой указано количество x_i нестандартных изделий в одной партии, а в другом столбце – количество n_i партий, содержащих x_i нестандартных изделий. Требуется при уровне значимости $\alpha = 0,05$ проверить гипотезу о том, что случайная величина X (число нестандартных изделий в одной партии) распределена по закону Пуассона. Процесс вычислений иллюстрируется в табл. 7.2.

Таблица 7.1

x_i	n_i
0	403
1	370
2	167
3	46
4	12
5	2

Сумма 1000

Таблица 7.2

x_i	n_i	$x_i n_i$
0	403	0
1	370	370
2	167	334
3	46	138
4	12	48
5	2	10

Выборочное среднее равно 0,9.

Примем в качестве оценки $\lambda = x_{cp} = 0,9$. Проверим нулевую гипотезу $H_0 =$ (Число нестандартных изделий в партии распределено по закону Пуассона), т. е. $P(x=k) = \frac{\lambda^k}{k!} e^{-\lambda} = \frac{0,9^k}{k!} e^{-0,9}$, $k = 0, 1, \dots, 5$ при уровне значимости 0,05.

Находим теоретические вероятности $p_k = \frac{0,9^k}{k!} e^{-0,9}$ и теоретические частоты $n'_k = np_k = 1000 p_k$. Результаты вычислений занесем в табл. 7.3.

Таблица 7.3

Результаты вычислений

k	n_k	p_k	n'_k	$(n_k - n'_k)^2$	$(n_k - n'_k)^2/n'_k$
0	403	0,40567	405,67	12,7449	0,031347
1	370	0,36591	365,91	16,7281	0,045716
2	167	0,16466	164,66	5,4756	0,033254
3	46	0,0494	049,4	43,56	0,881781
4	12	0,01111	011,11	0,7921	0,071296
5	2	0,002	2	0	0
				Сумма	1,063395

Из расчетной таблицы находим наблюдаемое значение критерия Пирсона $\chi^2 = 1,06$. Критическая точка для уровня значимости 0,05 при количестве степеней свободы $k = 6 - 2 = 4$ (число групп минус два) равна 9,5. Так как наблюдаемое значение критерия 1,06 меньше критического значения 9,5, следует принять нулевую гипотезу о распределении генеральной совокупности по закону Пуассона.

Контрольные вопросы

1. Какие ошибки возможны при проверке статистических гипотез?
2. Охарактеризуйте общую схему проверки гипотез.
3. На какие типы подразделяются критерии значимости?
4. Поясните проверку гипотез о среднем значении.
5. В чем состоит сравнение дисперсий двух совокупностей?
6. Что понимают под непараметрическими гипотезами?
7. Когда используются критерии согласия?

8. СЛУЧАЙНЫЕ ПРОЦЕССЫ

Случайные сигналы и помехи относятся к случайным явлениям природы, изучением основных закономерностей которых занимается теория вероятностей. Все случайные явления, изучаемые в теории вероятностей, можно разбить на три типа: случайные события, случайные величины и случайные процессы. Каждый из этих типов случайных явлений имеет свои особенности и характеристики.

Для математического описания сигналов и помех необходимо решить две задачи: к какому типу случайных явлений отнести случайный сигнал (помеху) в конкретной ситуации и как определить необходимые вероятностные характеристики? Помехи в системах связи описываются методами теории случайных процессов.

Функция называется случайной, если в результате эксперимента она принимает тот или иной вид, заранее неизвестно, какой именно. Случайным процессом называется случайная функция времени. Конкретный вид, который принимает случайный процесс в результате эксперимента, называется реализацией случайного процесса (рис. 8.1).

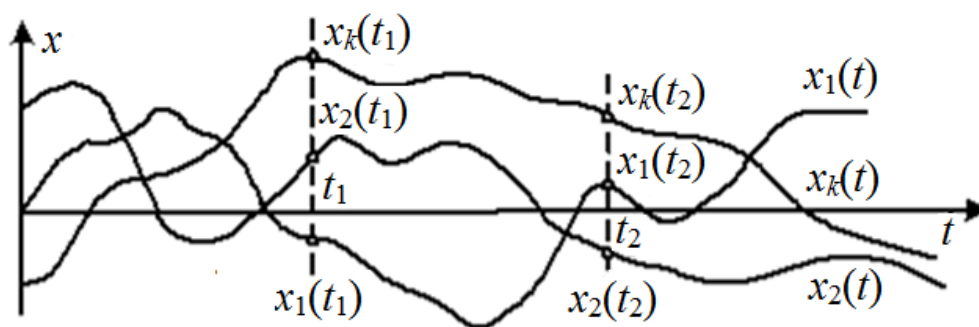


Рис. 8.1. Реализация случайного процесса

Случайный процесс $X(t)$ представляет собой функцию, которая отличается тем, что принимаемые ею значения в любые произвольные моменты времени по координате t являются случайными [7]. Строго с теоретических позиций случайный процесс $X(t)$ следует рассматривать как совокупность временных функций $x_k(t)$, имеющих определенную общую статистическую закономерность. При регистрации случайного процесса на определенном временном интервале осуществляется фиксирование единичной реализации $x_k(t)$ из бесчисленного числа возможных реализаций процесса $X(t)$. Эта единичная реализация называется выборочной функцией случайного процесса $X(t)$. Примеры выборочных функций модельного случайного процесса $X(t)$ приведены на рис. 8.2 [1]. В дальнейшем без дополнительных пояснений при рассмотрении различных параметров и характеристик случайных процессов для сопровождающих примеров будем использовать данную модель процесса.

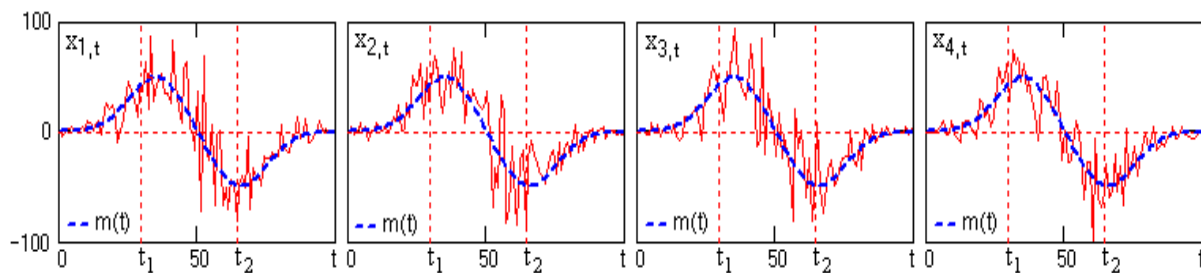


Рис. 8.2. Выборочные функции случайного процесса

С практической точки зрения выборочная функция представляет собой результат отдельного эксперимента, после которого данную реализацию $x_k(t)$ можно считать детерминированной функцией. Сам случайный процесс в целом должен анализироваться с позиции бесконечной совокупности таких реализаций, образующих *статистический ансамбль*. Полной статистической характеристикой такой системы является N -мерная плотность вероятностей $p(x_n; t_n)$. Однако как экспериментальное определение N -мерных плотностей вероятностей процессов, так и их использование в математическом анализе представляет значительные математические трудности. Поэтому на практике обычно ограничиваются одно- и двумерной плотностью вероятностей процессов.

8.1. Функциональные характеристики случайного процесса

Допустим, что случайный процесс $X(t)$ задан ансамблем реализаций $\{x_1(t), x_2(t), \dots, x_k(t), \dots\}$. В произвольный момент времени t_1 зафиксируем значения всех реализаций $\{x_1(t_1), x_2(t_1), \dots, x_k(t_1), \dots\}$ (рис. 8.2).

Совокупность этих значений представляет собой случайную величину $X(t_1)$ и является одномерным сечением случайного процесса $X(t)$.

Одномерная функция распределения вероятностей (x, t_i) определяет вероятность того, что в момент времени t_i значение случайной величины $X(t_i)$ не превысит значения x

$$F(x, t_i) = P\{X(t_i) \leq x\}.$$

Очевидно, что в диапазоне значений вероятностей от 0 до 1 функция $F(x, t)$ является неубывающей с предельными значениями $F(-\infty, t) = 0$ и $F(\infty, t) = 1$. При известной функции $F(x, t)$ вероятность того, что значение $X(t_i)$ в выборках будет попадать в определенный интервал значений $[a, b]$ будет определяться выражением

$$P\{a < X(t_i) \leq b\} = F(b, t_i) - F(a, t_i).$$

Одномерная плотность вероятностей $p(x, t)$ случайного процесса $X(t)$ характеризует распределение вероятностей реализации случайной величины $X(t_i)$ в произвольный момент времени t_i . Она представляет собой производную от функции распределения вероятностей $p(x, t_i) = dF(x, t_i)/dx$.

Моменты времени t_i являются сечениями случайного процесса $X(t)$ по пространству возможных состояний, и плотность вероятностей $p(x, t_i)$ представляет собой плотность вероятностей случайных величин $X(t_i)$ данных сечений. Произведение $p(x, t_i) dx$ равно вероятности реализации случайной величины $X(t_i)$ в бесконечно малом интервале dx в окрестности значения x , откуда следует, что плотность вероятностей также является неотрицательной величиной.

На рис. 8.3 приведены примеры распределения вероятностей и плотности вероятностей сечения случайного процесса $X(t)$ в точке t_1 . Функции вероятностей определены по $N = 1000$ выборок дискретной модели случайного процесса и сопоставлены с теоретическими распределениями при $N \rightarrow \infty$.

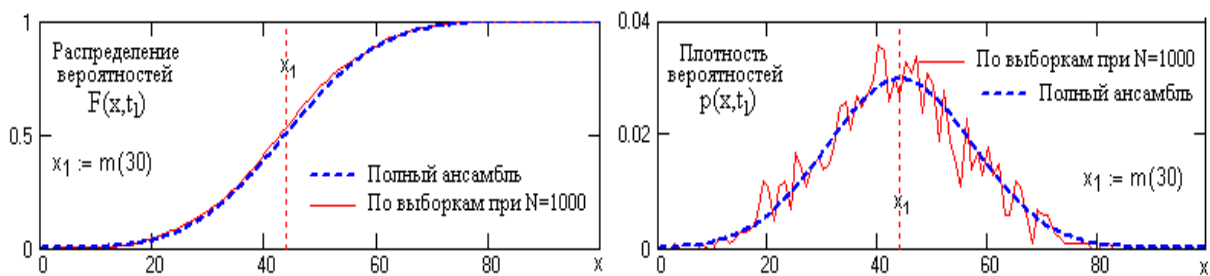


Рис. 8.3. Распределение вероятностей и плотность вероятностей сечения случайного процесса

При известной функции плотности вероятностей вероятность реализации значения $X(t_i)$ в произвольном интервале значений $[a, b]$ вычисляется по формуле

$$P(a < X(t_i) \leq b) = \int_a^b p(x, t_i) dx.$$

Функция плотности вероятностей должна быть нормирована к 1, так как случайная величина обязана принимать какое-либо значение из числа возможных, образующих полное пространство случайных величин:

$$\int_{-\infty}^{\infty} p(x, t) dx = 1.$$

По известной плотности распределения вычисляется и функция распределения вероятностей

$$F(x, t_i) = \int_{-\infty}^{\infty} p(x, t_i) dx = 1.$$

Случайные процессы и их функции характеризуются неслучайными функциями математического ожидания (среднего значения), дисперсии и корреляции.

Математическое ожидание (*mean value*) представляет собой статистическое усреднение случайной величины $X(t_i)$, под которым понимают усреднение по ансамблю реализаций в каком-либо фиксированном сечении t_i случайного процесса. Соответственно функция математического ожидания служит теоретической оценкой среднего взвешенного значения случайного процесса по временной оси

$$m_x(t) = M\{X(t)\} = \int_{-\infty}^{\infty} xp(x; t) dx.$$

Математическое ожидание $m_x(t)$ представляет собой неслучайную составляющую случайного процесса $X(t)$ и соответствует выборкам $N \rightarrow \infty$.

Функция дисперсии (*variance*) случайного процесса является теоретической оценкой среднего взвешенного значения квадрата разности $X(t) - m_x(t)$, которая называется флуктуационной частью процесса:

$$D_x(t) = M\{[X(t) - m_x(t)]^2\} = M\{X^2(t)\} - m_x^2(t) = \int_{-\infty}^{\infty} [x_0(t)]^2 p(x; t) dx,$$

где $x_0(t) = x(t) - m_x(t)$.

Функция среднего квадратического отклонения (*standard deviation*) служит амплитудной мерой разброса значений случайного процесса по временной оси относительно математического ожидания процесса

$$\sigma_x(t) = \sqrt{D_x(t)}.$$

Учитывая последнее выражение, дисперсия случайной величины обычно обозначается индексом σ_x^2 . На рис. 8.4 приведен пример флуктуационной составляющей случайного процесса $X(t)$ в одной из реализаций в сопоставлении со средним квадратическим отклонением $\pm\sigma$ случайных величин от математического ожидания $m(t)$.

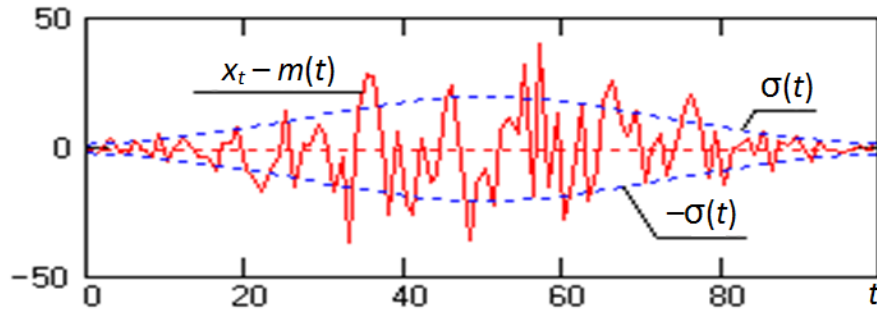


Рис. 8.4. Флуктуационная составляющая случайного процесса

Корреляционные функции случайных процессов. Одномерные законы плотности распределения вероятностей случайных процессов не несут каких-либо характеристик связи между значениями случайных величин для различных значений аргументов.

Двумерная плотность вероятностей $p(x_1, x_2; t_1, t_2)$ определяет вероятность совместной реализации значений случайных величин $X(t_1)$ и $X(t_2)$ в произвольные моменты времени t_1 и t_2 и в какой-то мере уже позволяет оценивать динамику развития процесса. Двумерная плотность вероятностей описывает двумерную случайную величину $\{X(t_i), X(t_j)\}$ в виде функции вероятности реализации случайной величины $X(t_i)$ в бесконечно малом интервале dx_i в окрестностях x_i в момент времени t_i при условии, что в момент времени t_j значение $X(t_j)$ будет реализовано в бесконечно малом интервале dx_j в окрестностях x_j :

$$p(x_i, x_j; t_i, t_j) dx_i dx_j = P\{|X(t_i) - x_i| \leq dx_i/2, |X(t_j) - x_j| \leq dx_j/2\}.$$

Характеристикой динамики изменения двумерной случайной величины $\{X(t_i), X(t_j)\}$ служит корреляционная функция, которая описывает случайный процесс в целом,

$$R_X(t_i, t_j) = M\{X(t_1) X(t_2)\}.$$

Корреляционная функция представляет собой статистически усредненное произведение значений случайного процесса $X(t)$ в моменты времени t_i и t_j по всем значениям временных осей t_i и t_j , а следовательно, тоже считается двумерной функцией. В терминах теории вероятностей корреляционная функция является вторым начальным моментом случайного процесса.

На рис. 8.5 приведены примеры реализаций двух случайных процессов, которые характеризуются одной и той же функцией математического ожидания и дисперсии. На рисунке видно, что хотя про-

странство состояний обоих процессов практически одно и то же, динамика развития процессов в реализациях существенно различается. Единичные реализации коррелированных процессов в произвольный момент времени могут быть такими же случайными, как и некоррелированных, а в пределе во всех сечениях оба процесса могут иметь один и тот же закон распределения случайных величин. Однако динамика развития по координате t (или любой другой независимой переменной) единичной реализации коррелированного процесса по сравнению с некоррелированным является более плавной, а следовательно, в коррелированном процессе имеется определенная связь между последовательными значениями случайных величин.

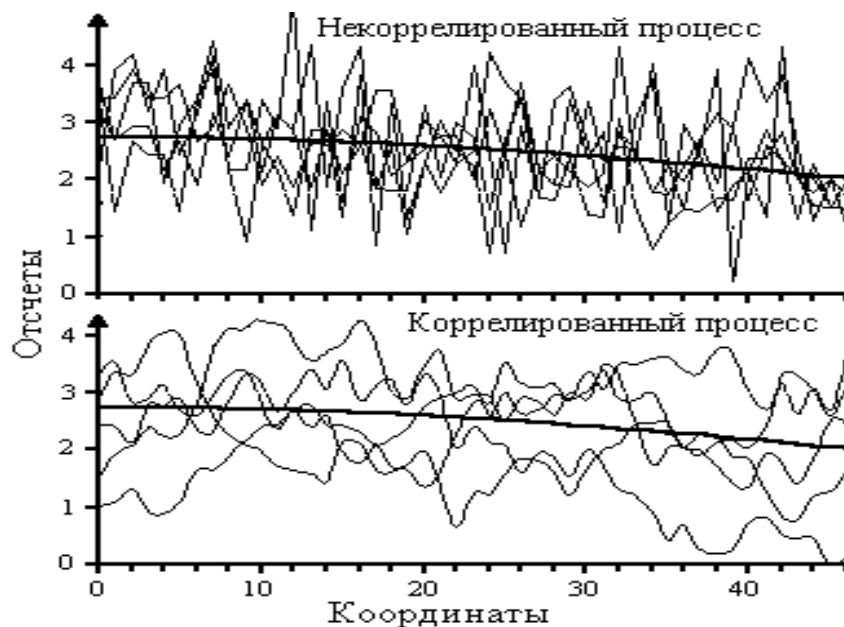


Рис. 8.5. Примеры реализаций двух случайных процессов

Оценка степени статистической зависимости мгновенных значений какого-либо процесса $X(t)$ в произвольные моменты времени t_1 и t_2 и производится функцией корреляции. По всему пространству значений случайного процесса $X(t)$ корреляционная функция определяется выражением

$$R_X(t_i t_j) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(t_i) x(t_j) p(x_i, t_j, x_i, t_j) d_{x_i} d_{x_j}.$$

На рис. 8.6 приведена форма модельного случайного процесса $X(t)$ в одной выборке со значительной и изменяющейся неслучайной

составляющей. Модель задана на интервале $0 - T$ ($T = 100$) в дискретной форме с шагом $\Delta t = 1$. Корреляционная функция вычислена по заданной плотности вероятностей модели.

При анализе случайных процессов второй момент времени t_j удобно задавать величиной сдвига τ относительно первого момента, который при этом может быть задан в виде координатной переменной

$$R_X(t, t + \tau) = M\{X(t)X(t + \tau)\}.$$

Функция, задаваемая этим выражением, обычно называется автокорреляционной функцией случайного процесса.

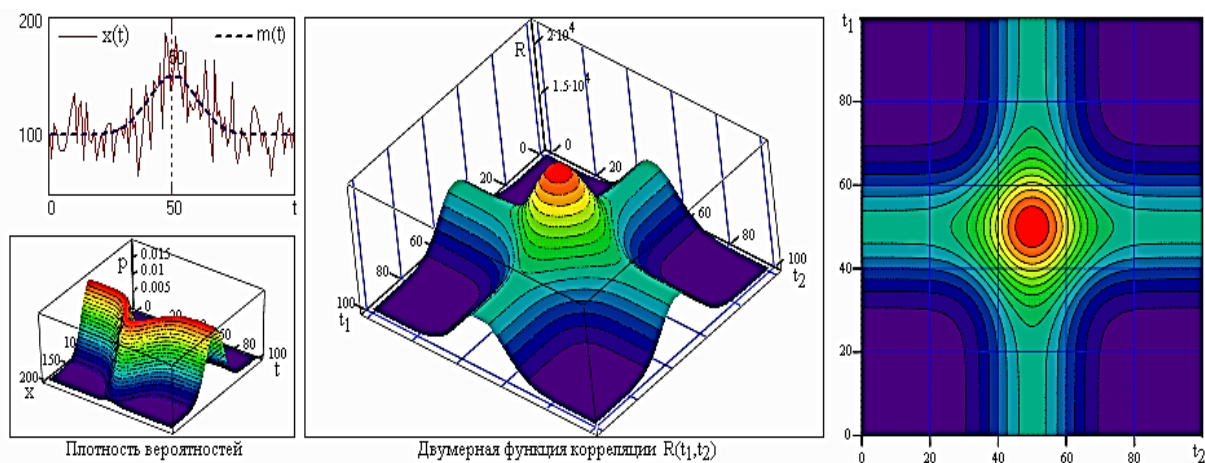


Рис. 8.6. Двумерная плотность вероятностей и корреляционная функция процесса $X(t)$

Ковариационные функции. Частным случаем корреляционной функции будет функция автоковариации (ФАК), которая широко используется при анализе сигналов. Она представляет собой статистически усредненное произведение значений центрированной случайной функции $X(t) - m_x(t)$ в моменты времени t_i и t_j и характеризует флуктуационную составляющую процесса

$$K_X(t_i, t_j) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [x(t_i) - m_x(t_i)] \{x(t_j) - m_x(t_j)\} p(x_i, t_i; x_j, t_j) d_{x_i} d_{x_j}.$$

В терминах теории вероятностей ковариационная функция является вторым центральным моментом случайного процесса. Для центрированных случайных процессов ФАК тождественна функции корреляции. При произвольных значениях m_x ковариационные и корреляционные функции связаны соотношением

$$K_X(t, t + \tau) = R_X(t, t + \tau) - m_x^2(t).$$

Нормированная функция автоковариации (функция корреляционных коэффициентов)

$$\rho_X(t, t + \tau) = K_X(t, t + \tau) / [\sigma(t)\sigma(t + \tau)].$$

При $\tau = 0$ значение ρ_X равно 1, а ФАК вырождается в дисперсию случайного процесса

$$K_X(t) = D_X(t).$$

Отсюда следует, что для случайных процессов и функций основными характеристиками являются функции математического ожидания и корреляции (ковариации). Особой необходимости в отдельной функции дисперсии не имеется. Примеры реализаций двух различных случайных процессов и нормированных ковариационных функций приведены на рис. 8.7.

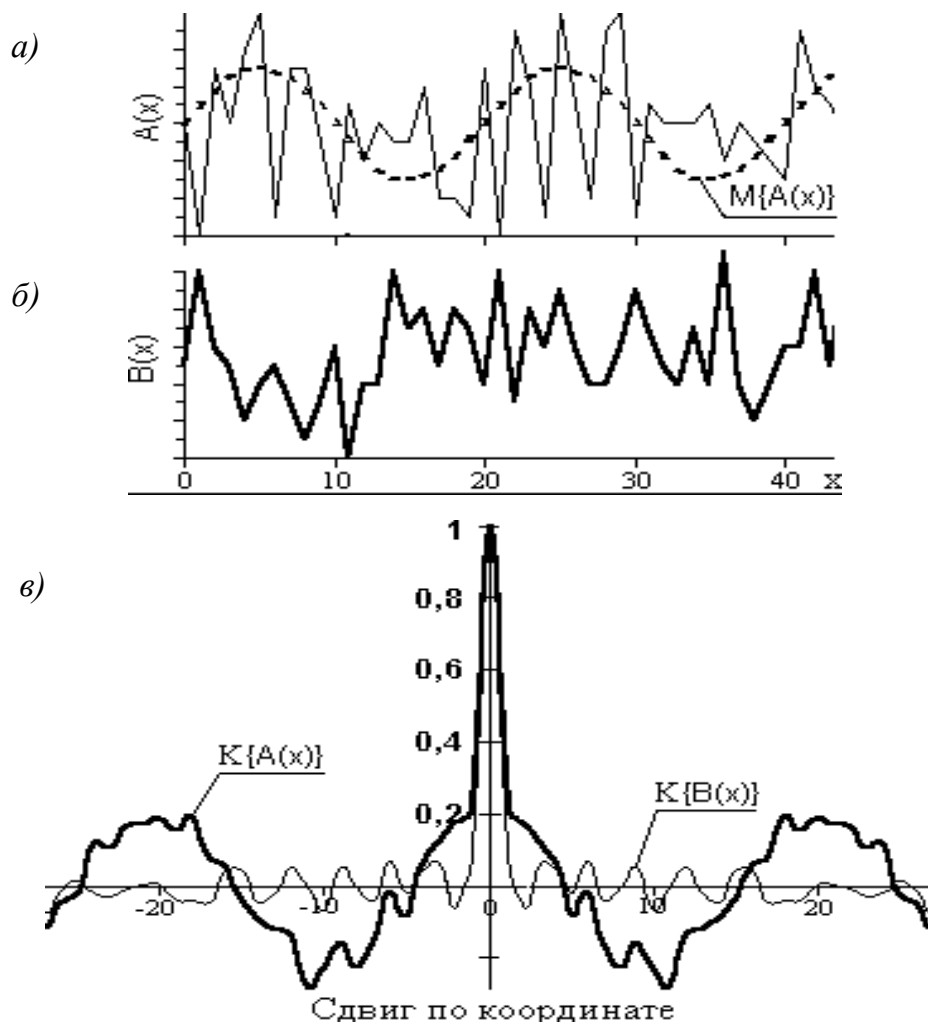


Рис. 8.7. Реализация и ковариационные функции случайных процессов: а – реализация случайного процесса $A(x)$; б – реализация случайного процесса $B(x)$; в – ковариационные функции реализации $A(x)$ и $B(x)$

8.2. Свойства функций автоковариации и автокорреляции

Функции автоковариации и автокорреляции обладают следующими свойствами:

1. Максимум функций наблюдается при $\tau = 0$. Это очевидно, так как при $\tau = 0$ вычисляется степень связи отсчетов с собой же, которая не может быть меньше связи разных отсчетов. Значение максимума функции корреляции равно средней мощности сигнала.

2. Функции автокорреляции и автоковариации являются четными: $R_X(\tau) = R_X(-\tau)$. Последнее также очевидно: $X(t)X(t + \tau) = X(t - \tau)X(t)$ при $t = t - \tau$. Иначе говоря, моменты двух случайных величин $X(t_1)$ и $X(t_2)$ не зависят от последовательности, в которой эти величины рассматриваются, и соответственно симметричны относительно своих аргументов: $R_X(t_1, t_2) = R_X(t_2, t_1)$, равно как и $K_X(t_1, t_2) = K_X(t_2, t_1)$.

3. При $\tau \rightarrow \infty$ значения ФАК для сигналов, конечных по энергии, стремятся к нулю, что прямо следует из физического смысла ФАК. Это позволяет ограничивать длину ФАК определенным максимальным значением τ_{\max} – радиусом корреляции, за пределами которого отсчеты можно считать независимыми. Интегральной характеристикой времени корреляции случайных величин обычно считают эффективный интервал корреляции, определяемый по формуле

$$T_k = 2 \int_0^{\infty} |\rho_x(\tau)| d\tau = (2 / K_x(0)) \int_0^{\infty} |K_x(\tau)| d\tau.$$

4. Отсчеты (сечения) случайных функций, отстоящие друг от друга на расстояние большее T_k , при инженерных расчетах считают некоррелированными.

Заметим, что для некоррелированных процессов при $t \rightarrow \infty$ значение T_k стремится к 2, что несколько противоречит физическому смыслу радиуса корреляции, который в этом случае должен был бы стремиться к 1. С учетом последнего эффективный интервал корреляции целесообразно определять по формуле

$$T_k = 2 \int_0^{\infty} |\rho_x(\tau)| d\tau - 1 = (2 / K_x(0)) \int_0^{\infty} |K_x(\tau)| d\tau - 1.$$

5. Если к случайной функции $X(t)$ прибавить неслучайную функцию $f(t)$, то ковариационная функция не изменяется. Обозначим новую случайную функцию как $Y(t) = X(t) + f(t)$. Функция математи-

ческого ожидания новой величины $\bar{y}(t) = \bar{x}(t) + f(t)$. Отсюда следует, что $Y(t) - \bar{y}(t) = X(t) - \bar{x}(t)$, и соответственно

$$K_y(t_1, t_2) = K_x(t_1, t_2).$$

6. Если случайную функцию $X(t)$ умножить на неслучайную функцию $f(t)$, то ее корреляционная функция $R_x(t_1, t_2)$ умножится на $f(t_1)f(t_2)$. Обоснование данного свойства проводится по методике, аналогичной предыдущему пункту.

7. При умножении функции случайного процесса на постоянное значение C значения ФАК увеличиваются в C^2 раз.

Пример. Случайный процесс определяется формулой $X(t) = X \cos(\omega t)$, где X – случайная величина. Найти основные характеристики этого процесса, если $M(X) = a$, $D(X) = \sigma^2$.

Решение. На основании свойств математического ожидания и дисперсии имеем

$$\begin{aligned} ax(t) &= M(X \cos(\omega t)) = \cos(\omega t) M(X) = a \cos(\omega t), \\ Dx(t) &= D(X \cos(\omega t)) = \cos^2(\omega t) D(X) = \sigma^2 \cos^2(\omega t). \end{aligned}$$

Корреляционную функцию найдем по формуле

$$\begin{aligned} K_x(t_1, t_2) &= M[(X \cos(\omega t_1) - a \cos(\omega t_1)) (X \cos(\omega t_2) - a \cos(\omega t_2))] = \\ &= \cos(\omega t_1) \cos(\omega t_2) M[(X - a)(X - a)] = \cos(\omega t_1) \cos(\omega t_2) D(X) = \\ &= \sigma^2 \cos(\omega t_1) \cos(\omega t_2). \end{aligned}$$

Нормированную корреляционную функцию найдем по формуле

$$\rho_x(t_1, t_2) = \sigma^2 \cos(\omega t_1) \cos(\omega t_2) / (\sigma \cos(\omega t_1))(\sigma \cos(\omega t_2)) \equiv 1.$$

Взаимные моменты случайных процессов второго порядка дают возможность оценить совместные свойства двух случайных процессов $X(t)$ и $Y(t)$ путем анализа произвольной пары выборочных функций $x_k(t)$ и $y_k(t)$.

Мера связи между двумя случайными процессами $X(t)$ и $Y(t)$ также устанавливается корреляционными функциями, а именно функциями взаимной корреляции и взаимной ковариации. В общем случае для произвольных фиксированных моментов времени $t_1 = t$ и $t_2 = t + \tau$ имеем

$$\begin{aligned} R_{XY}(t, t + \tau) &= M\{X(t)Y(t + \tau)\}, \\ K_{XY}(t, t + \tau) &= M\{(X(t) - m_x(t))(Y(t + \tau) - m_y(t + \tau))\}. \end{aligned}$$

Взаимные функции являются произвольными функциями (не обладают свойствами четности или нечетности) и удовлетворяют следующим соотношениям:

$$R_{xy}(-\tau) = R_{yx}(\tau),$$

$$|R_{xy}(\tau)|^2 \leq R_x(0)R_y(0).$$

Если один из процессов центрированный, то имеет место $R_{xy}(t) = K_{xy}(t)$. Нормированная взаимная ковариационная функция (коэффициент корреляции двух процессов), которая характеризует степень линейной зависимости между случайными процессами при данном сдвиге τ одного процесса по отношению ко второму, определяется выражением

$$\rho_{xy}(\tau) = K_{xy}(\tau)/(\sigma_x\sigma_y).$$

Статистическая независимость случайных процессов определяет отсутствие связи между значениями двух случайных величин X и Y . Это означает, что плотность вероятности одной случайной величины не зависит от того, какие значения принимает вторая случайная величина. Двумерная плотность вероятностей при этом должна представлять собой произведения одномерных плотностей вероятностей этих двух величин $p(x, y) = p(x)p(y)$, что является обязательным условием статистической независимости случайных величин. В противном случае между случайными величинами может существовать определенная статистическая связь как линейная, так и нелинейная. Мерой линейной статистической связи выступает коэффициент корреляции

$$r_{xy} = [M\{XY\} - M\{X\}M\{Y\}] / \sqrt{D(X)D(Y)}.$$

Значения r_{xy} могут изменяться в пределах от -1 до $+1$. В частном случае, если случайные величины связаны линейным соотношением $x = ay + b$, коэффициент корреляции равен ± 1 в зависимости от знака константы a . Случайные величины некоррелированы при $r_{xy} = 0$, при этом из выражения для r_{xy} следует

$$M\{XY\} = M\{X\}M\{Y\}.$$

Из статистической независимости величин следует их некоррелированность. Обратное не очевидно. Так, например, случайные величины $x = \cos\varphi$ и $y = \sin\varphi$, где φ – случайная величина с равномерным распределением в интервале $0 \dots 2\pi$, имеют нулевой коэффициент корреляции, и вместе с тем их зависимость очевидна.

8.3. Классификация случайных процессов

Случайные процессы различают по степени однородности их протекания во времени (по аргументу).

В общем случае значения функций математического ожидания, дисперсии и корреляции могут быть зависимыми от момента времени t , т. е. изменяться во времени. Такие процессы составляют *класс нестационарных процессов*.

