М. А. Ковалева
С. Б. Волошин

2019

Анализ данных

Учебное пособие

УДК 519
ББК 22.172
К 56

Рецензент(ы): И.И. Болотаева – к.т.н., доцент кафедры «Информационные технологии и системы» ФГБОУ ВО «Северо-Кавказский горно-металлургический институт (государственный технологический университет)».

Ковалева, Мария Александровна
Волошин, Сергей Борисович
К 56 Анализ данных. Учебное пособие – М.: Мир науки, 2019. – Сетевое издание. Режим доступа: https://izd-mn.com/PDF/32MNNPU19.pdf– Загл. с экрана.

ISBN 978-5-6043306-2-3

Учебное пособие предназначено для студентов направления подготовки 38.03.05 «Бизнес-информатика» и содержит перечень изучаемых тем в соответствии с рабочей программой дисциплины «Анализ данных», решения типовых задач с подробными пояснениями, с использованием R., варианты контрольной работы, вопросы к экзамену, таблицы, а также список основной и дополнительной литературы.

Дисциплина «Анализ данных» является базовой компонентой цикла математических и естественнонаучных дисциплин ФГОС ВО по направлению подготовки 38.03.05 «Бизнес информатика», профиль ИТ менеджмент в бизнесе.

Учебное пособие будет полезно преподавателям, аспирантам, научным сотрудникам, студентам, изучающих применение современных вычислительных технологий в анализе данных и эконометрике, для широкого круга читателей.

ISBN 978-5-6043306-2-3

Оглавление

Введение

Цели дисциплины.

Анализ данных – один из важнейших разделов математической статистики и информатики, представляющий собой комплекс методов и средств, позволяющих получить из определенным образом организованных данных информацию для принятия решений.

Целью изучения дисциплины «Анализ данных» является освоение теоретических основ и методов анализа данных, применяемых при решении прикладных (в том числе экономических) задач.

Основной принцип, лежащий в основе данной дисциплины, состоит в повышении уровня фундаментальной экономико-математической и статистической подготовки студентов с усилением ее прикладной экономической направленности.

Задачи дисциплины Задачи изучения дисциплины вытекают из требований к результатам освоения и условиям реализации основной образовательной программы и компетенций, установленных Федеральным государственным образовательным стандартом высшего образования по направлению «Бизнес-информатика». В ходе изучения дисциплины ставятся задачи:

1. Формирование навыков обработки, обобщения и анализа информации для оценки состояния, и выявления тенденций, закономерностей и конкретных особенностей развития социально- экономических и бизнеспроцессов.

2. Овладение современными методиками статистического моделирования при решении задач.

3. Освоение компьютерных технологий, применяемых в анализе данных.

Перечень планируемых результатов обучения дисциплине.

В совокупности с другими дисциплинами базовой и вариативной части общенаучного и профессионального циклов образовательной программы подготовки бакалавров по направлению 38.03.05 «Бизнесинформатика», профиль ИТ-менеджмент в бизнесе, дисциплина «Анализ данных» обеспечивает формирование следующих компетенций:

ПК-17 - способность использовать основные методы естественнонаучных дисциплин в профессиональной деятельности для теоретического и экспериментального исследования.

Знать основные понятия и методы теории вероятностей и математической статистики, необходимые для успешного решения математических, финансовых и экономических задач, связанных с анализом и обработкой результатов наблюдений, основные математические методы в контексте анализа данных.

Уметь решать типовые теоретиковероятностные и статистические задачи, применять соответствующие методы и знания высшей математики и теории вероятностей для решения математических и экономических задач, интерпретировать результаты решения задач, применять основные математические методы и инструментальные средства в профессиональной деятельности для решения прикладных задач и исследования объектов профессиональной деятельности; строить математические модели объектов профессиональной деятельности; использовать математические инструментальные средства для обработки, анализа и систематизации информации по теме исследования.

Владеть методикой построения, анализа и применения вероятностных и статистических моделей для статистической оценки состояния и прогноза развития анализируемых экономических явлений и процессов, навыками применения инструментов математического моделирования, методами статистического анализа и прогнозирования случайных процессов.

ПК-18 - Способность использовать соответствующий математический аппарат и инструментальные средства для обработки, анализа и систематизации информации по теме исследования.

Знать основные методы и средства решения задач анализа данных; иметь представление об основных тенденциях развития теории и практики данных и методах работы с ними, основные понятия и методы теории вероятностей и математической статистики, необходимые для успешного решения математических, финансовых и экономических задач, связанных с анализом и обработкой результатов наблюдений.

Уметь решать типовые теоретиковероятностные и статистические задачи, применять соответствующие методы и знания высшей математики и теории вероятностей для решения математических и экономических задач, интерпретировать результаты решения задач.

Владеть методикой построения, анализа и применения вероятностных и статистических моделей для статистической оценки состояния и прогноза развития анализируемых экономических явлений и процессов.

ПКП-2 - способность к управлению экономикой и финансами ИТ.

Знать экономические модели и методики финансовоэкономического анализа, применяемые аналитические инструментальные средства, основные понятия и методы теории вероятностей и математической статистики, необходимые для успешного решения прикладных математических, финансовых и экономических задач, связанных с анализом и обработкой результатов наблюдений.

Уметь решать типовые теоретиковероятностные и статистические задачи, применять соответствующие методы и знания высшей математики и теории вероятностей для решения математических и экономических задач, интерпретировать результаты решения задач.

Владеть методикой разработки систем математического обеспечения при решении научнотехнических и производственных задач различных профилей, собирать и анализировать информации по решаемой задаче, составлять ее математическое описание, обеспечивать накопление, анализ и систематизацию собранных данных с использованием современных методов автоматического сбора и обработки информации.

ПКП-3 - умения разрабатывать эффективные коммуникации между ИТ – персоналом и бизнес-пользователями.

Знать основные математические методы в контексте анализа данных.

Уметь применять основные математические методы в профессиональной деятельности для решения прикладных задач.

Владеть методами статистического анализа и прогнозирования случайных процессов.

Место дисциплины в структуре образовательной программы.

Дисциплина «Анализ данных» является является базовой компонентой цикла математических и естественнонаучных дисциплин ФГОС ВО по направлению подготовки 38.03.05 «Бизнесинформатика», профиль ИТменеджмент в бизнесе.

При изучении дисциплины закладывается фундамент теоретических основ и методов анализа данных, применяемых при решении прикладных задач.

Объём дисциплины и виды учебной работы:

Общая трудоёмкость дисциплины составляет 4 зачётных единиц.

Виды промежуточного контроля – контрольная работа.

Вид промежуточной аттестации – экзамен.

1 Введение в анализ данных

Анализ данных — широкое понятие. Сегодня существуют десятки его определений. В самом общем смысле *анализ данных* — это исследования, связанные с обсчетом многомерной системы данных, имеющей множество параметров. В процессе анализа данных исследователь производит совокупность действий с целью формирования определенных представлений о характере явления, описываемого этими данными. Как правило, для анализа данных используются различные математические методы.

Анализ данных нельзя рассматривать только как обработку информации после ее сбора. Анализ данных — это прежде всего средство проверки гипотез и решения задач исследователя.

Известное противоречие между ограниченными познавательными способностями человека и бесконечностью Вселенной заставляет нас использовать модели и моделирование, тем самым упрощая изучение интересующих объектов, явлений и систем.

Слово «модель» (лат. modelium) означает «мера», «способ», «сходство с какойто вещью».

Построение моделей — универсальный способ изучения окружающего мира, позволяющий обнаруживать зависимости, прогнозировать, разбивать на группы и решать множество других задач. Основная цель моделирования в том, что модель должна достаточно хорошо отображать функционирование моделируемой системы.

Модель — объект или описание объекта, системы для замещения (при определенных условиях, предположениях, гипотезах) одной системы (то есть оригинала) другой системой для лучшего изучения оригинала или воспроизведения каких-либо его свойств.

Моделирование — универсальный метод получения, описания и использования знаний. Применяется в любой профессиональной деятельности.

По виду моделирования модели делят:

1. на эмпирические — полученные на основе эмпирических фактов, зависимостей;

2. теоретические — полученные на основе математических описаний, законов;

3. смешанные, полуэмпирические — полученные на основе эмпирических зависимостей и математических описаний.

Таким образом, анализ данных тесно связан с моделированием. Отметим важные свойства любой модели.

1. Упрощенность. Модель отображает только существенные стороны объекта и, кроме того, должна быть проста для исследования или воспроизведения.
2. Конечность. Модель отображает оригинал лишь в конечном числе его отношений, и, кроме того, ресурсы моделирования конечны.
3. Приближенность. Действительность отображается моделью грубо или приближенно.
4. Адекватность. Модель должна успешно описывать моделируемую систему.
5. Целостность. Модель реализует некоторую систему (то есть целое).
6. Замкнутость. Модель учитывает и отображает замкнутую систему необходимых основных гипотез, связей и отношений.
7. Управляемость. Модель должна иметь хотя бы один параметр, изменениями которого можно имитировать поведение моделируемой системы в различных условиях.

Аналитический подход к моделированию

Модель в традиционном понимании представляет собой результат отображения одной структуры (изученной) на другую (малоизученную). Так, отображая физическую систему (объект) на математическую (например, математический аппарат уравнений), получим физико-математическую модель системы, или математическую модель физической системы.

Любая модель строится и исследуется при определенных допущениях, гипотезах. Делается это обычно с помощью математических методов.

Пример

Рассмотрим экономическую систему. Величина ожидаемого спроса s на будущий месяц (t + 1) рассчитывается на основе формулы s(t + 1) = [s(t) + s(t - 1) + s(t - 2)] / 3, то есть как среднее от продаж за предыдущие три месяца. Это простейшая математическая модель прогноза продаж. При построении этой модели были приняты следующие гипотезы.

Во-первых, годовая сезонность в продажах отсутствует.

Во-вторых, на величину продаж не влияют никакие внешние факторы: действия конкурентов, макроэкономическая ситуация и т. д.

Использовать такую модель легко: имея данные о продажах за предыдущие месяцы, по формуле мы получим прогноз на будущий месяц.

Такой подход к моделированию в литературе называют *аналитическим.*

Аналитический подход к моделированию базируется на том, что исследователь при изучении системы отталкивается от модели, в соответствии с рисунком 1. В этом случае он по тем или иным соображениям выбирает подходящую модель. Как правило, это теоретическая модель, закон, известная зависимость, представленная чаще всего в *функциональном виде* (например, уравнение, связывающее выходной параметр *y* с входными воздействиями x_1, x_2…). Варьирование входных параметров на выходе даст результат, который моделирует поведение системы в различных условиях.