Процесс называют *стационарным*, если плотность вероятностей процесса не зависит от начала отсчета времени и если на интервале его существования выполняются условия постоянства математического ожидания и дисперсии, а корреляционная функция является функцией только разности аргументов $\tau = t_2 - t_1$, (рис. 8.8), т. е.

$$m_X(t_1) = m_X(t_2) = m_X = \text{const},$$

$$D_X(t_1) = D_X(t_2) = D_X = \text{const},$$

$$R_X(t_1, t_1 + \tau) = R_X(t_2 - \tau, t_2) = R_X(\tau) = R_X(-\tau),$$

$$r_x(\tau) = R_X(\tau)/D_X,$$

$$r_x(0) = 1, |r_x(\tau)| \leq 1, r_x(-\tau) = r_x(\tau).$$

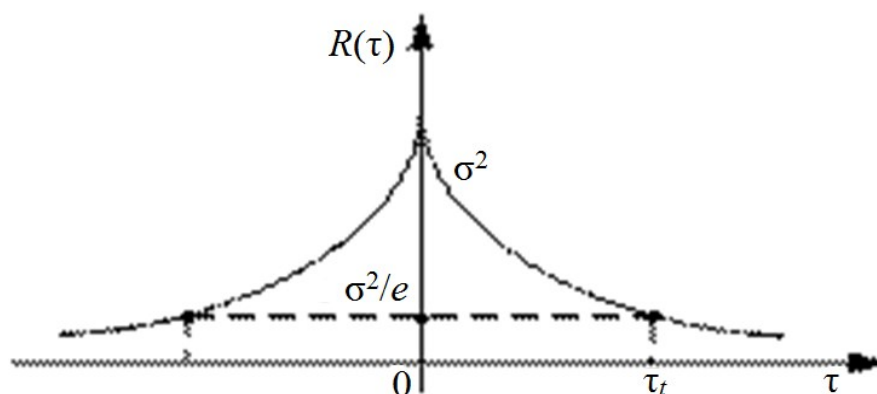


Рис. 8.8. Корреляционная функция стационарного процесса

Последние выражения свидетельствуют о четности корреляционной (а равно и ковариационной) функции и функции корреляционных коэффициентов. Из него вытекает также еще одно свойство смешанных моментов стационарных процессов:

$$|R_x(\tau)| \leq R_x(0), |K_x(\tau)| \leq K_x(0) = D_x.$$

Чем медленнее по мере увеличения значений τ убывают функции $R_x(\tau)$ и $r_x(\tau)$, тем больше эффективный интервал корреляции случайного процесса и тем медленнее изменяются во времени его реализации.

Среди стационарных процессов выделяют строго стационарные процессы, для которых постоянны во времени не только математическое ожидание, дисперсия и корреляция, но и все остальные моменты высших порядков (в частности, асимметрия и эксцесс).

Стационарные случайные процессы наиболее часто встречаются при решении физических и технических задач. Теория стационарных случайных функций разработана наиболее полно, и для ее использования обычно достаточно определения стационарности в широком смысле: случайная функция считается стационарной, если ее математическое ожидание постоянно, а корреляционная функция зависит только от одного аргумента. Случайные процессы, удовлетворяющие условиям стационарности на ограниченных, интересующих нас интервалах, также обычно относят к числу стационарных в широком смысле и называют квазистационарными.

Эргодические процессы. Строго корректно характеристики случайных процессов оцениваются путем усреднения по ансамблю реализаций в определенные моменты времени (по сечениям процессов). Но большинство стационарных случайных процессов обладает эргодическим свойством. Сущность его заключается в том, что по одной достаточно длинной реализации процесса можно судить обо всех его статистических свойствах так же, как по любому количеству реализаций. Другими словами, закон распределения случайных величин в таком процессе может быть одним и тем же как по сечению для ансамбля реализаций, так и по координате развития. Такие процессы получили название эргодических (ergodic). Для эргодических процессов имеет место

$$\begin{aligned}
 m_X(t) &= M\{x(t)\} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t) dt, \\
 D_X(t) &= M[x(t) - m_X(t)]^2 = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [x(t) - m_X(t)]^2 dt, \\
 R_X(\tau) &= M[x(t)x(t + \tau)] = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [x(t)x(t + \tau)] dt.
 \end{aligned} \tag{8.1}$$

Эргодичность – очень важное свойство случайных стационарных и только стационарных процессов. Математическое ожидание эргодического случайного процесса равно постоянной составляющей

любой его реализации, а дисперсия является мощностью его флуктуационной составляющей. Так как определение функций производится по ограниченным статистическим данным одной реализации и выступает только определенным приближением к соответствующим фактическим функциям процессов, целесообразно называть эти функции статистическими. Заметим, что, как это следует из (8.1), вычисление корреляционной функции подобно свертке (с делением на интервал реализации) и может записываться символически

$$R_x(\tau) = (1/T) \int_0^T x(t) x(t + \tau) dt.$$

Свойства эргодичности могут проявляться только по отношению к двум первым моментам случайного процесса, что вполне достаточно для использования соответствующих методик исследования процессов. Практическая проверка эргодичности процесса обычно производится проверкой выполнения условия Слуцкого

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T K(\tau) dt = 0.$$

Если ковариационная функция процесса стремится к нулю при возрастании значения аргумента (τ), то процесс относится к числу эргодических, по крайней мере, относительно моментов первого и второго порядков.

Примером случайного процесса может служить флуктуационный шум, наиболее характерный для большинства каналов электро-связи. Для количественных расчетов воздействия флуктуационного шума на сигнал необходимо знать основные вероятностные характеристики. Поскольку шум образуется как сумма большого числа отдельных независимых колебаний, он согласно центральной предельной теореме представляет собой стационарный эргодический случайный процесс с гауссовским (нормальным) распределением вероятности. Гауссовский процесс описывается формулой

$$w(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left[-\frac{(x - m)^2}{2\sigma^2} \right],$$

в которую входят два числовых параметра m и σ^2 , имеющие смысл математического ожидания и дисперсии. График плотности вероятности $w(x)$ представляет собой колоколообразную кривую с единствен-

ным максимумом в точке $x = m$ (рис. 8.9). Из графика видно, что с уменьшением σ кривая все более локализуется в окрестности точки $x = m$. Для флуктуационного шума обычно $M(X) = 0$.

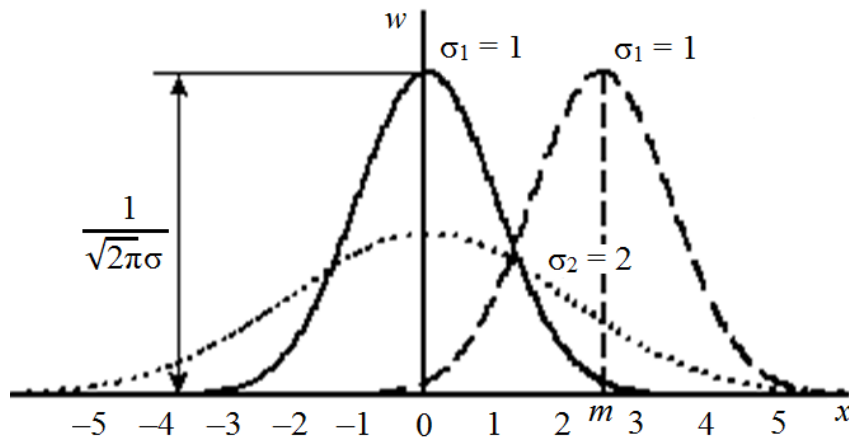


Рис. 8.9. График плотности вероятности гауссовского случайного процесса

Функция распределения вероятности для гауссовского случайного процесса имеет вид

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

После замены переменных $y = (x - m)/\sigma$ эта функция приводится к виду

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-m)/\sigma} \exp\left[-\frac{y^2}{2}\right] dy = 0,5 + \Phi_0\left(\frac{x-m}{\sigma}\right),$$

где $\Phi_0(z) = \frac{1}{\sqrt{2\pi}} \int_0^z \exp(-\frac{y^2}{2}) dy$ — интеграл вероятности.

Функция $\Phi_0(z)$ табулирована в математических справочниках. Заметим, что $\Phi_0(-z) = -\Phi_0(z)$, $\Phi_0(0) = 0$, $\Phi_0(\infty) = 0,5$. Для приближенных вычислений можно воспользоваться приближенным выражением

$$\Phi_0(z) \approx 0,5 - 0,65 \exp\left[-0,44(z + 0,75)^2\right].$$

Пример. Случайная функция задана выражением $Z(t) = X(t) + Y$, где $X(t)$ — стационарная эргодичная функция; Y — случайная величина, некоррелированная с $X(t)$. Эргодична ли функция $Z(t)$?

$$m_z(t) = m_x(x) + m_y, K_z(\tau) = K_x(\tau) + D_y.$$

Функция $Z(t)$ стационарна, но не эргодична, так как при $\tau \rightarrow \infty$ имеет место $K_z(\tau) \rightarrow D_y$.

8.4. Энергетический спектр случайного процесса

Другой важной характеристикой следует назвать энергетический спектр случайного процесса. Он определяется как преобразование Фурье от корреляционной функции [8]

$$G(\omega) = \int_{-\infty}^{\infty} R(\tau) e^{-j\omega\tau} d\tau.$$

Очевидно, справедливо и обратное преобразование

$$R(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega) e^{j\omega\tau} d\omega.$$

Энергетический спектр показывает распределение мощности случайного процесса, например помехи, на оси частот.

При анализе систем автоматического управления (САУ) очень важно определить характеристики случайного процесса на выходе линейной системы при известных характеристиках процесса на входе САУ. Предположим, что линейная система задана импульсной переходной характеристикой $h(\tau)$. Тогда выходной сигнал в момент времени t_1 определяется интегралом Дюамеля

$$x(t_1) = \int_{-\infty}^{\infty} h(\tau_1) g(t_1 - \tau_1) d\tau_1.$$

где $g(t)$ – процесс на входе системы.

Для нахождения корреляционной функции $R_x(t_1, t_2) = M[x(t_1)x(t_2)]$ запишем

$$x(t_2) = \int_{-\infty}^{\infty} h(t_2) g(t_2 - \tau_2) d\tau_2$$

и после перемножения найдем математическое ожидание

$$R_x(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(\tau_1) h(\tau_2) M\{g(t_1 - \tau_1) g(t_2 - \tau_2)\} d\tau_1 d\tau_2.$$

Таким образом, связь между корреляционными функциями входного и выходного случайных процессов устанавливается с помощью следующего двойного интеграла:

$$R_x(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(\tau_1)h(\tau_2)R_g(t_1 - \tau_1, t_2 - \tau_2)d\tau_1d\tau_2.$$

Более простое соотношение можно найти для энергетических спектров $G_s(w)$ и $G_x(w)$ входного и выходного сигналов при известной передаточной функции $W(jw)$ линейной системы.

Поскольку преобразование Фурье от импульсной характеристики дает передаточную функцию, окончательно находим связь между энергетическими спектрами процессов на входе и выходе линейной системы

$$G_x(\omega) = W(j\omega)W(-j\omega)G_g(\omega) = |W(j\omega)|^2 G_g(\omega).$$

8.5. Нормальный случайный процесс

Среди многообразия случайных процессов наиболее широко распространенным представляется нормальный случайный процесс [8]. Нормальный случайный процесс $X(t)$ в любом сечении его реализации $x(t)$ характеризуется плотностью вероятности

$$w(x) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-\frac{(x-m_x)^2}{2\sigma_x^2}}. \quad (8.2)$$

Для нормального случайного процесса все числовые характеристики, кроме математического ожидания и дисперсии, равны нулю. Поскольку дисперсия представляет собой значение автокорреляционной функции при $\tau = 0$, то нормальный случайный процесс полностью определяется математическим ожиданием m_x и автокорреляционной функцией $\rho_x(\tau)$. Поэтому (8.2) можно представить следующим образом:

$$w(x) = \frac{1}{\sqrt{2\pi B_x(0)}} e^{-\frac{(x-m_x)^2}{2B_x(0)}},$$

где $B_x(0) = \rho_x(0)$.

Двумерная плотность распределения вероятностей значений x_1 и x_2 нормального случайного процесса, разделенных интервалом времени τ , имеет вид

$$w(x_1, x_2) = \frac{1}{2\pi B_x(0)\sqrt{1-\rho_x(\tau)}} \exp \left\{ -\frac{1}{2B_x(0)[1-\rho_x^2(\tau)]} \left[(x_1 - m_x)^2 + \right. \right.$$

$$\left. + (x_2 - m_x)^2 - 2\rho_x(\tau)(x_1 - m_x)(x_2 - m_x) \right\},$$

где $\rho_x(\tau) = \frac{B_x(\tau)}{D_x}$ – нормированная автокорреляционная функция процесса.

Следует отметить, что для нормального случайного процесса некоррелированность двух значений, разделенных интервалом времени τ , означает и их статистическую независимость.

Контрольные вопросы

1. Что понимают под выборочной функцией случайного процесса?
2. Что определяет одномерная функция распределения вероятностей?
3. Что характеризует одномерная плотность вероятностей?
4. Что определяет функция математического ожидания случайного процесса?
5. Что характеризует функция дисперсии случайного процесса?
6. Дайте определение функции среднего квадратического отклонения случайного процесса.
7. Охарактеризуйте корреляционные функции случайных процессов.
8. Охарактеризуйте ковариационные функции случайных процессов.
9. Какими свойствами обладают функций автоковариации и автокорреляции?
10. Что понимают под нестационарным случайным процессом?
11. Что понимают под стационарным случайным процессом?
12. Охарактеризуйте эргодические случайные процессы.
13. Как образуется флуктуационный шум?
14. Как определяется энергетический спектр случайного процесса?
15. Как определяются характеристики случайного процесса на выходе линейной системы?
16. Чем характеризуется нормальный случайный процесс?

9. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Корреляционный анализ, разработанный К. Пирсоном и Дж. Юлом, является одним из методов статистического анализа взаимозависимости нескольких переменных – компонент случайного вектора.

Основным показателем взаимосвязи двух случайных величин следует рассматривать парный коэффициент корреляции, служащий мерой линейной статистической зависимости между этими величинами, когда статистическая связь между соответствующими признаками в генеральной совокупности линейна.

Указанное условие выполняется, если генеральная совокупность распределена по многомерному нормальному закону.

9.1. Функциональные и корреляционные связи между переменными

Экспериментальные данные представляют собой количественные характеристики каких-либо объектов или процессов. Они формируются под действием множества факторов, не все из которых доступны контролю. Неконтролируемые факторы могут принимать случайные значения из некоторого множества значений и тем самым обуславливать случайность данных, которые они определяют. Стохастическая природа экспериментальных данных определяет необходимость применения соответствующих статистических методов для их обработки и анализа.

Статистические распределения характеризуются наличием более или менее значительной вариации в величине измеренных переменных у отдельных единиц совокупности. Возникает вопрос о том, какие же причины формируют уровень переменной в данной совокупности и каков конкретный вклад каждой из них. Изучение зависимости вариации переменных от окружающих условий и составляет содержание теории корреляции.

Изучение действительности показывает, что вариация каждой переменной находится в тесной связи с вариацией других переменных, характеризующих исследуемую совокупность единиц. Вариация уровня производительности труда работников предприятий зависит

от степени совершенства применяемого оборудования, технологии, организации производства, труда и управления и других самых различных факторов.

При изучении конкретных зависимостей одни переменные выступают в качестве факторов, обуславливающих изменение других переменных. Переменные этой первой группы будем называть факторными переменными, а переменные, которые являются результатом влияния этих факторов, будем называть зависимыми (результативными).

Например, при изучении зависимости между производительностью труда рабочих и энерговооруженностью их труда уровень производительности труда будет результативной переменной, а энерговооруженность труда рабочих – факторной переменной.

Рассматривая зависимости между переменными, необходимо выделить, прежде всего, две категории зависимости: 1) функциональные и 2) корреляционные.

Функциональные связи характеризуются полным соответствием между изменением факторной переменной и изменением результативной величины, и каждому значению факторной переменной соответствуют вполне определенные значения результативной переменной. Функциональная зависимость может связывать результативную переменную с одним или несколькими факторными переменными. Так, величина начисленной заработной платы при повременной оплате труда зависит от количества отработанных часов.

В корреляционных связях между изменением факторной и результативной переменными нет полного соответствия, воздействие отдельных факторов проявляется лишь в среднем при массовом наблюдении фактических данных. Одновременное воздействие на изучаемую переменную большого количества самых разнообразных факторов приводит к тому, что одному и тому же значению факторной переменной соответствует целое распределение значений результативной переменной, поскольку в каждом конкретном случае прочие факторные переменные могут изменять силу и направленность своего воздействия.

При сравнении функциональных и корреляционных зависимостей следует иметь в виду, что при наличии функциональной зависимости между переменными можно, зная величину факторной переменной,

точно определить величину результативной переменной. При наличии же корреляционной зависимости устанавливается лишь тенденция изменения результативной переменной при изменении величины факторной переменной. В отличие от жесткости функциональной связи корреляционные связи характеризуются множеством причин и следствий и устанавливаются лишь их тенденции.

9.2. Задачи корреляционного анализа

Основная задача корреляционного анализа заключается в выявлении взаимосвязи между случайными переменными путем точечной и интервальной оценки парных (частных) коэффициентов корреляции, вычисления и проверки значимости множественных коэффициентов корреляции и детерминации. Кроме того, с помощью корреляционного анализа решаются следующие задачи:

- отбор факторов, оказывающих наиболее существенное влияние на результативную переменную, на основании измерения степени связи между ними;
- обнаружение ранее неизвестных причинных связей.

Корреляция непосредственно не выявляет причинных связей между переменными, но устанавливает численное значение этих связей и достоверность суждений об их наличии.

Выборочная ковариация является мерой взаимосвязи между двумя переменными. Ковариация между двумя переменными X и Y рассчитывается следующим образом:

$$\text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

где $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ – фактические значения случайных переменных x и y . Средние арифметические значения переменных

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Ковариация – это статистическая мера взаимодействия двух случайных переменных, таких, например, как доходности двух ценных бумаг. Положительное значение ковариации показывает, что доходности этих ценных бумаг имеют тенденцию изменяться в одну сторону.

Ковариация зависит от единиц, в которых измеряются переменные X и Y . Поэтому для измерения силы связи между двумя переменными используется другая статистическая характеристика, называемая коэффициентом корреляции.

При проведении корреляционного анализа вся совокупность данных рассматривается как множество переменных (факторов), каждая из которых содержит n наблюдений; x_{ik} – i -е наблюдение k -й переменной. Основными средствами анализа данных являются парные коэффициенты корреляции, частные коэффициенты корреляции и множественные коэффициенты корреляции.

Коэффициент парной корреляции

Для двух переменных X и Y теоретический коэффициент корреляции определяется следующим образом:

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sqrt{\sigma_x^2 \sigma_y^2}},$$

где σ_x^2 , σ_y^2 – дисперсии случайных переменных X и Y ; ρ_{xy} – их ковариация.

Парный коэффициент корреляции служит показателем тесноты связи лишь в случае линейной зависимости между переменными и обладает следующими основными свойствами:

1. Коэффициент корреляции принимает значение в интервале $(-1, +1)$, или $|\rho_{xy}| < 1$.

2. Коэффициент корреляции не зависит от выбора начала отсчета и единицы измерения, т. е. $\rho(\alpha_1 X + \beta; \alpha_2 Y + \beta) = \rho_{xy}$, где $\alpha_1, \alpha_2, \beta$ – постоянные величины, причем $\alpha_1 > 0, \alpha_2 > 0$.

3. Случайные величины X, Y можно уменьшать (увеличивать) в α раз, а также вычитать или прибавлять к значениям X и Y одно и то же число β – это не приведет к изменению коэффициента корреляции ρ .

При $\rho = \pm 1$ случайные величины X и Y связаны линейной зависимостью, т. е. $Y = \alpha X + \beta$. При $\rho = 0$ линейная корреляционная связь отсутствует. В практических расчетах коэффициент корреляции ρ генеральной совокупности обычно не известен. По результатам выборки может быть найдена его точечная оценка – выборочный коэффициент корреляции r , так как выборочная совокупность переменных X и Y случайна, то в отличие от параметра ρ r – случайная величина.

Оценкой коэффициента корреляции ρ является выборочный парный коэффициент корреляции

$$r_{x,y} = \frac{\text{cov}(X,Y)}{S_x S_y} = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_x S_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}},$$

где $S_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$, $S_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$ – оценки дисперсий величин X и Y .

Проверка значимости парного коэффициента корреляции

Для оценки значимости коэффициента корреляции применяется t -критерий Стьюдента. При этом фактическое значение этого критерия определяется по формуле

$$t_{\text{набл}} = \sqrt{\frac{r_{y,x}^2}{1 - r_{y,x}^2}} (n - 2).$$

Вычисленное по этой формуле значение $t_{\text{набл}}$ сравнивается с критическим значением t -критерия, которое берется из таблицы значений t -критерия Стьюдента с учетом заданного уровня значимости и числа степеней свободы.

Если $t_{\text{набл}} > t_{\text{кр}}$, то полученное значение коэффициента корреляции признается значимым, т. е. нулевая гипотеза, утверждающая равенство нулю коэффициента корреляции, отвергается. И таким образом делается вывод о том, что между исследуемыми переменными есть тесная статистическая взаимосвязь.

9.3. Многомерный корреляционный анализ

Коэффициенты парной корреляции используются для измерения силы линейных связей различных пар переменных из их множества [7]. Для множества m переменных n наблюдений получают матрицу коэффициентов парной корреляции R .

$$R = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1m} \\ & r_{22} & \dots & r_{2m} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & r_{mm} \end{bmatrix}. \quad (9.1)$$

Одной корреляционной матрицей нельзя полностью описать зависимости между величинами. В связи с этим в многомерном корреляционном анализе рассматриваются две задачи:

1. Определение тесноты связи одной случайной величины с совокупностью остальных $(m - 1)$ величин, включенных в анализ.
2. Определение тесноты связи между величинами при фиксировании или исключении влияния остальных k величин, при $k < (m - 2)$.

Эти задачи решаются с помощью коэффициентов множественной и частной корреляции соответственно.

Множественный коэффициент корреляции

Первая задача решается с помощью выборочного коэффициента множественной корреляции по формуле

$$R_{j,1,2,\dots,j-1,j+1,\dots,m} = \sqrt{1 - \frac{|R|}{R_{jj}}},$$

где $|R|$ – определитель корреляционной матрицы R (9.1); R_{jj} – алгебраическое дополнение элемента r_{jj} той же матрицы R .

Квадрат коэффициента множественной корреляции $R_{j,1,2,\dots,j-1,j+1,\dots,m}^2$ принято называть выборочным множественным коэффициентом детерминации, который показывает, какую долю вариации (случайного разброса) исследуемой величины X_j объясняет вариация остальных случайных величин X_1, X_2, \dots, X_m .

Коэффициенты множественной корреляции и детерминации – величины положительные, принимающие значения в интервале от 0 до 1. При приближении коэффициента R^2 к единице можно сделать вывод о тесноте взаимосвязи случайных величин, но не о ее направлении. Коэффициент множественной корреляции может только увеличиваться, если в модель включать дополнительные переменные, и не увеличится, если из имеющихся переменных производить исключение.

Значимость коэффициента множественной корреляции проверяется сравнением расчетного значения критерия Фишера

$$F_{\text{расч}} = \frac{R^2 / (n - m)}{(1 - R^2) / (m - 1)}$$

с табличным $F_{\text{табл}}$. Табличное значение критерия

определяется заданным уровнем значимости α и степенями свободы $k_1 = m - 1$ и $k_2 = n - m$. Коэффициент R^2 значимо отличается от нуля, если выполняется неравенство $F_{\text{расч}} > F_{\text{табл}}$.

Частный коэффициент корреляции

Если рассматриваемые случайные величины коррелируют друг с другом, то на величине коэффициента парной корреляции частично сказывается влияние других величин. В связи с этим возникает необходимость исследования частной корреляции между величинами при исключении влияния одной или нескольких других случайных величин.

Выборочный частный коэффициент корреляции определяется по формуле

$$r_{jk.1,2,\dots,m} = \frac{R_{jk}}{\sqrt{R_{jj}R_{kk}}}, \quad (9.2)$$

где R_{jk} , R_{jj} , R_{kk} – алгебраические дополнения к соответствующим элементам матрицы (9.1).

Частный коэффициент корреляции, так же как и парный коэффициент корреляции, изменяется от -1 до $+1$.

Пример вычисления коэффициентов парной, множественной и частной корреляции. В табл. 9.1 приведена информация об объемах продаж и затратах на рекламу одной фирмы, а также индекс потребительских расходов за ряд текущих лет.

Таблица 9.1

Объемы продаж и затраты на рекламу одной фирмы

Объем продаж Y , тыс. руб.	126	137	148	191	274	370	432	445	367	367	321	307	331	345
Затраты на рекламу X_1 , тыс. руб.	4	4,8	3,8	8,7	8,2	9,7	14,7	18,7	19,8	10,6	8,6	6,5	12,6	6,5
Индекс потребительских расходов X_2 , %	100	98,4	101,2	103,5	104,1	107	107,4	108,5	108,3	109,2	110,1	110,7	110,3	111,8

Требуется:

- построить диаграмму рассеяния (корреляционное поле) для переменных «объемы продаж» и «индекс потребительских расходов»;
- определить степень влияния индекса потребительских расходов на объемы продаж (вычислить коэффициент парной корреляции);
- оценить значимость вычисленного коэффициента парной корреляции;
- построить матрицу коэффициентов парной корреляции по трем переменным;
- рассчитать множественный коэффициент корреляции;
- найти оценки коэффициентов частной корреляции.

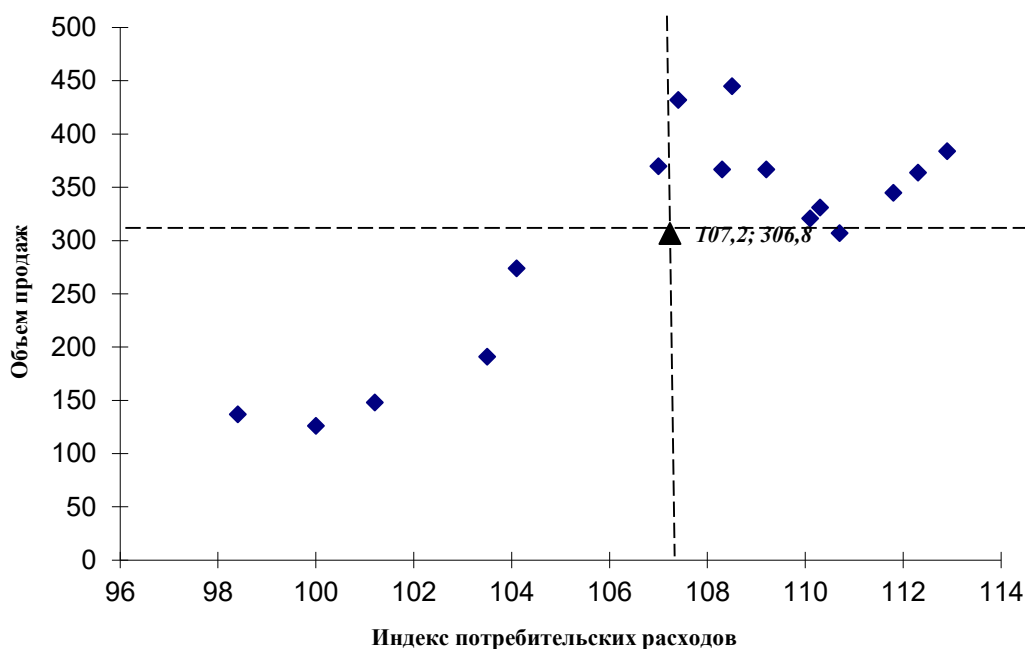


Диаграмма рассеяния (корреляционное поле)

Решение.

1. Вытянутость облака точек на диаграмме рассеяния вдоль наклонной прямой (см. рисунок) позволяет сделать предположение о том, что существует некоторая объективная тенденция прямой линейной связи между значениями переменных x – индексом потребительских расходов и y – объемом продаж.

2. Промежуточные расчеты при вычислении коэффициента корреляции между переменными x – индексом потребительских расходов и y – объемом продаж опускаем.

Средние значения случайных величин X и Y , которые считаются наиболее простыми показателями, характеризующими последовательности x_1, x_2, \dots, x_{16} и y_1, y_2, \dots, y_{16} , рассчитывают соответственно по формулам

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 107,2, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = 306,8.$$

Дисперсия характеризует степень разброса значений x_1, x_2, \dots, x_{16} (y_1, y_2, \dots, y_{16}) вокруг своего среднего \bar{x} (\bar{y} соответственно)

$$S_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{305,474}{15} = 20,36,$$

$$S_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2 = \frac{158718,438}{15} = 10581,23.$$

Стандартные ошибки случайных величин X и Y рассчитывают соответственно по формулам

$$S_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = 4,51; \quad S_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = 102,87.$$

Коэффициент корреляции равен (табл. 9.2)

$$r_{x,y} = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_x S_y} = \frac{\frac{1}{15} 5681,99}{4,51 \cdot 102,87} = 0,816.$$

3. Оценим значимость коэффициента корреляции. Для этого считаем значение t – статистики по формуле

$$t_{\text{расч}} = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,816 \sqrt{14}}{\sqrt{1-0,666}} = 5,282.$$

Табличное значение критерия Стьюдента равно $t_{\text{табл}} (\alpha = 0,1; k = n - 2 = 14) = 1,76$. Сравнивая числовые значения критериев, видно, что $t_{\text{расч}} > t_{\text{табл}}$, т. е. полученное значение коэффициента корреляции значимо.

Таким образом, индекс потребительских расходов оказывает весьма высокое влияние на объемы продаж.

4. Матрица R коэффициентов парной корреляции, вычисленная для трех факторов, имеет вид, приведенный в табл. 9.2.

Таблица 9.2

Коэффициент парной корреляции

Параметр	Объем продаж Y	Затраты на рекламу X_1	Индекс потребительских расходов X_2
Объем продаж Y	1	0,646	0,816
Затраты на рекламу X_1	0,646	1	0,273
Индекс потребительских расходов X_2	0,816	0,273	1

5. Вычисление множественного коэффициента корреляции y с x_1 и x_2

$$R_{j,1,2,\dots,1,\dots,m} = \sqrt{1 - \frac{|R|}{R_{jj}}} = \sqrt{1 - \frac{0,1304}{0,9253}} = 0,9269,$$

где $|R|$ – определитель корреляционной матрицы R равен 0,1304;

R_{jj} – алгебраическое дополнение 1-го диагонального элемента r_{jj} той же матрицы R

$$R_{jj} = (-1)^2 \begin{bmatrix} 1 & 0,273 \\ 0,273 & 1 \end{bmatrix} = 0,9253.$$

6. Вычисление коэффициентов частной корреляции

$$r_{jk.1,2,\dots} = -\frac{R_{jk}}{\sqrt{R_{jj}R_{kk}}};$$

$$r_{12(3)} = \frac{R_{12}}{\sqrt{R_{11}R_{22}}} = -\frac{0,423}{\sqrt{0,925 \cdot 0,334}} = -0,706,$$

где R_{12} – алгебраическое дополнение элемента r_{12} матрицы R ; R_{22} – алгебраическое дополнение 2-го диагонального элемента r_{22}

$$R_{12} = (-1)^3 \begin{bmatrix} 0,646 & 0,273 \\ 0,816 & 1 \end{bmatrix} = -0,423;$$

$$R_{22} = (-1)^4 \begin{bmatrix} 1 & 0,816 \\ 0,816 & 1 \end{bmatrix} = 0,334.$$

Коэффициенты частной корреляции можно вычислить, используя коэффициенты парной корреляции:

$$r_{12(3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} = \frac{0,646 - 0,816 \cdot 0,273}{\sqrt{(1-0,816^2)(1-0,273^2)}} = 0,706;$$

$$r_{13(2)} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} = \frac{0,816 - 0,646 \cdot 0,273}{\sqrt{(1-0,646^2)(1-0,273^2)}} = 0,871.$$

Контрольные вопросы

1. Назовите основные показатели взаимосвязи двух случайных величин.
2. Чем характеризуются функциональные и корреляционные связи между переменными?
3. Какова основная задача корреляционного анализа?
4. Как определяется мера взаимосвязи между двумя переменными?
5. Что является показателем тесноты связи между переменными?
6. Назовите основные свойства парного коэффициента корреляции.
7. В чем суть проверки значимости парного коэффициента корреляции?
8. Каково назначение матрицы коэффициентов парной корреляции?
9. Какие задачи рассматриваются в многомерном корреляционном анализе?
10. Как определяется множественный коэффициент корреляции и оценивается его значимость?
11. Как определяется коэффициент частной корреляции?

10. ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ – метод в математической статистике, направленный на поиск зависимостей в экспериментальных данных путем исследования значимости различий в средних значениях. В отличие от t -критерия позволяет сравнивать средние значения трех и более групп.

В отличие от корреляционного анализа в дисперсионном анализе исследователь исходит из предположения, что одни переменные выступают как влияющие (именуемые факторами, или независимыми переменными), а другие (результативные признаки, или зависимые

переменные) – подвержены влиянию этих факторов. Хотя такое допущение и лежит в основе математических процедур расчета, оно, однако, требует осторожности при выводах о причине и следствии.

Откуда произошло название «дисперсионный анализ»?

Может показаться странным, что процедура сравнения средних называется дисперсионным анализом. В действительности это связано с тем, что при исследовании статистической значимости различия между средними двух (или нескольких) групп мы на самом деле сравниваем (т. е. анализируем) выборочные дисперсии. Фундаментальная концепция дисперсионного анализа предложена Фишером в 1920 году. Возможно, более естественным был бы термин «анализ суммы квадратов», или анализ вариации, но в силу традиции употребляется термин «дисперсионный анализ».