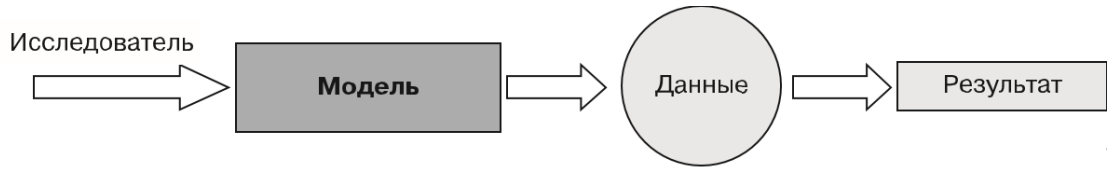


Рисунок 1 – Движение от модели к результату

Результат моделирования может соответствовать действительности, а может и не соответствовать. В последнем случае исследователю ничего не остается, кроме как выбрать другую модель или другой метод ее исследования. Новая модель, возможно, будет более адекватно описывать рассматриваемую систему.

При аналитическом подходе не модель «подстраивается» под действительность, а мы пытаемся подобрать существующую аналитическую модель таким образом, чтобы она адекватно отражала реальность.

Модель всегда исследуется каким-либо методом (численным, качественным и т. п.). Поэтому выбор метода моделирования часто означает выбор модели.

Информационный подход к моделированию

При использовании в бизнесе традиционного аналитического подхода неизбежно возникнут проблемы из-за несоответствия между методами анализа и реальностью, которую они призваны отражать. Существуют трудности, связанные с формализацией бизнес-процессов. Здесь факторы, определяющие явления, столь многообразны и многочисленны, их взаимосвязи так «переплетены», что почти никогда не удается создать модель, удовлетворяющую таким же условиям. Простое наложение известных аналитических методов, законов, зависимостей на изучаемую картину реальности не принесет успеха.

В сложности и слабой формализации бизнес-процессов главным образом «виноват» человеческий фактор, поэтому бывает трудно судить о характере закономерностей априори (а иногда и апостериори, после реализации какого-либо математического метода). С одинаковым успехом описывать эти закономерности могут различные модели. Использование разных методов для решения одной и той же задачи нередко приводит исследователя к противоположным выводам. Какой метод выбрать? Получить ответ на

подобный вопрос можно, лишь глубоко проанализировав как смысл решаемой задачи, так и свойство используемого математического аппарата.

Поэтому в последние годы получил распространение *информационный подход* к моделированию, ориентированный на использование данных. Его цель — освобождение аналитика от рутинных операций и возможных сложностей в понимании и применении современных математических методов.

При информационном подходе реальный объект рассматривается как «черный ящик», имеющий ряд входов и выходов, между которыми моделируются некоторые связи. Иными словами, известна только структура модели (например, нейронная сеть, линейная регрессия), а сами параметры модели «подстраиваются» под данные, которые описывают поведение объекта. Для корректировки параметров модели используется обратная связь — отклонение результата моделирования от действительности, а процесс настройки модели часто носит итеративный (то есть цикличный) характер, в соответствии с рисунком2.

Таким образом, *при информационном подходе отправной точкой являются данные, характеризующие исследуемый объект, и модель «подстраивается» под действительность*.

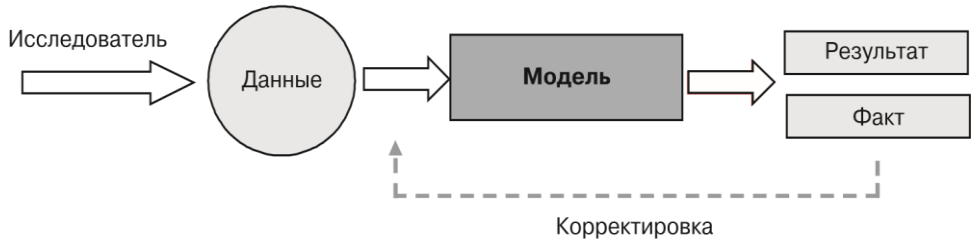


Рисунок 2 – Построение модели от данных

Если при аналитическом подходе мы можем выбрать модель, даже не имея никаких экспериментальных данных, характеризующих свойства системы, и начать ее использовать, то при информационном подходе без данных невозможно построить модель, так как ее параметры полностью определяются ими.

Модели, полученные с помощью информационного подхода, учитывают специфику моделируемого объекта, явления, в отличие от аналитического подхода. Для бизнес-процессов последнее качество очень важно, поэтому информационный подход лег в основу большинства современных промышленных технологий и методов анализа данных: Knowledge Discovery in Databases, Data Mining, машинного обучения.

Однако концепция «моделей от данных» требует тщательного подхода к качеству исходных данных, поскольку ошибочные, аномальные и зашумленные данные могут привести к моделям и выводам, не имеющим никакого отношения к действительности. Поэтому в информационном моделировании важную роль играют консолидация данных, их очистка и обогащение.

Модель, построенная на некотором множестве данных, описывающих реальный объект или систему, может оказаться не работающей на практике, поэтому в информационном моделировании используются специальные приемы: разделение данных на обучающее и тестовое множества, оценка обучающей и обобщающей способностей модели, проверка предсказательной силы модели.

В дальнейшем, говоря об анализе данных, мы будем предполагать использование именно информационного подхода. Поскольку данные могут быть представлены в различной форме, круг рассмотрения будет ограничен областью структурированных данных. Инструментальной поддержкой процесса построения моделей на основе информационного подхода выступают современные технологии анализа данных KDD и Data Mining, а средством построения прикладных решений в области анализа — аналитические платформы.

Процесс анализа

В информационном подходе к анализу данных, помимо модели, присутствуют еще три важные составляющие: эксперт, гипотеза и аналитик.

Эксперт — специалист в предметной области, профессионал, который за годы обучения и практической деятельности научился эффективно решать задачи, относящиеся к конкретной предметной области.

Эксперт — ключевая фигура в процессе анализа. По-настоящему эффективные аналитические решения можно получить не на основе одних лишь компьютерных программ, а в результате сочетания лучшего из того, что могут человек и компьютер. Эксперт выдвигает гипотезы (предположения) и для проверки их достоверности либо просматривает некие выборки различными способами, либо строит те или иные модели.

В крупных проектах по созданию прикладных аналитических решений участвуют, как правило, несколько экспертов, а также аналитик.

Аналитик — специалист в области анализа и моделирования. Аналитик на достаточном уровне владеет какими-либо инструментальными и программными средствами анализа данных, например, методами Data Mining. Кроме того, в обязанности аналитика входят функции систематизации данных, опроса мнений экспертов, координации действий всех участников проекта по анализу данных.

Аналитик играет роль «мостика» между экспертами, то есть является связующим звеном между специалистами разных уровней и областей. Он собирает у экспертов различные гипотезы, выдвигает требования к данным, проверяет гипотезы и вместе с экспертами анализирует полученные результаты. Аналитик должен обладать системными знаниями, так как помимо задач анализа на его плечи часто ложатся технические вопросы, связанные с базами данных, интеграцией с источниками данных и производительностью.

Поэтому в дальнейшем главным лицом в анализе данных мы будем считать аналитика, предполагая, что он тесно сотрудничает с экспертами предметных областей.

Несмотря на то что существует множество аналитических задач, можно выделить две основные группы методов их решения, в соответствии с рисунок 3.

– извлечение и визуализация данных;
– построение и использование моделей.

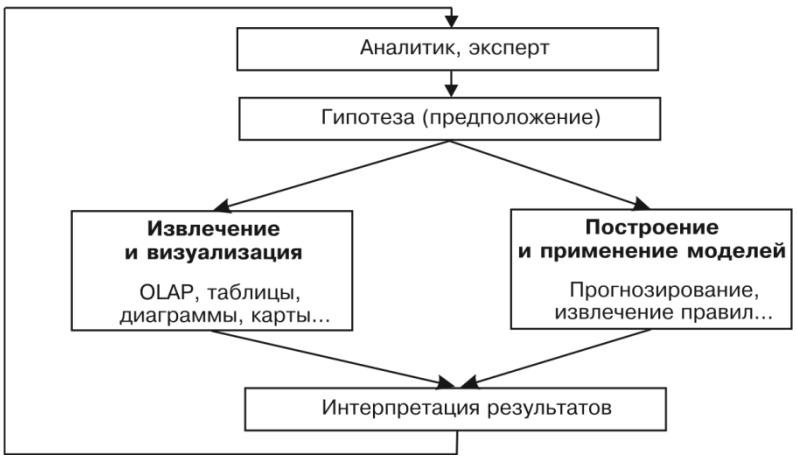


Рисунок 3 – Общая схема анализа

Экспериментальные методы широко используются как в науке, так и в промышленности, однако нередко с весьма различными целями. Обычно основная цель научного исследования состоит в том, чтобы показать статистическую значимость эффекта воздействия определенного фактора на изучаемую зависимую переменную. В условиях промышленного эксперимента основная цель обычно заключается в извлечении максимального количества объективной информации о влиянии изучаемых факторов на производственный процесс с помощью наименьшего числа дорогостоящих наблюдений.

Если в научных приложениях методы дисперсионного анализа используются для выяснения реальной природы взаимодействий, проявляющейся во взаимодействии факторов высших порядков, то в промышленности учет эффектов взаимодействия факторов часто считается излишним в ходе выявления существенно влияющих факторов.

Указанное отличие приводит к существенному различию методов, применяемых в науке и промышленности. Если просмотреть классические учебники по дисперсионному анализу, то обнаружится, что в них, в основном, обсуждаются планы с количеством факторов не более пяти (планы же с более чем шестью факторами обычно оказываются бесполезными). Основное внимание в данных рассуждениях сосредоточено на выборе общезначимых и устойчивых критериев значимости. Однако если обратиться к стандартным учебникам по экспериментам в промышленности, то окажется, что в них обсуждаются, в основном, многофакторные планы (например, с 16-ю или 32-мя факторами), в которых нельзя оценить эффекты взаимодействия, и основное внимание сосредоточивается на том получении несмещенных оценок главных эффектов (или, реже, взаимодействий второго порядка) с использованием наименьшего числа наблюдений.

Возможные подходы к статистическому анализу данных

Развитие теории и практики статистической обработки данных шло в двух параллельных направлениях. Первое включает методы математической статистики, предусматривающие возможность классической вероятностной интерпретации анализируемых данных и полученных статистических выводов (вероятностный подход). Второе направление содержит статистические методы, которые априори не опираются на вероятностную природу обрабатываемых данных, т.е. остаются за рамками научной дисциплины «математическая статистика» (логико-алгебраический подход). Ко второму подходу исследователь вынужден обращаться лишь тогда, когда условия сбора исходных данных не укладываются в рамки статистического ансамбля, т.е. в ситуации, когда не имеется практической или хотя бы принципиально мысленно представимой возможности многократного тождественного воспроизведения основного комплекса условий, при которых производились измерения анализируемых данных.

Типы реальных ситуаций

Выделяют три типа реальных ситуаций: с высокой работоспособностью вероятностно-статистических методов; с допустимостью вероятностно-статистических приложений (при этом нарушатся требования сохранения неизменными условия эксперимента); с недопустимостью вероятностно-статистических приложений (в этом случае идея многократного повторения одного и того же эксперимента в неизменных условиях является бессодержательной).

Примеры подходов к статистическому анализу данных

ПримерИсследуется массовое производство. Контролируется брак на изделиях. Результаты фиксируются в выборке:

$$x_1,\dots,x_N \qquad (1.1)$$

где $X_i=1$, если изделие дефектно, а иначе – $X_i=0$. Если производство отлажено и действует в стационарном режиме, то ряд наблюдений (1.1) естественно интерпретировать как ограниченную выборку из соответствующей бесконечной (генеральной) совокупности, которую бы мы имели, если бы осуществляли сплошной контроль изделий. В подобных ситуациях имеется принципиальная возможность многократного повторения наблюдения в рамках одинаковых условий. Такие ситуации могут быть описаны вероятностными моделями. Ряд (1.1) интерпретируется как случайная выборка из генеральной совокупности, т.е. как экспериментальные значения анализируемой случайной величины. Заметим, что в

теории вероятностей под случайным явлением понимают явление, относящееся к классу повторяемых, обладающих свойством статистической устойчивости при повторении однородных опытов. Здесь для статистической обработки применяются классические математико-статистические методы. Если основные свойства и характеристики генеральной совокупности не известны исследователю, то они оцениваются по соответствующим свойствам и характеристикам выборок с помощью этих методов.

ПримерИсследуется совокупность средних городов России (с численностью [100; 500] тысяч человек) для выяснения типов городов, сходных или однородных по структуре уровня образования жителей, половозрастному составу и характеру занятости. Подробный анализ большого числа городов практически не реален, поэтому в фиксированном пространстве небольшого числа интегральных параметров города разделяются на типы, выделяются эталоны, а для них проводят подробный анализ с целью выявления наиболее характерных черт и закономерностей в социально-экономическом облике средних по величине типичных городов.

Так для N средних городов (например, для России их оказалось 74) $X_1 \ldots X_N$ были зарегистрированы 32 параметра

$$\begin{pmatrix} X_1 \\ \cdots \\ X_n \end{pmatrix} = \begin{bmatrix} x_1^{(1)} & \ldots & x_1^{(32)} \\ \ldots & \ldots & \ldots \\ x_n^{(1)} & \ldots & x_n^{(32)} \end{bmatrix},$$

где $x^{(i)}$ - параметры, характеризующие среднее число жителей, приходящихся на 1000 человек населения города. Причем x^i, i=1…4 – параметры, характеризующие уровень образования (высшее, незаконченное высшее, среднее специальное, среднее); x^i, i=5…16 – 12 параметров, характеризующих половозрастной состав; x^i, i=17…21 – 5 параметров для описания социального характера занятости населения; x^i, i=22…32 – параметры, характеризующие занятость в материальном или нематериальном производстве и источники доходов.

Если допустить, что геометрическая близость двух точек – городов X_i и X_j в соответствующем 32-мерном пространстве означает их однородность (сходство) по анализируемым признакам и является основанием для их отнесения к одному типу, то для решения задачи надо привлечь методы кластер-анализа и снижения размерности. Математический аппарат этих методов предполагает вычисление средних, дисперсий, ковариаций, но эти характеристики описывают уже природу и структуру только реально анализируемых данных, т.е. статистически обследованную совокупность из n анализируемых городов.

Сравнение подходов к статистическому анализу данных

Основные отличительные особенности подходов на примере задачи классификации представим схематично в таблице 1.

Таблица 1–Отличительные особенности подходов

Составляющие	Первое направление	Второе направление
Цели исследования	Выделение классов, как инвариантов в потоке выборочных объектов	Выяснение распределения данных в системе
Объекты и признаки.	Независимы	Зависимость предполагается, ее нужно обнаружить
Выделяемые классы	Характеризуются эталоном и не пересекаются	Четко не выделяются, т.е. пересекаются
Аппарат исследования	Вероятностный - преобразование пространства признаков (даже в одномерную ось)	Логико-комбинаторный

Первое направление развития анализа данных, ориентированное на технические области знания, отстаивает идею простоты используемых моделей. В рамках этого направления неудовлетворительные результаты объясняют отсутствием информативных признаков.

Второе направление развития анализа данных ориентировано на социально-экономическую и социологическую информацию. При ее обработке появилось много новых идей, в частности, идея поэтапной группировки и коллектива решающих правил. Разработаны методы многомерного шкалирования, экспертных оценок.

В отличие от первого примера во втором примере невозможно: интерпретировать исходные данные в качестве случайной выборки генеральной совокупности (в связи с неприятием главной идеи понятия статистического ансамбля: идея многократного повторения одного и того же эксперимента в неизменных условиях теряет смысл); использовать вероятностную модель для построения и выбора наилучших методов статистической обработки; дать вероятностную интерпретацию выводам, основанным на статистическом анализе исходных данных.

Но в обоих случаях выбор наилучшего из всех возможных методов обработки данных производится в соответствии с некоторыми функционалами качества метода. Способ обоснования выбора этого функционала, а также его интерпретация различны. В первом случае выбор основан на допущении о вероятностной природе исходных данных и интерпретация тоже. Во втором случае исследователь не пользуется априорными сведениями о вероятностной природе исходных данных и при обосновании выбора оптимального критерия качества опирается на соображения содержательного (физического) плана - как именно и для чего получены данные. Когда критерий выбран, в обоих случаях используются методы решения экстремальных задач. На этапе осмысления и интерпретации каждый из подходов имеет свою специфику.

При выборе типа модели следует понимать, что всякая модель является упрощенным (математическим) представлением изучаемой действительности. Мера адекватности модели и действительности является решающим фактором работоспособности используемых затем методов обработки. А так как ни одна модель не может идеально соответствовать реальной ситуации, то желательна многократная обработка исходных данных для разных вариантов модели.

При решении задач анализа данных выделяется два этапа: исследовательский и эксплуатационный. Первый – этап выработки и проверки статистических гипотез. Второй – использование полученных гипотез для построения вывода. Поэтому с целью оптимизации и внедрения диалоговых систем необходим исследовательский этап, когда пользователь получает инструментальные психологические навыки и знания о системе. Как показывает практика и анализ публикаций, наибольший эффект в результате внедрения диалоговых систем (поддержки пользователя) достигается при выборе одного из двух требований:

- пользователь является профессионалом, диалоговые средства – рабочим инструментом;
- диалоговые средства ориентированы на поддержание простейших форм диалога и достижение простых ясно формализуемых целей.

Случайные события и случайные величины, их числовые характеристики

Все процессы, происходящие в природе, являются результатом взаимодействия многих факторов. Для того чтобы изучить эти процессы и в дальнейшем ими управлять, необходимо выяснить, какую роль в рассматриваемом процессе играет каждый фактор в отдельности. Все эти факторы необходимо выразить в каких–либо количественных оценках. Для этого используют математические методы, и чтобы получить необходимые числовые данные, нужно произвести серию наблюдений.

Однако даже самый тщательно подготовленный эксперимент не позволяет выделить

интересующий нас фактор в чистом виде. Мы не в силах изолировать многие посторонние факторы: изучая химические реакции, мы никогда не имеем дела с чистыми веществами; изучая электронные процессы, не можем вести их в абсолютном вакууме и т.д. Наконец, нужно вспомнить о различных помехах, связанных с окружающей обстановкой – ведь даже шум идущего по улице автомобиля сказывается на проводимом в лаборатории эксперименте.

Следовательно, каждое наблюдение дает нам лишь результат взаимодействия основного изучаемого фактора с многочисленными посторонними. Некоторые их этих факторов можно учесть, так как они сами по себе достаточно изучены. Учет других факторов (например, наличие примесей в веществах) может быть очень громоздким. Он сильно затягивает эксперимент, делает его неоправданно дорогим. Наконец, многие факторы (помехи) бывают настолько неожиданными, что их вообще нельзя учесть. Сюда следует отнести и те факторы, о которых на данном этапе развития науки вообще ничего не известно.

Вывод: полное и точное описание какого–либо процесса возможно лишь в том случае, если известны все факторы, влияющие на этот процесс. Иными словами, такое описание вообще невозможно.

К счастью, оно и не нужно!

Так все применяемые в эксперименте измерительные приборы обладают некоторым пределом точности – минимальной разницей в значениях двух величин, которую они в состоянии обнаружить. Этот предел обычно указывается на приборах, изготовленных в заводских условиях. Например, аналитические весы, взвешивающие с точностью до 0,1 мг, не смогут различить такие веса, как 12,52 и 12,54 мг, и в обоих случаях покажут 12,5 мг. В результате все дальнейшие вычисления, связанные с этими данными, также будут содержать некоторую неточность, даже если пользоваться абсолютно точными и полными формулами, описывающими исследуемый процесс.

Таким образом, в наших наблюдениях всегда допускается некоторая «законная» неточность, величину которой можно рассчитать заранее. Благодаря этому мы можем не учитывать те посторонние факторы, действие которых намного меньше этой неточности. По этой же причине удается избежать детального исследования многочисленных непредвиденных (случайных) помех. Хотя действие каждой из них может оказаться вполне заметным, как правило, уравновешивают друг друга, лишь изредка давая заметный суммарный эффект.

Установлено, что случайные, непредвиденные события подчиняются некоторым общим неслучайным закономерностям.

К категории случайных явлений можно отнести все те явления, точное предсказание протекания которых в каждом отдельном случае оказывается невозможным. Однако, если вместо того, чтобы рассматривать каждое из случайных явлений в отдельности, мы обратимся к совокупности большого их числа, то окажется, что средние результаты обнаруживают своего рода устойчивость.

В теории вероятностей рассматриваются три класса случайных явлений: случайные события; случайные величины; случайные функции.

Случайные события

Событие, которое при заданном комплексе факторов может либо произойти, либо не произойти, называется **случайным событием.**

Примеры: 1. Выпадение герба при бросании монеты.

2. Попадание в цель при выстреле.

Различные события мы будем обозначать буквами *A, B*, …

Событие называют **достоверным**, если оно непременно должно произойти.

Событие называют **невозможным**, если оно заведомо не наступит. Пример: ток в разомкнутой цепи.

Пусть A – некоторое событие. Под событием, противоположным ему будем понимать событие, состоящее в том, что A не наступило, обозначим его $\overline{A}$. События A и B

называют**несовместными**, если наступление одного из них исключает возможность наступления другого. Например: появление четного и нечетного чисел одновременно при бросании кости невозможно.

Событие называют **единственно возможным**, если в результате каждого испытания, хотя бы одно из них,наверное,наступит. Эти события образуют полную группу событий.

Вероятность события

Статистическое определение вероятности

Допустим, что имеется возможность неограниченного повторения испытаний, в каждом из которых при сохранении неизменных условий отмечается появление или не появление некоторого события A. Пусть при достаточно большом числе n испытаний интересующее нас событие A наступило m раз. Отношение $\mu = \frac{m}{n}$ – называется **частостью** (частотой) события A. В ряде случаев при очень большом числе испытаний эта частота сохраняет почти постоянную величину, причем колебания ее становятся тем меньше, чем больше число испытаний.