Цель дисперсионного анализа – проверка статистической значимости различия между средними (для групп или переменных). Эта проверка проводится с помощью разбиения суммы квадратов на компоненты, т. е. с помощью разбиения общей дисперсии (вариации) на части, одна из которых обусловлена случайной ошибкой (т. е. внутригрупповой изменчивостью), а вторая связана с различием средних значений. Последняя компонента дисперсии затем используется для анализа статистической значимости различия между средними значениями. Если это различие значимо, нулевая гипотеза отвергается и принимается альтернативная гипотеза о существовании различия между средними.

Зависимые и независимые переменные. Переменные, значения которых определяются с помощью измерений в ходе эксперимента (например, балл, набранный при тестировании), называются зависимыми переменными. Переменные, которыми можно управлять при проведении эксперимента (например, методы обучения или другие критерии, позволяющие разделить наблюдения на группы или классифицировать) называются факторами, или независимыми переменными.

Ограничение метода: независимые признаки могут измеряться по номинальной, порядковой или метрической шкале, зависимые – только по метрической. Для проведения дисперсионного анализа выделяют несколько градаций факторных признаков, а все элементы выборки группируют в соответствии с этими градациями.

Формулировка гипотез в дисперсионном анализе. Нулевая гипотеза: «Средние величины результативного признака во всех условиях действия фактора (или градациях фактора) одинаковы». Альтернативная гипотеза: «Средние величины результативного признака в разных условиях действия фактора различны».

Дисперсионный анализ можно подразделить на несколько категорий в зависимости:

- от количества рассматриваемых независимых факторов;
- количества результативных переменных, подверженных действию факторов;
- характера, природы получения и наличия взаимосвязи сравниваемых выборок значений.

При наличии одного фактора, влияние которого исследуется, дисперсионный анализ именуется однофакторным и распадается на две разновидности:

- анализ несвязанных (т. е. различных) выборок. Например, одна группа респондентов решает задачу в условиях тишины, вторая – в шумной комнате. В этом случае, к слову, нулевая гипотеза звучала бы так: «среднее время решения задач такого-то типа будет одинаково в тишине и в шумном помещении», т. е. не зависит от фактора шума;
- анализ связанных выборок, т. е. двух замеров, проведенных на одной и той же группе респондентов в разных условиях. Тот же пример: в первый раз задача решалась в тишине, второй – сходная задача – в условиях шумовых помех. На практике к подобным опытам следует подходить с осторожностью, поскольку в действие может вступить неучтенный фактор «научаемость», влияние которого исследователь рискует приписать изменению условий, а именно шуму.

В случае если исследуется одновременное воздействие двух или более факторов, мы имеем дело с многофакторным дисперсионным анализом, который также можно подразделить по типу выборки.

Если же воздействию факторов подвержено несколько переменных, речь идет о многомерном анализе. Проведение многомерного дисперсионного анализа предпочтительнее одномерного только в том случае, когда зависимые переменные не являются независимыми друг от друга и коррелируют между собой.

Обобщенно задача дисперсионного анализа состоит в том, чтобы из общей вариативности признака выделить три частные вариативности:

- обусловленную действием каждой из исследуемых независимых переменных (факторов);
- обусловленную взаимодействием исследуемых независимых переменных;
- случайную, обусловленную всеми неучтенными обстоятельствами.

Для оценки вариативности, обусловленной действием исследуемых переменных и их взаимодействием, вычисляется отношение соответствующего показателя вариативности и случайной вариативности. Показателем этого соотношения выступает F – критерий Фишера.

$$F_{\text{эмпА}} = \frac{\text{вариативность, обусловленная действием переменной А}}{\text{случайная вариативность}};$$

$$F_{\text{эмпВ}} = \frac{\text{вариативность, обусловленная действием переменной В}}{\text{случайная вариативность}};$$

$$F_{\text{эмпАВ}} = \frac{\text{вариативность, обусловленная взаимодействием А, В}}{\text{случайная вариативность}}.$$

Чем в большей степени вариативность признака обусловлена действием влияющих факторов или их взаимодействием, тем выше эмпирические значения критерия F . В формулу расчета критерия F входят оценки дисперсий, и, следовательно, этот метод относится к разряду параметрических.

10.1. Однофакторный дисперсионный анализ

Дисперсионный анализ, рассматривающий только одну независимую переменную, называется однофакторным дисперсионным анализом [10].

Однофакторный дисперсионный анализ (ANOVA – analysis of variance) используется для сравнения средних значений для трех и более выборок (групп). Каждая выборка (группа) соответствует одной из градаций независимой переменной (фактора). Фактор имеет несколько значений – уровней фактора.

Пусть, например, качество программного продукта определяется с помощью k различных тестов и необходимо исследовать, влияет ли фактор «тест» на результат проверки. Если тестов два, то проверка гипотезы о средних показателях тестов проводится рассмотренными ранее методами проверки статистических гипотез о равенстве сред-

них с использованием t -критерия Стьюдента. Если тестов более двух, то гипотеза о равенстве средних показателей тестов проверяется с использованием методов дисперсионного анализа.

Проверяется нулевая гипотеза $H_0: m_1 = m_2 = \dots = m_k$ об отсутствии влияния на результативный признак X (результат тестирования) фактора A (тест), имеющего k уровней $A_j, j = 1, 2, \dots, k$.

Основная идея дисперсионного анализа состоит в том, чтобы сопоставить дисперсию за счет воздействия фактора A с дисперсией, обусловленной случайными причинами (остаточная дисперсия). Если различие между ними несущественно, то влияние фактора A на признак X незначительно. Если же различие между факторной и остаточной дисперсиями значимо, то это говорит о влиянии фактора A на рассматриваемый признак X .

Предполагается, что случайная величина X имеет нормальное распределение с математическим ожиданием m_j , зависящим от уровня фактора A_j и постоянной дисперсией σ^2 . В качестве исходных данных используются выборочные значения величины X , полученные для каждого уровня фактора A ; число элементов выборки на каждом уровне равно n , тогда общее число наблюдений равно nk . Обозначим через x_{ij} результат i -го наблюдения ($i = 1, 2, \dots, n$) за j -м фактором.

Выборочное среднее, соответствующее j -у уровню фактора A (групповое среднее), вычисляется по формуле $\bar{X} = \frac{1}{n} \sum_{i=1}^n x_{ij}$, а общее

среднее по выражению $\bar{x} = \frac{1}{nk} \sum_{j=1}^k \sum_{i=1}^n x_{ij} = \frac{1}{k} \sum_{j=1}^k \bar{x}_j$.

Общая сумма квадратов – это сумма квадратов отклонений наблюдаемых значений x_{ij} от общего среднего

$$Q = \sum_{j=1}^k \sum_{i=1}^n (x_{ij} - \bar{x})^2.$$

Факторная сумма квадратов, обусловленная влиянием фактора A , – это сумма квадратов отклонений групповых средних от общей средней

$$Q_A = n \sum_{j=1}^k (\bar{x}_j - \bar{x})^2.$$

Остаточная сумма квадратов характеризует рассеяние внутри группы

$$Q_e = \sum_{j=1}^k \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2.$$

На практике эта сумма находится из основного тождества дисперсионного анализа, в соответствии с которым $Q = Q_A + Q_e$. Соответствующее число степеней свободы равно $\nu = nk - 1$; $\nu_A = k - 1$; $\nu_e = k(n - 1)$, а дисперсии равны $s^2 = Q/\nu$, $s_A^2 = Q_A/\nu_A$, $s_e^2 = Q_e/\nu_e$.

Если нулевая гипотеза о равенстве средних справедлива, то эти дисперсии являются несмещенными оценками дисперсий генеральной совокупности. Значительное превышение дисперсии s_A^2 над дисперсией s_e^2 можно объяснить различием средних в группах. Поэтому для проверки нулевой гипотезы, которая имеет распределение Фишера с уровнем значимости α и числом степеней свободы $(k - 1)$ и $k(n - 1)$,

используется отношение этих средних $F = \frac{s_A^2}{s_e^2} = \frac{Q_A / (k - 1)}{Q_e / k(n - 1)}$.

Нулевая гипотеза не противоречит результатам наблюдений на заданном уровне значимости α , если $F < F_{1-\alpha, (k-1), k(n-1)}$.

В этом случае считается, что фактор А не оказывает существенного влияния на показатель Х.

Условия применения F-статистики:

- генеральные совокупности, из которых формируются выборки, должны быть нормально распределены;
- выборки должны быть независимы;
- дисперсии генеральных совокупностей должны быть равны.

Результаты расчета дисперсионного анализа сводятся в табл. 10.1.

Таблица 10.1

Результаты расчета дисперсионного анализа

Источник дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия	Статистика Фишера
Фактор А	Q_A	ν_A	s_A^2	F
Остаток	Q_e	ν_e	s_e^2	–
Общий	Q	ν	s^2	–

Пример. К закаленному стеклу предъявляются высокие требования к допускам на отклонения гнутых изделий от заданной формы. Форму и размеры гнутых изделий проверяют по контрольному шаблону [11]. Контролируется также поперечная кривизна по отклонению образующей линии от цилиндрической поверхности.

Необходимо оценить влияние конфигурации вырабатываемых закаленных автомобильных стекол на дисперсию отклонения образующей линии от цилиндрической поверхности в квадратных миллиметрах.

Конфигурация стекла определяется стороной остекления, фактор А. Кодировалась числами: левое – 0, правое – 1, $k = 2$. Объем выборки при анализе составил $n = 313$ измерений.

В табл. 10.2 приведены результаты дисперсионного анализа влияния конфигурации стекла на отклонение образующей цилиндра.

Таблица 10.2

Результаты дисперсионного анализа

Источник дисперсии	Сумма квадратов, мм ²	Число степеней свободы	Дисперсия, мм ²	Статистика Фишера
Фактор А	3,74	1	3,74	68
Остаток	34,32	624	0,055	–
Общий	38,06	625	0,0609	–

Для уровня значимости 0,05, числа степеней свободы $\nu_A = 2 - 1 = 1$ и $\nu_e = 624$ квантиль распределения Фишера равен $F_{\text{табл}} = 3,86$.

Так как выборочное значение статистики оказалось больше критического, нулевая гипотеза отвергается и принимается альтернативная: влияние конфигурации стекла на отклонение образующей цилиндра существенно.

10.2. Двухфакторный дисперсионный анализ

Мир по своей природе сложен и многомерен. Ситуации, когда некоторое явление полностью описывается одной переменной, чрезвычайно редки. Например, если мы пытаемся научиться выращивать большие помидоры, следует рассматривать факторы, связанные с генетической структурой растений, типом почвы, освещенностью, температурой и т. д. Таким образом, при проведении типичного эксперимента приходится иметь дело с большим количеством факторов. Ос-

новная причина, по которой использование дисперсионного анализа предпочтительнее повторного сравнения двух выборок при разных уровнях факторов с помощью серий t -критерия, заключается в том, что дисперсионный анализ существенно эффективен и для малых выборок более информативен.

Принципиальной разницы между многофакторным и однофакторным дисперсионным анализом нет. Многофакторный анализ не меняет общую логику дисперсионного анализа, а лишь несколько усложняет ее, поскольку, кроме учета влияния на зависимую переменную каждого из факторов по отдельности, следует оценивать их совместное действие [12]. Таким образом, то новое, что вносит в анализ данных многофакторный дисперсионный анализ, касается в основном возможности оценить межфакторное взаимодействие. Тем не менее по-прежнему остается возможность оценивать влияние каждого фактора в отдельности. В этом смысле процедура многофакторного дисперсионного анализа (в варианте ее компьютерного использования) более экономична, поскольку всего за один запуск решает сразу две задачи: оценивается влияние каждого из факторов и их взаимодействие.

Рассмотрим многофакторный анализ на примере двухфакторного [Там же]. Двухфакторный дисперсионный анализ позволяет проверить эффекты влияния обоих факторов на зависимую переменную одновременно, а не по отдельности. Кроме этого можно проверить гипотезу об эффекте взаимодействия между двумя независимыми переменными (рис. 10.1).

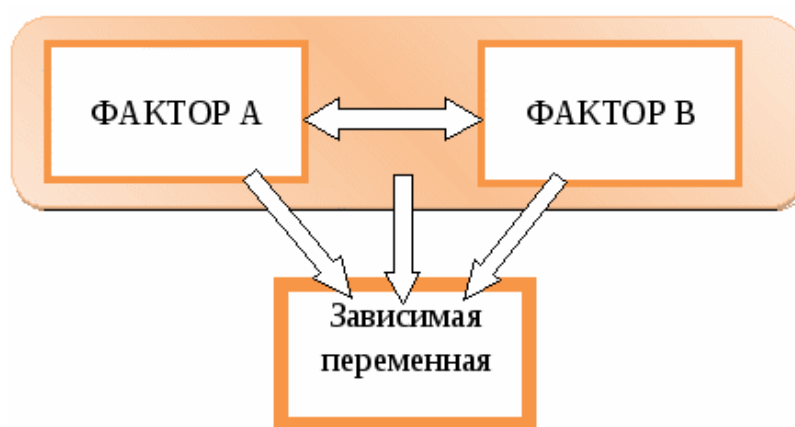


Рис. 10.1. Двухфакторный дисперсионный анализ

Например, компания хочет проверить эффективность своей рекламы (табл. 10.3). Выбран продукт и созданы два типа рекламных роликов: серьезный и смешной. Реклама размещается в рабочие и выходные дни. Выбраны 16 потенциальных зрителей, которые распределяются по группам случайным образом:

Группа 1: смешной ролик, рабочий день.

Группа 2: смешной ролик, выходной день.

Группа 3: серьезный ролик, рабочий день.

Группа 4: серьезный ролик, выходной день.

Эта схема 2×2 , так как каждая переменная состоит из двух уровней. После того как каждый зритель просмотрел ролик, его просят оценить эффективность рекламы (привлекательность, ясность, краткость ролика и т. д.) по двадцатибалльной шкале.

Необходимо на уровне значимости α выяснить зависимость оценок от указанных факторов, используя двухфакторный дисперсионный анализ.

Таблица 10.3

Оценки эффективности рекламы

Тип ролика	День	
	Рабочий	Выходной
Смешной	6, 10, 11, 9	15, 18, 14, 16
Серьезный	8, 13, 12, 10	19, 20, 13, 17

Исследуемые группы называют эффектами обработки (*treatment groups*). Двухфакторный дисперсионный анализ позволит проверить эффекты влияния типа ролика и типа дня одновременно, а не по отдельности, а также гипотезу об эффекте взаимодействия между двумя переменными. Наличие значимого эффекта будет означать, что тип ролика по-разному влияет на эффективность рекламы в зависимости от типа дня.

Схема двухфакторного дисперсионного анализа имеет несколько нулевых гипотез: одна для каждой независимой переменной и одна для взаимодействия.

H_0 : тип ролика и день не имеют эффекта взаимодействия на эффективность рекламы.

H_1 : тип ролика и день имеют эффект взаимодействия на эффективность рекламы.

H_0 : эффективность рекламы не зависит от типа ролика.

H_1 : эффективность рекламы зависит от типа ролика.

H_0 : эффективность рекламы не зависит от типа дня.

H_1 : эффективность рекламы зависит от типа дня.

Результаты вычислений могут быть представлены в виде табл. 10.4.

Таблица 10.4

Результаты анализа

Фактор	Сумма квадратов	Степень свободы	Дисперсия	F
Фактор А	SS_A	$a - 1$	$MS_A = \frac{SS_A}{a - 1}$	F_A
Фактор В	SS_B	$b - 1$	$MS_B = \frac{SS_B}{b - 1}$	F_B
Взаимодействие А и В	$SS_{A \times B}$	$(a - 1)(b - 1)$	$MS_{A \times B} = \frac{SS_{A \times B}}{(a - 1)(b - 1)}$	$F_{A \times B}$
Ошибка	SS_{error}	$ab(n - 1)$	$MS_{error} = \frac{SS_{error}}{ab(n - 1)}$	–
Общий	SS	n	MS	–

Использованы следующие обозначения:

SS_A – сумма квадратов для фактора А;

SS_B – сумма квадратов для фактора В;

$SS_{A \times B}$ – сумма квадратов для взаимодействия факторов;

SS_{error} – сумма квадратов для ошибки;

a – количество уровней фактора А;

b – количество уровней фактора В;

n – количество объектов в каждой группе.

Общая изменчивость в двухфакторном дисперсионном анализе может быть разложена следующим образом (рис. 10.2).



Рис. 10.2. Распределение изменчивости

Статистическая проверка гипотезы о наличии различий осуществляется на основании F -статистики:

$$F_A = \frac{MS_A}{MS_{error}}, F_{\text{крит}}(\alpha, a-1, ab(n-1));$$

$$F_B = \frac{MS_B}{MS_{error}}, F_{\text{крит}}(\alpha, b-1, ab(n-1));$$

$$F_{A \times B} = \frac{MS_{A \times B}}{MS_{error}}, F_{\text{крит}}(\alpha, (a-1)(b-1), ab(n-1)).$$

Условия применения двухфакторного анализа:

- генеральные совокупности, из которых извлечены выборки, должны быть нормально распределены;
- выборки должны быть независимыми;
- дисперсии генеральных совокупностей, из которых извлекались выборки, должны быть равными;
- группы должны иметь одинаковый объем выборки.

Пример. Для использования двухфакторного дисперсионного анализа [12] необходимо выяснить, оказывают ли влияние тип потребляемого бензина и тип автомобиля на расход топлива. Для этого будут использованы два типа бензина – обычный и высокооктановый, а для каждой группы – два типа автомобилей: с двумя и четырьмя ведущими колесами. Для каждой группы будут использованы по два автомобиля, всего восемь (табл. 10.5).

Алгоритм решения задачи:

1. Сформулировать гипотезы.
2. Найти критическое значение для каждого значения F -критерия при заданном α , например, $\alpha = 0,05$.
3. Заполнить итоговую таблицу, чтобы получить фактические значения критерия.
4. Принять решение.

Таблица 10.5

Пробег автомобиля в милях на галлон

Топливо	Тип автомобиля	
	Два колеса	Четыре колеса
Обычное	26,7	28,6
	25,2	29,3
Высокооктановое	32,3	26,1
	32,8	24,2

Формулировка гипотез

Для взаимодействия типа топлива и типа автомобиля:

H_0 : тип топлива и тип автомобиля не оказывают эффекта взаимодействия на потребление бензина.

H_1 : тип топлива и тип автомобиля оказывают эффект взаимодействия на потребление бензина.

Для типов топлива:

H_0 : для двух типов топлива нет разницы между средним потреблением бензина.

H_1 : для двух типов топлива существует разница между средним потреблением бензина.

Для типов автомобилей:

H_0 : для автомобилей с двумя и четырьмя ведущими колесами нет разницы в среднем потреблении бензина.

H_1 : для автомобилей с двумя и четырьмя ведущими колесами существует разница в среднем потреблении бензина.

Каждая независимая переменная, или фактор, имеет два уровня (принимает два значения).

Фактор А – тип топлива: обычное и высокооктановое, $a = 2$.

Фактор В – тип автомобиля: также имеет два значения, $b = 2$.

Число объектов в каждой группе, $n = 2$.

Степени свободы для каждого фактора:

- фактор А: $df_A = a - 1 = 2 - 1 = 1$;
- фактор В: $df_B = b - 1 = 2 - 1 = 1$;
- взаимодействие (А×В): $df_{A \times B} = (a - 1)(b - 1) = (2 - 1)(2 - 1) = 1$;
- ошибка внутри группы: $df_{error} = ab(n - 1) = 2 \cdot 2(2 - 1) = 4$.

Критические значения:

- $F_{критА}(0,05; 1; 4) = 7,71$;
- $F_{критВ}(0,05; 1; 4) = 7,71$.

Если факторы имеют различное число градаций, критические значения будут различными (табл. 10.6).

Таблица 10.6

Результаты дисперсионного анализа

Фактор	Сумма квадратов	Степень свободы df	Дисперсия	F
Топливо А	3,92	1	3,92	4,752
Автомобиль В	9,68	1	9,68	11,733
Взаимодействие А и В	54,08	1	54,08	65,552
Ошибка (внутри группы)	3,3	4	0,825	–
Общая	70,98	7	–	–

Поскольку $F_B = 11,733$, $F_{AB} = 65,522$, что превышает критический уровень 7,71, нулевые гипотезы об отсутствии влияния эффекта взаимодействия и типа автомобиля отвергаются. Можно сделать вывод о том, что тип автомобиля и сочетание типа топлива и типа автомобиля оказывают существенное влияние на потребление топлива.

Анализ взаимодействия

Влияние каждого фактора называют основными, или главными, эффектами. Если нет значимого эффекта взаимодействия, основные эффекты можно интерпретировать независимо друг от друга. Однако если значимый эффект взаимодействия существует, надо более внимательно интерпретировать основные эффекты. Чтобы интерпретировать результаты двухфакторного дисперсионного анализа, можно использовать график, на который наносятся средние значения каждой группы.

Приведем пример, рассматривающий влияние типа бензина и типа автомобиля на расход топлива. В табл. 10.7 даны средние значения пробега.

Таблица 10.7

Средний пробег автомобиля в милях на галлон топлива

Топливо	Тип автомобиля	
	Два колеса	Четыре колеса
Обычное	25,95	28,95
Высокооктановое	32,55	25,15

На графике рис. 10.3 прямые, соединяющие соответствующие средние, пересекаются. В случае такого пересечения и при значительном эффекте взаимодействия это взаимодействие называется беспорядочным. В случае беспорядочного взаимодействия не следует интерпретировать основные эффекты без учета эффекта взаимодействия.

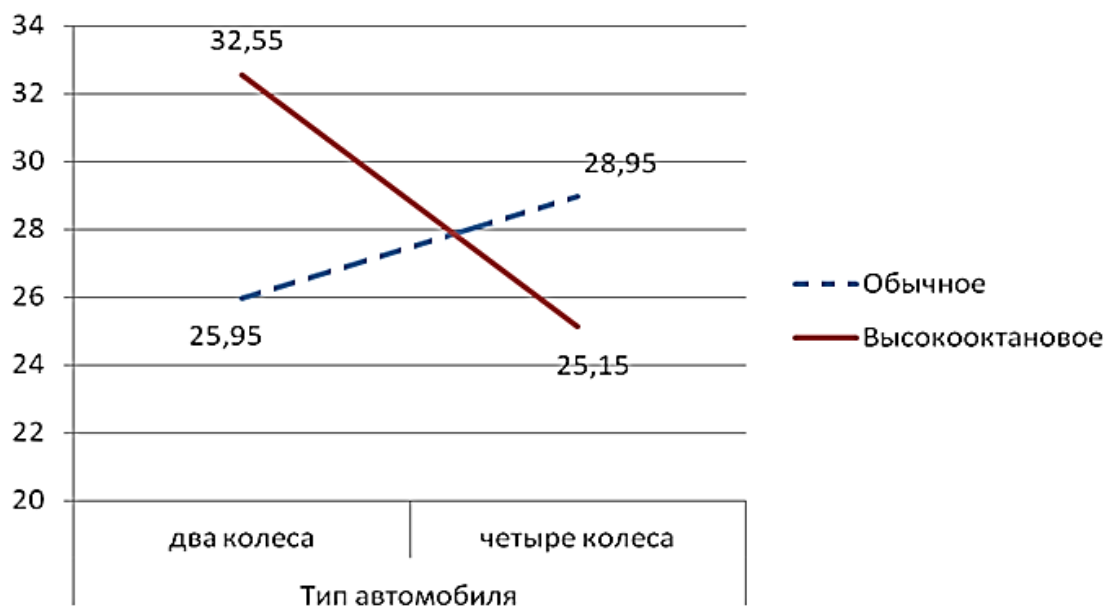


Рис. 10.3. Беспорядочное взаимодействие

Другой возможный тип взаимодействия – порядковое взаимодействие (рис. 10.4). Если значение F -критерия для взаимодействия оказывается значимым и прямые не пересекаются, тогда взаимодействие называется порядковым, и основные эффекты можно интерпретировать отдельно друг от друга.

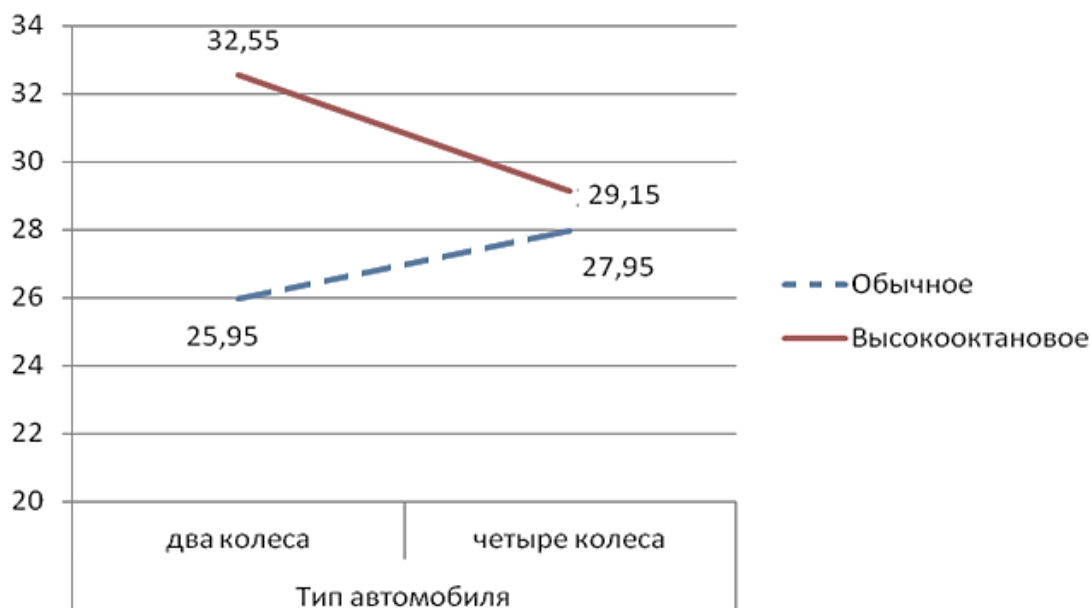


Рис. 10.4. Порядковое взаимодействие

Наконец, когда нет значительного эффекта взаимодействия, прямые на графике будут параллельными или почти параллельными (рис. 10.5). В подобной ситуации основные эффекты можно интерпретировать независимо друг от друга, поскольку не существует значимого взаимодействия. На рисунке приведен график двух переменных, когда эффект взаимодействия незначителен, прямые практически параллельны.

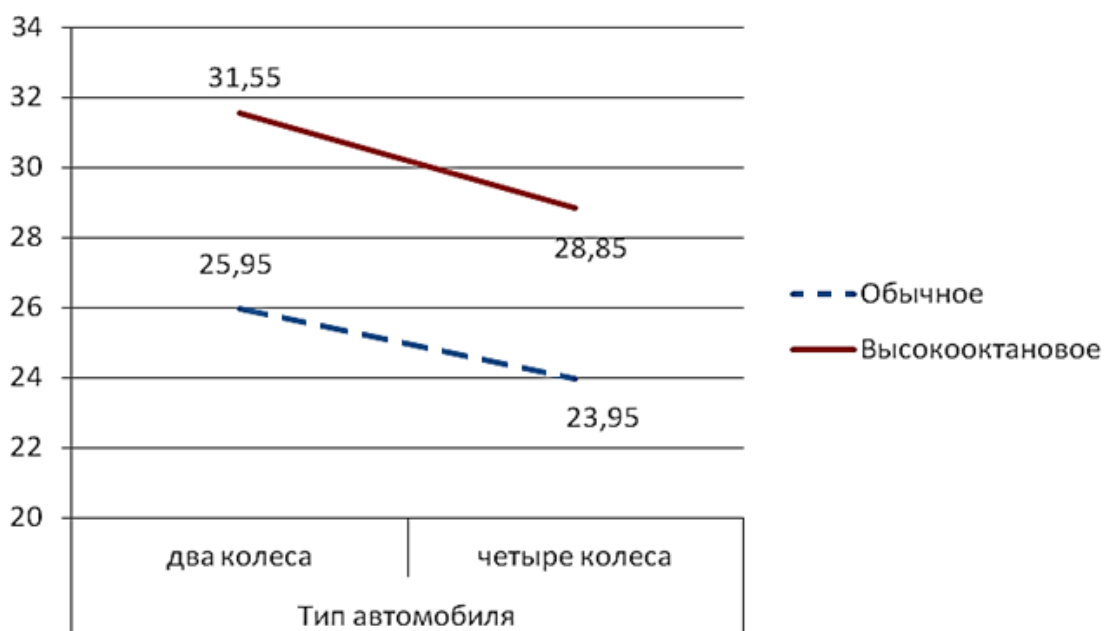


Рис. 10.5. Отсутствие значимого взаимодействия

Контрольные вопросы

1. В чем отличие дисперсионного анализа от корреляционного?
2. Что является целью дисперсионного анализа?
3. Сформулируйте гипотезы в дисперсионном анализе.
4. В чем заключается содержание задачи дисперсионного анализа?
5. Когда используют однофакторный дисперсионный анализ?
6. Какие условия применения однофакторного дисперсионного анализа можно назвать?
7. Поясните процедуру выполнения однофакторного дисперсионного анализа.
8. В чем идея дисперсионного анализа?
9. Какие расчеты выполняются в однофакторном дисперсионном анализе?
10. Объясните разницу между однофакторным и многофакторным дисперсионными анализами.
11. Что позволяет проверить двухфакторный дисперсионный анализ?
12. Что понимают под нулевыми гипотезами в двухфакторном дисперсионном анализе?
13. Как может быть разложена общая изменчивость в двухфакторном дисперсионном анализе?
14. Каковы условия применения двухфакторного дисперсионного анализа?
15. Для чего проводится анализ взаимодействия в двухфакторном дисперсионном анализе?

11. РЕГРЕССИОННЫЙ АНАЛИЗ

Регрессионный анализ предназначен для описания зависимости исследуемой переменной от различных факторов и отображения их взаимосвязи в форме регрессионной модели.

Задача исследователя по поиску математической модели заключается в отыскании связи между переменными, имеющими количественную меру. Так как результаты наблюдений – величины случайные, то говорят о связи средних значений исследуемых величин.

На математическом языке задача регрессионного анализа формулируется следующим образом [13]: нужно получить некоторое представление о функции отклика $\hat{y} = \varphi(x_1, x_2, \dots, x_k)$, где \hat{y} – параметр системы или процесса, подлежащий анализу; x_1, x_2, \dots, x_k – факторные переменные, которыми можно варьировать при анализе, прогнозировании и принятии решений по управлению.

Геометрический образ, соответствующий функции отклика, называют поверхностью отклика, которая может быть линейной и нелинейной.

Приступая к исследованию системы или процесса, экспериментатор может располагать априорной информацией о виде функции отклика, в этом случае ему необходимо оценить неизвестные параметры модели. В ряде случаев аналитическое выражение функции отклика бывает неизвестным, поэтому приходится ограничиваться представлением ее полиномом $\hat{y} = \beta_0 + \sum \beta_i x_i + \sum \beta_{ij} x_i x_j + \sum \beta_{ii} x_i^2 + \dots$ с коэффициентами регрессии $\beta_0, \beta_i, \beta_{ij}, \beta_{ii}, \dots$.

Используя результаты наблюдений, можно определить только выборочные коэффициенты регрессии $b_0, b_i, b_{ij}, b_{ii}, \dots$, которые являются только оценками для теоретических коэффициентов регрессии $\beta_0, \beta_i, \beta_{ij}, \beta_{ii}, \dots$.

Уравнение регрессии, полученное на основе наблюдений, запишем в следующем виде: $\hat{y} = b_0 + \sum b_i x_i + \sum b_{ij} x_i x_j + \sum b_{ii} x_i^2 + \dots$, где \hat{y} – значение зависимой (результатирующей) переменной, предсказанное уравнением регрессии.

Еще Ньютон умел представлять функции степенными рядами, а Гаусс предложил метод наименьших квадратов для оценки коэффициентов регрессии по результатам наблюдений. Сегодня мы располагаем строгой теорией регрессионного анализа, базирующейся на современных теоретико-вероятностных представлениях [14]. Эта теория позволяет значительно глубже понять и оценить результаты, получаемые методом наименьших квадратов.

11.1. Вычисление коэффициентов регрессии

Допустим, что мы располагаем N результатами наблюдений над величиной y , зависящей от факторных переменных x_1, x_2, \dots, x_k . Положим, что результаты наблюдений можно представить полиномом,

например второй степени. Задача заключается в том, чтобы по результатам наблюдений определить коэффициенты регрессии, число которых будет равно C_{2+k}^2 . Количество наблюдений выбирается из соотношения $N \geq C_{2+k}^2$.

Будем считать, что выполняются предпосылки регрессионного анализа:

- 1) результаты наблюдений $y_1, y_2, y_3, y_4, \dots, y_N$ представляют собой независимые, нормально распределенные случайные величины;
- 2) дисперсии $\sigma^2(y_u)$, $u = 1, 2, 3, \dots, N$ равны друг другу, т. е. выборочные оценки $s^2(y_u)$ однородны;
- 3) факторные переменные x_1, x_2, \dots, x_k измеряются с пренебрежимо малой ошибкой по сравнению с ошибкой в определении y .

Метод наименьших квадратов можно применять и в том случае, когда величина y не подчиняется нормальному закону распределения. При этом нельзя говорить о том, насколько эффективным будет применение метода наименьших квадратов, особенно при выборках малого объема.

Для упрощения расчетов введем фиктивную переменную $x_0 = 1$ и заменим члены второго порядка линейными, введя обозначения $x_i^2 = x_{k+i}$ для $i = 1, 2, \dots, k$, $x_i x_j = x_{2k+i}$ для $i, j = 1, 2, 3, \dots, k$, $i \neq j$. Аналогичным образом линейными членами можно заменить члены любого порядка.