Вероятностью события называют характеризующее его число, около которого колеблется частота появления события при сохранении неизменных условий опыта.

Классическое определение вероятности

Имеется система конечного числа событий $A_1, A_2, \ldots, A_n$.

1) Эти события попарно несовместны, т.е. для любых двух событий появление одного исключает появление другого.

2) События $A_1, A_2, \ldots, A_n$ – единственно возможны.

3) События равновозможны, т.е. не существует никаких объективных причин, вследствие которых одно из них могло бы наступать чаще, чем какое–либо другое.

Пусть имеется событие A, которое наступает при появлении некоторых из наших «элементарных» событий $A_1, A_2, \ldots, A_m$ и не наступает при появлении других. Будем говорить в таком случае, что те из «элементарных» событий A_i, при наступлении которых наступает также событие A, благоприятствуют событию A.

Допустим, что из общего числа n рассматриваемых событий $A_1, A_2, \ldots, A_n$ событию A благоприятствует m из них. Тогда **вероятностью события *A*** называется отношение числа событий, благоприятствующих событию A, к общему числу всех равновозможных событий.

Если $P(A)$ – вероятность A, то $P(A) = \frac{m}{n}$.

Пример: бросание игральной кости. (A_1 – выпадение единицы, A_2 – выпадение двойки и т.д.).

$P(A_1) = P(A_2) = \ldots = P(A_6) = \frac{1}{6}$, так как $m = 1$, $n = 6$.

Если событие A означает появление четного числа очков, то ему благоприятствуют события A_2, A_4, A_6, состоящие в появлении 2–х, 4–х, 6–и очков. Таким образом, для события A: m = 3 и $P(A) = \frac{3}{6} = \frac{1}{2}$.

Свойства вероятности

1. Вероятность любого события A подчинена условиям

$0 \le P(A) \le 1$, так как $0 \le m \le n$.

2. Вероятность достоверного события E:

$P(E) = 1$, так как $m = n$.

3. Вероятность невозможного события U:

$P(U) = 0$, так как $m = 0$.

Сложение и умножение вероятностей

Пусть A и B – два несовместных события. Тогда вероятность того, что осуществится хотя бы одно из этих двух событий, равна сумме их вероятностей, т.е. $P(A$ или $B)=P(A)+P(B)$ – это так называемая теорема сложения.

Следствие теоремы: если события A_1, A_2, …, A_n несовместны и единственно возможны, то

$$P(A_1)+P(A_2)+\ldots+P(A_n)=1.$$

То есть они образуют полную группу событий. В частности $P(A)+P(\overline{A})=1$.

Условная вероятность. Умножение вероятностей

ПримерНа складе – 400 электрических лампочек, изготовленных на двух заводах (75 % – на первом, 25 % – на втором). Пусть на 1 заводе – 83 % соответствует стандарту, на 2 заводе – 63 %. Определим вероятность того, что случайно взятая со склада лампочка соответствует требованиям стандарта.

Решение. Всего на 1–м заводе изготавливают $400\cdot 0{,}75\cdot 0{,}83=249$ стандартных лампочек. На втором заводе – $400\cdot 0{,}25\cdot 0{,}63=63$. Т.е. всего 249+63=312 стандартных лампочек. Так как выбор любой лампочки следует считать равновозможным, то имеем 312 благоприятствующих случаев из 400, и поэтому

$$P(B)=\frac{312}{400}=0{,}78,$$

где событие B состоит в том, что выбранная на любом из заводов лампочка стандартна.

Но если известно, что выбранная лампочка изготовлена на первом заводе (событие A_1), то вероятность того, что она стандартна, будет не 0,78, а 0,83.

Такого рода вероятность, т.е. вероятность события B при условии, что имеет место событие A, называют **условной вероятностью** события B при условии наступления события A и обозначается $P_A(B)$.

Вероятность совмещения событий A и B равна произведению вероятности одного из событий на условную вероятность другого в предположении, что первое имело место

$$P(A \text{ и } B) = P(A)P_A(B).$$

Это так называемая теорема умножения.

Совмещение событий A и B – это наступление каждого из них, т.е. наступление как события A, так и события B.

Так как события A и B равноправны, то поменяв их местами:

$$P(A \text{ и } B) = P(A)P_B(A) \text{ или}$$
$$P(A)P_A(B) = P(B)P_B(A).$$

Пример В продукции предприятия признаются годными (событие A) 96 % изделий. К первому сорту (событие B) оказываются принадлежащими 75 изделий из каждой сотни годных. Определить вероятность того, что произвольно взятое изделие будет годным и принадлежит к первому сорту.

Искомая вероятность есть вероятность совмещения событий A и B.Поусловию:$P(A) = 0{,}96$,$P_A(B) = 0{,}75$,тогда $P(A \text{ и } B) = 0{,}96\cdot 0{,}75=0{,}72$.

Два события называются **независимыми**, если вероятность одного из них не изменяется в результате того, наступило или не наступило другое.

Пример: повторное бросание монеты ($P = 1/2$ независимо от того, выпал или не выпал герб в первом случае). Условие независимости событий A и B:

$$P_A(B) = P_{\overline{A}}(B) = P(B) \text{ или}$$
$$P_B(A) = P_{\overline{B}}(A) = P(A).$$

Для независимых событий теорема умножения формулируется следующим образом. Если события A и B независимы, то вероятность их совмещения равна произведению вероятностей этих событий $P(A \text{ и } B) = P(A)P(B)$.

Полная вероятность

Пусть события $H_1, H_2, \ldots, H_n$ образуют полную группу событий и при наступлении каждого из них (H_i) событие A может наступить с некоторой условной вероятностью $P_{H_i}(A)$. Какова будет при этом вероятность наступления события A?

В соответствии с теорией умножения найдем, что вероятность наступления A при условии $H_1, H_2, \ldots, H_n$:

$$P(H_1 \text{ и } A) = P(H_1)\, P_{H_1}(A)$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots..$$
$$P(H_n \text{и } A) = P(H_n)\ P_{H_n}(A).$$

По теореме сложения (события $H_1, H_2, \ldots$ несовместны):

$$P(A) = P(H_1)P_{H_1}(A) + \ldots + P(H_n)P_{H_n}(A) \text{ или}$$

$P(A) = \sum_{i=1}^{n} P(H_i)P_{H_i}(A)$ – это так называемая формула полной вероятности.

ПримерПри разрыве снаряда образуются крупные, средние и мелкие осколки, причем число крупных осколков составит 0,1 их общего числа; средних – 0,3; мелких – 0,6.

При попадании в танк крупный осколок прибивает броню с вероятностью 0,9; средний – 0,3; мелкий – 0,1.

Какова вероятность того, что попавший в броню осколок пробьет ее?

$P(H_1) = 0{,}1$ $\quad P_{H_1}(A) = 0{,}9$
$P(H_2) = 0{,}3$ $\quad P_{H_2}(A) = 0{,}3$
$P(H_3) = 0{,}6$ $\quad P_{H_3}(A) = 0{,}1$

$$P(A) = 0{,}1 \cdot 0{,}9 + 0{,}3 \cdot 0{,}3 + 0{,}6 \cdot 0{,}1 = 0{,}24.$$

Случайные величины

Под **случайной величиной** понимается величина, принимающая в результате опыта какое–либо числовое значение из множества возможных значений.

Примеры:

1) Число выстрелов, произведенных до первого попадания в цель (любое целое положительное число).

2) Расстояние от центра мишени до точки попадания (любое положительное число или 0).

Случайная величина, принимающая конечное число или последовательность различных значений, называется **дискретной случайной величиной** (пример 1).

Случайная величина, принимающая все значения из некоторого интервала, называется **непрерывной случайной величиной** (пример 2).

Чтобы охарактеризовать дискретную случайную величину, прежде всего необходимо указать возможные ее значения. Однако этого недостаточно: нужно еще знать, насколько часто принимаются различные значения этой величины, что лучше всего характеризовать вероятностью отдельных ее значений. Иначе говоря, для случайной величины X следует указывать не только ее значения x_1, x_2,…, x_n, но и вероятности событий $X= x_i$, $p_i= P(X= x_i)$, (i= 1,2,…,n) состоящих в том, что случайная величина X приняла значение x_i.

Если перечислены все возможные значения X, то события $X = x_i$ не только несовместны, но и единственно возможны, так что сумма заданных вероятностей p_i должна равняться единице.

Соотношение, устанавливающее связь между значениями случайной величины и вероятностями этих значений, называется **законом распределения случайной величины**.

Для дискретной случайной величины закон распределения удобно записывать в виде таблицы, причем $\sum_i p_i = 1$.

x_1	x_2	…	x_i	…	x_n
p_1	p_2	…	p_i	…	p_n

Иногда этот закон задают в виде графика: по оси абсцисс откладывают возможные значения случайной величины X, а по оси ординат – соответствующие значения вероятностей. Получаемая при этом ломаная линия называется многоугольником распределения.

Пример: X – число очков, выпадающих на игральной кости. Возможные значения 1,…,6 равновероятны.

Закон распределения:

x_i	1	2	3	4	5	6
p_i	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Многоугольник распределения, рисунок 4:

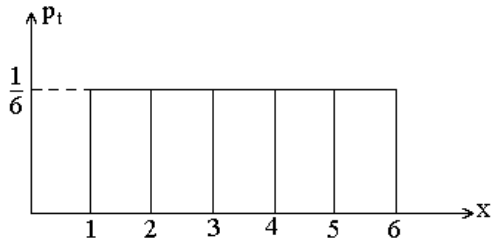


Рисунок 4 – Многоугольник распределения

Функции распределения случайной величины

Закон распределения не всегда можно задать таблицей. Например, для непрерывной случайной величины невозможно перечислить все ее значения. Поэтому ее характеризуют не вероятностями отдельных значений, как дискретную, а вероятностями того, что случайная величина принимает значения из определенного интервала, т.е. вероятностями неравенств вида $-\alpha < X < \beta$. (Можно и для дискретных).

Обычно говорят о вероятности неравенства $-\infty < X < x$, т.е. вероятности того, что случайная величина принимает значение, меньшее x. Эта вероятность $P(X<x)$ является, очевидно, функцией x. Обозначим ее $F(x)$

$$F(x) = P(X < x).$$

Функцию $F(x)$ называют **интегральным законом распределения или функцией распределения случайной величины *X***.

Свойства функции распределения

Пусть X – случайная величина, x_1 и x_2 – две произвольные точки, причем $x_1<x_2$. Сравним значения функции $F(x)$ в этих точках. Так как событие $X<x_1$ влечет событие $X<x_2$, то ясно, что

$$P(X < x_1) \le P(X < x_2), \text{ или}$$

по определению функции распределения,

$$F(x_1) \le F(x_2).$$

Таким образом, функция распределения для любой случайной величины всегда является монотонно неубывающей.

Очевидно $F(-\infty) = 0, \quad F(+\infty) = 1$, откуда следует, что $F(+\infty) - F(-\infty) = 1$. Так как$F(x)$ – монотонна и заключена между 0 и 1, то график функции $y= F(x)$ имеет две горизонтальные асимптоты: $y = 0$ при $x \to -\infty$ и $y = 1$ при $x \to +\infty$.

Если все значения случайной величиныпринадлежат интервалу (a, b), то слева от точки «a» имеем $F(x) = 0$, а справа от «b» – функция $F(x) = 1$, рисунок 5.

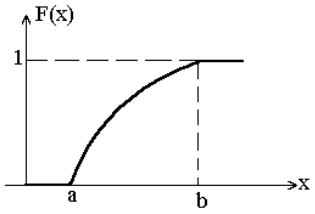


Рисунок 5 – Функция распределения случайной величины X

Функция распределения для дискретной случайной величины

Функция распределения для дискретной случайной величины

$$F(x) = P(X < x) = P(-\infty < X < x) = \sum_{x_i < x} p_i ,$$

где суммирование распространяется на значения x_i, удовлетворяющие неравенству $x_i<x$. В промежутке между двумя последовательными значениями X функция $F(x)$ постоянна. При переходе же аргумента x через возможное значение случайной величины x_i функция $F(x)$ скачком возрастает на величину $p_i= P(X = x_i)$, так что x_iбудет точкой разрыва первого рода функции $F(x)$, поэтому функция распределения для дискретной случайной величины будет ступенчатой функцией, рисунок 6.

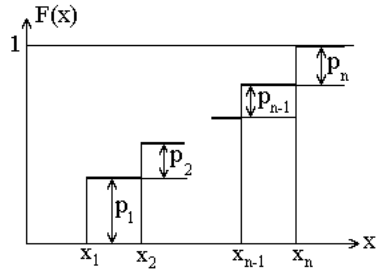


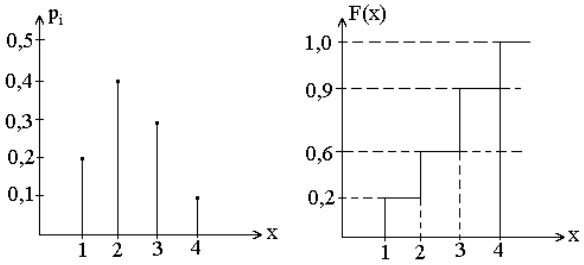
Рисунок 6 - Функция распределения для дискретной случайной величины

Т.о., для конечно–значной величины функция распределения $F(x)$ равна сумме вероятностей всех допустимых значений, меньших, чем x.

Пример. Смешаны шарики (100) различной крупности, среди которых 20 шариков диаметром 1 см; 40 шариков – 2 см; 30 шариков – 3 см и 10 шариков – 4 см.

Определить вероятности извлечения шарика определенной крупности и построить F(x), если x – крупность шарика.

Ответ:



Функция распределения непрерывной случайнойвеличины

Пусть X – непрерывная случайная величина с известной функцией распределения $F(x)$.

Найдемвероятноститого,чтоXзаключенавпределах$\alpha<X<\beta$. Пользуясь теоремой сложения вероятностей:

$$P(X<\beta)=P(X<\alpha)+P(\alpha<X<\beta),$$

откуда

$$P(\alpha<X<\beta)=P(X<\beta)-\mathbf{P(X<\alpha)}.$$

Исходя из определения функции распределения:

$$P(\alpha<X<\beta)=F(\beta)-F(\alpha).$$

Определим вероятность $P(X=\alpha)$:

$$P(X=\alpha)=\lim_{\beta\to\alpha}P(\alpha<X<\beta)=\lim_{\beta\to\alpha}[F(\beta)-F(\alpha)].$$

Если функция $F(x)$ непрерывна, то последний предел равен нулю: $P(X=\alpha)=0$.

Т.е., если функция распределения случайной величины непрерывна, то вероятность того, что случайная величина примет заранее заданное значение, равна нулю, но это событие не является невозможным.

Плотность вероятности случайной величины

Предположим, что X – непрерывная случайная величина, ее функция распределения непрерывна и дифференцируема.

Производная функции распределения $\varphi(x)=F'(x)$ называется дифференциальным законом распределения или плотностью вероятности случайной величины X.

Ее вероятностный смысл:

$$\varphi(x) = F'(x) = \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x},$$

где числитель $F(x+\Delta x) - F(x) = P(x < X < x+\Delta x)$,

таким образом

$$\varphi(x) = \lim_{\Delta x \to 0} \frac{P(x < X < x+\Delta x)}{\Delta x}, \quad (1.2)$$

т.е. плотность вероятности случайной величины X в точке x равна пределу отношения вероятности попадания величины X в интервал $(x, x+\Delta x)$ к Δx, когда Δx стремится к нулю.

Зная $\varphi(x)$, можно вычислить вероятность $\alpha < X < \beta$

$$P(\alpha < X < \beta) = F(\beta) - F(\alpha) = \int_{\alpha}^{\beta} F'(x)dx = \int_{\alpha}^{\beta} \varphi(x)dx. \quad (1.3)$$

Геометрическое истолкование этой формулы представлен на рисунке 7.

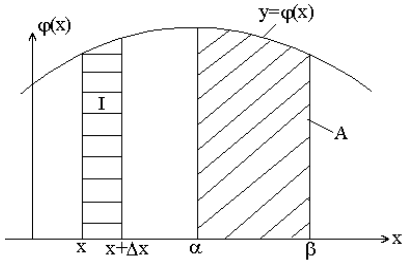


Рисунок 7 – Плотность вероятности случайной величины

Из (1.2) вытекает, что с точностью до бесконечно малых величин высшего порядка по сравнению с Δx справедливо равенство:

$$P(x < X < x+\Delta x) \approx \varphi(x)\Delta x.$$

Это произведение геометрически изображается площадью прямоугольника I.

Вероятность того, что случайная величина примет значения внутри участка (α, β), выражается площадью криволинейной трапеции A – в этом состоит геометрический смысл выражения (1.3).

Известно, что $\varphi(x) \geq 0$, так как $F(x)$ – монотонно неубывающая функция, а из формулы следует (1.2) $\int_{-\infty}^{\infty} \varphi(x)dx = 1$.

Легко установить выражение функции распределения через плотность вероятности. $F(x)$ – есть первообразная от $\varphi(x)$, причем такая, которая обращается в нуль при $x=-\infty$. Поэтому

$$F(x)=\int_{-\infty}^{x}\varphi(t)dt\,.$$

Физическая интерпретация:

$F(x)$ – Представим единицу массы, распределенную вдоль прямой так, чтобы количество массы, сосредоточенное во всех точках прямой, для которых $X \le x$, т.е. находится слева от точки x, равнялось $F(x)$.

Тогда $\varphi(x)$ будет представлять собой плотность единичной массы в рассматриваемой точке.

Числовые характеристики случайных величин– это такие характеристики, которые одновременно связаны с функциями распределения, но которые могут быть найдены даже тогда, когда $F(x)$ или $\varphi(x)$ неизвестны, и которыми удобно пользоваться при решении прикладных задач. К числу таких характеристик относятся средние значения и моменты случайных величин.

Средние значения случайных величин

Предположим, что X – дискретная случайная величина, которая в результате эксперимента принимала значения x_1, x_2,…, x_nс вероятностями p_1, p_2,…, p_n, $\sum_{i=1}^{n} p_i = 1$. Тогда средним значением или математическим ожиданием величины X называется сумма $\tilde{X} = M[X] = \sum_{i=1}^{n} x_i p_i$, т.е. средневзвешенное значение величины X, где весами служат вероятности p_i.

Пример. Определить среднее значение ошибки регулирования ε, если на основании большого числа опытов установлено, что вероятность ошибки p_i равна:

ε, %	0,1	0,15	0,2	0,25	0,3
p_i	0,2	0,2	0,3	0,15	0,15

Решение:

$$M[\varepsilon] = 0{,}1\cdot 0{,}2 + 0{,}15\cdot 0{,}2 + 0{,}2\cdot 0{,}3 + 0{,}25\cdot 0{,}15 + 0{,}3\cdot 0{,}15 = 0{,}19\ \%.$$

В том случае, если g(X) является функцией X (причем вероятность того, что $X = x_i$ равна p_i), то среднее значение функции определяется как

$$\tilde{g} = M[g(X)] = \sum_{i=1}^{n} g(x_i) p_i.$$

Предположим, что X – случайная величина с непрерывным распределением и характеризуется плотностью вероятности $\varphi(x)$. Тогда вероятность того, что X заключена между x и $x + \Delta x$:

$$P(x < X < x + \Delta x) \approx \varphi(x)\Delta x\,.$$

Величина X при этом приближенно принимает значение x. В пределе при $\Delta x \to 0$, можно предположить, что приращение Δx численно равно дифференциалу dx.

Произведя замену Δx = dx, получаем точную формулу для расчета среднего значения X:

$$\tilde{X} = M[X] = \int_{-\infty}^{\infty} x\varphi(x)dx = \int_{-\infty}^{\infty} xdF(x).$$

Аналогично для g(X): $\tilde{g} = M[g(X)] = \int_{-\infty}^{\infty} g(x)\varphi(x)dx.$

Как правило, недостаточно бывает знать только среднее значение (математическое ожидание) случайной величины. Для оценки меры случайности величины (для оценки разброса конкретных значений X относительно математического ожидания $M[X]$) вводится понятие дисперсии случайной величины. Дисперсия – среднее значение квадрата отклонения каждого конкретного значения X от математического ожидания. Чем больше дисперсия $D[X]$, тем больше случайности разброса величины от математического ожидания. Если случайная величина дискретная, то

$$D[X] = M[(X - M[X])]^2 = \sum_{i=1}^{n} (x_i - M[X])^2 p_i.$$

Для непрерывной случайной величины дисперсию можно записать аналогично:

$$D[X] = \int_{-\infty}^{\infty} (x - M[X])^2 \varphi(x)dx.$$

Дисперсия $D[X]$ хорошо описывает разброс величины, но при этом есть один недостаток: размерность $D[X]$ не соответствует размерности X. Чтобы избавиться от этого недостатка, часто в конкретных приложениях рассматривают не $D[X]$, а положительное значение $\sqrt{D[X]} = \beta$, которое называется **средним квадратическим отклонением**.

Свойства математического ожидания

1. Математическое ожидание неслучайной величины равно самой этой величине $M[C] = C$.
2. Неслучайный множитель C можно выносить за знак математического ожидания $M[CX] = CM[X]$.
3. Математическое ожидание суммы случайных величин равно сумме математических ожиданий этих случайных величин.
4. $$M[X_1 + X_2] = M[X_1] + M[X_2].$$
5. Математическое ожидание произведения независимых случайных величин равно произведению математических ожиданий этих величин (условие независимости случайных величин).
6. $$M[X_1 \times X_2] = M[X_1] \times M[X_2].$$

Свойства дисперсии

1. Дисперсия неслучайной величины C равна нулю: $D[C]=0$.