В новой системе обозначений полином любой степени записывается однородным линейным уравнением

$$\hat{y} = b_0 + b_1 x_{1u} + \dots + b_2 x_{2u} + \dots + b_m x_{mu}.$$

Чтобы найти коэффициенты регрессии методом наименьших квадратов, необходимо минимизировать сумму квадратов отклонений, рассчитанных по уравнению регрессии данных от результатов наблюдений:

$$Q = \sum (y_u - (b_0 x_0 + b_1 x_1 + \dots + b_2 x_2 + \dots + b_m x_m))^2.$$

Приведенная функция по отношению к искомым коэффициентам регрессии $b_0, b_1, b_2, \dots, b_m$ является непрерывной и одноэкстремальной. Минимум этой функции определяется решением системы нормальных уравнений, которая получается приравниванием нулю частных производных этой функции по переменным $b_0, b_1, b_2, \dots, b_m$.

Для упрощения системы обозначений и облегчения вывода интересных нас формул обратимся к матричной алгебре. Представим результаты наблюдений в матричной форме,

$$\mathbf{X} = \begin{bmatrix} x_{01} & x_{11} & \dots & x_{m1} \\ x_{02} & x_{12} & & x_{m2} \\ \dots & \dots & \dots & \dots \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}.$$

Искомые коэффициенты регрессии запишем в виде матрицы (вектора)

$$\mathbf{B} = \begin{bmatrix} b_0 \\ b_1 \\ \dots \\ b_m \end{bmatrix}.$$

В матричной форме систему нормальных уравнений запишем следующим образом: $(\mathbf{X} \times \mathbf{X})\mathbf{B} = \mathbf{X} \times \mathbf{Y}$.

Найдем матрицу $(\mathbf{X} \times \mathbf{X})^{-1}$, обратную матрице $\mathbf{X} \times \mathbf{X}$. Умножив слева обе части записанного матричного уравнения на $(\mathbf{X} \times \mathbf{X})^{-1}$, получим

$$(\mathbf{X} \times \mathbf{X})^{-1}(\mathbf{X} \times \mathbf{X})\mathbf{B} = \mathbf{B} = (\mathbf{X} \times \mathbf{X})^{-1} \times \mathbf{X} \times \mathbf{Y}.$$

Отсюда следует, что интересующие нас коэффициенты регрессии определяются выражением

$$\mathbf{B} = (\mathbf{X} \times \mathbf{X})^{-1} \times \mathbf{X} \times \mathbf{Y}. \quad (11.1)$$

Для получения решений матрица коэффициентов нормальных уравнений $\mathbf{X} \times \mathbf{X}$ должна быть невырожденной, что обеспечивается линейной независимостью переменных x_1, x_2, \dots, x_m .

Из формулы 11.1 следует, что коэффициенты регрессии не могут быть определены независимо друг от друга. Если мы в процессе анализа отбросим какую-либо независимую переменную или изменим порядок полинома, то все вычисления надо производить заново. Такая неопределенность в оценке коэффициентов регрессии затрудняет их интерпретацию. В этом случае приходится рассматривать уравнение регрессии как интерполяционную формулу, пригодную для оценки некоторого промежуточного значения y по результатам остальных n значений y_1, y_2, \dots, y_n . При таком использовании уравнения регрессии

перераспределение численных значений для коэффициентов регрессии, связанное с изменением порядка приближения (структуры регрессионного уравнения), не будет вызывать каких-либо недоумений [13].

От неопределенности, связанной с неоднозначной числовой оценкой коэффициентов регрессии, можно избавиться, если эксперименты планировать по некоторой схеме, т. е. ставить активный эксперимент над исследуемой системой, что далеко не всегда возможно реализовать.

11.2. Статистический анализ уравнения регрессии

После вычисления коэффициентов регрессии нужно провести статистический анализ полученного уравнения.

Определим остаточную дисперсию, характеризующую рассеяние точек относительно найденного уравнения регрессии. Значения зависимой переменной, предсказанные уравнением регрессии, вычисляются с помощью уравнения, записанного в матричной форме:

$$\hat{Y} = X \times B, \quad (11.2)$$

где \hat{Y} – вектор-столбец значений, предсказанных уравнением регрессии (11.2).

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ y_n \end{bmatrix}.$$

Тогда остаточную дисперсию можно определить по формуле $s_R^2 = (Y \times Y - B \times X \times Y) / (n - m - 1)$, $f_R = (n - m - 1)$ – число степеней свободы.

Если априори есть достаточно оснований для выбора степени полинома (для анализа была использована вторая степень полинома, а в общем случае может быть любая), то остаточную дисперсию s_R^2 можно рассматривать как оценку дисперсии $\sigma^2\{y\}$, характеризующую ошибку, содержащуюся в наблюдениях зависимой переменной y .

Теперь остается найти дисперсии, характеризующие ошибки в определении коэффициентов регрессии и ковариации, определяющие статистическую зависимость между коэффициентами регрессии.

Дисперсия ошибки в определении коэффициентов регрессии вычисляется с использованием диагональных элементов обратной матрицы

$$\sigma^2\{b_i\} = c_{ii} \sigma^2\{y\}, \quad i = 1, 2, \dots, m,$$

где c_{ii} – диагональный элемент матрицы $(X \times X)^{-1}$.

Ковариация между коэффициентами регрессии рассчитывается по формуле

$$\text{cov}\{b_i b_j\} = c_{ij} \sigma^2\{y\}, \quad i \neq j = 1, 2, \dots, m.$$

Диагональные элементы матрицы $(X \times X)^{-1}$ определяют дисперсии коэффициентов регрессии, недиагональные элементы – ковариации соответствующих им коэффициентов регрессии. Поэтому матрица $(X \times X)^{-1}$ называется матрицей ошибок, или корреляционной матрицей.

Зная дисперсии ошибок в определении коэффициентов регрессии, можно установить доверительные границы для каждого из коэффициентов регрессии (в предположении диагональности корреляционной матрицы [13]) $b_i \pm t \sigma\{b_i\}$, где t – табличное значение критерия Стьюдента, выбираемое для уровня значимости α и числа степеней свободы f_R , которое связано с дисперсией S_R^2 .

Можно также проверить нуль-гипотезу равенства нулю коэффициентов регрессии. В этом случае последовательно вычисляют t значения для каждого коэффициента регрессии

$$t_i = |b_i / \sigma\{b_i\}|, \quad i = 1, 2, \dots, m, \quad (11.3)$$

которые затем сравнивают с табличными значениями t , выбираемыми для уровня значимости α и числа степеней свободы f_R .

Если расчетное значение t_i окажется больше табличного t , то нуль-гипотеза отбрасывается как маловероятная и принимается альтернативная гипотеза о неравенстве нулю проверяемого коэффициента регрессии.

Факторы с незначимыми коэффициентами регрессии должны исключаться из структуры уравнения. При проверке по критерию (11.3) могут оказаться незначимыми несколько коэффициентов регрессии. Отбрасывание незначимых факторов должно проводиться последовательно по одному, начиная с фактора с малым значением t_i . После отбрасывания фактора необходимо вновь рассчитать коэффициенты уравнения регрессии и повторно провести оценку значимости

полученных коэффициентов. В уравнении должны оставаться факторы со значимыми коэффициентами регрессии. Количество оставшихся факторов k в уравнении регрессии не больше m .

В качестве величины, характеризующей вклад k коэффициентов регрессии в уравнение, содержащее $(k + 1)$ членов, вводят множественный коэффициент корреляции R

$$R = (1 - \sum(y - \hat{y})^2 / \sum(y - y_{\text{cp}})^2)^{1/2} = (S_k / \sum(y - y_{\text{cp}})^2)^{1/2},$$

где y_{cp} – среднее арифметическое значение зависимой переменной; S_k – сумма квадратов, относящаяся к k -коэффициентам регрессии.

Дисперсия, относящаяся к k -коэффициентам регрессии, рассчитывается по формуле

$$s_k^2 = S_k / f_k, \quad f_k = k.$$

Величина R может изменяться в пределах от 0 до +1. Если уравнение регрессии полностью описывает результаты наблюдений, то $\sum(y - \hat{y})^2 = 0$ и $R = 1$. Если вклад, вносимый k -коэффициентами регрессии, равен нулю $\sum(y - \hat{y})^2 = \sum(y - y_{\text{cp}})^2$, то $R = 0$.

Коэффициент множественной корреляции R может интерпретироваться как мера линейной связи между наблюдаемыми значениями зависимой переменной Y и множеством независимых переменных X , полученных после линеаризации параболического уравнения.

Значимость множественного коэффициента корреляции определяется F -отношением $F = s_k^2 / S_R^2$.

Если расчетное значение F -отношения выше табличного значения F_T , выбранного для уровня значимости α и числа степеней свободы f_k и f_R соответственно для числителя и знаменателя, то сформулированная гипотеза о значимости множественного коэффициента корреляции R считается принятой.

11.3. Проверка выполнения предпосылок методом наименьших квадратов (МНК)

Выполнение предпосылок МНК проверяется на основе анализа остаточной компоненты. Анализ остатков позволяет получить представление, насколько хорошо подобрана сама модель и правильно выбран метод оценки коэффициентов. Согласно общим предположениям регрессионного анализа остатки должны вести себя как независимые (в действительности почти независимые) одинаково распреде-

ленные случайные величины. В классических методах регрессионного анализа предполагается также нормальный закон распределения остатков.

Исследование остатков полезно начинать с изучения их графика. График может показать наличие какой-то зависимости, не учтенной в модели. Скажем, при подборе простой линейной зависимости между Y и X график остатков может показать необходимость перехода к нелинейной модели (квадратичной, полиномиальной, экспоненциальной) или включения в модель периодических компонент.

График остатков хорошо показывает резко отклоняющиеся от модели наблюдения – выбросы. Подобным аномальным наблюдениям надо уделять особо пристальное внимание, так как их присутствие может грубо исказить значения оценок. Устранение эффектов выбросов может проводиться либо с помощью удаления этих точек из анализируемых данных (эта процедура называется цензурированием), либо с помощью применения методов оценивания параметров, устойчивых к подобным грубым отклонениям.

Автокорреляция случайной составляющей нарушает одну из предпосылок нормальной линейной модели регрессии. Наличие (отсутствие) автокорреляции в отклонениях проверяют с помощью критерия Дарбина – Уотсона. Установить наличие автокорреляции остатков можно вычислив первый коэффициент автокорреляции

$$r(1) = \left(\sum_{i=2}^n \varepsilon_i \varepsilon_{i-1} \right) / \sum_{i=1}^n \varepsilon_i^2.$$

Для принятия решения о наличии или отсутствии автокорреляции в исследуемом ряду фактическое значение коэффициента автокорреляции $r(1)$ сопоставляется с табличным (критическим) значением для 5%-ного уровня значимости (вероятности допустить ошибку при принятии нулевой гипотезы о независимости уровней ряда). Если фактическое значение коэффициента автокорреляции меньше табличного, то гипотеза об отсутствии автокорреляции в ряду может быть принята, а если фактическое значение больше табличного – делают вывод о наличии автокорреляции в ряду динамики.

Для обнаружения гетероскедастичности обычно используют тесты, в которых делаются различные предположения о зависимости между дисперсией случайного члена и объясняющей переменной.

При малом объеме выборки для оценки гетероскедастичности может использоваться метод Голдфельда – Квандта.

Данный тест применяется для проверки такого типа гетероскедастичности, когда дисперсия остатков возрастает пропорционально квадрату фактора. При этом делается предположение, что случайная составляющая e распределена нормально.

Чтобы оценить нарушение гетероскедастичности по тесту Голдфельда – Квандта, необходимо выполнить следующие шаги. Упорядочение n наблюдений по мере возрастания переменной x .

1. Разделение совокупности на две группы n_1 и n_2 (соответственно с малыми и большими значениями фактора x) и определение по каждой из групп уравнений регрессии.

2. Определение остаточной суммы квадратов для первой регрессии $S_{1e} = \sum_{i=1}^{n_1} (y_i - y_{i1})^{-2}$ и второй регрессии $S_{2e} = \sum_{i=1}^{n_2} (y_i - y_{i2})^{-2}$.

3. Вычисление отношений $\frac{S_{2e}}{S_{1e}}$ (или $\frac{S_{1e}}{S_{2e}}$). В числителе должна быть большая сумма квадратов.

Полученное отношение имеет F распределение со степенями свободы $k_1 = n_1 - m$ и $k_2 = n_2 - m$, (m – число оцениваемых параметров в уравнении регрессии). Если $F_{\text{набл}} = \frac{S_{1e}}{S_{2e}} > F_{\text{кр}(\alpha, k_1, k_2)}$, то гетероскедастичность имеет место.

11.4. Оценка влияния отдельных факторов на зависимую переменную на основе модели

Важную роль при оценке влияния факторов играют коэффициенты регрессионной модели. Однако непосредственно с их помощью нельзя сопоставить факторы по степени их влияния на зависимую переменную из-за различия единиц измерения и разной степени колеблемости. Для устранения таких различий при интерпретации применяются средние частные коэффициенты эластичности $\mathcal{E}(j)$ и бета-коэффициенты $\beta(j)$, которые рассчитываются соответственно по формулам

$$\mathcal{E}_j = b_j \frac{x_j}{y}, \quad j = 1, 2, \dots, k,$$

$$\beta_j = b_j \frac{S_{xj}}{S_y}, \quad j = 1, 2, \dots, k,$$

где S_{xj} – среднее квадратическое отклонение фактора j ; S_y – среднее квадратическое отклонение зависимой переменной y .

Коэффициент эластичности показывает, на сколько процентов изменяется зависимая переменная при изменении фактора j на один процент. Однако он не учитывает степень колеблемости факторов.

Бета-коэффициент показывает, на какую часть величины среднего квадратического отклонения S_y изменится зависимая переменная Y с изменением соответствующей независимой переменной X_j на величину своего среднее квадратического отклонения при фиксированном на постоянном уровне значении остальных независимых переменных.

Указанные коэффициенты позволяют упорядочить факторы по степени влияния факторов на зависимую переменную. Долю влияния фактора в суммарном влиянии всех факторов можно оценить по величине дельта-коэффициентов $\Delta(j)$

$$\Delta_j = r_{y,x_j} \beta_j / R^2,$$

где r_{y,x_j} – коэффициент парной корреляции между фактором j ($j = 1, \dots, m$) и зависимой переменной.

11.5. Построение точечных и интервальных прогнозов на основе регрессионной модели

Одна из важнейших целей моделирования заключается в прогнозировании поведения исследуемого объекта. Обычно термин «прогнозирование» используется в тех ситуациях, когда требуется предсказать состояние системы в будущем. Для регрессионных моделей он имеет, однако, более широкое значение. Как уже отмечалось, данные могут не иметь временной структуры, но и в этих случаях вполне может возникнуть задача оценки значения зависимой переменной для некоторого набора независимых объясняющих переменных, которых нет в исходных наблюдениях. Именно в этом смысле – как построение оценки зависимой переменной и следует понимать прогнозирование.

При использовании построенной модели для прогнозирования делается предположение о сохранении в период прогнозирования существовавших ранее взаимосвязей переменных.

Для того чтобы определить область возможных значений результативного показателя, при рассчитанных значениях факторов следует учитывать два возможных источника ошибок: рассеивание наблюдений относительно линии регрессии и ошибки, обусловленные математическим аппаратом построения самой линии регрессии. Ошибки первого рода измеряются с помощью характеристик точности, в частности, величиной S_y . Ошибки второго рода обусловлены фиксацией численного значения коэффициентов регрессии, в то время как они в действительности являются случайными, нормально распределенными величинами.

Для линейной модели регрессии доверительный интервал рассчитывается следующим образом. Оценивается величина отклонения от линии регрессии (обозначим ее U)

$$u = S_{\varepsilon} t_{\alpha} \sqrt{V_{\text{пр}}} = S_{\varepsilon} t_{\alpha} \sqrt{1 + X_{\text{пр}}^T (X^T X)^{-1} X_{\text{пр}}},$$

где $X_{\text{пр}}^T = (1, X_{1\text{пр}}, X_{2\text{пр}}, \dots, X_{k\text{пр}})$.

Применение классического регрессионного анализа для обработки результатов многофакторных наблюдений может приводить к разочарованию [13]. Здесь нужно учитывать следующие обстоятельства.

1. При обработке результатов наблюдений (пассивный многофакторный эксперимент) трудно оценить ошибку наблюдений, следовательно, нельзя строго проверить гипотезу об адекватности представления результатов наблюдений выбранной математической моделью.

2. Невозможно построить строгий критерий для отбрасывания аномальных данных, содержащих грубые ошибки.

3. Независимые переменные x_1, x_2, x_3, \dots или хотя бы часть из них часто оказываются попарно коррелированными, поэтому соответствующие эффекты невозможно разделить.

4. В хорошо организованных процессах и протекающих явлениях независимые переменные могут варьироваться в очень узком интервале значений. В этом случае исследователь находится перед неразрешимой задачей: ему нужно описать многофакторный процесс по

результатам измерений, находящихся в окрестности одной точки в многомерном факторном пространстве.

Пример. Имеются условные данные о деятельности крупнейших компаний (табл. 11.1).

Таблица 11.1

Деятельность крупных компаний

№ п/п	Чистый доход Y , млрд у. е.	Обходный капитал $X1$, млрд у. е.	Использованный капитал $X2$, млрд у. е.	Численность служащих $X3$, тыс. чел.	Рыночная капитализация компании $X4$, млрд у. е.
1	0,9	31,3	18,9	43	40,9
2	1,7	13,4	13,7	64,7	40,5
3	0,7	4,5	18,5	24	38,9
4	1,7	10	4,8	50,2	38,5
5	2,6	20	21,8	106	37,3
6	1,3	15	5,8	96,6	26,5
7	4,1	137,1	99	347	37
8	1,6	17,9	20,1	85,6	36,8
9	6,9	165,4	60,6	745	36,3
10	0,4	2	1,4	4,1	35,3
11	1,3	6,8	8	26,8	35,3
12	1,9	27,1	18,9	42,7	35
13	1,9	13,4	13,2	61,8	26,2
14	1,4	9,8	12,6	212	33,1
15	0,4	19,5	12,2	105	32,7
16	0,8	6,8	3,2	33,5	32,1
17	1,8	27	13	142	30,5
18	0,9	12,4	6,9	96	29,8
19	1,1	17,7	15	140	25,4
20	1,9	12,7	11,9	59,3	29,3
21	-0,9	21,4	1,6	131	29,2
22	1,3	13,5	8,6	70,7	29,2
23	2	13,4	11,5	65,4	29,1
24	0,6	4,2	1,9	23,1	27,9
25	0,7	15,5	5,8	80,8	27,2

Расчеты выполним с применением стандартного пакета программ Statgraphics Plus-Untitled StatFolio. Прежде всего находим коэффициенты парной корреляции для всех переменных – они приведены в матрице R .

$$R = \begin{bmatrix} 1,0 & 0,848 & 0,763 & 0,829 & 0,2689 \\ 0,848 & 1,0 & 0,897 & 0,911 & 0,248 \\ 0,763 & 0,897 & 1,0 & 0,712 & 0,348 \\ 0,829 & 0,911 & 0,712 & 1,0 & 0,115 \\ 0,268 & 0,248 & 0,3484 & 0,115 & 1,0 \end{bmatrix}.$$

Сильная корреляционная зависимость выявлена между следующими переменными: $r_{x1, x2} = 0,911$; $r_{x1, x3} = 0,897$. В разрабатываемой модели не должны присутствовать сильно коррелированные факторы, чего можно добиться либо исключением из числа влияющих факторов независимой переменной $X1$, либо одновременным исключением переменных $X2$ и $X3$. Исключаем фактор $X1$. В качестве факторов выберем независимые переменные $X2$, $X3$ и $X4$.

Рассмотрим точность описания результатов наблюдений линейным уравнением регрессии. Рассчитанные параметры уравнения регрессии со статистическими оценками приведены в табл. 11.2.

Таблица 11.2

Параметры уравнения регрессии

Parameter	Estimate	Standard Error	<i>t</i> -Statistic	<i>P</i> -Value
CONSTANT	-0,394447	1,15102	-0,342692	0,7352
$X2$	0,0206713	0,0115195	1,79446	0,0871
$X3$	0,00583944	0,00151125	3,86398	0,0009
$X4$	0,0289384	0,0356879	0,810875	0,4265

Расчетные *t*-значения для коэффициентов регрессии при факторах $X4$ и $X2$ меньше табличного значения, выбранного для уровня значимости $\alpha = 0,05$ и числа степеней свободы $f_R = 21$, равного 2,07. Коэффициенты статистически незначимы. Последовательно проведем исключение из структуры модели фактора $X4$, а при незначимости и фактора $X2$. После исключения фактора $X4$ оставшиеся коэффициенты регрессии стали значимыми. Результаты повторного регрессионного анализа приведены в табл. 11.3.

Коэффициент детерминации $R^2 = 0,748$, а коэффициент множественной корреляции $R = 0,86$. Значимость множественного коэффициента корреляции определяется с помощью *F*-отношения.

Расчетное значение *F*-отношения равно 32,72. Табличное значение F_T -критерия для уровня значимости $\alpha = 0,05$ и числа степеней сво-

боды $f_k = 2$ и $f_R = 22$ равно 3,44. Так как расчетное значение F -отношения больше табличного, то можно говорить о значимости (адекватности) полученного линейного уравнения регрессии.

Таблица 11.3

Параметры уточненного уравнения регрессии

Parameter	Estimate	Standard Error	t -Statistic	P -Value	
CONSTANT	0,524672	0,198571	2,64224	0,0149	
X_2	0,0242418	0,0105616	2,29528	0,0316	
X_3	0,00559141	0,00146841	3,80781	0,0010	
Analysis of Variance					
Source	Sum of Squares	Df	Mean Square	F -Ratio	P -Value
Model	37,2118	2	18,6059	32,72	0,0000
Residual	12,5082	22	0,568555		

24

Total (Corr.) 49,72

R-squared = 0,748

R-squared (adjusted for d.f.) = 0,725

Standard Error of Est. = 0,754

Mean absolute error = 0,553

Durbin-Watson statistic = 2,20

Исследуемый экономический процесс может быть описан линейным уравнением $\hat{y} = 0,524672 + 0,0242418 X_2 + 0,00559141 X_3$. Точность полученного уравнения регрессии можно оценить величиной стандартной ошибки, равной $s_R = 0,754$, и средней абсолютной ошибкой, равной 0,553.

Можно было бы выбрать и другие структуры регрессионного уравнения, адекватно описывающие результаты наблюдений. Например, отбросить переменные x_2 и x_3 , сильно коррелированные с x_1 , или добавить к линейному уравнению члены, характеризующие парные взаимодействия типа $x_i x_j$, либо квадратичные члены типа x_i^2 .

После отбрасывания коррелированных переменных x_2 и x_3 получаем уравнение парной регрессии $\hat{y} = 0,756 + 0,0315 x_1$, адекватно описывающее результаты наблюдений. Точность полученного уравнения регрессии характеризуется стандартной ошибкой, равной $s_R = 0,78$, и средней абсолютной ошибкой, равной 0,576.

Формально оценивая результаты регрессионного анализа, можно говорить лишь о наличии статистической связи между переменными, но нельзя ничего сказать о том, фактически какой характер носит эта связь. Поэтому не имеет смысла придавать какое-либо значение индивидуальным коэффициентам регрессии.

Любое из вышеполученных адекватных уравнений может быть использовано как интерполяционная формула. Линейное уравнение парной регрессии имеет преимущество только в смысле своей простоты.

Если мы ищем математическую модель процесса или явления, чтобы использовать ее в дальнейшем для управления этим процессом, то такая неопределенность в результатах исследований в значительной степени лишает их смысла. Неудачной здесь представляется сама постановка задачи [13]. Результаты наблюдений (пассивного эксперимента), протекающих в условиях больших помех, при сильных ограничениях, наложенных на интервалы варьирования независимых переменных, не содержат информации о математической модели процесса. Отсюда не следует, что нужно полностью отказаться от обработки данных наблюдений, а нужно более внимательно относиться к использованию регрессионного анализа при математическом описании исследуемых процессов и явлений.

Контрольные вопросы

1. Для чего предназначен регрессионный анализ?
2. Назовите ограничения в применении метода наименьших квадратов.
3. В чем заключается статистический анализ уравнения регрессии?
4. Как оценивается влияние отдельных факторов на зависимую переменную?
5. Как используются многофакторные модели для прогнозирования развития явлений и процессов?
6. Сформулируйте предпосылки метода наименьших квадратов в регрессионном анализе.
7. Для чего выполняется оценка чувствительности зависимой переменной к изменению факторных переменных?
8. Какие выводы можно сделать, формально оценивая результаты регрессионного анализа?

12. НЕЛИНЕЙНАЯ РЕГРЕССИЯ

Различают два вида моделей нелинейной регрессии:

– нелинейная относительно включенных в структуру моделей влияющих факторов, но линейная по оцениваемым параметрам

$$y = a + vx + cx^2 + \dots + dx^3;$$

– нелинейная по оцениваемым параметрам $y = e^{a+bx}$.

Приступая к исследованию системы или процесса, экспериментатору в ряде случаев аналитическое выражение функции отклика бывает неизвестно. В этом случае приходится ограничиваться представлением ее полиномом $\hat{y} = \beta_0 + \sum \beta_i x_i + \sum \beta_{ij} x_i x_j + \sum \beta_{ii} x_i^2 + \dots$ с коэффициентами регрессии $\beta_0, \beta_i, \beta_{ij}, \beta_{ii}, \dots$.

Используя результаты наблюдений, можно определить только выборочные коэффициенты регрессии $b_0, b_i, b_{ij}, b_{ii}, \dots$, которые являются оценками для теоретических коэффициентов регрессии $\beta_0, \beta_i, \beta_{ij}, \beta_{ii}, \dots$.

Уравнение регрессии, полученное на основе наблюдений, запишем в следующем виде:

$$\hat{y} = b_0 + \sum b_i x_i + \sum b_{ij} x_i x_j + \sum b_{ii} x_i^2 + \dots,$$

где \hat{y} – значение зависимой (результатирующей) переменной, предсказанное уравнением регрессии.

Для упрощения расчетов коэффициентов регрессии заменим члены второго порядка линейными, введя обозначения $x_i^2 = x_{k+i}$ для $i = 1, 2, \dots, k$, $x_i x_j = x_{2k+i}$ для $i, j = 1, 2, 3, \dots, k$, $i \neq j$. Аналогичным образом линейными членами можно заменить члены любого порядка.

В новой системе обозначений полином любой степени запишем однородным линейным уравнением

$$\hat{y} = b_0 + b_1 x_{1u} + \dots + b_2 x_{2u} + \dots + b_m x_{mu}.$$

Коэффициенты регрессии находим методом наименьших квадратов $B = (XX)^{-1}XY$.

В экономике наиболее часто встречаются следующие виды нелинейных моделей [15, 16):

1. Парабола $y = a + bx + cx^2 + e$.

Этими моделями описываются потребление товара от уровня дохода семьи, затраты от объема выпуска продукции, урожайность от количества внесенного удобрения и др.

2. Равносторонняя гипербола $y = a + b/x + e$.

Кривая Филипса – процент прироста зарплаты от нормы безработицы ($b > 0$), кривая Энгеля – зависимость доли расходов на товары длительного пользования от общих сумм расходов (доходов) ($b < 0$), трудоемкость продукции от масштабов производства, валовый доход от уровня занятости.

3. Полулогарифмическая кривая $y = a + b \ln x + e$. Модель предложили Уоркинг (1943 г.) и Лизер (1964 г.).

4. Модель $y = a + b\sqrt{x} + e$. Используется для описания урожайности, трудоемкости сельскохозяйственного производства.

5. Степенная функция $y = ax^b e$. Описывается эластичность спроса (спрашиваемое количество) от цены (b -коэффициент эластичности), ставка межбанковского кредита (%) от срока его предоставления (дни).

6. Экспоненциальная функция $y = \exp(a + bx)e$.

Описанные выше нелинейные модели решаются приведением их к линейной форме с использованием метода наименьших квадратов.

7. Существует вид внутренне нелинейных моделей, которые с помощью преобразований невозможно привести к линейной форме, например: $y = a + bx^c + e$.

Обратная функция $y = 1/(a + bx + e)$. Этой моделью описывается рентабельность продукции от трудоемкости.

Показательная функция $y = ab^x e$.

При линеаризации нелинейных функций происходит преобразование зависимой переменной y , поэтому следует проверять выполнение предпосылок МНК:

- случайный характер остатков « e »;
- нулевое значение среднего остатка, независимость от x_i ;
- одинаковая дисперсия остатка « e » для всех значений x_i (гомоскедастичность);
- отсутствие автокорреляции остатков;
- остатки описываются нормальным законом распределения.

Выполнение предпосылок МНК проверяется на основе анализа остаточной компоненты. Методика проверки описана в п. 11.3.

Влияние отдельных факторов на зависимую переменную оценивается по коэффициенту эластичности. Коэффициент эластичности для нелинейных функций рассчитывается по формуле $\mathcal{E} = y'(x) \frac{x}{y(x)}$, где $y'(x)$ – производная нелинейной функции по x .

Коэффициенты эластичности для некоторых нелинейных функций приведены в табл. 12.1.

Таблица 12.1

Коэффициенты эластичности для нелинейных функций

Вид функции	Производственная функция $y'(x)$	Коэффициент эластичности \mathcal{E}
Полином $y = a + bx$	b	$bx/(a + bx)$
Парабола $y = a + bx + cx^2$	$b + 2cx$	$(b + 2cx)x/(a + bx + cx^2)$
Гипербола $y = a + b/x$	$-b/x^2$	$-b/x^2x/(a + b/x)$
Показательная функция $y = ab^x$	$\ln(bab^x)$	$x \ln b$
Степенная функция $y = ax^b$	abx^{b-1}	b
Обратная функция $y = 1/(ax + b)$	$-b/(a + bx)^2$	$-bx/(a + bx)$

Коэффициент эластичности не является величиной постоянной, он зависит от величины x и представляет экономический интерес, но возможны случаи, когда он не имеет экономического смысла. Например, изменение зарплаты с ростом стажа работы на 1 %.

Если между анализируемыми (экономическими) явлениями существуют нелинейные соотношения, то они выражаются с помощью соответствующих нелинейных функций [15, 16, 17].

Пример. По семи предприятиям легкой промышленности региона получена информация, характеризующая зависимость объема выпуска продукции (Y , млн руб.) от объема капиталовложений (X , млн руб.).

Y	64	56	52	48	50	46	38
X	64	68	82	76	84	96	100

Требуется:

1. Для характеристики Y от X построить следующие модели:
 - линейную (для сравнения с нелинейными);
 - степенную;
 - показательную;
 - гиперболическую.
2. Оценить каждую модель, определив:
 - индекс корреляции;
 - среднюю относительную ошибку;
 - коэффициент детерминации;
 - F -критерий Фишера.
3. Составить сводную таблицу вычислений, выбрать лучшую модель, дать интерпретацию рассчитанных характеристик.
4. Рассчитать прогнозные значения результативного признака по лучшей модели, если объем капиталовложений составит 89,573 млн руб.
5. Результаты расчетов отобразить на графике.

Решение:

1. Построение линейной модели парной регрессии

Определим линейный коэффициент парной корреляции по следующей формуле:

$$r_{Y,X} = \frac{\sum (y - \bar{y})(x - \bar{x})}{\sqrt{\sum (y - \bar{y})^2 \sum (x - \bar{x})^2}} = \frac{-593,714}{\sqrt{397,71 \cdot 1077,71}} = -0,907.$$

Можно сказать, что связь между объемом капиталовложений X и объемом выпуска продукции Y обратная, достаточно сильная.

Уравнение линейной регрессии имеет вид $\hat{y} = a + bx$. Значения параметров a и b линейной модели равны

$$b = \frac{\overline{yx} - \bar{y} \bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{4033,14 - 50,57 \cdot 81,43}{6784,57 - 81,43^2} = -0,55;$$

$$a = \bar{y} - b\bar{x} = 50,57 + 0,55 \cdot 81,43 = 95,36.$$

Уравнение линейной регрессии имеет вид $\hat{y} = 95,36 - 0,55x$.

С увеличением объема капиталовложений на 1 млн руб. объем выпускаемой продукции уменьшится в среднем на 550 тыс. руб. Это свидетельствует о неэффективности работы предприятий, и необходимо принять меры для выяснения причин и устранения этого недостатка.

Рассчитаем коэффициент детерминации $R^2 = R_{YX}^2 = 0,822$. Вариация результата Y (объема выпуска продукции) на 82,2 % объясняется вариацией фактора X (объемом капиталовложений). Оценку значимости уравнения регрессии проведем с помощью F -критерия Фишера

$$F = \frac{r_{Y,X}^2}{1 - r_{Y,X}^2} (n - 2) = \frac{0,822}{1 - 0,822} (7 - 2) = 23,09.$$

$$F > F_{\text{табл}} = 6,61 \text{ для } \alpha = 0,05 ; k_1 = m = 1, k_2 = n - m - 1 = 5.$$

Уравнение регрессии с вероятностью 0,95 в целом статистически значимое, так как $F > F_{\text{табл}}$. Определим среднюю относительную ошибку

$$\bar{E}_{\text{отн}} = \frac{1}{n} \sum \frac{|E_i|}{y} 100 \% = \frac{1}{n} \sum \left| \frac{y - \hat{y}}{y} \right| 100 \% = \frac{39,798}{7} = 5,685 \%$$

В среднем расчетные значения \hat{y} для линейной модели отличаются от фактических значений на 5,68 %.

2. Построение степенной модели парной регрессии.

Уравнение степенной модели имеет вид $\hat{y} = ax^b$. Для построения этой модели необходимо провести линеаризацию переменных. Для этого проведем логарифмирование обеих частей уравнения

$$\lg \hat{y} = \lg a + b \lg x.$$

Обозначим $Y = \lg \hat{y}$, $X = \lg x$, $A = \lg a$. Тогда уравнение примет вид $Y = A + b X$ – линеаризованное уравнение регрессии. Преобразуем данные для расчетов, которые сведены в табл. 12.2.

Таблица 12.2

Расчетные данные

№ п/п	$Y(t)$	$\text{Lg}(Y)$	$X(t)$	$\text{Lg}(x)$
1	64,0	1,806	64	1,806
2	56,0	1,748	68	1,833
3	52,0	1,716	82	1,914
4	48,0	1,681	76	1,881
5	50,0	1,699	84	1,924
6	46,0	1,663	96	1,982
7	38,0	1,580	100	2,000
Сумма	354	11,893	570	13,340
Среднее значение	50,5714	1,699	81,429	1,906

Рассчитаем параметры модели, используя данные таблицы

$$b = \frac{\overline{YX} - \overline{Y}\overline{X}}{\overline{X^2} - \overline{X}^2} = \frac{3,2339 - 1,699 \cdot 1,9057}{3,6361 - 1,9057 \cdot 1,9057} = -0,8921;$$

$$A = \overline{Y} - b \overline{X} = 1,699 + 0,8921 \cdot 1,9057 = 3,3991.$$

Линеаризованное уравнение регрессии будет иметь вид $y = 3,3991 - 0,8921x$.

Перейдем к исходным переменным x и y , выполнив потенцирование параметров данного уравнения, $\overline{y} = 10^{3,399} \cdot x^{-0,892}$.

Получим уравнение степенной модели регрессии $\overline{y} = 2506 \cdot 915 \times \times x^{-0,892}$.

$$\text{Определим индекс корреляции } \rho_{YX} = \sqrt{1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \overline{y})^2}} = 0,914.$$

Связь между показателем y и фактором x можно считать достаточно сильной. Коэффициент детерминации равен 0,836

$$R^2 = \rho_{YX}^2 = 0,914^2 = 0,836.$$

Вариация результата y (объема выпуска продукции) на 83,6 % объясняется вариацией фактора x (объемом капиталовложений),

$$\text{Рассчитаем } F\text{-критерий Фишера } F = \frac{R^2}{1 - R^2} (n - 2) = \frac{0,836}{1 - 0,836} \cdot 5 = 25,5.$$

$$F > F_{\text{табл}} = 6,61 \text{ для } \alpha = 0,05, k_1 = m = 1, k_2 = n - m - 1 = 5.$$

Уравнение регрессии с вероятностью 0,95 в целом статистически значимо, так как $F > F_{\text{табл}}$. Средняя относительная ошибка модели

$$\overline{E}_{\text{отн}} = \frac{1}{n} \sum \frac{|E_i|}{y} 100 \% = \frac{1}{n} \sum \left| \frac{y - \hat{y}}{y} \right| 100 \%,$$

$$\overline{E}_{\text{отн}} = \frac{42,32}{7} = 6,04 \%.$$

В среднем расчетные значения \hat{y} для степенной модели отличаются от фактических значений на 6,04 %.

3. Построение показательной функции

Уравнение показательной кривой имеет вид $\hat{y} = ab^x$. Для построения этой модели необходимо произвести линеаризацию переменных. Для этого выполним логарифмирование обеих частей уравнения $\lg \hat{y} = \lg a + x \lg b$. Обозначим $Y = \lg \hat{y}$, $B = \lg b$, $A = \lg a$. Получим линейное уравнение регрессии: $Y = A + Bx$. Рассчитаем его параметры, используя данные табл. 12. 3.

Таблица 12.3

Параметры линейного уравнения регрессии

t	y	Y	x	Yx	x^2	$Y - \bar{Y}$	$(Y - \bar{Y})^2$	$x - \bar{x}$	$(x - \bar{x})^2$	\hat{y}	$(y - \hat{y})^2$	ε_i	$\left \frac{\varepsilon_i}{y_i} \right \cdot 100\%$
1	64	1,8062	64	115,60	4096	0,1072	0,0115	-17,43	303,76	60,6	11,464	3,3859	5,290
2	56	1,7482	68	118,88	4624	0,0492	0,0024	-13,43	180,33	58	3,9632	-1,991	3,555
3	52	1,7160	82	140,71	6724	0,0170	0,0003	0,57	0,33	49,7	5,4221	2,3285	4,478
4	48	1,6812	76	127,77	5776	-0,017	0,0003	-5,43	29,47	53,1	25,804	-5,08	10,583
5	50	1,6990	84	142,71	7056	0,0000	0,0000	2,57	6,61	48,6	2,0031	1,4153	2,831
6	46	1,6628	96	159,62	9216	-0,036	0,0013	14,57	212,33	42,5	11,933	3,4544	7,509
7	38	1,5798	100	157,98	10000	-0,119	0,0142	18,57	344,90	40,7	7,3132	-2,704	7,117
Сум- ма	354	11,8931	570	963,28	4749	-	0,0300	-	1077,7	-	67,903	0,8093	41,363
Ср. зна- чение	50,57	1,6990	81,4	137,61	6785	-	-	-	-	-	-	-	5,909

$$B = \frac{\overline{Yx} - \bar{Y}\bar{x}}{x^2 - \bar{x}^2} = \frac{137,61 - 1,699 \cdot 81,43}{6785 - 81,43 \cdot 81,43} = -0,0048;$$

$$A = \bar{Y} - B\bar{x} = 1,699 + 0,0048 \cdot 81,43 = 2,09.$$

Уравнение будет иметь вид $Y = 2,09 - 0,0048x$. Перейдем к исходным переменным x и y , выполнив потенцирование данного уравнения

$$\hat{y} = 10^{2,09} (10^{-0,0048})^x = 123,03 \cdot 0,989^x.$$

Определим индекс корреляции

$$\rho_{YX} = \sqrt{1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}} = \sqrt{1 - \frac{67,9}{397,71}} = 0,91.$$

Связь между показателем y и фактором x можно считать достаточно сильной. Индекс детерминации $R^2 = \rho_{YX}^2 = 0,91^2 = 0,828$. Вариация результата Y (объема выпуска продукции) на 82,8 % объясняется вариацией фактора X (объемом капиталовложений). Рассчитаем F -критерий Фишера $F = \frac{R^2}{1 - R^2} (n - 2) = 24,06$.

$$F > F_{\text{табл}} = 6,61 \text{ для } \alpha = 0,05; k1 = m = 1, k2 = n - m - 1 = 5.$$

Уравнение регрессии с вероятностью 0,95 в целом статистически значимое, так как $F > F_{\text{табл}}$.

$$\text{Средняя относительная ошибка } \bar{E}_{\text{отн}} = \frac{41,363}{7} = 5,909 \text{ \%}.$$

В среднем расчетные значения \hat{y} для показательной функции отличаются от фактических данных на 5,9 %.

4. Построение гиперболической функции

Уравнение гиперболической функции имеет вид $\hat{y} = a + b/x$.

Произведем линейризацию модели путем замены $X = 1/x$. В результате получим линейное уравнение $\hat{y} = a + bX$. Рассчитаем его параметры по данным табл. 12.4.

Таблица 12.4

Параметры линейного уравнения

t	y	x	X	yX	X^2	$Y - \bar{Y}$	$(Y - \bar{Y})^2$	\hat{y}	ε_i	$(y - \bar{y})^2$	$ \varepsilon_i / y_i \cdot 100\%$
1	64	64	0,015	1,000	0,000244	13,4	180,33	61,5	2,48	6,195	3,889
2	56	68	0,014	0,823	0,000216	5,43	29,47	58,2	-2,22	4,963	3,978
3	52	82	0,012	0,634	0,000148	1,43	2,04	49,3	2,74	7,508	5,270
4	48	76	0,013	0,631	0,000173	-2,57	6,61	52,7	-4,69	22,07	9,789
5	50	84	0,011	0,595	0,000141	-0,57	0,3265	48,2	1,777	3,1591	3,555
6	46	96	0,0104	0,4792	0,000108	-4,57	20,90	42,9	3,093	9,5648	6,723
7	38	100	0,0100	0,3800	0,000100	-12,5	158,04	41,4	-3,41	11,69	8,997
Сумма	354	-	0,0880	4,5437	0,001132	-	397,71	354,2	-0,24	65,159	42,202
Ср. значение	50,57	-	0,0126	0,6491	0,000161	-	-	-	-	-	6,029

$$b = \frac{\overline{yX} - \bar{y}\bar{X}}{\overline{X^2} - \bar{X}^2} = \frac{0,6491 - 50,57 \cdot 0,0126}{0,0001618 - 0,0126 \cdot 0,0126} = 3571,95;$$

$$a = \bar{y} - b\bar{X} = 50,57 - 3571,9 \cdot 0,0126 = 5,7.$$

Получим следующее уравнение гиперболической модели:

$$\hat{y} = 5,7 + 3571,9/x.$$

Определим индекс корреляции

$$\rho_{YX} = \sqrt{1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}} = \sqrt{1 - \frac{65,159}{397,71}} = 0,914.$$

Связь между показателем y и фактором x можно считать достаточно сильной. Индекс детерминации

$$R^2 = \rho_{YX}^2 = 0,914^2 = 0,835.$$

Вариация результата Y (объема выпуска продукции) на 83,5 % объясняется вариацией фактора X (объемом капиталовложений).

Рассчитаем F -критерий Фишера

$$F = \frac{R^2}{1 - R^2} (n - 2) = \frac{0,835}{1 - 0,835} 5 = 25,3.$$

$$F > F_{\text{табл}} = 6,61 \text{ для } \alpha = 0,05 ; k_1 = m = 1, k_2 = n - m - 1 = 5.$$

Уравнение регрессии с вероятностью 0,95 в целом статистически значимое, так как $F > F_{\text{табл}}$. Средняя относительная ошибка

$$\bar{E}_{\text{отн}} = \frac{1}{n} \sum \frac{|E_i|}{y} 100 \% = \frac{42,202}{7} = 6,029 \%$$

В среднем расчетные значения \hat{y} для гиперболической модели отличаются от фактических значений на 6,029 %.

Для выбора лучшей модели построим сводную таблицу результатов (табл. 12.5).

Таблица 12.5

Результаты расчетов

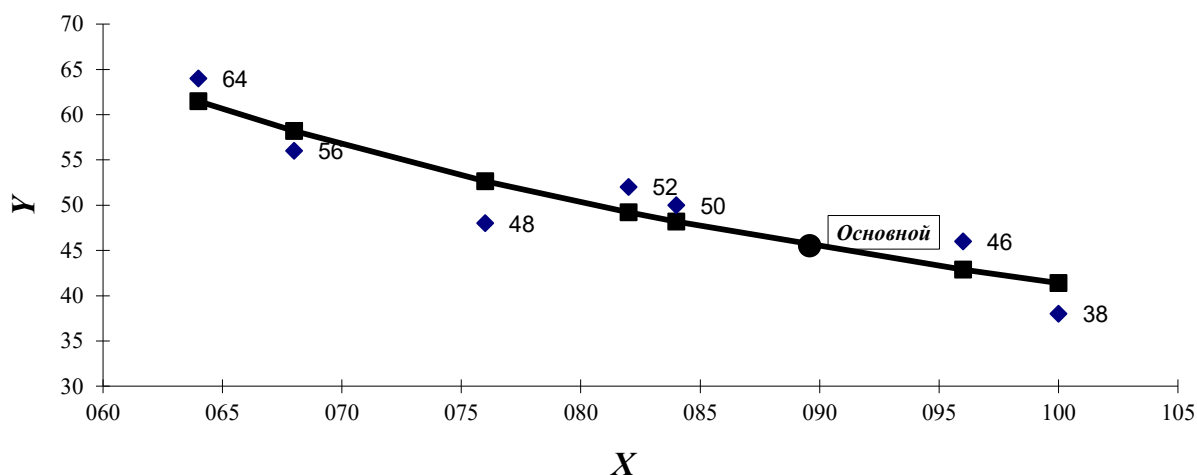
Модель	Коэффициент детерминации R^2	F -критерий Фишера	Индекс корреляции ρ_{yx} (r_{yx})	Средняя относительная ошибка $E_{\text{отн}}$
Линейная	0,822	23,09	0,907	5,685
Степенная	0,828	24,06	0,910	6,054
Показательная	0,828	24,06	0,910	5,909
Гиперболическая	0,835	25,30	0,914	6,029

Все модели имеют примерно одинаковые характеристики, но большее значение F -критерия Фишера и большее значение коэффициента детерминации R^2 имеет гиперболическая модель. Ее можно взять в качестве лучшей для построения прогноза.

Прогнозное значение результативного признака (объема выпуска продукции) определим по уравнению гиперболической модели, подставив в него планируемую (заданную по условию) величину объема капиталовложений:

$$\hat{Y}_{\text{пр}} = 5,7 + 3571,9/X_{\text{пр}} = 5,7 + 3571,9/89,573 = 45,542, \text{ млн руб.}$$

Фактические, расчетные и прогнозные значения по лучшей модели отобразим на графике (см. рисунок).



Прогноз по лучшей модели

Контрольные вопросы

1. Назовите виды моделей нелинейной регрессии.
2. Какие виды нелинейных моделей часто встречаются в экономике?
3. Почему при линеаризации нелинейных функций проверяется выполнение предпосылок МНК?

13. КОМПОНЕНТНЫЙ АНАЛИЗ

При исследовании сложных систем может оказаться, что набор параметров избыточен и для описания наиболее важных свойств достаточно иметь несколько параметров. Это кажется разумным: если среди многочисленных параметров есть такие, которые находятся в сильной связи друг с другом (а это вполне естественное предположение), то часть из них можно отбросить. Например, если один из параметров выражается через другие, то его можно просто выкинуть без потери информации.

Но как выяснить, какой именно набор параметров хорошо описывает наш набор данных, но при этом имеет небольшую избыточность? Иными словами, как уменьшить размерность пространства, в котором присутствуют данные, потеряв при этом минимум информации?

Способы решения этой задачи называются методами уменьшения размерности (dimensionality reduction). Метод главных компонент (principal components analysis, PCA) – один из них. Он очень простой и при этом довольно популярный, так что имеет смысл с ним познакомиться подробнее.

Компонентный анализ является методом определения структурной зависимости между случайными переменными. Идея метода заключается в замене сильно коррелированных переменных новыми переменными (главными компонентами), между которыми корреляция отсутствует. В результате его использования получается сжатое описание малого объема, несущее почти всю информацию, содержащуюся в исходных данных. Главные компоненты получаются из исходных переменных путем целенаправленного вращения, т. е. как линейные комбинации исходных переменных. Вращение производится таким образом, чтобы главные компоненты были ортогональны и имели максимальную дисперсию среди возможных линейных комбинаций исходных переменных. При этом переменные некоррелированы между собой и упорядочены по убыванию дисперсии (первая компонента имеет наибольшую дисперсию). Кроме того, общая дисперсия после преобразования остается без изменений.

Ход рассуждений при выполнении поиска главных компонент заключается в следующем. Мы предполагаем наличие некоррелированных переменных $Z_j (j = 1 \dots k)$, каждая из которых представляется нам комбинацией основных переменных (суммирование по $i = 1 \dots k$)

$$Z_j = \sum A_{ji} X_i$$

и, кроме того, обладает дисперсией, такой что

$$D(Z_1) \geq D(Z_2) \geq \dots \geq D(Z_k).$$

Поиск коэффициентов A_{ji} (их называют весом j -й компоненты в содержании i -й переменной) сводится к решению матричных уравнений. Суть метода весьма интересна и на ней стоит задержаться.

Как известно из векторной алгебры, диагональная матрица $[2 \times 2]$ может рассматриваться как описание двух точек (точнее, вектора) в двумерном пространстве, а такая же матрица размерами $[k \times k]$ – как описание k точек k -мерного пространства.

Замена реальных, хотя и нормированных переменных X_i на точно такое же количество переменных Z_j означает не что иное, как поворот k осей многомерного пространства.

Перебирая поочередно оси, мы находим вначале ту из них, где дисперсия вдоль оси наибольшая. Затем делаем пересчет дисперсий для оставшихся $k - 1$ осей и снова находим «ось-чемпион» по дисперсии и т. д.

Образно говоря, мы заглядываем в куб (трехмерное пространство) по очереди по трем осям и вначале ищем то направление, где видим наибольший «туман» (наибольшая дисперсия говорит о наибольшем влиянии чего-то постороннего); затем «усредняем» картинку по оставшимся двум осям и сравниваем разброс данных по каждой из них – находим «среднячка» и «аутсайдера». Теперь остается решить систему, чтобы отыскать матрицу коэффициентов (весов) $A[k \times k]$.

Если коэффициенты A_{ji} найдены, то можно вернуться к основным переменным, поскольку доказано, что они однозначно выражаются в виде суммирования по $j = 1 \dots k$: $X_i = \sum A_{ji} Z_j$.

Отыскание матрицы весов $A[k \times k]$ требует использования ковариационной матрицы $C[k \times k]$, которая вычисляется следующим образом. Пусть мы провели по n наблюдений за каждым из k измеряемых показателей эффективности некоторой системы и данные этих наблюдений представили в виде матрицы (табл. 13.1).

Таблица 13.1

Матрица исходных данных $E[n \times k]$

E_{11}	E_{12}	...	E_{1i}	...	E_{1k}
E_{21}	E_{22}	...	E_{2i}	...	E_{2k}
...
E_{j1}	E_{j2}	...	E_{ji}	...	E_{jk}
...
E_{n1}	E_{n2}	...	E_{ni}	...	E_{nk}

Предполагаем, что на эффективность системы влияют и другие ненаблюдаемые, но легкоинтерпретируемые (объяснимые по смыслу, причине и механизму влияния) величины – факторы.

Наблюдаем, что чем больше n и чем меньше число факторов m (а может, их и нет вообще!), тем больше надежда оценить их влияние на интересующий нас показатель E .

Столь же легко понять необходимость условия $m < k$, объяснимого на простом примере аналогии. Если мы исследуем некоторые

предметы с использованием всех пяти человеческих чувств, то наивно надеяться на обнаружение более пяти «новых», легкообъяснимых, но неизмеряемых признаков у таких предметов, даже если мы «испытываем» очень большое их количество.

Вернемся к исходной матрице наблюдений $E[n \times k]$ и отметим, что перед нами, по сути дела, совокупности по n наблюдений над каждой из k случайных величин E_1, E_2, \dots, E_k . Именно эти величины «подозреваются» в связях друг с другом, или во взаимной коррелированности.

Из метода оценок таких связей следует, что мерой разброса случайной величины E_i служит ее дисперсия, определяемая суммой квадратов всех зарегистрированных значений этой величины $\sum(E_{ij})^2$ и ее средним значением (суммирование ведется по столбцу).

Если применить замену переменных в исходной матрице наблюдений, т. е. вместо E_{ij} использовать стандартизированные случайные величины $X_{ij} = \frac{E_{ij} - M(E_i)}{S(E_i)}$, то преобразуем исходную матрицу в новую $X[n \times k]$ (табл. 13.2).

Таблица 13.2

Матрица преобразованных данных $X[n \times k]$

X_{11}	X_{12}	...	X_{1i}	...	X_{1k}
X_{21}	X_{22}	...	X_{2i}	...	X_{2k}
...
X_{j1}	X_{j2}	...	X_{ji}	...	X_{jk}
...
X_{n1}	X_{n2}	...	X_{ni}	...	X_{nk}

Отметим, что все элементы новой матрицы $X[n \times k]$ окажутся безразмерными, нормированными величинами и, если некоторое значение X_{ij} составит, к примеру, +2, то это будет означать только одно – в строке j наблюдается отклонение от среднего по столбцу i на два среднеквадратичных отклонения (в большую сторону).

Выполним теперь следующие операции.

1. Просуммируем квадраты всех значений столбца 1 и разделим результат на $(n - 1)$, получим дисперсию (меру разброса) случайной величины X_1 , т. е. D_1 . Повторяя эту операцию, найдем таким же образом дисперсии всех наблюдаемых (но уже нормированных) величин.

2. Просуммируем произведения элементов соответствующих строк (от $i = 1$ до $i = n$) для столбцов 1, 2 и также разделим на $(n - 1)$, то теперь получим коэффициент ковариации C_{12} случайных величин X_1, X_2 , который служит мерой их статистической связи.

3. Если повторить предыдущую процедуру для всех пар столбцов, то в результате получим еще одну квадратную матрицу $C[k \times k]$, которую принято называть ковариационной (табл. 13.3).

Эта матрица имеет на главной диагонали дисперсии случайных величин X_i , а в качестве остальных элементов – ковариации этих величин ($i = 1 \dots k$).

Таблица 13.3

Ковариационная матрица $C[k \times k]$

D_1	C_{12}	C_{13}	C_{1k}
C_{21}	D_2	C_{23}	C_{2k}
...
C_{j1}	C_{j2}	...	D_j	...	C_{jk}
...
C_{k1}	C_{k2}	...	C_{ki}	...	D_k

Метод главных компонент характеризуется тем, что дает всегда единственное решение задачи. Правда, трактовка этого решения своеобразна.

1. Решаем задачу о наличии ровно столько факторов, сколько наблюдаемых переменных, т. е. вопрос о нашем согласии на меньшее число латентных факторов невозможно поставить.

2. В результате решения, теоретически всегда единственного, а практически связанного с громадными вычислительными трудностями при разных физических размерностях основных величин, получим ответ примерно такого вида: фактор такой-то (например, привлекательность продавцов при анализе дневной выручки магазинов) занимает третье место по степени влияния на основные переменные. Этот ответ обоснован – дисперсия этого фактора оказалась третьей по величине среди всех прочих. Больше ничего получить в этом случае нельзя. Другое дело, что этот вывод оказался нам полезным или мы его игнорируем – это наше право решать, как использовать системный подход.

Перед тем как выполнить компонентный анализ, проводится анализ независимости исходных признаков. Проверяется значимость матрицы парных корреляций с помощью критерия Уилкса (табл.13.4).

Таблица 13.4

Корреляционная матрица факторных переменных

X	x_1	x_2	x_3
x_1	1	0,985	0,931
x_2	0,985	1	0,914
x_3	0,931	0,914	1

Выдвигается гипотеза: $H_0: \hat{R}$ незначима и альтернативная $H_1: \hat{R}$ значима.

Рассчитывается статистика, которая распределена по закону χ^2 с $[n(n-1)]/6$ степенями свободы $\gamma_n = -\left(N - \frac{1}{6}(2n+5)\right) \ln|\hat{R}|$, где $N = n_1 + n_2 + n_3 = 10 + 10 + 10 = 30$; $|R|$ – детерминант (определитель) корреляционной матрицы, равный 0,0039; $n = 10$ – количество строк в выборке.

Сравнивается расчетное значение $\gamma_n = 143,4$ с табличным $\chi_{кр}^2 = 7,26$, полученным для уровня значимости $\alpha = 0,05$ и $[n(n-1)]/6 = 15$ степенями свободы. Расчетное значение критерия больше табличного значения $\gamma_n > \chi_{кр}^2$, гипотеза H_0 отвергается и принимается альтернативная $H_1: \hat{R}$ значима. Следовательно, имеет смысл проводить компонентный анализ. Затем проверяется гипотеза о диагональности ковариационной матрицы.

Выдвигается нулевая гипотеза:

$$H_0: cov(x_i x_j) = 0, \forall i \neq j \text{ и } H_1: cov(x_i x_j) \neq 0.$$

Рассчитывается статистика $\gamma_n = \left(n - \frac{2n+11}{6}\right) \ln|\hat{R}| = 137,83$, которая распределяется по закону χ^2 с $n \frac{n-1}{2}$ степенями свободы.

Сравнивается расчетное значение $\gamma_n = 137,83$ с табличным значением $\chi_{кр}^2 = 30,61$, полученным для уровня значимости $\alpha = 0,05$ и $n \frac{n-1}{2} = 45$ степенями свободы.

Расчетное значение критерия больше табличного значения $\gamma_n > \chi_{кр}^2$, гипотеза H_0 отвергается и принимается альтернативная H_1 : ковариационная матрица недиагональная, что подтверждает мультиколлинеарность данных, следовательно, имеет смысл проводить компонентный анализ.

Анализ данных (см. табл. 13.4) действительно выявил значимую коррелированность переменных $x_1 - x_3$, что также подтверждает целесообразность проведения компонентного анализа.

Пример. Имеются данные, описывающие зависимость результирующей переменной y от факторных переменных $x_1 - x_3$ (табл. 13.5).

Таблица 13.5

Зависимость y от факторных переменных

x_1	x_2	x_3	y
1,1	1,1	1,2	26,2
1,4	1,5	1,1	25,9
1,7	1,8	2	32,5
1,7	1,7	1,8	31,7
1,8	1,9	1,8	31,7
1,8	1,8	1,9	33,6
1,9	1,8	2	34,2
2	2,1	2,1	34,4
2,3	2,4	2,5	35,5
2,5	2,5	2,4	36,5

Требуется выделить главные компоненты и построить уравнение регрессии на главных компонентах. Компонентный анализ проводим с использованием ППП Statgraphics Plus. Для получения данных компонентного анализа вызываем меню Special, подменю Multivariate Methods, программу Principal Components. Результаты решения приведены в табл. 13.6 – 13.7.

Таблица 13.6

Главные компоненты

Number	Component Eigenvalue	Percent of Variance	Cumulative Percentage
1	2,888	96,26	96,26
2	0,0985	3,28	99,54
3	0,0137	0,45	100,00

Каждая из главных компонент $Z_j (j = 1, 2, 3)$ определяется комбинацией основных переменных x_i (суммирование по $i = 1, 2, 3$) $Z_j = \sum A_{ji} x_i$.

Таблица 13.7

Матрицы весов $A[3 \times 3]$

X	Component 1	Component 2	Component 3
x_1	0,583212	-0,320839	0,746275
x_2	0,580007	-0,478737	-0,659093
x_3	0,568732	0,817235	-0,0931156

$$Z_1 = -0,583x_1 + 0,58x_2 + 0,568x_3.$$

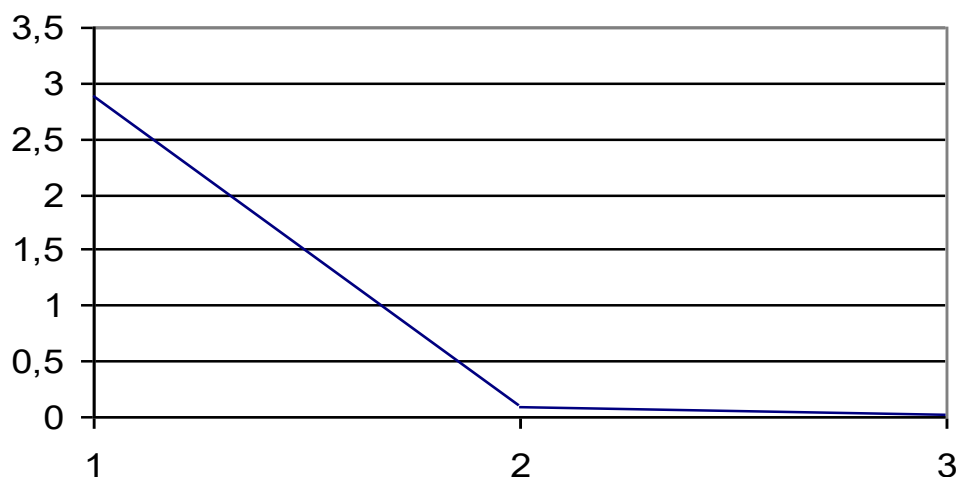
Используя найденные коэффициенты A_{ji} , можно вернуться к основным переменным, поскольку они однозначно выражаются в виде суммирования по $j = 1 \dots k$ главных компонент $x_i = \sum A_{ji} Z_j$. Так, первая переменная x_1 вычисляется по формуле

$$x_1 = 0,583Z_1 - 0,32Z_2 + 0,746Z_3.$$

Теперь, когда получена информация, сколько дисперсии объяснил каждый главный компонент (см. табл. 13.6), можно вернуться к вопросу о том, на каком числе главных компонент следует остановить процедуру. По своей природе это решение зависит от позиции исследователя. Однако имеются некоторые общеупотребительные рекомендации, и на практике следование им дает наилучшие результаты [18].

Первый принцип определения числа факторов носит название критерия Кайзера. Согласно этому критерию предлагается выделять главные компоненты до тех пор, пока дисперсия, объясненная ими, превышает единицу. По существу, это означает, что отбрасывать следует те компоненты, каждый из которых «слабее» (в смысле объясненной дисперсии), чем каждая из исходных переменных. (Напомним, после стандартизации дисперсия каждой исходной переменной равна единице.) Этот критерий предложен Кайзером (Kaiser) в 1960 году и применяется наиболее широко. Как видно из табл. 13.6, на основе этого критерия предлагается сохранить два главных компонента. Действительно, третий фактор объясняет 0,0137 единицы дисперсии, т. е. меньше единицы.

Второй подход носит название критерия каменистой осыпи. Он впервые предложен Кеттелем (Cattell) в 1966 году и является графическим методом. Суть этого критерия такова. Давайте изобразим собственные значения, представленные во втором столбце табл. 13.6, в виде графика (см. рисунок).



Критерий каменистой осыпи: собственные значения по убыванию номеров

Кеттель предложил выделять такое число главных компонент, после которого, судя по графику, убывание собственных значений слева направо максимально замедляется. Предполагается, что справа от этой точки находится только «факториальная осыпь». В соответствии с этим критерием в нашем примере следует оставить не три, а только две главные компоненты.

На практике возникает еще один важный критерий: следует оставлять такое число главных компонент, при котором они хорошо интерпретируются из содержательных соображений. Поэтому обычно исследуется несколько решений с бóльшим или меньшим числом главных компонент и затем выбирается одно наиболее «осмысленное».

На уровне информативности 95 % и выше выделяется одна главная компонента. Она имеет наибольшую дисперсию, равную 96,26 %. Использование второй главной компоненты не приводит к существенному увеличению дисперсии (всего на 3,28 %). Опция Data Table рассчитывает значения главных компонент для всех опытных данных (табл. 13.8).

Таблица 13.8

Значения главных компонент

Row	Component 1	Component 2	Component 3
1	-2,98326	0,227223	0,030408
2	-2,10734	-0,662453	-0,0371572
3	-0,107196	0,384307	-0,150787
4	-0,502301	0,137551	0,0519013
5	-0,073535	-0,176436	-0,0850164
6	-0,088807	0,122606	0,0554835
7	0,182803	0,224771	0,220295
8	0,880063	-0,024394	-0,0985813
9	2,24715	0,112702	-0,108564
10	2,55243	-0,345876	0,122018

Используя значения первой главной компоненты, строим регрессионное уравнение $y = 32,22 + 2,00Z_1$. Первая главная компонента Z_1 адекватно описывает зависимую переменную y . Коэффициент детерминации $R^2 = 89,34 \%$, статистически значим при уровне значимости 0,05. Стандартная ошибка модели равна 1,24.

Использование второй главной компоненты Z_2 позволяет увеличить точность регрессионной модели: $y = 32,22 + 2,00Z_1 + 2,7Z_2$.

Коэффициент детерминации равен $R^2 = 94,9 \%$, статистически значим при уровне значимости 0,05. Стандартная ошибка модели 0,92. Влияние третьей главной компоненты статистически незначимо.

Контрольные вопросы

1. Назовите подходы к решению задач, в которых используются статистические данные.
2. Сформулируйте назначение компонентного анализа.
3. В чем заключается идея метода компонентного анализа?
4. Как получают главные компоненты из исходных переменных?
5. Объясните получение ковариационной матрицы из таблицы с исходными данными.
6. Поясните решения, получаемые методом главных компонент.
7. Для чего проводится анализ независимости исходных признаков перед компонентным анализом?

8. Как определяется число главных компонент по критерию Кайзера?

9. В чем суть определения числа главных компонент по критерию «каменистой осыпи»?

14. МЕТОДЫ АНАЛИЗА БОЛЬШИХ СИСТЕМ. ФАКТОРНЫЙ АНАЛИЗ

Теория систем большей частью основывает свои практические методы на платформе математической статистики. Можно выделить три подхода к решению задач, в которых используются статистические данные.

1. Алгоритмический подход, при котором мы имеем статистические данные о некотором процессе и по причине слабой изученности процесса его основную характеристику (например, эффективность экономической системы) и вынуждены сами строить «разумные» правила обработки данных, базируясь на своих собственных представлениях об интересующем нас показателе.

2. Аппроксимационный подход, когда есть полное представление о связи этого показателя с имеющимися у нас данными, но неясна природа возникающих ошибок – отклонений от этих представлений.

3. Теоретико-вероятностный подход, когда требуется глубокое проникновение в суть процесса для выяснения связи показателя со статистическими данными.

В настоящее время все эти подходы достаточно строго научно обоснованы и «снабжены» апробированными методами практических действий. Но существуют ситуации, когда исследователя интересует не один, а несколько показателей процесса и, кроме того, возможно наличие нескольких влияющих на процесс воздействий – факторов, которые являются ненаблюдаемыми, скрытыми, или латентными.