2. Дисперсия произведения неслучайного множителя C на случайную величину равна произведению C^2 на дисперсию случайной величины.

$$D[CX] = C^2 D[X].$$

3. Дисперсия суммы независимых случайных величин X_1 и X_2 равна сумме дисперсий слагаемых $D[X_1 + X_2] = D[X_1] + D[X_2]$.

Моменты случайной величины

Пусть X – непрерывная случайная величина. Если ν – целое положительное число, а функция x^ν интегрируема на интервале $(-\infty; +\infty)$, то среднее значение

$$\alpha_\nu = M[X^\nu] = \int_{-\infty}^{\infty} x^\nu dF(x) = \int_{-\infty}^{\infty} x^\nu \varphi(x) dx, \nu = 0, 1, \ldots, n$$

называется **начальным моментом** порядка ν случайной величины X.

Очевидно, что момент нулевого порядка

$$\alpha_o = \int_{-\infty}^{\infty} x^o dF(x) = \int_{-\infty}^{\infty} dF(x) = \int_{-\infty}^{\infty} \varphi(x) dx = 1,$$

а начальный момент первого порядка

$$\alpha_1 = \int_{-\infty}^{\infty} x^1 dF(x) = \int_{-\infty}^{\infty} x\varphi(x) dx = M[X] = \tilde{X},$$

есть математическое ожидание самой случайной величины X.

Момент второго порядка

$$\alpha_2 = \int_{-\infty}^{\infty} x^2 dF(x) = \int_{-\infty}^{\infty} x^2 \varphi(x) dx$$

есть математическое ожидание квадрата случайной величины X.

Аналогично находят α_2, α_3 и т.д.

Если $X^o = X - M[X]$ – центрированная случайная величина, то представляет интерес рассмотрение центральных моментов порядка ν, где $\nu = 0, 1, \ldots, n$:

$$\beta_\nu = M\left[(X - M[X])^\nu\right] = \int_{-\infty}^{\infty} (X - M[X])^\nu dF(x) =$$

$$\int_{-\infty}^{\infty} (X - M[X])^\nu \varphi(x) dx$$

$$\beta_0 = \int_{-\infty}^{\infty} (X - M[X])^0 dF(x) = 1$$

$$\beta_1 = \int_{-\infty}^{\infty} (X - M[X])^1 dF(x) = \int_{-\infty}^{\infty} x dF(x) - \int_{-\infty}^{\infty} M[X] dF(x) = M[X] - M[X] \int_{-\infty}^{\infty} dF(x) = 0$$

$$\beta_2 = \int_{-\infty}^{\infty}(X - M[X])^2 dF(x) =$$

$$\int_{-\infty}^{\infty} x^2 dF(x) - 2M[X]\int_{-\infty}^{\infty} x dF(x) + (M[X])^2 \int_{-\infty}^{\infty} dF(x) = D[X]$$

Есть связь между начальными и центральными моментами.

Так $\beta_o = \alpha_o$

$\beta_2 = \alpha_2 - M^2[X]$

$\beta_3 = \alpha_3 - 3M[X]\alpha_2^2 + 2M^3[X]$ и т.п.

Примеры законов распределения случайной величины

Рассмотрим примеры распределения случайной величины.

ПримерРавномерное распределение дискретной случайной величины.

При бросании игральной кости может выпасть 1,2,3,… или 6, рисунок 8 . Здесь величина X принимает значения $x_i = i$ с вероятностями соответственно $p_i = \frac{1}{6}$ (i = 1, 2, 3…, 6). Ввиду равенства всех вероятностей можно говорить о равномерном распределении случайной величины X.

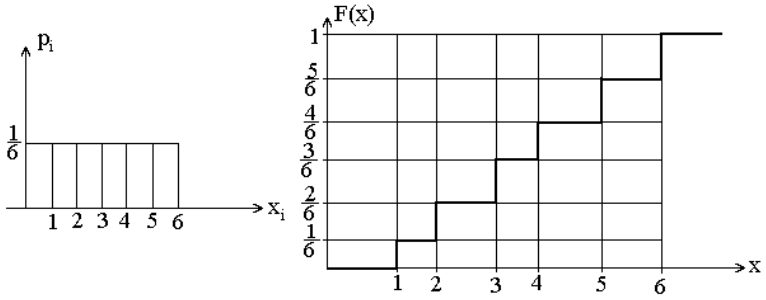


Рисунок 8 – Распределении случайной величины X

Рассчитаем для этой случайной величины математическое ожидание $M[X]$ и дисперсию $D[X]$:

$$M[X] = \sum_{1}^{6} x_i p_i = \frac{1}{6}\sum_{i=1}^{6} x_i = \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = 3{,}5$$

$$D[X] = \sum_{i=1}^{6}(x_i - M[X])^2 p_i = \sum_{1}^{6}\left(x_i^2 - 2x_i M[X] + M[X]^2\right)p_i =$$

$$\sum_{i=1}^{6} x_i{}^2 p_i - 2M[X]\sum_{i=1}^{6} x_i p_i + (M[X])^2 \sum_{i=1}^{6} p_i =$$

$$p_i \sum x_i^2 - 2M[X]M[X] + (M[X])^2 =$$

$$\frac{1}{6}\sum_{i=1}^{6} x_i^2 - (3{,}5)^2 = 15{,}17 - 12{,}25 = 2{,}92$$

При этом $\sigma = \sqrt{2{,}92} \approx 1{,}7$.

Пример Равномерное распределение непрерывной случайнойвеличины.

Предположим, что случайная величина имеет равномерное и непрерывное распределение. Причем ее плотность вероятности $\varphi(x) = 0$ для всех значений, кроме интервала (a, b), на котором она постоянна. Постоянное значение обозначим через A. Тогда можно записать

$$\int_{-\infty}^{\infty} \varphi(x)dx = 1; \int_{a}^{b} \varphi(x)dx = 1; \quad \int_{a}^{b} Adx = 1; \quad A\int_{a}^{b} dx = 1;$$

$$\text{откуда} \quad A(b-a) = 1;$$

или $A = \frac{1}{b-a}$. Поэтому плотность равномерного распределения задается формулой, в соответсвии:

$$\varphi(x) = \begin{cases} 0; & при \ -\infty < x < a \\ \frac{1}{b-a}; & при \ a < x < b \\ 0; & при \ b < x < \infty \end{cases}.$$

и соответственно в виде графика, рисунок 9:

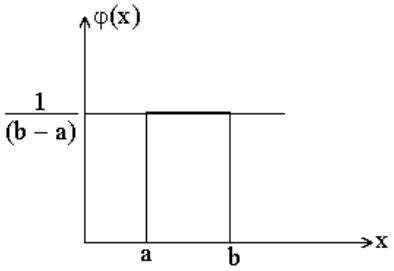


Рисунок 9 – Плотность равномерного распределения

Случайная величина называется непрерывной, если ее функция распределения $F(x)$ непрерывна на всей числовой оси, а плотность вероятности $\varphi(x)$ существует и непрерывна всюду, кроме дискретного множества точек. Для нахождения функции распределения $F(x)$ воспользуемся формулой

$$F(x) = \int_{-\infty}^{x} \varphi(t)dt.$$

При $x \le a$ $\varphi(x) = 0$. Тогда $F(x) = 0$.

Для $a < x < b$ получим

$$F(x)=\int_{-\infty}^{x}\varphi(t)dt=\int_{-\infty}^{a}\varphi(t)dt+\int_{a}^{x}\varphi(t)dt=0+\int_{a}^{b}\frac{1}{b-a}dt=\frac{x-a}{b-a}.$$

Наконец при $x \geq b$ получим:

$$F(x)=\int_{-\infty}^{a}\varphi(t)dt+\int_{a}^{b}\varphi(t)dt+\int_{b}^{x}\varphi(t)dt=0+\int_{a}^{b}\frac{1}{b-a}dt+0=\frac{b-a}{b-a}=1.$$

Таким образом, интегральный закон равномерного распределения случайной величины задается формулой

$$F(x)=\begin{cases}0, при -\infty < x < a;\\ \dfrac{x-a}{b-a}, при\ a < x < b;\\ 1, при\ b < x < \infty\end{cases}$$

и соответственно в виде графика, , рисунок 10:

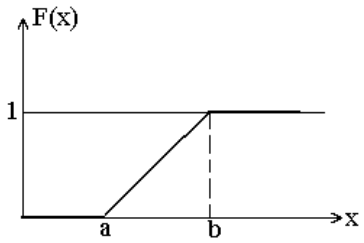


Рисунок 10 – Интегральный закон равномерного распределения случайной величины

Числовые характеристики непрерывной случайной величины

$$M[X]=\int_{-\infty}^{\infty}x\varphi(x)dx=\int_{a}^{b}\frac{x}{b-a}dx=\frac{1}{b-a}\frac{x^2}{2}\Bigg|_{a}^{b}=\frac{b^2-a^2}{2(b-a)}=\frac{b+a}{2}.$$

$$D[X]=\beta_2=\alpha_2-(M[X])^2;$$

$$\alpha_2=\int_{a}^{b}\frac{x^2}{b-a}dx=\frac{b^3-a^3}{3(b-a)}=\frac{a^2+ab+b^2}{3};$$

$$M[X]=\frac{b+a}{2};\ D[X]=\frac{a^2+ab+b^2}{3}-\frac{(b+a)^2}{4}=\frac{(b-a)^2}{12}.$$

$$\sigma_x=\sqrt{D[X]}=\frac{b-a}{2\sqrt{3}}.$$

Нормальный закон распределения (закон Гаусса)

Среди законов распределения, которым подчиняются встречающиеся на практике случайные величины, чаще всего приходиться иметь дело с нормальным законом

распределения. Это предельный закон, к которому приближаются многие другие законы распределения при определенных условиях. Если случайную величину можно рассматривать как результат суммарного воздействия многих независимых факторов, то закон распределения такой случайной величины будет близок к нормальному.

Для этого закона плотность вероятности задается формулой:

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

Выясним геометрический смысл параметров «a» и «σ» (a – математическое ожидание; σ^2 – дисперсия, σ – среднеквадратическое отклонение).

Из формулы видно, что кривая $y = \varphi(x)$ достигает максимума при $x = a$, причем максимальное значение $y_{\max}=\frac{1}{\sigma\sqrt{2\pi}}$. С ростом σ величина максимального значения уменьшается, а так как площадь, ограниченная всей кривой и осью абсцисс, равна единице, то с ростом σ кривая как бы растягивается вдоль оси ох и наоборот. Приведены графики $y = \varphi(x)$ при различных «a», но при одном и том же σ. На другом – при $a = 0$, но различных σ.