Обратим внимание на понятие «латентный», скрытый, непосредственно неизмеримый фактор. Конечно, нет прибора и эталона понятий вежливости, образованности, выносливости и т. п. Но это не мешает нам самим «измерить» их, применив соответствующую шкалу для таких признаков, разработав тесты для оценки данных свойств по этой шкале. Так в чем же тогда «ненаблюдаемость»? А в том, что в

процессе массового эксперимента (обязательно) мы не можем непрерывно сравнивать все эти признаки с эталонами, и приходится брать предварительные, усредненные, полученные совсем не в «рабочих» условиях данные.

Можно отойти от экономики и обратиться к спорту. Кто будет спорить, что результат спортсмена при прыжках в высоту зависит от фактора «сила толчковой ноги». Этот фактор можно измерить в обычных физических единицах (ньютонх или килограммах), но когда? Не во время же прыжка на соревнованиях! А ведь именно в это рабочее время фиксируются статистические данные, накапливается материал для исходной матрицы.

Наиболее интересным и полезным в плане понимания сущности факторного анализа – метода решения задач в этих ситуациях – является пример использования наблюдений при эксперименте, который ведет природа. Ни о каком планировании здесь не может идти речь – нам приходится довольствоваться пассивным экспериментом.

Удивительно, но и в этих тяжелых условиях теория систем предлагает методы выявления таких факторов отсеивания, слабо проявляющих себя, оценки значимости полученных зависимостей, показателей работы системы от этих факторов.

14.1. Факторный анализ

Фактор (*factor* – лат.) – делатель, творец чего-нибудь, т. е. движущая сила, фактическая причина какого-нибудь процесса, обуславливающая его или определяющая его характер [19]. В экономической информатике факторы – это причины, воздействующие на изучаемый экономический показатель. Одни из них непосредственно связаны между собой, другие – косвенно.

Например, на величину валовой продукции непосредственное влияние оказывают такие факторы, как численность рабочих и уровень производительности труда. Влияют также субъективные или косвенные факторы – внутренние (руководство тем или иным производственным коллективом, организация производства, финансов, экономическая или организационная подготовленность исполнителей и т. д.). Без всестороннего и тщательного изучения факторов невозможно сделать обоснованные выводы о результатах деятельности, выявить резервы производства, обосновать планы и управленческие решения.

Факторный анализ – это методика комплексного и системного изучения и измерения воздействия факторов на величину резуль- тивного показателя. В результате анализа факторы получают количе- ственную и качественную оценку. Каждый показатель может, в свою очередь, выступать в роли факторного и резуль- тивного.

Факторный анализ служит для выявления и обоснования дей- ствия различных признаков и их комбинаций на исследуемый процесс путем снижения их размерности. Такая задача решается, как правило, путем «сжатия» исходной информации и выделения из нее наиболее «существенной», т. е. описание объектов меньшим числом обобщен- ных признаков, называемых факторами.

14.2. Сущность факторного анализа

Пусть для каждого конкретного объекта измерены четыре ха- рактеристики, которые обусловлены действием двух факторов F_1 и F_2 [20]. Фактор F_1 действует на все четыре характеристики объекта, а фактор F_2 – лишь на два признака X_2 и X_3 (рис. 14.1).

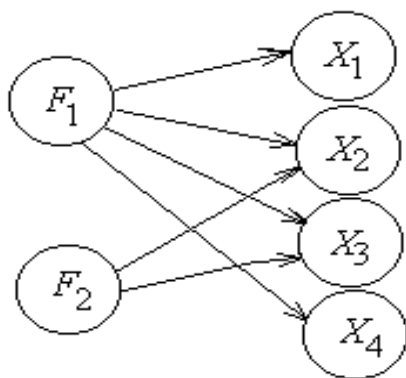


Рис. 14.1. Схема модели факторного анализа

Значит, значения признаков X_1 и X_2 определяются только фактором F_1 , а при- знаки X_2 и X_3 – совокупным действием факторов F_1 и F_2 . Но вначале неизвестно ни количество действующих факторов, ни их взаимосвязь с измеренными признака- ми. Необходимо исследовать интенсив- ность влияния факторов F_1 и F_2 на при-

знаки X_i и выделить в значениях X_i те части, которые обусловлены действием каждого из факторов F_1 и F_2 в отдельности.

Для решения этой задачи предполагают, что X_i линейно зависят от $F_m (m = 1, 2)$. Для рассматриваемого случая имеем

$$x_i = b_{i1}F_1 + b_{i2}F_2, i = 1, 2, 3, 4,$$

где b_{i1}, b_{i2} – коэффициенты, называемые факторными нагрузками.

Если рассмотреть метод на основании приведенного выше при- мера, когда имеется N рассматриваемых объектов, для каждого из ко- торых определено значение четырех признаков, то в четырехмерном графическом пространстве с осями координат X_1, X_2, X_3, X_4 это может

быть представлено как облако из N точек. Если это четырехмерное пространство рассеять плоскостью, в которой находятся координатные оси, отвечающие признакам X_1 и X_2 , то в сечении мы увидим облако точек, которое в условиях взаимосвязи признаков X_1 и X_2 друг с другом представляет собой эллипс рассеяния.

Перед проведением факторного анализа исходные значения признаков выборочной совокупности необходимо стандартизировать (центрировать и нормировать) с помощью преобразования

$$z_{jt} = (X_{jt} - \bar{X}_j) / \sigma_j,$$

где X_{jt} – исходное значение j -го признака t -го объекта; \bar{X}_j – среднее значение j -го признака; σ_j – стандартное отклонение j -го признака.

Центр эллипса рассеяния стандартизированных значений будет находиться в точке начала координат, как показано на рис. 14.2.

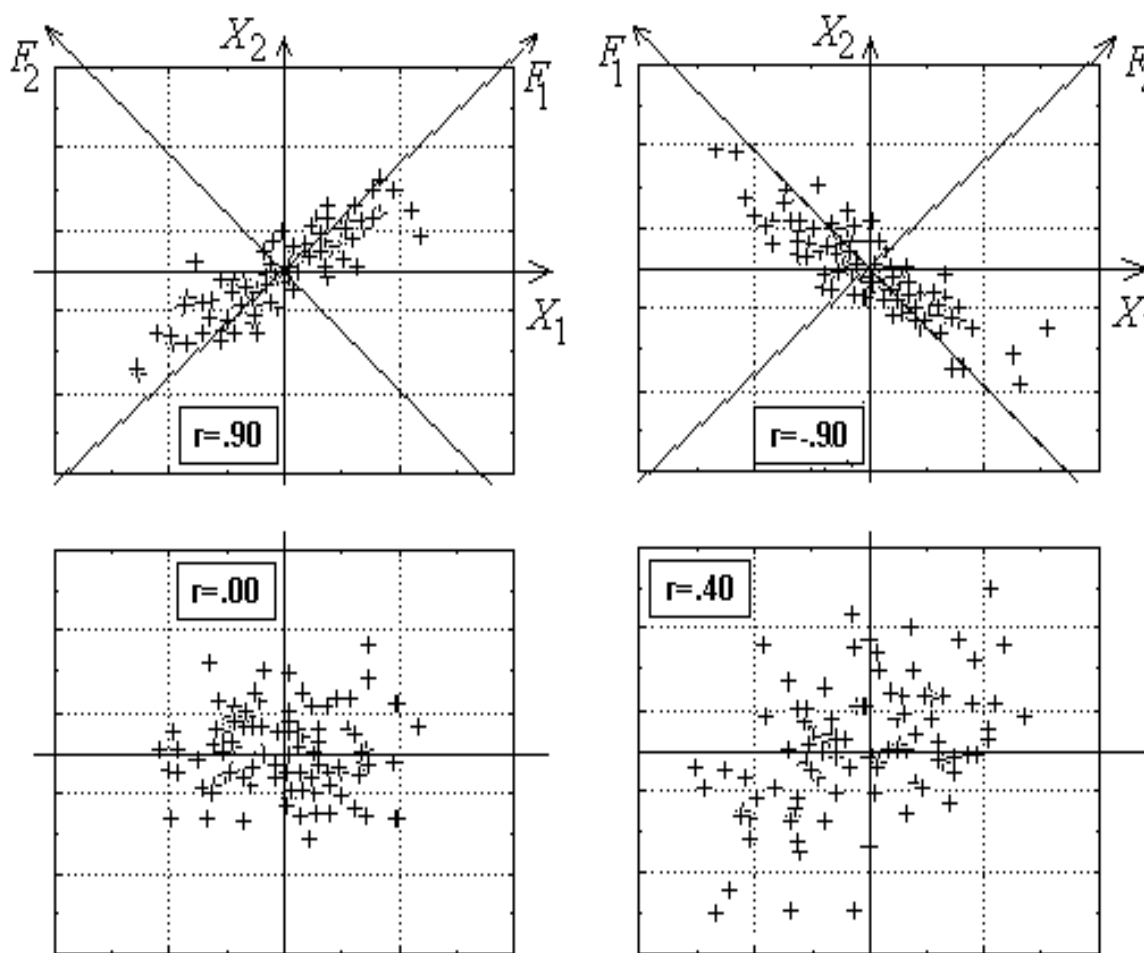


Рис. 14.2. Эллипсы рассеяния в пространстве двух стандартизированных переменных

Форма этого эллипса (сжатость – вытянутость) будет определяться величиной коэффициента корреляции X_1 с X_2 , т. е. $r_{X_1X_2}$. Чем больше модуль $r_{X_1X_2}$, тем более вытянут эллипс и при $r_{X_1X_2} = 1$ он превращается в прямую линию, а при $r_{X_1X_2} = 0$ – в круг. Если провести оси эллипса F_1 и F_2 , то по мере увеличения модуля $r_{X_1X_2}$ происходит уменьшение степени разброса точек наблюдений вдоль одной оси эллипса (на рисунке – ось F_2) и увеличение разброса вдоль другой оси эллипса (на рисунке – ось F_1).

Если перейти от исходной координатной системы X_1, X_2 к новой F_1, F_2 , оси которой ориентированы вдоль осей эллипса рассеяния, то очевидно, что в новой системе координат значения переменной $x_{ji} = (x_{1i}, x_{2i})$ вдоль оси F_2 будут иметь меньшую дисперсию, чем в исходной системе вдоль оси X_2 , а значения этой переменной вдоль оси F_1 , наоборот, будут иметь бóльшую дисперсию, чем в исходной системе вдоль оси X_1 .

Поэтому переменная F_1 несет в себе больше информации о выборке, чем F_2 . При этом чем сильнее связаны между собой признаки X_1 и X_2 , тем бóльшим становится удельный вес той из новых переменных, которая ориентируется вдоль главной оси эллипса рассеяния.

Следовательно, в случае многомерного пространства появляется возможность ранжирования переменных (признаков) по их дисперсии в соответствии с их вкладом (значимостью) в общую характеристику изучаемого объекта, т. е. по уменьшению дисперсии значений признаков вдоль новых координатных осей F_j .

Трудно представить, как выглядит в многомерном пространстве облако точек выборочной многомерной совокупности. По аналогии с рассмотренным выше двумерным случаем можно предполагать, что оно представляет собой эллипсоид с несколькими разновеликими ортогональными осями. Поэтому в условиях взаимозависимости признаков для более компактного представления информации переходят к новой ортогональной системе координат (ориентированной по главным осям этого эллипсоида). Переменные этой новой системы – главные компоненты F_j ($j = 1, 2, \dots, m$ и $m \ll k$) – концентрируют в себе основную информацию об исходной выборке и снижают размерность исходного признакового пространства ($m \ll k$). Это процедура перехода к новой системе координат F_j .

Указанный переход не затрагивает геометрической структуры взаимного расположения точек наблюдений x_{ji} . Характер их распределения сохраняется. Поэтому суммарная дисперсия остается прежней, т. е. $\sigma_{X_1}^2 + \sigma_{X_2}^2 = \sigma_{F_1}^2 + \sigma_{F_2}^2$ или в общем виде $\sum_j \sigma_{X_j}^2 = \sum_j \sigma_{F_j}^2$. Фак-

торные нагрузки b_{ij} в уравнении $x_i = \sum_{j=1}^m b_{ij} F_j$ представляют собой коэффициенты корреляции между исходными X_j и новыми F_j переменными $b_{ij} = r(x_i, F_j)$.

Факторный анализ может быть [21]:

- разведочным – он осуществляется при исследовании скрытой факторной структуры без предположения о числе факторов и их нагрузках;

- подтверждающим, предназначенным для проверки гипотез о числе факторов и их нагрузках.

Условия применения факторного анализа. Практическое выполнение факторного анализа начинается с проверки его условий. Обязательные условия факторного анализа предполагают:

- все признаки должны быть количественными;
- число наблюдений должно быть не менее чем в два раза больше числа переменных;
- выборка должна быть однородная;
- исходные переменные должны быть распределены симметрично;
- факторный анализ осуществляется по коррелирующим переменным.

При использовании методов факторного анализа решаются следующие задачи:

- отыскание скрытых, но объективно существующих закономерностей исследуемого процесса, определяемых воздействием внутренних и внешних причин;
- описание изучаемого процесса значительно меньшим числом факторов по сравнению с первоначально взятым количеством признаков;
- выявление первоначальных признаков, наиболее тесно связанных с основными факторами;
- прогнозирование процесса на основе уравнения регрессии, построенного по полученным факторам.

14.3. Последовательность факторного анализа

Особенность факторного анализа заключается в неопределенности решения его основных проблем. Нет четких критериев качества их решения, есть лишь рекомендации, которыми руководствуется исследователь в своем стремлении содержательно интерпретировать получаемые результаты. Поэтому факторный анализ – это пошаговая процедура, где на каждом шаге исследователь принимает решение о дальнейших преобразованиях данных. Главным ориентиром на этом пути остается возможность получения содержательной интерпретации конечных результатов.

Таким образом, нет ничего удивительного в том утверждении, что факторный анализ, а значит, и системный анализ в современных условиях – больше искусство, чем наука. Здесь менее важно владеть «навыками» и крайне важно понимать как силу, так и ограниченные возможности этого метода.

Есть и еще одно обстоятельство, затрудняющее профессиональную подготовку в области факторного анализа, – необходимость быть профессионалом в «технологическом» плане, в нашем случае в предметной области.

Но стать профессионалом высокого уровня вряд ли можно, не имея хотя бы представлений о возможностях анализировать и эффективно управлять системами на базе решений, найденных с помощью факторного анализа.

Не следует обольщаться обещаниями популяризаторов метода факторного анализа, не следует верить мифам о его могуществе и универсальности. Этот метод «на вершине» только по одному показателю – своей сложности, как по сущности, так и по сложности практической реализации даже при всеобщем использовании компьютерных программ.

Весь процесс факторного анализа можно представить как выполнение шести этапов [22]:

1. Выбор исходных данных.
2. Предварительное решение проблемы числа факторов.
3. Факторизация матрицы интеркорреляций.
4. Вращение факторов и их предварительная интерпретация.
5. Принятие решения о качестве факторной структуры.
6. Вычисление факторных коэффициентов и оценок.

Исследователь в зависимости от своих целей решает, сколько раз повторить эту последовательность, какие из этапов будут пропущены и насколько глубоко будет проработан каждый из них.

Этап 1-й. Выбор исходных данных.

Модель факторного анализа разрабатывалась для метрических данных. Поэтому первое требование к исходным данным – представление всех признаков в метрической шкале (не обязательно с одинаковыми средними и дисперсиями).

Этап 2-й. Решение проблемы числа факторов.

На этом этапе матрица интеркорреляций исходных признаков обрабатывается с использованием анализа главных компонент. Применяются критерий отсеивания Р. Кеттелла и критерий Кайзера.

Эти критерии не являются жесткими, поэтому далее проверяется несколько гипотез о числе факторов. Начинать при этом рекомендуется с максимально возможного числа факторов с учетом обоих критериев, постепенно уменьшая их число.

Этап 3-й. Факторизация матрицы интеркорреляций.

Выбирается метод факторизации, желательно главных осей, наименьших квадратов или максимального правдоподобия. Задается число факторов в соответствии с проверяемой гипотезой. Результатом данного этапа будет матрица факторных нагрузок (факторная структура) до вращения, которая не подлежит интерпретации.

Полезной информацией на этом этапе могут стать суммарная доля дисперсии (информативность) факторов и значения общностей переменных. Суммарная доля дисперсии – показатель того, насколько полно выделяемые факторы могут представить данный набор признаков, а этот набор – выделяемые факторы. Общность переменной – показатель ее «участия» в факторном анализе, насколько она влияет на факторную структуру. Переменные с наименьшими общностями – ближайшие кандидаты на исключение из анализа в дальнейшем.

Этап 4-й. Вращение факторов и их предварительная интерпретация.

Выбирается один из аналитических методов вращения факторов, обычно варимакс-вращение (*Varimax normalized*). Существуют и другие методы вращения. В результате вращения достигается факторная структура, наиболее доступная для интерпретации при данном соотношении переменных и факторов. Интерпретация факторов произво-

дится по таблице факторных нагрузок после вращения. По каждой переменной (строке) выделяется наибольшая по абсолютной величине нагрузка как доминирующая.

Этап 5-й. Принятие решения о качестве факторной структуры.

Формальное требование к факторной структуре сформулировал Л. Терстоун еще в 1930-х годах, назвав его принципом простой структуры. Геометрически принцип простой структуры означает, что все переменные располагаются на осях факторов, т. е. каждая переменная имеет близкие к нулю нагрузки по всем факторам, кроме одного. Поэтому основным критерием остается возможность хорошей содержательной интерпретации каждого фактора по двум и более исходным переменным. Приближение к простой структуре связано с невосполнимой потерей исходной эмпирической информации. И каждый раз исследователь должен решать, насколько целесообразна эта потеря в свете стоящих перед ним задач.

Этап 6-й. Вычисление факторных коэффициентов и оценок.

Оценки факторных коэффициентов являются коэффициентами линейного уравнения, связывающего значение фактора и значения исходных переменных. Они показывают, с каким весом входят исходные значения каждой переменной в оценку фактора. Факторные коэффициенты можно использовать для вычисления факторных оценок для новых объектов, не включенных ранее в факторный анализ.

Пример. Имеются данные, описывающие зависимость результативного показателя y от контролируемых переменных x_1 – x_3 (табл. 14.1).

Таблица 14.1

Зависимость y от контролируемых переменных

x_1	x_2	x_3	y
1,1	1,1	1,2	26,2
1,4	1,5	1,1	25,9
1,7	1,8	2	32,5
1,7	1,7	1,8	31,7
1,8	1,9	1,8	31,7
1,8	1,8	1,9	33,6
1,9	1,8	2	34,2
2	2,1	2,1	34,4
2,3	2,4	2,5	35,5
2,5	2,5	2,4	36,5

Требуется изучить воздействие факторов на величину результирующего показателя. Факторный анализ проводим с использованием ППП Statgraphics Plus. Для получения данных факторного анализа вызываем меню Special, подменю Multivariate Methods, программу Factor Analysis.

Модель факторного анализа разрабатываем для метрических данных (см. табл. 14.1). На втором этапе матрицу интеркорреляций исходных признаков обрабатываем с использованием анализа главных компонент.

Analysis Summary

Data variables:

x_1
x_2
x_3

Factor Analysis			
Factor Number	Percent of Eigenvalue	Cumulative Variance	Percentage
1	2,88782	96,261	96,261
2	0,0985219	3,284	99,545
3	0,0136627	0,455	100,000

Initial Variable Communalities

x_1 1,0
x_2 1,0
x_3 1,0

Применяя критерий отсеивания Р. Кеттелла и Кайзера, проверяем гипотезу о числе факторов (рис. 14.3).

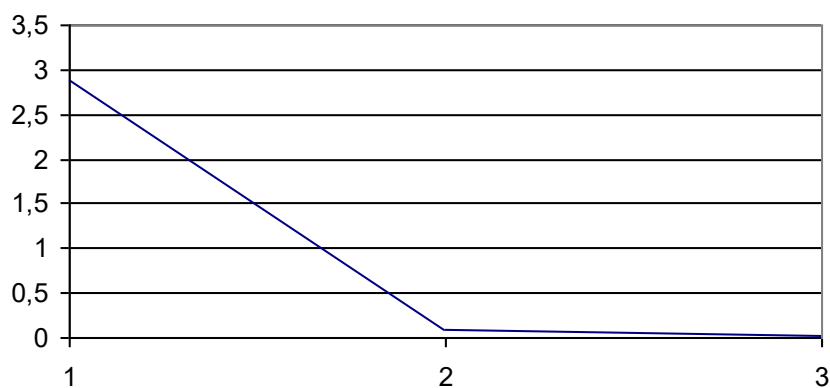


Рис. 14.3. Критерий «каменистой осыпи»: собственные значения расположены по убыванию номеров

Исследование начнем с максимально возможного числа факторов, равного двум, затем при необходимости будем постепенно уменьшать их число.

На третьем этапе выбираем метод факторизации из списка главных осей, наименьших квадратов или максимального правдоподобия. Задаем число факторов в соответствии с проверяемой гипотезой.

Выбираем метод главных компонент (установлен по умолчанию), задаем число факторов, равное двум. Для задания параметров вызываем контекстное меню *Analysis Options*. Устанавливаем число факторов *Number of Factors*, равное двум. В меню *Tabular Options* устанавливаем режим *Rotation statistics*. Результатом данного этапа будет матрица факторных нагрузок, строки соответствуют переменным x_1 , x_2 и x_3 , а столбцы – факторам $F1$ и $F2$. Значения в ячейках таблицы b_{ik} – это факторная нагрузка переменной x_i по фактору k .

Factor Loading Matrix After Varimax Rotation

	Factor 1	Factor 2	bik
x_1	0,810422	0,579316	0,810422
x_2	0,839084	0,538519	0,830084
x_3	0,555904	0,831175	0,831175
x_2		x_3	

Интерпретация факторов производится по матрице факторных нагрузок после вращения в следующем порядке. По каждой переменной (строке) выделяется наибольшая по абсолютной величине нагрузка как доминирующая. После просмотра всех строк – переменных начинают просмотр столбцов – факторов. По каждому фактору выписывают наименования (обозначения) переменных, имеющих наибольшие нагрузки по этому фактору, выделенных на предыдущем шаге. При этом обязательно учитывают знак факторной нагрузки переменной. Если знак отрицательный, это отмечают как противоположный полюс переменной. После такого просмотра всех факторов каждому из них присваивают наименование, обобщающее по смыслу включенные в него переменные. В нашем примере это x_2 и x_3 .

На четвертом этапе выбирают один из аналитических методов вращения факторов *Varimax* (задано по умолчанию). Выводим ре-

зультаты решений, нажав иконку меню Save results и указав вид выводимых данных Factor matrix.

Factor Loading Matrix After Varimax Rotation

	Factor 1	Factor 2
x_1	0,810422	0,579316
x_2	0,839084	0,538519
x_3	0,555904	0,831175

Расчеты собственного значения каждого фактора и общности каждой переменной выполняют в табличной форме (табл. 14.2). Исходные данные для расчетов даны в таблице факторной структуры после вращения.

Собственное значение (Eigenvalue) каждого фактора λ_k равно сумме квадратов факторных нагрузок всех переменных по фактору k (по столбцу). Общность каждой переменной h_i^2 равна сумме квадратов факторных нагрузок переменной i по всем факторам.

Таблица 14.2

Собственные значения факторов и общность переменных

Переменная	Factor 1	Factor 2	h^2
x_1	0,810422	0,579316	0,992391
x_2	0,839084	0,538519	0,994065
x_3	0,555904	0,831175	0,999881
Eigenvalue	1,669875	1,316462	2,972766
Доля дисперсии	0,556625	0,438821	0,990922

Суммарная информативность двух факторов, равная сумме собственных значений, деленной на количество переменных, составляет 0,990922. Иными словами, выделенные факторы объясняют 99,09 % суммарной дисперсии признаков, что считается хорошим результатом.

Все признаки однозначно соотносятся по факторным нагрузкам с двумя факторами $F1$ и $F2$. Можно заключить, что полученная факторная структура является достаточно простой и можно приступить к интерпретации факторов. Перед интерпретацией по каждой переменной

ной (по строке) отмечается наибольшая по абсолютной величине факторная нагрузка.

Фактор 1 имеет информативность 55,66 %. Его положительный полюс определяется положительными полюсами переменных x_1 , x_2 и x_3 .

Фактор 2 имеет информативность 43,88 %, положительный полюс определяется положительными полюсами переменных x_1 , x_2 и x_3 .

На пятом этапе принимается решение о качестве факторной структуры. Качество факторной структуры определяется степенью приближения к простой структуре. В рассматриваемом примере структура простая, содержит два фактора.

На шестом этапе вычисляют факторные коэффициенты и их оценки. Оценки факторных коэффициентов являются коэффициентами линейного уравнения, связывающего значение фактора и значения исходных переменных. Они показывают, с каким весом входят исходные значения каждой переменной в оценку фактора. Выводим результаты решений, нажав иконку меню Save results и указав вид выводимых данных Rotadet factor matrix.

Factor Loading Matrix After Varimax Rotation

	Factor 1	Factor 2
x_1	0,810422	0,579316
x_2	0,839084	0,538519
x_3	0,555904	0,831175

Variable	Estimated Communality
Col_1	0,992391
Col_2	0,994065
Col_3	0,999882

$$F1 = 0,81x_1 + 0,839x_2 + 0,555x_3;$$
$$F2 = 0,579x_1 + 0,538x_2 + 0,831x_3.$$

Вычисленные оценки факторов как новые переменные являются независимыми, отражающими реальную структуру взаимосвязей исходных признаков и наиболее полно передающими исходную эмпирическую информацию. Программа выдает значения факторных переменных для исходной выборки.

Table of Factor Scores
Factor

Row	1	2
1	-3,85223	-3,29645
2	-2,55048	-2,52245
3	-0,21644	-0,02983
4	-0,66921	-0,53164
5	-0,0572	-0,12414
6	-0,1387	-0,07084
7	0,186542	0,258227
8	1,12756	0,982496
9	2,84283	2,54993
10	3,32731	2,7847

Используя эти данные, добавив столбец с данными y , строим уравнение регрессии, описывающее зависимость результирующей переменной y от факторных переменных $F1$ и $F2$.

Регрессионная статистика	
Множественный R	0,974175
R -квадрат	0,949018
Нормированный R -квадрат	0,934451
Стандартная ошибка	0,920961
Наблюдения	10

Дисперсионный анализ				
	df	SS	MS	F
Регрессия	2	110,5188	55,25941	65,15135
Остаток	7	5,937188	0,84817	
Итого	9	116,456		

	Коэффициенты	Стандартная ошибка	t -статистика	P -значение
Y -пересечение	32,22	0,291234	110,6328	1,3E-12
Переменная $F 1$	-4,80755	2,060565	-2,33312	0,052373
Переменная $F 2$	7,242423	2,339809	3,095306	0,017437

$$y = 32,22 - 4,8F1 + 7,24F2.$$

Факторы F_1 и F_2 адекватно описывают результирующую переменную y . Коэффициент детерминации уравнения $R^2 = 94,9 \%$. Факторные коэффициенты можно использовать для вычисления факторных оценок для новых объектов, не включенных ранее в факторный анализ.

Контрольные вопросы

1. Назовите подходы к решению задач, использующие статистические данные.
2. В чем назначение факторного анализа?
3. В чем сущность факторного анализа?
4. Какие задачи решаются методом факторного анализа?
5. Поясните значение факторной нагрузки b_{ij} в уравнении
$$x_i = \sum_{j=1}^m b_{ij} F_j.$$
6. Какие условия применения факторного анализа вы можете назвать?
7. Объясните этапы процесса факторного анализа.

15. МОДЕЛИ ВРЕМЕННЫХ РЯДОВ И СТАТИСТИЧЕСКИЕ ОЦЕНКИ ВЗАИМОСВЯЗИ ВРЕМЕННЫХ РЯДОВ

15.1. Модели временных рядов

Модели, построенные по данным, характеризующим функционирование системы или процесс за ряд последовательных равноотстоящих моментов времени, называются моделями временных рядов, в дальнейшем – временными рядами. Простейшей является модель аддитивного случайного процесса, имеющая вид [23, 24]

$$Y_t = U_t + V_t + e_t, \quad (15.1)$$

где U_t – трендовая компонента; V_t – сезонная компонента; e_t – случайная компонента; t – время наблюдения, $t = 1, 2, 3, \dots$.

Для построения модели (15.1) необходимо получить оценки каждой компоненты. Для выделения составляющих компонент используются процедуры фильтрации, регрессионного и корреляционного анализов.

Относительно трендовой составляющей U_t предполагают, что она представляет некоторую гладкую функцию, описываемую полиномом небольшой степени. Для этого чаще всего используются следующие функции времени:

- линейная $U_t = a + bt$;
- парабола второго и реже более высокого порядков
 $U_t = a + b_1t + b_2t^2 + b_3t^3 + \dots + b_nt^n$;
- экспонента $U_t = e^{a+bt}$ и др.

Параметры тренда определяются методом наименьших квадратов, в качестве независимой переменной выступает время $t = 1, 2, 3, \dots$, а в качестве зависимой переменной – уровни временного ряда Y_t . Критерием отбора наилучшей формы тренда является значение коэффициента детерминации R^2 .

Пример. Имеются данные о выработке продукции (автомобильных стекол) за 18 месяцев работы производственного участка (табл. 15.1) [25].

Таблица 15.1

Данные о выработке продукции по месяцам

Месяц	Выработка продукции	Месяц	Выработка продукции
1	596488	10	ẽ
2	615925	11	568649
3	612846	12	420148
4	634217	13	452529
5	659835	14	447319
6	615392	15	456579
7	708291	16	505584
8	580846	17	484261
9	509008	18	453356

Требуется:

1. Построить график динамики выработки продукции.
2. Отобрать наилучшую форму тренда.
3. Выделить сезонную компоненту.
4. Построить аддитивную модель.

Решение. Решение проводим, применяя программу ППП MS EXCEL. С использованием Мастера диаграмм строим график динамики выработки продукции (рис. 15.1).

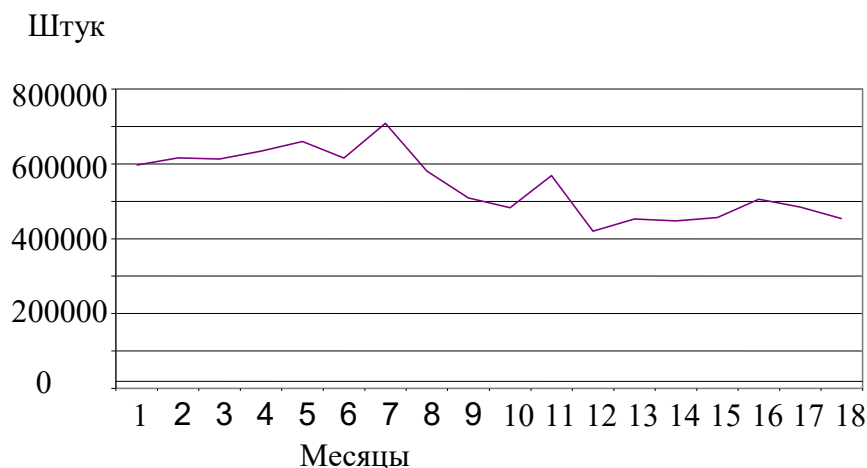


Рис. 15.1. График выработки продукции по месяцам

График характеризует убывающую тенденцию выработки продукции с периодическими колебаниями. Проведем подбор тренда путем наложения линий тренда. Одновременно установим режим отображения уравнения регрессии, описывающего тренд, и коэффициента детерминации. В табл. 15.2 приведены характеристики подбираемых линий тренда.

Таблица 15.2

Подбор линий тренда МНК

Вид тренда	Коэффициент детерминации, %	Уравнение тренда
Линейный	61	$U_t = 665390 - 12707 t$
Парабола	61,5	$U_t = -50,31t^2 - 11751t + 662203$
Экспонента	62,4	$U_t = 672830e^{-0,0235 t}$

Все три вида тренда адекватно описывают характер изменения выработки продукции во времени. Коэффициенты детерминации статистически значимы при уровне значимости 0,05, расчетные значения критерия Фишера превышают табличные данные. Для математического описания тренда выбираем более простое линейное уравнение.

Для выделения сезонной компоненты совместно со случайной составляющей $(V_t + e_t)$ из исходного ряда Y_t вычитаем трендовую компоненту U_t . При этом получаем центрированный временной ряд

$$(V_t + e_t) = Y_t - U_t.$$

График центрированного временного ряда отображен на рис. 15.2.

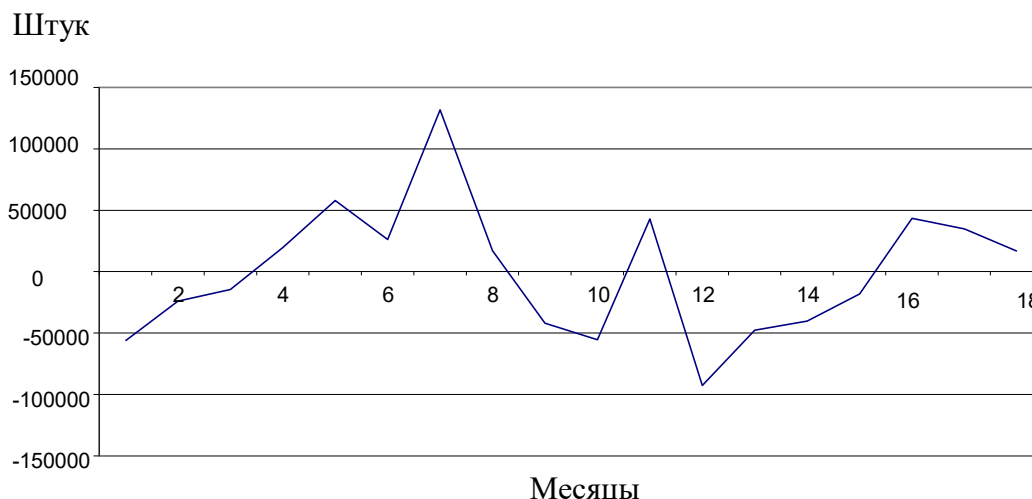


Рис. 15.2. График компонент $(V_t + e_t)$ в динамическом ряду выработки продукции

Для определения периода циклической компоненты V_t вычисляем автокорреляционную функцию центрированного временного ряда (рис. 15.3). На графике просматривается периодическая составляющая с периодом $(13 - 1) = 12$ месяцев и временным сдвигом $(12 - 3) = 9$ месяцев. Амплитуда гармоники может быть приближенно оценена с помощью дисперсии центрированного временного ряда. Из условия аддитивности модели вытекает баланс дисперсий центрированного ряда

$$S^2(V_t + e_t) = S^2(V_t) + S^2(e_t),$$

где $S^2(V_t + e_t)$ – оценка дисперсии центрированного временного ряда; $S^2(V_t)$ – оценка дисперсии сезонной (гармонической) компоненты, равная квадрату амплитуды гармоники; $S^2(e_t)$ – оценка дисперсии случайной компоненты.

Если пренебречь дисперсией случайной компоненты, то за амплитуду гармонической составляющей можно принять (оценка сверху) стандартное отклонение центрированного ряда. В рассматриваемом примере это будет

$$A_{V_t} = S(V_t) = 53660.$$

Коэффициент
корреляции



Рис. 15.3. Автокорреляционная функция централизованного временного ряда

Амплитуда гармоники может быть уточнена по критерию минимума случайной компоненты временного ряда. На графике рис. 15.4 приведены совмещенные компоненты $(V_t + e_t)$ и гармоническая компонента V_t с уточненной амплитудой, равной 50000:

$$V_t = 50000 \sin((2\pi/12)t + 2\pi(2,85/4)).$$

Для выделения случайной компоненты e_t из централизованного временного ряда $(V_t + e_t)$ вычитаем гармоническую компоненту V_t .

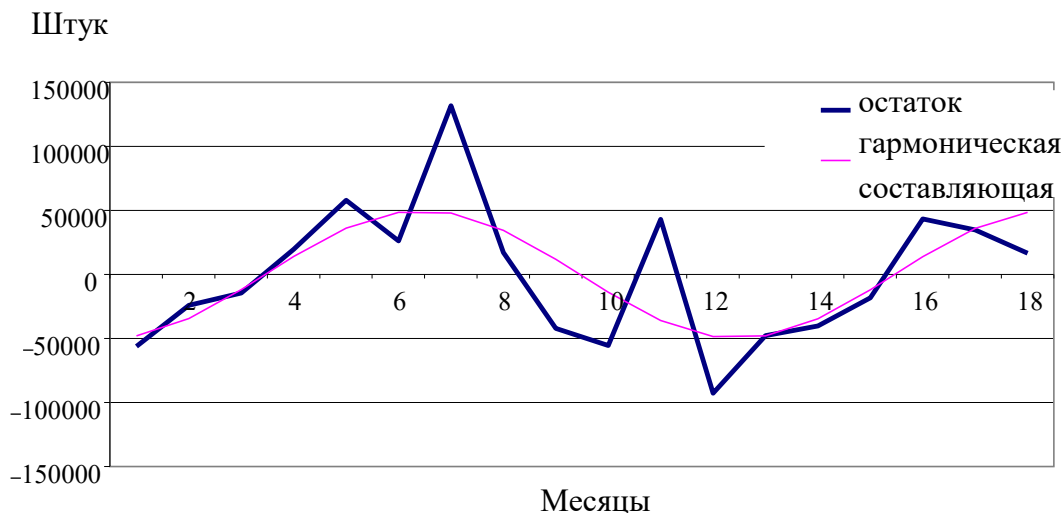


Рис. 15.4. График централизованного ряда $(V_t + e_t)$ с наложением гармонической компоненты $V_t = 50000 \sin((2\pi/12)t + 2\pi(2,85/4))$

График случайной компоненты приведен на рис. 15.5.

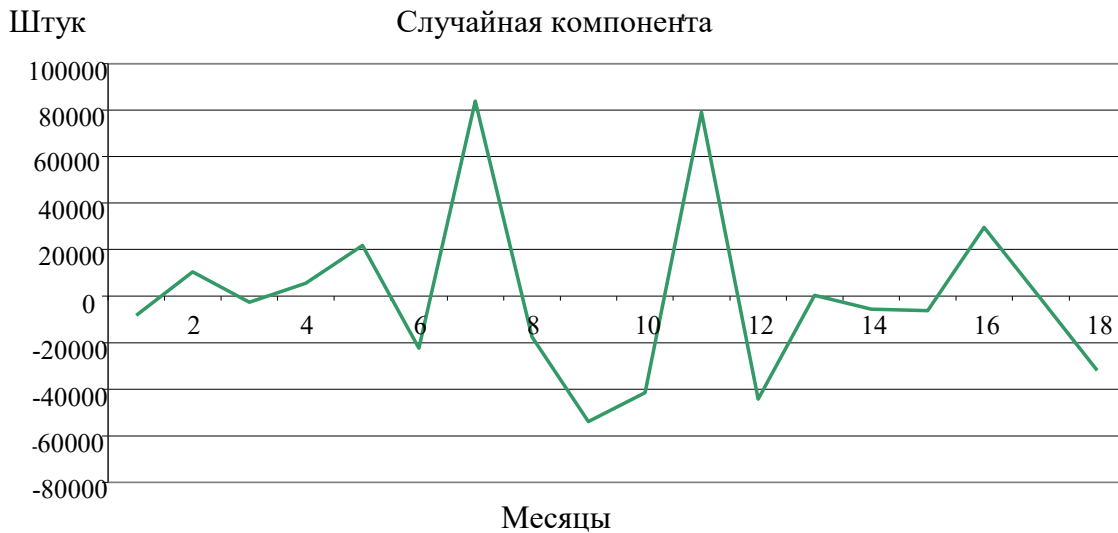


Рис. 15.5. График случайной компоненты временного ряда выработки продукции

Случайная компонента e_t имеет следующие параметры:

- среднее значение равно $-226,3$ шт./мес., что статистически незначимо при уровне значимости $0,05$;
- оценка дисперсии равна $13,71 \times 10^8$ шт./мес².

После подстановки в исходное уравнение (15.1) всех компонент, временной ряд выработки продукции, уровни которых представлены в табл. 15.1, описывается следующей аддитивной моделью:

$$Y_t = -12707t + 665390 + 50000 \sin((2\pi/12)t + 2\pi(2,85/4)) + e_t. \quad (15.2)$$

Адекватность модели (15.2) оцениваем по результатам анализа случайной компоненты e_t . Проверяем выполнение предпосылок методом наименьших квадратов.

Случайность остатков модели определяем по числу точек перегиба

$$p = 11 > p_k = 9,$$

где $p_k = [2(n-2)/3 - 2\sqrt{\frac{16n-29}{90}}]$.

Для определения независимости значений уровней случайной компоненты можно воспользоваться первым коэффициентом автокорреляции

$$r(1) = \left(\sum_{i=2}^n \varepsilon_i \varepsilon_{i-1} \right) / \sum_{i=1}^n \varepsilon_i^2.$$

Для принятия решения о наличии или отсутствии автокорреляции в исследуемом ряду фактическое значение коэффициента автокорреляции $r(1)$ сопоставляется с табличным (критическим) значением для 5%-го уровня значимости (вероятность допустить ошибку при принятии нулевой гипотезы о независимости уровней ряда). Если фактическое значение коэффициента автокорреляции меньше табличного, то гипотеза об отсутствии автокорреляции в ряду принимается. Если фактическое значение больше табличного – делают вывод о наличии автокорреляции в ряду динамики.

Для обнаружения гетероскедастичности обычно используют три теста, в которых делаются различные предположения о зависимости между дисперсией случайного члена и объясняющей переменной: тест ранговой корреляции Спирмена, тест Голдфельда – Квандта и тест Глейзера [Доугерти].

При малом объеме выборки для оценки гетероскедастичности может использоваться тест Голдфельда – Квандта. Он применяется для проверки такого типа гетероскедастичности, когда дисперсия остатков возрастает пропорционально квадрату фактора. При этом делается предположение, что случайная составляющая e_t распределена нормально. Алгоритм применения теста Голдфельда – Квандта для оценки гетероскедастичности описан в п. 11.3 данного пособия.

В рассматриваемом примере все предпосылки МНК выполняются, что подтверждает адекватность разработанной модели (15.2). Оценим точность разработанной модели. Для этого вычисляем среднюю абсолютную и среднюю относительную ошибки. Расчеты показали следующие результаты:

– средняя абсолютная ошибка разработанной модели равна 25877,8 шт./мес.;

– средняя относительная ошибка равна 4,7 %.

Приведем интерпретацию результатов исследований с учетом особенностей анализируемого производственного процесса. В рассматриваемом временном интервале работа участка характеризуется нестабильностью. Среднее абсолютное уменьшение выработки изделий в течение ме-

сяца составляет $\Delta u_{\text{ср}} = 12707$ шт. Темп уменьшения выработки изделий в последнем месяце составил величину $(12707/449371)100\% = 2,83\%$.

Сезонная компонента V_t отражает увеличение выработки изделий в зимние месяцы года (декабрь – январь) и уменьшение в летние месяцы (июнь – июль) на величину, примерно равную 50000 шт./мес. Одной из причин могут быть колебания спроса.

15.2. Статистические оценки взаимосвязи двух временных рядов

Изучение причинно-следственных зависимостей переменных, представленных в виде временных рядов, является сложной задачей моделирования. Каждый уровень временного ряда в общем случае может описываться следующей моделью:

$$Y_t = U_t + V_t + e_t,$$

где U_t – трендовая компонента; V_t – сезонная компонента; e_t – случайная компонента; t – время наблюдения, $t = 1, 2, 3, \dots$.

Наличие этих компонент может привести к серьезным проблемам при проведении корреляционно-регрессионного анализа данных временных рядов. Поэтому на предварительном этапе анализа необходимо выявить структуру изучаемых временных рядов. Для этого строим совмещенные графики анализируемых рядов и проводим визуальный анализ. И если в одном из временных рядов (результатная переменная) тенденция изменения может быть следствием того, что другая переменная (факторная) тоже содержит такую же тенденцию или противоположную направленность, то это может быть причиной наличия коинтеграции временных рядов.

Под коинтеграцией понимается причинно-следственная зависимость в уровнях двух или более временных рядов, которая выражается в совпадении или противоположной направленности их тенденций и случайной колеблемости.

Пример 1. Оценить тесноту связи временных рядов среднедушевого располагаемого дохода $x(t)$ и среднедушевого расхода на конечное потребление $y(t)$ в условный период. Исходные данные для расчетов даны в табл. 15.3 [23].

Таблица 15.3

Исходные данные для расчетов

Год, t	$y(t)$	$x(t)$	$y_p(t)$	$e(t)$	$\Delta e(t)$	$y(t)$	$x(t)$
1	6698	7264	6524,16	173,836	–	–	–
2	6740	7382	6632,98	107,023	–66,813	1780,131	2003,008
3	6931	7583	6818,33	112,672	5,649	1940,03	2116,629
4	7089	7718	6942,82	146,182	33,51	1956,595	2102,789
5	7384	8140	7331,96	52,0365	–94,1455	2134,596	2424,821
6	7703	8508	7671,31	31,6868	–20,3497	2235,148	2480,33
7	8005	8822	7960,87	44,1329	12,4461	2300,929	2521,826
8	8163	9114	8230,13	–67,1337	–111,267	2235,298	2581,309
9	8506	9399	8492,95	13,0547	80,1884	2461,299	2650,083
10	8737	9606	8683,83	53,1705	40,1158	2438,307	2646,041
11	8842	9875	8931,89	–89,8868	–143,057	2372,252	2761,757
12	9022	10111	9149,51	–127,513	–37,6262	2474,499	2798,563
13	9425	10414	9428,92	–3,9235	123,5895	2744,209	2926,805
14	9752	11013	9981,29	–229,289	–225,366	2772,788	3301,433
15	9602	10832	9814,38	–212,381	16,908	2380,644	2676,874
16	9711	10906	9882,62	–171,619	40,762	2600,719	2884,904
17	10121	11192	10146,4	–25,3531	146,2659	2930,005	3116,107
18	10425	11406	10343,7	81,3077	106,6608	2930,4	3118,324
19	10744	11851	10754	–10,0473	–91,355	3024,288	3404,857
20	10867	12039	10927,4	–60,4107	–50,3634	2911,068	3263,335
21	10746	12005	10896,1	–150,058	–89,6473	2698,987	3090,121
22	10770	12156	11035,3	–265,302	–115,244	2812,587	3266,298
23	10782	12146	11026,1	–244,08	21,222	2806,815	3144,482
24	11179	12349	11213,3	–34,276	209,804	3194,929	3354,887
25	11617	13029	11840,3	–223,335	–189,059	3338,951	3884,566
26	12015	13258	12051,5	–36,5067	186,8283	3412,612	3610,026
27	12336	13552	12322,6	13,3823	49,889	3438,893	3734,451
28	12568	13545	12316,2	251,837	238,4547	3433,192	3509,744
29	12903	13890	12634,3	268,697	16,86	3596,396	3859,928
30	13027	14030	12763,4	263,597	–5,1	3472,329	3744,455
31	13051	14154	12877,7	173,25	–90,347	3404,507	3764,785
32	12889	13987	12723,8	165,249	–8,001	3224,735	3505,963

На рис. 15.6 приведены графики изменения во времени среднедушевого располагаемого дохода $x(t)$ и среднедушевого расхода на конечное потребление $y(t)$ в условный период.

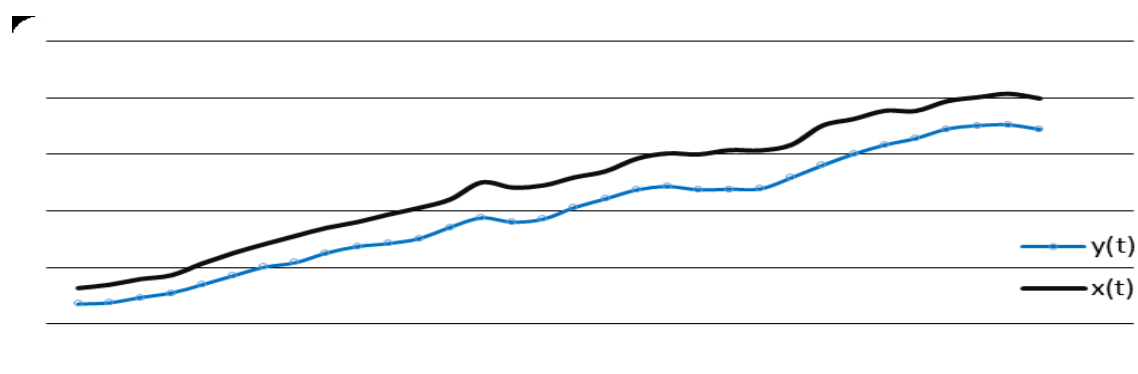


Рис. 15.6. Взаимосвязь временных рядов среднедушевого располагаемого дохода $x(t)$ и среднедушевого расхода на конечное потребление $y(t)$

Визуальный анализ показывает, что тенденции изменения этих временных рядов совпадают. Для проверки гипотезы отсутствия коинтеграции между рядами воспользуемся критерием Энгеля – Грангера [23]. Для этого рассчитаем уравнение регрессии вида

$$\Delta e(t) = f(e(t-1)), \quad (15.3)$$

где $e(t-1)$, $t = 2, 3, \dots, 32$ – остаток регрессионной модели (15.3); $\Delta e(t)$, $t = 2, 3, \dots, 32$ – первые разности остатков.

Результаты расчета остатков приведены в табл. 15.3. Параметры уравнения регрессии (15.3), рассчитанные с помощью программы STATGRAPHICS Plus, приведены ниже.

Multiple Regression Analysis

Dependent variable: $\Delta e(t)$

Parameter	Estimate	Error	Standard t Statistic	P-Value
CONSTANT	-1,73	18,94	-0,091	0,93
$e(t-1)$	-0,27	0,127	-2,15	0,04

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	51265,8	1	51265,8	4,62	0,04

Residual	322068,0	29	11105,8
Total (Corr.)	373334,0	30	

R-squared = 13,7 percent

R-squared (adjusted for d.f.) = 10,75 percent

Standard Error of Est. = 105,38

Mean absolute error = 82,83

Durbin-Watson statistic = 2,03

Уравнение регрессии имеет вид

$$\Delta e(t) = -1,73 - 0,27e(t-1).$$

Расчетное значение t -критерия значимости коэффициента регрессии при остатке $e(t-1)$ по модулю равно 2,15, превышает критическое значение $t_{кр} = 1,94$. С вероятностью 95 % можно отклонить нуль гипотезу и сделать вывод о коинтеграции анализируемых временных рядов.

Если в результате проведенного анализа будет обнаружено отсутствие коинтеграции между рядами либо на предварительном анализе совмещенных графиков в структуре изучаемых временных рядов обнаруживается тренд либо циклические колебания, то перед проведением дальнейших исследований взаимозависимости необходимо устранить тренд и циклическую компоненту из уровней каждого ряда. Наличие этих компонент может привести к завышению истинных показателей тесноты связи изучаемых временных рядов, если оба ряда будут содержать циклические колебания одинаковой периодичности, либо к занижению – в случае если только один из рядов будет содержать циклическую составляющую или периодичности колебаний циклических составляющих будут различными. Методика исключения трендовой составляющей и циклической компоненты рассмотрена в п. 15.1.

Дальнейший анализ взаимосвязи рядов проводят с использованием не исходных уровней, а центрированных рядов, получаемых путем вычитания из исходного ряда составляющих тренда и циклической компоненты. Содержательная интерпретация параметров модели, рассчитанной по центрированным рядам, затруднительна. Ее можно использовать только для прогнозирования.

Пример 2. Расходы на конечное потребление и совокупный доход в течение восьми лет в условных единицах приведены в табл. 15.4 [24].

Таблица 15.4

Расходы на конечное потребление и совокупный доход

Год	Расходы на конечное потребление	Совокупный доход
1	7	10
2	8	12
3	8	11
4	10	12
5	11	14
6	12	15
7	14	17
8	16	20

По табличным данным строим совмещенный график временных рядов (рис. 15.7).

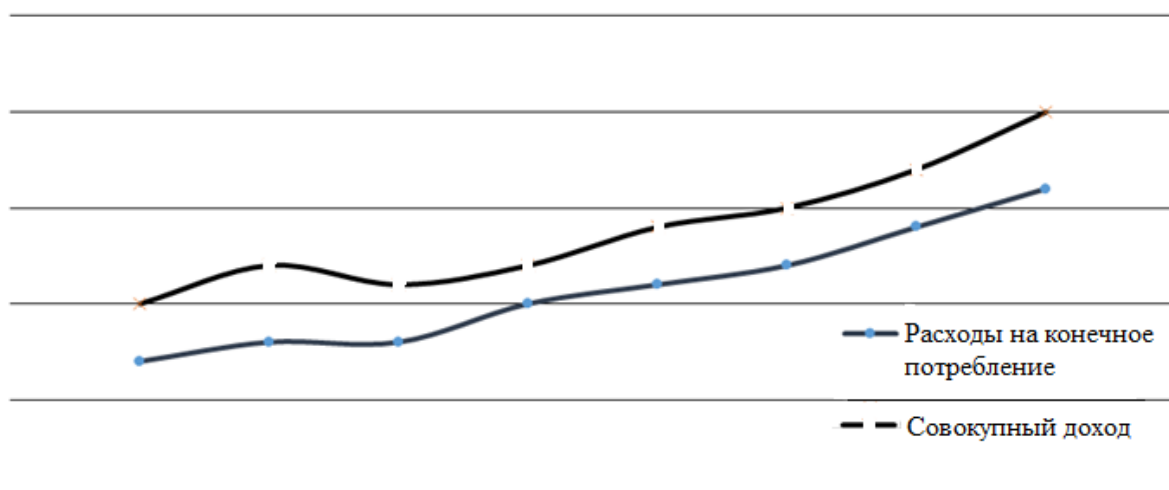


Рис. 15.7. Взаимосвязь временных рядов расхода на конечное потребление $y(t)$ и совокупный доход $x(t)$

На графике видно наличие тренда в анализируемых временных рядах. Корреляционно-регрессионный анализ, проведенный по исходным данным, дает следующие результаты:

$$y(t) = -2,047 + 0,922x(t), \quad R^2 = 95,5\%, \quad r = 0,98.$$

Можно предположить, что полученные результаты (большое значение коэффициента парной корреляции $r = 0,98$) содержат ложную корреляцию, так как в каждом из рядов содержится трендовая компонента.

Выделим трендовые компоненты из исходных рядов. Как видно из графиков рис. 15.8, тренд можно описать полиномом второго порядка.

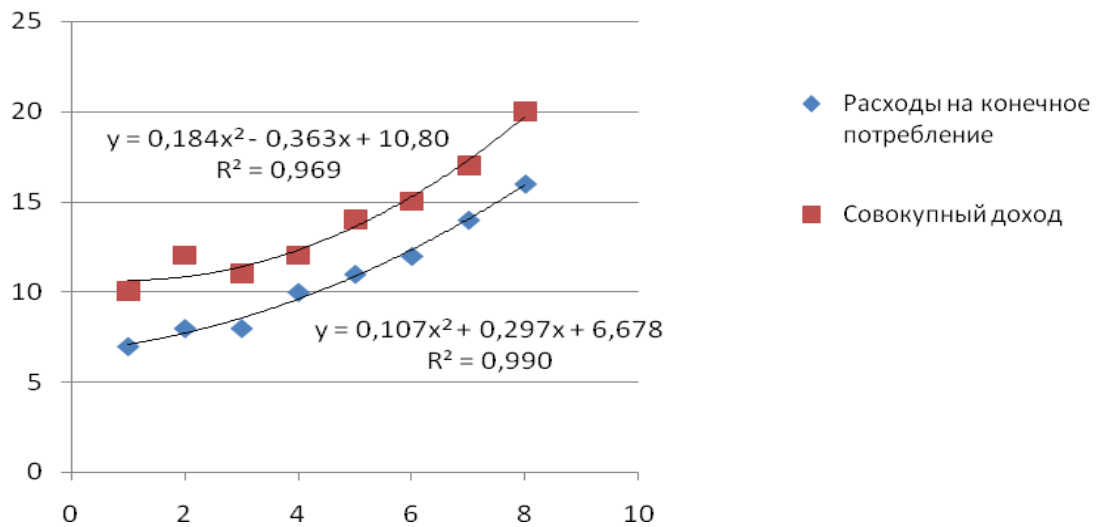


Рис. 15.8. Выделение тренда во временных рядах расхода на конечное потребление и совокупного дохода

Результаты построения модели регрессии по центрированным рядам $y_0(t)$ и $x_0(t)$ приведены ниже

$$y_0(t) = 0,0026 + 0,269 x_0(t).$$

Multiple Regression Analysis

Dependent variable: $y_0(t)$

Parameter	Estimate	Error	Standard T Statistic	P-Value
constant	0,0026	0,103	0,024	0,98
$x_0(t)$	0,269	0,188	1,43	0,20

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	0,176	1	0,176	2,06	0,20
Residual	0,514	6	0,0857		
Total (Corr.)	0,69	7			

R-squared = 25,52 percent

R-squared (adjusted for d.f.) = 13,10 percent

Standard Error of Est. = 0,29

Mean absolute error = 0,1767

Durbin-Watson statistic = 2,82

Регрессионная модель получилась неадекватной, так как расчетное значение критерия Фишера $F = 2,06$ меньше табличного значения для уровня значимости 0,05, числа степеней свободы 1; 6 ($F_{1;6} = 5,99$). Коэффициент корреляции между центрированными рядами незначимый, равен $r = 0,5$.

Связь между временными рядами на конечное потребление и совокупным доходом отсутствует. Уточненный анализ дал противоположные результаты по сравнению с тем, который мог получиться при неучете тренда в исходных временных рядах.

Контрольные вопросы

1. Какой вид имеет модель аддитивного случайного процесса?
2. Назовите функции, используемые для аппроксимации трендовой составляющей временного ряда.
3. Какие функции используются для аппроксимации периодической составляющей временного ряда?
4. Как выполняется оценка случайной компоненты временного ряда?
5. Как оценивается точность модели временного ряда?
6. В чем состоит оценка выполнения предпосылок МНК?
7. Какие причины, вызывающие коинтеграции анализируемых временных рядов, можно назвать?
8. Какие методы обнаружения тренда и циклических колебаний в исходных данных вы знаете?

16. ПРОГНОЗИРОВАНИЕ ВРЕМЕННЫХ РЯДОВ

16.1. Основное содержание прогнозирования процессов

Исследование динамики показателей работы организации и процессов, выявление и характеристика основных тенденций развития и моделей взаимосвязи дают основание для прогнозирования – определения ожидаемых показателей.

Основным содержанием прогнозирования процессов является качественный и количественный анализ реальных процессов, выявление объективных условий, факторов и тенденций развития на основе трех основных принципов разработки прогнозов [24]:

- 1) системности;
- 2) адекватности;
- 3) альтернативности.

Системность прогнозов означает, что процесс рассматривается, с одной стороны, как единое целое, с другой – как совокупность относительно самостоятельных подходов к прогнозированию процесса. Реализация этого принципа предполагает создание моделей, соответствующих каждому отдельному составляющему процесса, и вместе с тем позволяет построить целостную картину возможного развития объекта в будущем. Например, проблему прогнозирования качества вырабатываемого листового стекла в производстве можно решить, смоделировав все стадии производства [26]. Сначала моделировать процесс подготовки шихты с учетом качества поставляемых сырьевых материалов, затем с учетом особенностей конструкции стекловаренной печи и технологических режимов однородность сваренной стекломассы. Далее моделируются стадия формования ленты стекла на расплаве олова и отжиг вырабатываемой ленты в печи отжига и, наконец, способ резки и раскроя ленты стекла на требуемые форматы прямоугольной формы.

Адекватность прогнозирования предполагает соответствие и максимальное приближение разработанной модели к реальному процессу. Модель считается адекватной, если результаты расчетов по модели будут находиться в области рассеивания результатов реального процесса.

Альтернативность прогнозирования связана с возможностями развития процесса по разным направлениям при разных взаимосвязанных и структурных соотношениях. Реализация этого принципа состоит в выявлении разных вариантов развития процесса и выбор того, который может быть реализован в рассматриваемых условиях.

Модель прогнозирования представляет собой модель исследуемого процесса, протекающего в объекте, записанную в аналитической форме в виде алгоритма, компьютерной программы, позволяющей получать информацию о возможных состояниях объекта в будущем.

Важное место в прогнозировании занимают статистические методы прогноза. Применение прогнозирования предполагает, что закономерность развития, действующая в прошлом внутри ряда динамики, сохранится и в прогнозируемом будущем. Теоретической основой

распространения тенденции на будущее является свойство инерционности, которое позволяет выявить сложившиеся взаимосвязи между уровнями динамического ряда, а также между группой взаимосвязанных рядов динамики.

Надежность прогноза возрастает для сопоставимых рядов динамики, полученных на основе использования единой методологии. Точность прогноза зависит от периода упреждения: чем короче период упреждения, тем более надежные и точные результаты дает прогнозирование. За короткий период не успевают сильно измениться условия работы объекта и характер его динамики.

Случайные процессы представляют семейство случайных величин y_t , зависящих от параметра времени t . Модель случайного процесса может быть представлена в следующем виде: $y_t = x_t + \xi_t$, $t = 1, 2, \dots, n$, где x_t – детерминированная неслучайная компонента процесса; ξ_t – стохастическая случайная компонента процесса.

Основные характеристики стационарных случайных процессов:

- математическое ожидание;
- дисперсия;
- автокорреляционная функция.

Характеристики стационарных случайных процессов не зависят от времени, в то время как характеристики нестационарных случайных процессов являются функциями временного аргумента.

16.2. Методы прогнозирования временных рядов

Для математических методов прогнозирования характерны подбор и обоснование математической модели исследуемого процесса, а также способ определения ее неизвестных параметров. Среди математических методов выделяют методы экстраполяции ввиду их простоты. Методологическая предпосылка экстраполяции состоит в признании преимущественной связи между прошлым, настоящим и будущим.

В настоящее время разработана большая группа экстраполяционных методов прогнозирования временных рядов:

1. Методы, основанные на построении корреляционно-регрессионных моделей. При этом строится модель, включающая набор переменных, от которых зависит поведение функции. Прогноз отличается невысокой точностью, используется при прогнозировании показателей конкретных объектов.

2. Методы авторегрессии, учитывающие взаимосвязь членов временного ряда. Применяются для прогнозирования, когда невозможно выделить стабильные причинно-следственные связи. Модель временного ряда имеет вид $y_t = a_0 + a_1y_{t-1} + \dots + a_ny_{t-n}$.

3. Методы, основанные на разложении временного ряда на компоненты: главная тенденция, сезонные колебания, случайная составляющая.

4. Методы, позволяющие учесть неравнозначность исходных данных: метод авторегрессии с последующей адаптацией коэффициентов уравнения, метод взвешенных отклонений.

5. Метод прямой экстраполяции, при котором используются различные трендовые модели. Такие модели применяются для краткосрочного прогнозирования временных рядов, например, на небольшое число шагов и т. д.

Построение и анализ коррелограммы позволяет оценить характер и тенденцию изменения во времени прогнозируемого процесса. Если анализируемый ряд имеет тренд и колебания вокруг него или существует явная зависимость между прошлым и будущим ряда, коррелограмма при тенденции анализируемого ряда к росту будет отражать убывание положительных коэффициентов корреляции с увеличением временного сдвига (рис. 16.1) [24].

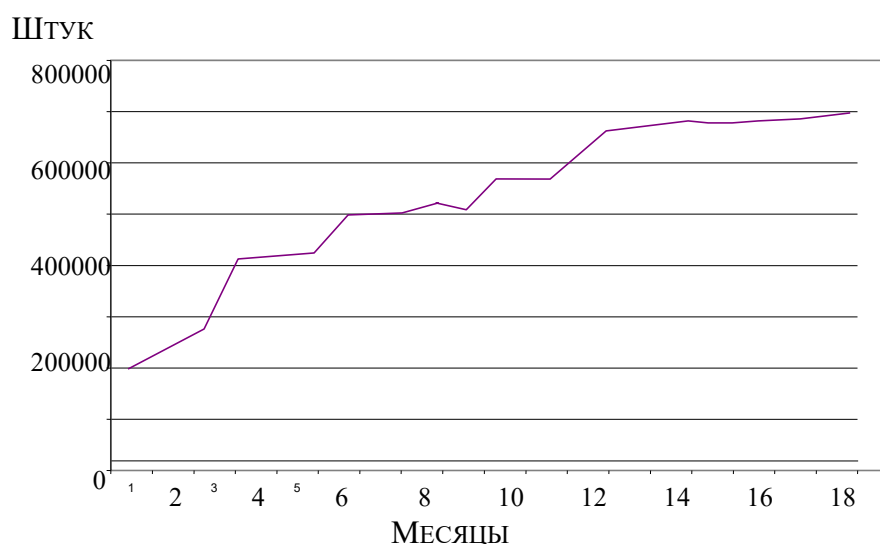


Рис. 16.1. Выработка продукции по месяцам года

Для анализируемого процесса важно оценить характер убывания корреляционной функции к нулю. При линейном характере убывания или степенном характере анализируемые ряды имеют «долговременную память». К таким рядам можно отнести ряды урожайности сельскохозяйственных культур, стоков рек, износ огнеупоров стекловаренной печи в процессе многолетней эксплуатации и др.

Если убывание автокорреляционной функции быстрое, носит экспоненциальный характер, то такие ряды имеют «кратковременную память» и могут быть описаны более сложными моделями автокорреляции – скользящего среднего (модели Бокса – Дженкинса). Более сложным случаем считается колебательный затухающий характер корреляционной функции (рис. 16.2).



Рис. 16.2. Автокорреляционная функция процесса

Наиболее часто используются простейшие алгоритмы прогнозирования:

- по среднему абсолютному приросту при линейной тенденции развития показателя во времени;
- по среднему темпу роста, когда тенденция ряда характеризуется показательной кривой;
- аналитическое описание линии тренда, когда на показатель оказывают влияние множество факторов и ее рассматривают в виде временной функции;

– по корреляционным связям между показателями ряда на ограниченном по времени интервале наблюдения;

– по среднему уровню ряда динамики в случае стационарного характера изменения во времени анализируемого показателя и др.

Алгоритм выбирается по характеру линии тренда:

– прогнозирование по среднему абсолютному приросту проводится по формуле $y_{\text{пр}} = y + (\Delta y)t$, где Δy – средний абсолютный прирост анализируемого показателя; t – период упреждения (прогноз);

– прогнозирование по среднему темпу роста $T_p y_{\text{пр}} = y T_p^t$, где T_p – средний темп роста показателя; t – период упреждения (прогноз); y – последний уровень ряда динамики;

– прогнозирование средним значением уровня ряда $y_{\text{пр}} = y_{\text{ср}}$, где $y_{\text{ср}}$ – среднее значение уровня анализируемого ряда динамики.

Для стационарных случайных процессов прогнозирование можно выполнить с использованием корреляционных зависимостей между последними пятью уровнями исследуемого ряда динамики по формуле

$$y_{\text{пр}} = (y_0 (1 + y_1 y_4 + y_2 y_3 + y_3 y_2 + y_4 y_1)) / (1 + y_4^2 + y_3^2 + y_2^2 + y_1^2),$$

где y_0, y_1, y_2, y_3, y_4 – уровни динамического ряда с показателями функционирования объекта; y_4 соответствует последнему значению уровня ряда.