При $|x|\to\infty$ имеет место предел, когда $\varphi = 0$ (по формуле). Разность $(x - a)$ содержится в формуле в квадрате, т.е. график функции симметричен относительно прямой $x = a$.

Контрольные вопросы:

1. Что называется случайным событием? Дайте определения достоверного и невозможного событий.
2. Какие события называются: несовместными, равновозможными и противоположными?
3. Что называют пространством элементарных исходов?
4. Дайте определение суммы событий. Приведите примеры сумм событий.
5. Дайте определение произведения событий. Приведите примеры произведения двух событий.
6. Сформулируйте теоремы сложения вероятностей для совместных и несовместных событий.
7. Сформулируйте определение зависимых и независимых событий. Приведите формулы умножения вероятностей для зависимых и независимых событий.
8. Дайте определение условной вероятности. Сформулируйте теорему о полной вероятности.
9. Дайте определение случайной величины и закона её распределения. Перечислите типы случайных величин. Что называют рядом распределения дискретной случайной величины?
10. Дайте определение математического ожидания для дискретной и непрерывной случайных величин. Перечислите свойства математического ожидания.
11. Дайте определение плотности распределения случайной величины. Укажите основные свойства функции плотности распределения.
12. Как определяется система двух случайных величин (двумерная случайная величина). Как определяется закон распределения двумерной случайной величины.

2. Точечное оценивание параметров. Регрессионный и корреляционный анализ

Выборочный метод

Применение математической статистики к обработке наблюдений оказывается возможным благодаря использованию выборочного метода.

Выборочный метод в самой общей форме выглядит следующим образом. Имеется некоторая большая совокупность объектов, называемая **генеральной совокупностью**. Из этой совокупности извлекается n объектов, которые образуют выборку; число n называется **объемом выборки**. Эти n объектов подвергаются детальному исследованию, по результатам которого требуется описать всю генеральную совокупность или какие–нибудь ее свойства, характеристики.

Получается следующая схема производства наблюдений: имеется случайная величина X и в результате n независимых испытаний получаются n ее допустимых значений. Если все допустимые значения случайной величины X считать генеральной совокупностью, то полученные при наблюдениях n значений образуют выборку. По этой выборке мы и должны определить распределение случайной величины X (т.е. распределение генеральной совокупности).

При наблюдениях получают числа $x_1, x_2,\ldots, x_n$ (элементы выборки). Их можно считать полной совокупностью значений некоторой конечнозначной случайной величины X_n и конечное распределение X_n называют **выборочным** (эмпирическим) распределением.

Доказано: с вероятностью равной единице максимальная разность между функциями распределения случайных величин X_n и X при $n\to\infty$ стремится к нулю. Практически это означает, что при достаточно большом объеме выборки функцию распределения генеральной совокупности можно приближенно заменять выборочной функцией распределения.

Среднее и дисперсия выборки

Пусть $M[X]$ – математическое ожидание случайной величины X. Это число нам неизвестно. Мы проводим наблюдения и при большом объеме выборки n можно вместо $M[X]$ рассматривать математическое ожидание X_n. Погрешность при этом будет тем меньше, чем больше объем выборки n.

Математическое ожидание выборки есть просто среднее арифметическое элементов выборки:

$$M[X_n]=\frac{x_1+x_2+\ldots+x_n}{n}.$$

Будем называть средним выборки $\bar{x}_в$. Если сгруппировать итоги наблюдений, то можно записать

$$\bar{x}_в=\frac{\sum_{i=1}^{к} n_i x_i}{n},$$

где x_i – варианта выборки; n_i – частота варианты x_i; n – объем выборки.

Таким образом, в качестве истинного результата можно брать $\bar{x}_в$. Такой выбор вносит определенные погрешности, которые тем меньше, чем больше n.

Наиболее достоверной оценкой измеряемой случайной величины является ее *среднее арифметическое* или *среднее взвешанное значение*.

Среднее арифметическое значение определяется тогда, когда все варианты (значения случайной величины) имеют одну и ту же частоту, равную единице (нет одинаковых значений случайной величины), что характерно для малых выборок.

Если варианты имеют различные частоты, что характерно для больших выборок, то рассчитывают среднее взвешанное значение случайной величины по формуле:

$$\overline{x} = \frac{\overline{x}_1 m_1 + \overline{x}_2 m_2 + \ldots + \overline{x}_k m_k}{m_1 + m_2 + \ldots + m_k},$$

где –*x* – значение варианта (случайной величины) в середине *i*-го интервала вариационного ряда; m_i– частота (число вариантов случайной величины), соответствующая *i*-му интервалу; *k* – число интервалов разбиения.

Наряду со средним взвешенным значением случайной величины в качестве характеристик вариационного ряда, дающих информацию о законе распределения, используют *медиану и моду.*

Медиана($m_{0,5}$) – это значение случайной величины, которое делит вариационный ряд или площадь, ограниченную кривой распределения, на две равные части. При нечетном объеме выборки медиана равна

$$m_{0,5} = x_{\overline{m}},$$

а при четном объеме

$$m_{0,5} = \frac{(x_m + x_{m+1})}{2},$$

где x_m– значение средней по порядку вариационного ряда случайной величины. (Например, если в вариационном ряду 51 – значение случайной величины, то $m_{0,5}$ будет равна значению 26).

Модой *m*0 называют варианту, которая имеет наибольшую частоту, т. е. соответствует вершине распределения (это наиболее вероятное значение случайной величины).

Оценивают моду по формуле:

$$m_0 = h^{н}_{m_0} + h\frac{m_{m_0} - m_{m_0-1}}{2m_{m_0} - m_{m_0-1} - m_{m_0+1}},$$

где h^{H}_{m0} – нижняя граница модального интервала, т. е. интервала, имеющего наибольшую частоту;

h – длина интервала разбиения (шаг);

m_{m0}– частота модального интервала;

m_{m-1} – частота интервала, предшествующего модальному интервалу;

m_{m+1} – частота интервала, следующего за модальным интервалом.

Весьма важной характеристикой нормального распределения является *степень разброса* (рассеивания) отдельных частей случайной величины относительно ее среднего значения.

Для оценки степени разброса пользуются несколькими показателями, из которых наиболее широко распространены следующие:

- *размах* (*R*), представляющий собой разность между наибольшим (*x*max) и наименьшим (*x*min) значениями вариант;
- *дисперсия* (*D*) – это среднее арифметическое значение квадра$

тов отклонений отдельных вариант от их средней арифметической.

Дисперсия $D[X]$ приближенно равна дисперсии $D[X_n]$.

$$D[X_n] = \frac{(x_1 - \bar{x}_в)^2 + (x_2 - \bar{x}_в)^2 + ... + (x_n - \bar{x}_в)^2}{n},$$

т.е. $D[X] \approx D[X_n]$. Это равенство было бы еще более надежным, если бы в формуле для $D[X_n]$ вместо $\bar{x}_в$ стоял непосредственно истинный результат $M[X]$. Обычно получаем заниженную оценку рассеяния значения генеральной совокупности. В связи с этим $D[X_n]$ называется **смещенной** оценкой дисперсии $D[X]$. Чтобы получить несмещенную оценку дисперсии требуется рассмотреть величину

$$S^2 = \frac{n}{n-1} D[X_n],$$

которая является несмещенной оценкой дисперсии.

Переход к несмещенной оценке S^2 важен в основном для малых выборок, ибо разница между S^2 и $D[X_n]$ при больших n незаметна.

Таким образом, среднее выборки

$$\bar{x}_в = \frac{1}{n} \sum_{i=1}^{n} x_i ,$$

а несмещенная оценка дисперсии выборки

$$S^2 = \frac{1}{n-1} \sum (x_i - \bar{x}_в)^2 .$$

В практических вычислениях для дисперсии S^2 часто удобна формула

$$S^2 = \frac{1}{n-1} \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right].$$

Величина S (корень квадратный из выборочной дисперсии) называется **средним квадратическим отклонением выборки** или **выборочным стандартом.**

Почему в формуле дисперсии n заменили на n– 1? Это связано с тем, что входящая в формулу величина $\bar{x}_в$ сама зависит от элементов выборки. Если бы в формуле еще одна величина была функцией элементов выборки, то пришлось бы взять n – 2 и т.д.

Каждая величина, зависящая от элементов выборки и участвующая в формуле выборочной дисперсии, называется **связью**. Эта разность показывает, какое количество элементов выборки можно произвольно изменять, не нарушая связей и называется **числом степеней свободы**. Таким образом, знаменатель выборочной дисперсии всегда равен разности между объемом выборки и числом связей, наложенных на эту выборку.

Отбраковка резко выделяющихся результатов (промахов)

Среди значений случайных величин, включенных в выборку, иногда присутствуют значения, которые весьма существенно отличаются от других. Такие значения появляются, как правило, вследствие грубых ошибок субъективного происхождения или так называемых промахов.

Промахи, как правило, обусловлены следующими причинами:

- неправильным использованием измерительной техники;

- ошибками в отчетах по измерительным приборам;
- ошибками в записях экспериментальных данных;
- ошибками в вычислениях при обработке результатов измерений.

Естественно, что в связи с этим возникает задача выявления и исключения таких сомнительных измерений, иначе они будут искажать результаты статистического анализа и сделанные по нему выводы.

Для этого используют различные правила и критерии. Рассмотрим наиболее употребительные из них.

Определение минимально необходимого числа замеров

Объем выборки определяется исходя из следующих условий:

- объема экспериментальных исследований;
- сроков, в которые будут проведены предполагаемые эксперименты;
- финансовые затраты, сопровождающие проведение экспериментальных исследований;
- требуемой точности и надежности предполагаемых результатов.

Очевидно, что нужно стремиться к тому, чтобы объем выборки был минимально необходимым и в то же время вполне достаточным для получения результатов с желаемой точностью и надежностью. При этом *точность и надежность* в значительной мере определяются изменчивостью изучаемого свойства или показателя, которая оценивается среднеквадратичным отклонением или коэффициентом вариации (для разнородных величин).

Это кажущееся противоречие разрешается следующим образом:

- сначала производится оценочная серия измерений,
- по результатам оценочной серии измерений рассчитываются необходимые точечные оценки,
- делается окончательный расчет необходимого числа замеров по одной из следующих методик.
-

Связь между случайными величинами. Корреляция

До сих пор изучали наблюдения над одной случайной величиной. Между тем для выяснения тех или иных причинно–следственных связей в окружающей природе необходимо вести одновременные наблюдения над целым рядом случайных величин, чтобы по полученным данным изучать взаимоотношения этих величин. Ограничимся пока двумя случайными величинами X и Y.

В математическом анализе зависимость между двумя величинами выражается понятием функции $y = f(x)$, где каждому допустимому значению одной переменной соответствует одно и только одно значение другой переменной. Такая зависимость называется функциональной, она обнаруживается с помощью строгих логических доказательств и не нуждается в опытной проверке. Если y = const при изменении x, то говорят, что y не зависит от x.