16.3. Оценка адекватности и точности трендовых моделей прогноза

Для анализа рядов динамики и их прогнозирования в простейшем случае можно использовать офисные информационные технологии, реализованные в электронной таблице EXCEL. В более сложных случаях используются модели на нечетких множествах, нейронных сетях и др.

Трендовая модель временного ряда считается адекватной, если она правильно отражает систематические компоненты временного ряда. Это эквивалентно требованию, чтобы остаточная компонента $\xi_t = y_t - y_{\text{пр}}$, $t = 1, 2, \dots, n$ удовлетворяла свойствам случайной компоненты временного ряда: случайность колебаний, нормальность закона распределения, независимость уровней случайной компоненты.

Исследование остатков ξ_t полезно начинать с изучения их графика. Он может показать наличие какой-то зависимости, не учтенной в модели, необходимость перехода к нелинейной модели (квадратичной, полиномиальной, экспоненциальной) или включения в модель периодических компонент.

График остатков показывает резко отклоняющиеся от модели наблюдения – выбросы. Подобным аномальным наблюдениям надо уделять особое внимание, так как их присутствие может грубо искажать значения оценок. Устранение эффектов выбросов может проводиться либо с помощью удаления этих точек из анализируемых данных (эта процедура называется цензурированием), либо с помощью применения методов оценивания параметров, устойчивых к подобным грубым отклонениям.

Независимость остатков можно проверить расчетом первого коэффициента автокорреляции

$$r(1) = \left(\sum_{i=2}^n \varepsilon_i \varepsilon_{i-1} \right) / \sum_{i=1}^n \varepsilon_i^2.$$

Для принятия решения о наличии или отсутствии автокорреляции в исследуемом ряду фактическое значение коэффициента автокорреляции $r(1)$ сопоставляется с табличным (критическим) значением для 5%-ного уровня значимости (вероятность допустить ошибку при принятии нулевой гипотезы о независимости уровней ряда). Если фактическое значение коэффициента автокорреляции меньше табличного, то гипотеза об отсутствии автокорреляции в ряду может быть принята, а если фактическое значение больше табличного, делают вывод о наличии автокорреляции в ряду динамики, т. е. зависимости остатков.

Для обнаружения гетероскедастичности обычно используют три теста, в которых делаются различные предположения о зависимости между дисперсией случайного члена и объясняющей переменной: тест ранговой корреляции Спирмена, тест Голдфелда – Квандта и тест Глейзера [Доугерти].

При малом объеме выборки для оценки гетероскедастичности может использоваться метод Голдфелда – Квандта. Данный тест применяется для проверки такого типа гетероскедастичности, когда

дисперсия остатков возрастает пропорционально квадрату фактора. При этом делается предположение, что случайная составляющая ξ_t распределена нормально. Оценка нарушения гомоскедастичности по тесту Голдфелда – Квандта описывалась ранее.

Нормальность закона распределения остатков приближенно может быть оценена проверкой статистической значимости расчетных коэффициентов асимметрии A_c и эксцесса ε_c остатков, которые рассчитываются по формулам

$$A_c = \frac{1/n \sum_{t=1}^n \xi_t^3}{\sqrt{\left(1/n \sum_{t=1}^n \xi_t^2\right)^3}}, \quad \sigma_{A_c} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}},$$

где σ_{A_c} – стандартная ошибка вычисления коэффициента асимметрии;

$$\varepsilon_c = \frac{1/n \sum_{t=1}^n \xi_t^4}{\left(1/n \sum_{t=1}^n \xi_t^2\right)^2} - 3, \quad \sigma_{\varepsilon_c} = \sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}},$$

где σ_{ε_c} – стандартная ошибка вычисления коэффициента эксцесса.

При одновременном выполнении неравенств $|A_c| < 1,5 \sigma_{A_c}$, $|\varepsilon_c + 6/(n+1)| < 1,5 \sigma_{\varepsilon_c}$ гипотеза о нормальности закона распределения ξ_t принимается. При соотношении неравенств ($\geq 2\sigma$) гипотеза нормальности распределения остатков отвергается.

Адекватность модели прогноза оценивается по значению коэффициента детерминации с использованием критерия Фишера. Значение коэффициента детерминации рассчитывается по формуле

$$R^2 = 1 - \frac{\sum_{t=1}^n (y_t - y_t^M)^2}{\sum_{t=1}^n (y_t - \bar{y})^2},$$

где y_t – фактическое значение уровня временного ряда; y_t^M – значение уровня временного ряда, рассчитанное по модели; \bar{y} – среднее арифметическое значение временного ряда.

Значимость коэффициента детерминации R^2 проверяется сравнением расчетного значения критерия Фишера с критическим значением

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} \geq F_{T(\alpha,k,n-k-1)},$$

где k – количество факторных переменных в функции тренда; α – уровень значимости, принимается равным 0,05; n – количество уровней в анализируемом временном ряду; $F_{T(\alpha,k,n-k-1)}$ – табличное значение критерия Фишера.

При выполнении неравенства $F > F_T$ коэффициент детерминации принимается значимым, модель адекватно описывает анализируемый временной ряд. Точность прогнозирования модели временного ряда оценивается с использованием ретроспективных данных, не использованных при построении модели временного ряда. Средняя квадратичная ошибка прогнозирования вычисляется по формуле

$$S_{\xi} = \sqrt{\left(\frac{1}{m-k} \sum_{t=1}^m (y_t - y_{\text{пр}})^2 \right)},$$

где $(y_t - y_{\text{пр}})$ – ошибка прогнозирования в точках $t = 1, 2, \dots, m$; k – число параметров в модели прогноза, при временном аргументе t значение $k = 1$.

Контрольные вопросы

1. Что представляет собой модель прогнозирования?
2. Какие методы используются для прогнозирования временных рядов?
3. Назовите наиболее часто используемые алгоритмы прогнозирования временных рядов.
4. Как оценивается точность прогнозирования временного ряда?
5. Что является теоретической основой распространения тенденции изменения временного ряда на будущее?
6. Как зависит точность прогнозирования от интервала упреждения?
7. Как оценивается адекватность модели прогноза?

Параметры приведенной формы модели δ_{ij} могут быть оценены по методу наименьших квадратов. По этим параметрам затем можно рассчитать структурные коэффициенты модели b_{ij} и a_{ij} . Для существования однозначного соответствия между параметрами структурной и приведенной форм необходимо выполнение условия идентификации.

Структурные формы модели могут быть:

- идентифицируемые;
- неидентифицируемые;
- сверхидентифицируемые.

Для того чтобы СФМ была идентифицируема, необходимо чтобы каждое уравнение системы было идентифицируемо. В этом случае число параметров СФМ равно числу параметров приведенной формы.

Если хотя бы одно уравнение СФМ неидентифицируемо, то вся модель считается неидентифицируемой. В этом случае число коэффициентов приведенной формы модели меньше, чем число коэффициентов СФМ.

Модель сверхидентифицируема, если число приведенных коэффициентов больше числа структурных коэффициентов. В этом случае можно получить два и более значений одного структурного коэффициента на основе коэффициентов приведенной формы модели. В сверхидентифицируемой модели хотя бы одно уравнение сверхидентифицируемо, а остальные уравнения идентифицируемы.

Если обозначить число эндогенных переменных в i -м уравнении СФМ через H , а число predetermined переменных, которые содержатся в системе, но не входят в данное уравнение, через D , то условие идентифицируемости модели может быть записано в виде следующего счетного правила:

- если $D + 1 < H$ – уравнение неидентифицируемо;
- если $D + 1 = H$ – уравнение идентифицируемо;
- если $D + 1 > H$ – уравнение сверхидентифицируемо.

Счетное правило является необходимым, но недостаточным условием идентификации. Кроме этого правила для идентифицируемости уравнения должно выполняться дополнительное условие.

Отметим в системе эндогенные и экзогенные переменные, отсутствующие в рассматриваемом уравнении, но присутствующие в системе. Из коэффициентов при этих переменных в других уравнениях составим матрицу. При этом если переменная стоит в левой части уравнения, то коэффициент надо брать с обратным знаком. Если определитель полученной матрицы не равен нулю, а ранг не меньше, чем количество эндогенных переменных в системе без одного, то достаточное условие идентификации для данного уравнения выполнено.

Поясним это на примере следующей структурной модели.

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2. \end{cases} \quad (17.1)$$

Проверим каждое уравнение системы на выполнение необходимого и достаточного условия идентификации.

В первом уравнении три эндогенных переменных: $y_{1,2}$ и y_3 ($H = 3$). В нем отсутствуют экзогенные переменные x_3 и x_4 ($D = 2$). Необходимое условие идентификации $D + 1 = H$ выполнено.

Для проверки на достаточное условие составим матрицу из коэффициентов при переменных x_3 и x_4 (табл. 17.1). В первом столбце таблицы показано, что коэффициенты при экзогенных переменных x_3 и x_4 взяты из уравнений 2 и 3 системы (17.1). Во втором уравнении эти переменные присутствуют, и коэффициенты при них равны a_{23} и a_{24} соответственно. В третьем уравнении эти переменные отсутствуют, т. е. коэффициенты при них равны нулю. Так как вторая строка матрицы состоит из нулей, определитель матрицы равен нулю. Значит, достаточное условие не выполнено, и первое уравнение нельзя считать идентифицируемым.

Таблица 17.1

Матрица, составленная из коэффициентов при переменных x_3 и x_4

Уравнение	Переменная	
	x_3	x_4
2	a_{23}	a_{24}
3	0	0

Во втором уравнении две эндогенные переменные: y_1 и y_2 ($H = 2$). В нем отсутствует экзогенная переменная x_1 ($D = 1$). Необходимое условие идентификации $D + 1 = H$ выполнено.

Для проверки на достаточное условие составим матрицу из коэффициентов при переменных y_3 и x_1 , которые отсутствуют во втором уравнении (табл. 17.2).

Таблица 17.2

Матрица, составленная из коэффициентов при переменных y_3 и x_1

Уравнение	Переменная	
	y_3	x_1
1	b_{13}	a_{11}
3	-1	a_{31}

В третьем уравнении при переменной y_3 коэффициент равен -1 , так как эта переменная стоит в левой части уравнения. Действительно, третье уравнение можно записать в виде $0 = b_{31}y_1 + b_{32}y_2 - 1 \cdot y_3 + a_{31}x_1 + a_{32}x_2$, и тогда равенство $b_{33} = -1$ становится очевидным.

В общем случае СФМ может быть представлена в виде матрицы коэффициентов при переменных. В этом случае третье уравнение может быть задано вектором $(b_{31}, b_{32}, -1, a_{31}, a_{32}, 0, 0)$, а вся система одновременных уравнений (17.1) будет представлена матрицей

$$\begin{pmatrix} -1 & b_{12} & b_{13} & a_{11} & a_{12} & 0 & 0 \\ b_{21} & -1 & 0 & 0 & a_{22} & a_{23} & a_{24} \\ b_{31} & b_{32} & -1 & a_{31} & a_{32} & 0 & 0 \end{pmatrix}.$$

Определитель представленной в табл. 17.2 матрицы не равен нулю, а ранг матрицы равен 2. Значит, достаточное условие выполнено, и второе уравнение идентифицируемо.

В третьем уравнении три эндогенные переменные: y_1 , y_2 и y_3 ($H = 3$). В нем отсутствуют экзогенные переменные x_3 и x_4 ($D = 2$). Необходимое условие идентификации $D + 1 = H$ выполнено.

Для проверки на достаточное условие составим матрицу из коэффициентов при переменных x_3 и x_4 , которые отсутствуют в третьем уравнении (табл. 17.3). Согласно таблице определитель матрицы равен нулю (первая строка состоит из нулей). Значит, достаточное условие не выполнено, и третье уравнение нельзя считать идентифицируемым.

Таблица 17.3

Матрица, составленная из коэффициентов при переменных x_3 и x_4

Уравнение	Переменная	
	x_3	x_4
1	0	0
2	a_{23}	a_{24}

В эконометрических моделях иногда используются балансовые тождества переменных (например, вида $y_3 = y_1 + y_2 + x_1$). Коэффициенты при переменных при этом не требуют оценок, и уравнение не надо исследовать на идентификацию, но в проверке на идентификацию всей системы эти уравнения участвуют. Присутствующие иногда в моделях свободные и остаточные члены ($a_{01}, a_{02}, a_{03}, \dots, \varepsilon_1, \varepsilon_2, \varepsilon_3, \dots$) не влияют на решение вопроса об идентификации.

Косвенный метод наименьших квадратов

При оценивании коэффициентов структурной модели используется ряд методов. С этими методами можно ознакомиться в рекомендованной литературе [17]. Рассмотрим косвенный метод наименьших квадратов (КМНК), который применяется в случае идентифицируемой структурной модели, на примере следующей идентифицируемой модели, содержащей две эндогенные и две экзогенные переменные:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{22}x_2 + \varepsilon_2. \end{cases} \quad (17.2)$$

Для построения модели мы располагаем информацией, представленной в табл. 17.4

Таблица 17.4

Фактические данные для построения модели

n	y_1	y_2	x_1	x_2
1	33,0	37,1	3	11
2	45,9	49,3	7	16
3	42,2	41,6	7	9
4	51,4	45,9	10	9
5	49,0	37,4	10	1
6	49,3	52,3	8	16
Сумма	270,8	263,6	45	62
Среднее значение	45,133	43,930	7,500	10,333

Структурную модель (17.2) преобразуем в приведенную форму модели

$$\begin{cases} y_1 = d_{11}x_1 + d_{12}x_2 + u_1, \\ y_2 = d_{21}x_1 + d_{22}x_2 + u_2, \end{cases} \quad (17.3)$$

где u_1 и u_2 – случайные ошибки.

Для каждого уравнения приведенной формы при расчете коэффициентов d применяем метод наименьших квадратов. Для упрощения расчетов можно работать с отклонениями от средних уровней $y = y - y_{\text{ср}}$ и $x = x - x_{\text{ср}}$ ($y_{\text{ср}}$ и $x_{\text{ср}}$ – средние значения). Преобразованные таким образом данные табл. 17.4 сведены в табл. 17.5. Здесь же показаны промежуточные расчеты, необходимые для определения коэффициентов d_{ik} .

Для нахождения коэффициентов d_{ik} первого приведенного уравнения можно использовать следующую систему нормальных уравнений:

$$\begin{cases} \Sigma y_1x_1 = d_{11}\Sigma x_1^2 + d_{12}\Sigma x_1x_2, \\ \Sigma y_1x_2 = d_{11}\Sigma x_1x_2 + d_{12}\Sigma x_2^2. \end{cases} \quad (17.4)$$

Таблица 17.5

Преобразованные данные для построения приведенной формы модели

n	y_1	y_2	x_1	x_2	y_1x_1	x_1^2	x_1x_2	y_1x_2	y_2x_1	y_2x_2	x_2^2
1	-12,133	-6,784	-4,500	0,667	54,599	20,250	-3,002	-8,093	30,528	-4,525	0,445
2	0,767	5,329	-0,500	5,667	-0,383	0,250	-2,834	4,347	-2,664	30,198	32,115
3	-2,933	-2,308	-0,500	-1,333	1,467	0,250	0,667	3,910	1,154	3,077	1,777
4	6,267	1,969	2,500	-1,333	15,668	6,250	-3,333	-8,354	4,922	-2,625	1,777
5	3,867	-6,541	2,500	-9,333	9,667	6,250	-23,333	-36,091	-16,353	61,048	87,105
6	4,167	8,337	0,500	5,667	2,084	0,250	2,834	23,614	4,168	47,244	32,115
Сумма	0,002	0,001	0,000	0,002	83,102	33,500	-29,001	-20,667	21,755	134,417	155,334

Подставляя рассчитанные в табл. 17.5 значения сумм в (17.4), получим

$$\begin{aligned} 83,102 &= 33,5d_{11} - 29,001d_{12}, \\ -20,667 &= -29,001d_{11} + 155,334d_{12}. \end{aligned}$$

Решение этих уравнений дает значения $d_{11} = 2,822$ и $d_{12} = 0,394$. Первое уравнение приведенной формы модели примет вид

$$y_1 = 2,822x_1 + 0,394x_2 + u_1. \quad (17.5)$$

Для нахождения коэффициентов d_{2k} второго приведенного уравнения из (17.3) можно использовать следующую систему нормальных уравнений:

$$\begin{cases} \Sigma y_2 x_1 = d_{21} \Sigma x_1^2 + d_{22} \Sigma x_1 x_2, \\ \Sigma y_2 x_2 = d_{21} \Sigma x_1 x_2 + d_{22} \Sigma x_2^2. \end{cases}$$

Подставляя рассчитанные в табл. 17.5 значения сумм, получим

$$\begin{aligned} 21,755 &= 33,5d_{21} - 29,001d_{22}, \\ 134,417 &= -29,001d_{21} + 155,334d_{22}. \end{aligned}$$

Решение этих уравнений дает значения $d_{21} = 1,668$ и $d_{22} = 1,177$. Второе уравнение приведенной формы модели примет вид

$$y_2 = 1,668x_1 + 1,177x_2 + u_2. \quad (17.6)$$

Для перехода от приведенной к структурной форме модели найдем x_2 из второго уравнения приведенной формы модели (17.6).

$$x_2 = (y_2 - 1,668x_1)/1,177.$$

Подставим это выражение в первое уравнение приведенной модели (17.5) и найдем структурное уравнение

$$\begin{aligned} y_1 &= 2,822x_1 + 0,394(y_2 - 1,668x_1)/1,177 = \\ &= 2,822x_1 + 0,335y_2 - 0,558x_1 = 0,335y_2 + 2,264x_1. \end{aligned} \quad (17.7)$$

Таким образом, $b_{12} = 0,335$; $a_{11} = 2,264$.

Найдем x_1 из первого уравнения приведенной формы модели (17.5)

$$x_1 = (y_1 - 0,394x_2)/2,822.$$

Подставим это выражение во второе уравнение приведенной модели (17.6), найдем структурное уравнение

$$\begin{aligned} y_2 &= 1,177x_2 + 1,668(y_1 - 0,394x_2)/2,822 = \\ &= 1,177x_2 + 0,591y_1 - 0,233x_2 = 0,591y_1 + 0,944x_2. \end{aligned} \quad (17.8)$$

Таким образом, $b_{21} = 0,591$; $a_{22} = 0,944$.

Свободные члены структурной формы находим из уравнений (17.7 и 17.8) с учетом того, что $y_{1,ср} = 45,133$; $y_{2,ср} = 43,93$; $x_{1,ср} = 7,5$; $x_{2,ср} = 10,333$ (см. табл. 17.4):

$$\begin{aligned} a_{01} &= y_{1,ср} - b_{12}y_{2,ср} - a_{11}x_{1,ср} = \\ &= 45,133 - 0,335 \cdot 43,93 - 2,264 \cdot 7,5 = 13,436, \\ a_{02} &= y_{2,ср} - b_{21}y_{1,ср} - a_{22}x_{2,ср} = \\ &= 43,93 - 0,591 \cdot 45,133 - 0,944 \cdot 10,333 = 7,502. \end{aligned}$$

Окончательный вид структурной модели

$$\begin{aligned} y_1 &= a_{01} + b_{12}y_2 + a_{11}x_1 + \varepsilon_1 = 13,436 + 0,335y_2 + 2,264x_1 + \varepsilon_1, \\ y_2 &= a_{02} + b_{21}y_1 + a_{22}x_2 + \varepsilon_2 = 7,502 + 0,591y_1 + 0,944x_2 + \varepsilon_2. \end{aligned}$$

Контрольные вопросы

1. Какие существуют виды систем линейных уравнений?
2. Какие структурные формы модели вы знаете?
3. Каковы необходимые условия идентифицируемости модели?
4. Сформулируйте достаточное условие идентифицируемости модели.
5. Как выполняется оценивание коэффициентов структурной модели косвенным методом наименьших квадратов?

ЗАКЛЮЧЕНИЕ

Анализ данных – область математики и информатики, занимающаяся построением и исследованием наиболее общих математических методов и вычислительных алгоритмов извлечения знаний из экспериментальных данных; процесс исследования, фильтрации, преобразования и моделирования данных с целью извлечения полезной информации и принятия решений. Анализ данных имеет множество аспектов и подходов, охватывает разные методы в различных областях науки и деятельности.

Любые методы обработки данных так или иначе используются для структурирования и анализа существующей информации. Методов по анализу информации много.

Интеллектуальный анализ данных – это особый метод анализа данных, который фокусируется на моделировании и открытии данных, а не на их описании. Бизнес-аналитика охватывает анализ данных, который полагается на агрегацию. В статистическом смысле некоторые разделяют анализ данных на описательную статистику, исследовательский анализ данных и проверку статистических гипотез. Исследовательский анализ данных занимается открытием новых характеристик данных, а проверка статистических гипотез – подтверждением или опровержением существующих гипотез. Прогнозный анализ фокусируется на применении статистических или структурных моделей для предсказания или классификации, а анализ текста применяет статистические, лингвистические и структурные методы для

извлечения и классификации информации из текстовых источников, принадлежащих к неструктурированным данным. Все это разновидности анализа данных.

Интеграция данных – это предшественник анализа данных, а сам анализ данных тесно связан с визуализацией данных и их распространением. Термин «анализ данных» иногда используется как синоним к моделированию данных.

Очевидно, невозможно в рамках одного учебного пособия описать все методы анализа данных. Поэтому авторами учебного пособия была сделана попытка осветить основные, наиболее широко применяемые методы анализа данных. Рассмотренные примеры дают возможность разобраться в особенностях использования статистических методов при анализе количественных данных. Количественные методы исследования помогают анализировать явления и процессы с опорой на количественные показатели. Полученные количественные характеристики позволяют выявить общие закономерности и устранить случайные незначительные отклонения в данных.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Луньков, А. Д. Интеллектуальный анализ данных : учеб. пособие для студентов спец. 080801.65 «Прикладная информатика (в экономике)» / А. Д. Луньков, А. В. Харламов ; Саратов. гос. соц.-эконом. ун-т. – Саратов, 2012. – 92 с. – ISBN 978-5-4345-0113-2.
2. Вентцель, Е. С. Теория вероятностей : учеб. для вузов / Е. С. Вентцель. – М. : КноРус, 2010. – 658 с. – ISBN 978-5-406-00476-0.
3. Макаров, Р. И. Изучение и анализ заболеваемости работников стекольного производства [Электронный ресурс] / Р. И. Макаров, Е. Р. Хорошева // Алгоритмы, методы и системы обработки данных. – 2018. – № 1(30). – С. 46 – 51. – URL: amisod.ru (дата обращения: 18.06.2020).
4. Прохоров, Ю. В. Лекции по теории вероятностей и математической статистике [Электронный ресурс] : учебник / Ю. В. Прохоров, Л. С. Пономаренко. – М. : Изд-во Моск. гос. ун-та, 2012. (Классический университетский учебник). – URL: <http://www.studentlibrary.ru/book/ISBN9785211062344.html> (дата обращения: 18.06.2020).
5. Смирнов, Н. В. Краткий курс математической статистики для технических приложений / Н. В. Смирнов, И. В. Дунин-Барковский. – М. : Физматгиз, 1959. – 436 с.
6. Проверка статистических гипотез [Электронный ресурс]. – URL: <http://www.studfiles.ru/preview/1582412/> (дата обращения: 18.06.2020).
7. Теория сигналов и линейных систем. Случайные процессы и сигналы [Электронный ресурс]. – URL: <http://www.bourabai.kz/signals/ts171.htm> (дата обращения: 18.06.2020).
8. Медиченко, М. П. Радиотехнические цепи и сигналы : учеб. пособие / М. П. Медиченко, В. П. Литвинов. – М. : Изд-во МГОУ, 2011. – 156 с. – ISBN 928-5-7045-8676-2.
9. Дубров, А. М. Многомерные статистические методы : учебник / А. М. Дубров, В. С. Мхитарян, Л. И. Трошин. – М. : Финансы и статистика, 2003. – 352 с. – ISBN 5-279-01945-3.
10. Клячкин, В. Н. Статистические методы в управлении качеством: компьютерные технологии : учеб. пособие / В. Н. Клячкин. – М. : Финансы и статистика, 2007. – 302 с. – ISBN 978-5-279-03046-0.
11. Информационные технологии в управлении качеством автомобильного стекла : учеб. пособие / Р. И. Макаров [и др.] ; Владим. гос. ун-т. – Владимир : Изд-во Владим. гос. ун-та, 2010. – 276 с. – ISBN 978-5-9984-0038-4.

12. Дисперсионный анализ [Электронный ресурс]. – URL: <http://www.studfiles.ru/preview/1582408/> (дата обращения: 18.06.2020).
13. Налимов, В. В. Статистические методы планирования экстремальных экспериментов / В. В. Налимов, Н. А. Чернова. – М. : Наука, 1965. – 340 с.
14. Линник, Ю. В. Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений / Ю. В. Линник. – 2-е изд., стер. – М. : Физматгиз, 1962. – 349 с.
15. Экономико-математические методы и прикладные модели : учеб. пособие для вузов / В. В. Федосеев [и др.]. – М. : ЮНИТИ, 1999. – 391 с.
16. Эконометрика : учебник / под ред. И. И. Елисеевой. – М. : Финансы и статистика, 2001. – 342 с. – ISBN 5-279-01955-0.
17. Практикум по эконометрике : учеб. пособие / под ред. И. И. Елисеевой. – М. : Финансы и статистика, 2001. – 189 с. – ISBN 5-279-01955-0.
18. Маркетинговые исследования [Электронный ресурс]. – URL: http://studme.org/1566072110865/marketing/osnovnye_rezultaty_primene_niya_metoda_glavnyh_komponent (дата обращения: 18.06.2020).
19. Факторный анализ [Электронный ресурс]. – URL: <http://www.bourabai.kz/tpoi/factor.htm> (дата обращения: 18.06.2020).
20. Факторный анализ [Электронный ресурс]. – URL: <http://www.studfiles.ru/preview/1582409> (дата обращения: 18.06.2020).
21. Факторный анализ [Электронный ресурс]. – URL: <https://ru.wikipedia.org/wiki/%D0%A4%D0%B0%D0%BA%D1%82%D0%BE%D1%80%D0%BD> (дата обращения: 18.06.2020).
22. Анализ главных компонент и факторный анализ [Электронный ресурс]. – URL: http://gym42.ru/stat/Book/Data/page_2_8_4.htm (дата обращения: 18.06.2020).
23. Эконометрика : учебник / под ред. И. И. Елисеевой. – 2-е изд., перераб. и доп. – М. : Финансы и статистика, 2006. – 576 с. – ISBN 5-279-02786-3.
24. Яновский, Л. П. Введение в эконометрику : учеб. пособие / Л. П. Яновский, А. Г. Буховец ; под ред. Л. П. Яновского. – 2-е изд., доп. – М. : КОНКУРС, 2007. – 256 с. – ISBN 978-5-85971-270-0.
25. Управление качеством автомобильного стекла / Р. И. Макаров [и др.] ; Владим. гос. ун-т. – Владимир : Изд-во Владим. гос. ун-та, 2009. – 275 с. – ISBN 978-5-89368-965-5.
26. Управление качеством листового стекла (флоат способ) : учеб. пособие / Р. И. Макаров [и др.]. – М. : Изд-во Ассоц. строит. вузов, 2004. – 152 с. – ISBN 5-93093-261-1.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ	3
1. АНАЛИЗ ДАННЫХ	6
1.1. Понятие интеллектуального анализа данных	6
1.2. Data Mining как часть рынка информационных технологий	7
1.3. Набор данных и их атрибутов	8
1.4. Задачи Data Mining	11
1.5. Основы анализа данных	12
1.6. Задача визуализации	21
1.7. Основные этапы интеллектуального анализа	23
1.8. Инструментальные средства анализа данных	24
Контрольные вопросы	25
2. СЛУЧАЙНЫЕ СОБЫТИЯ	26
2.1. Испытание. Поле событий. Операции над событиями	26
2.2. Операции над событиями	27
2.3. Частота и вероятность	28
2.4. Основные аксиомы теории вероятностей	29
2.5. Случайные события	32
Контрольные вопросы	36
3. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ. ЗАКОНЫ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН	37
3.1. Определение случайной величины	37
3.2. Законы распределения дискретных случайных величин	40
3.3. Законы распределения непрерывных случайных величин	43
Контрольные вопросы	46
4. НЕПРЕРЫВНЫЕ СЛУЧАЙНЫЕ ВЕЛИЧИНЫ. ЗАКОНЫ РАСПРЕДЕЛЕНИЯ	47
4.1. Нормальное распределение	47
4.2. Распределение χ -квадрат	51
4.3. Распределение Стьюдента	51
4.4. Распределение Фишера	52
Контрольные вопросы	53
5. МНОГОМЕРНОЕ РАСПРЕДЕЛЕНИЕ ДИСКРЕТНЫХ И НЕПРЕРЫВНЫХ СЛУЧАЙНЫХ ВЕЛИЧИН	53
5.1. Двумерное дискретное распределение	53
5.2. Двумерное непрерывное нормальное распределение	55
5.3. Многомерное распределение	57
5.4. Закон больших чисел	59

5.5. Основные предельные законы теории вероятностей	60
Контрольные вопросы	62
6. СТАТИСТИЧЕСКИЕ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ ПО МАЛЫМ ВЫБОРКАМ	62
6.1. Получение точечных оценок параметров генеральной совокупности по выборке	64
6.2. Интервальное оценивание параметров генеральной совокупности по выборке	67
Контрольные вопросы	69
7. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ. ОСНОВНЫЕ ПОНЯТИЯ, ИСПОЛЬЗУЕМЫЕ ПРИ ПРОВЕРКЕ ГИПОТЕЗ	70
7.1. Уровень значимости и мощность критерия. Ошибки при проверке гипотез	70
7.2. Статистические критерии	72
7.3. Критерий значимости при биномиальном распределении	74
7.4. Односторонние и двусторонние критерии	75
7.5. Некоторые типичные задачи проверки параметрических гипотез	77
7.6. Непараметрические гипотезы. Критерии согласия.....	83
7.7. Критерий Пирсона	83
Контрольные вопросы	86
8. СЛУЧАЙНЫЕ ПРОЦЕССЫ	86
8.1. Функциональные характеристики случайного процесса	88
8.2. Свойства функций автоковариации и автокорреляции	95
8.3. Классификация случайных процессов	98
8.4. Энергетический спектр случайного процесса	102
8.5. Нормальный случайный процесс	103
Контрольные вопросы	104
9. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ	105
9.1. Функциональные и корреляционные связи между переменными	105
9.2. Задачи корреляционного анализа	107
9.3. Многомерный корреляционный анализ	109
Контрольные вопросы	115
10. ДИСПЕРСИОННЫЙ АНАЛИЗ	115
10.1. Однофакторный дисперсионный анализ	118
10.2. Двухфакторный дисперсионный анализ	121
Контрольные вопросы	130

11. РЕГРЕССИОННЫЙ АНАЛИЗ	130
11.1. Вычисление коэффициентов регрессии	131
11.2. Статистический анализ уравнения регрессии	134
11.3. Проверка выполнения предпосылок методом наименьших квадратов (МНК)	136
11.4. Оценка влияния отдельных факторов на зависимую переменную на основе модели	138
11.5. Построение точечных и интервальных прогнозов на основе регрессионной модели	139
Контрольные вопросы	144
12. НЕЛИНЕЙНАЯ РЕГРЕССИЯ	145
Контрольные вопросы	154
13. КОМПОНЕНТНЫЙ АНАЛИЗ	154
Контрольные вопросы	163
14. МЕТОДЫ АНАЛИЗА БОЛЬШИХ СИСТЕМ. ФАКТОРНЫЙ АНАЛИЗ	164
14.1. Факторный анализ	165
14.2. Сущность факторного анализа	166
14.3. Последовательность факторного анализа	170
Контрольные вопросы	178
15. МОДЕЛИ ВРЕМЕННЫХ РЯДОВ И СТАТИСТИЧЕСКИЕ ОЦЕНКИ ВЗАИМОСВЯЗИ ВРЕМЕННЫХ РЯДОВ	178
15.1. Модели временных рядов	178
15.2. Статистические оценки взаимосвязи двух временных рядов	185
Контрольные вопросы	191
16. ПРОГНОЗИРОВАНИЕ ВРЕМЕННЫХ РЯДОВ	191
16.1. Основное содержание прогнозирования процессов	191
16.2. Методы прогнозирования временных рядов	193
16.3. Оценка адекватности и точности трендовых моделей прогноза	196
Контрольные вопросы	199
17. СИСТЕМЫ ЛИНЕЙНЫХ УРАВНЕНИЙ	200
Контрольные вопросы	208
ЗАКЛЮЧЕНИЕ	209
БИБЛИОГРАФИЧЕСКИЙ СПИСОК	211

Учебное издание

МАКАРОВ Руслан Ильич
ХОРОШЕВА Елена Руслановна

МЕТОДЫ АНАЛИЗА ДАННЫХ

Учебное пособие

Редактор А. П. Володина
Технический редактор Ш. В. Абдуллаев
Корректор О. В. Балашова
Компьютерная верстка Е. А. Герасиной
Выпускающий редактор А. А. Амирсейидова

Подписано в печать 20.12.21.
Формат 60×84/16. Усл. печ. л. 12,56. Тираж 50 экз.

Заказ

Издательство

Владимирского государственного университета
имени Александра Григорьевича и Николая Григорьевича Столетовых.
600000, Владимир, ул. Горького, 87.