Гораздо сложнее обстоит дело с понятием зависимости случайных величин: если при изменении x изменилось y, мы не можем сказать, является ли это изменение результатом зависимости y от x или это результат влияния случайных факторов. Здесь имеет место связь особого рода, при которой с изменением одной величины меняется распределение другой – такая связь называется **стохастической.**

Выявление стохастической связи и оценка ее силы представляют задачу математической статистики.

Рассматривая свойства дисперсии, мы указали, что дисперсия суммы двух независимых величин равна сумме дисперсий этих величин. Поэтому если для двух случайных величин X и $У$ окажется, что

$$D[X+У] \neq D[X]+D[У],$$

то это служит верным признаком наличия зависимости между X и $У$, т.е. корреляции.

Из этого неравенства вытекает (доказано), что справедливо следующее неравенство:

$$M[(X-M[X])(У-M[У])] \neq 0,$$

где $M[(X-M[X])(У-M[У])]$ называют **корреляционным моментом.**

Корреляционный момент зависит от единиц измерения величин X и У. Поэтому на практике чаще используется безразмерная величина, которая называется **коэффициентом корреляции**.

$$\rho = \frac{M[(X-M[X])(У-M[У])]}{\sqrt{D[X]D[У]}}.$$

Свойства коэффициента корреляции

1. Коэффициент корреляции независимых или некоррелированных величин равен нулю.

2. Коэффициент корреляции не меняется от прибавления к X или У каких–либо постоянных (неслучайных) слагаемых, от умножения их на положительные числа.

3. Если одну из случайных величин, не меняя другой, умножить на -1, то на -1 умножится и коэффициент корреляции.

4. Численно коэффициент корреляции заключен в пределах $-1 \leq \rho \leq 1$. Если коэффициент корреляции отличен от нуля, то он своей величиной характеризует не только наличие, но и силу стохастической связи между X и $У$. Чем больше абсолютная величина ρ, тем сильней корреляция между X и $У$. Максимальная корреляция соответствует $|\rho|=1$. Это возможно, когда между случайными величинами существует строгая функциональная связь.

5. Если $\rho > 0$, то величины X и $У$ с точностью до случайных погрешностей одновременно возрастают или убывают, если же $\rho < 0$, то с возрастанием одной величины другая убывает.

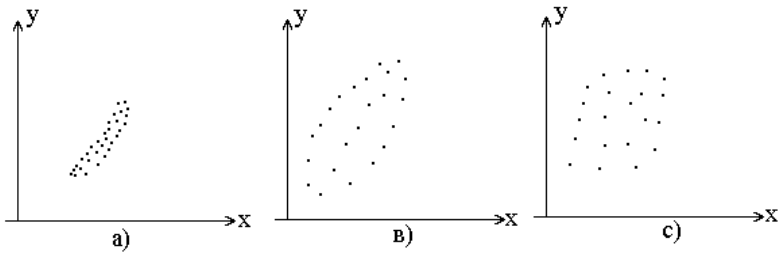
Но это справедливо только для линейной зависимости $У$ от X. Т.е. зависимость между X и $У$ может быть строго функциональной (например, квадратичной) без следа случайности, а коэффициент корреляции все еще будет меньше 1. Таким образом, коэффициент корреляции есть показатель того, насколько связь между случайными величинами близка к строгой линейной зависимости. Он одинаково отмечает и слишком большую долю случайности, и слишком большую криволинейность этой связи.

Если заранее, из общих соображений, можно предсказать линейную зависимость, то ρ является достаточным показателем тесноты связи между X и $У$.

Для случайных величин (большинство именно таких), подчиняющихся нормальному закону, равенство $\rho = 0$ означает одновременно и отсутствие всякой зависимости.

Определение коэффициента корреляции по данным наблюдений

Допустим, что проведено n испытаний и при каждом отмечались значения двух случайных величин. В результате получается n пар выборочных значений (x_1, y_1), (x_2, y_2),…, (x_n, y_n). Для наглядности эти пары значений можно рассматривать как координаты точек на плоскости. Образовавшаяся совокупность точек сразу же дает представление о силе корреляции, рисунок 11.



a – сильная корреляция;в – слабая корреляция;с – отсутствие корреляции

Рисунок 11 –Координаты точек на плоскости выборочных значений

Выборочный коэффициент корреляции r вычисляется по той же формуле, что и генеральный коэффициент ρ, только здесь берутся выборочные математические ожидания (средние) и дисперсии. Если через $\bar{x}$ и $\bar{y}$ обозначить средние значения для x_iи y_i:

$$\bar{x} = \frac{1}{n}\sum x_i, \quad \bar{y} = \frac{1}{n}\sum y_i ,$$

то выборочный корреляционный момент равен

$$\frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y}),$$

откуда

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y},$$

где через S_x^2 и S_y^2 обозначены выборочные дисперсии

$$S_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2, \quad S_y^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2.$$

Удобнее при вычислениях пользоваться следующими выражениями:

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n},$$

$$(n-1)S_x^2 = \sum x_i^2 - \frac{1}{n}(\sum x_i)^2 ,$$

$$(n-1)S_y^2 = \sum y_i^2 - \frac{1}{n}(\sum y_i)^2 .$$

При достаточно большом объеме выборки n выборочный коэффициент корреляции r приближенно равен генеральному коэффициенту ρ. Однако оценить возникающую при этом погрешность очень трудно. Это и не обязательно, так как точное значение ρ в расчетах почти не используется и нужно нам лишь как показатель силы связи.

В связи со случайностью выборки выборочный коэффициент корреляцииr может быть отличен от нуля, даже если между наблюдаемыми величинами нет корреляции.

Следовательно, для проверки гипотезы об отсутствии корреляции необходимо проверять, значительно ли *r* отличается от нуля.

Частная линейная корреляция

Одним из методов исследования линейной связи между тремя (или несколькими) признаками, измеряемыми единовременно у некоторых элементов, является линейная корреляция, применяемая в случае равнозначности признаков, когда неудобно делить их на независимые и зависимые случайные переменные (как это было сказано раньше). При этом различают частную линейную и множественную линейную корреляции.

Обозначим эти признаки *X, Y, Z.*

Например: исследуют свойства некоторого сорта стали: *X* – предел текучести, *Y* – предел прочности, *Z* – предел упругости.

Рассматривая каждый раз только по два из трех признаков, можно в качестве меры линейной зависимости вычислить эмпирические простые коэффициенты корреляции r_{xy}, r_{xz}, r_{yz} (аналогично расчетам, соответствующим парной корреляции). При рассмотрении более двух признаков, однако, для получения безупречного статистического решения простых коэффициентов корреляции оказывается уже недостаточно. Так, r_{xz} выражает зависимость между *X* и *Z*. Но она может возникнуть и по той причине, что оба признака в большей или меньшей мере подвержены воздействию третьего признака – *Y*.

Чтобы исключить влияние третьей случайной величины на две другие, вводят эмпирические частные коэффициенты корреляции, обозначенные через $r_{xy\cdot z}, r_{xz\cdot y}, r_{zy\cdot x}$.

Буквы перед точкой указывают, между какими признаками изучается зависимость, а буква после точки – влияние какого признака исключается.

Разумеется $r_{zy\cdot x} = r_{yz\cdot x}$; $r_{xz\cdot y} = r_{zx\cdot y}$; $r_{xy\cdot z} = r_{yx\cdot z}$.

Принятые обозначения легко распространить на частные коэффициенты корреляции для числа признаков больше трех и выражать линейную связь только между двумя признаками, исключая каждый раз воздействие остальных.

Частные коэффициенты корреляции рассчитываются по простым коэффициентам корреляции согласно формулам

$$r_{xy\cdot z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}},$$

$$r_{xz\cdot y} = \frac{r_{xz} - r_{xy}r_{zy}}{\sqrt{(1-r_{xy}^2)(1-r_{zy}^2)}},$$

$$r_{zy\cdot x} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{zx}^2)(1-r_{yx}^2)}}.$$

Второе и третье уравнения выводятся из первого путем циклической перестановки букв *x, y, z*. Подобно простым коэффициентам корреляции частные коэффициенты могут принимать значения, заключенные между –1 и +1.

Частные коэффициенты корреляции в большей или меньшей степени отклоняются от простых коэффициентов корреляции в зависимости от того, какое воздействие третий, исключаемый, признак оказывает на два оставшихся.

Множественная линейная корреляция

Для ответа на вопрос, зависит ли один из признаков одновременно от двух других (или *y* от x_1 и x_2 при взаимном влиянии x_1 и x_2), вводятся эмпирические множественные коэффициенты корреляции $r_{x\cdot yz}, r_{y\cdot xz}, r_{z\cdot xy}$. Они являются мерой линейной связи между одним

из признаков (буква индекса перед точкой) и совокупностью других признаков. Эти коэффициенты заключены в переделах –1 и +1.

Для практических расчетов обычно применяют следующую формулу:

$$r_{x\cdot yz} = \sqrt{\frac{r_{xy}^2 + r_{xz}^2 - 2z_{xy}r_{xz}r_{yz}}{1 - r_{yz}^2}}.$$

Регрессия

Корреляционный анализ служит установлению значимости (неслучайности) изменения наблюдаемой случайной величины в процессе испытаний. Следующей, еще более высокой ступенью должно явиться выяснение точных количественных характеристик изменения случайной величины. Подобно тому, как совокупность значений случайной величины описывалась набором неслучайных параметров, так и стохастическую, т.е. содержащую элемент случайности, связь нужно научиться выражать через строгие функциональные (неслучайные) соотношения.

Т.е. требуется установить зависимость некоторой случайной величины *Y* от параметра *X*, т.е. *y* от *x*, которая называется **регрессией**.Имеется выборка (x_1, y_1), (x_2, y_2),…, (x_n, y_n) и нужно найти уравнение приближенной регрессии, которое запишем как $y = f(x)$.

В качестве принципа приближенности обычно используют принцип наименьших квадратов, который формулируется так:

Пусть задан некоторый класс функций $f(x)$, накладывающих на выборку одинаковое число связей. Тогда наилучшее уравнение приближенной регрессии дает та функция из рассматриваемого класса, для которой сумма квадратов

$$S = \sum [y_i - f(x_i)]^2$$

имеет наименьшее значение.

Принцип наименьших квадратов позволяет полностью вычислить уравнение приближенной регрессии заданного типа (с неопределенными коэффициентами). Например уравнение регрессии имеет вид:

$$y = \alpha + \beta x + \gamma x^2 + \ldots$$

Составляется сумма $S = \sum [y_i - f(x_i)]^2$, где функция $f(x)$ записана со всеми неопределенными коэффициентами α, β, γ… Величину *S* можно теперь рассматривать как функцию от этих коэффициентов. Задача состоит в том, чтобы найти набор коэффициентов α, β, γ…, минимизирующий величину *S*. Известно, что необходимым условием минимума функции многих переменных $S(\alpha, \beta, \gamma\ldots)$ является выполнение равенств вида:

$$\frac{\partial S}{\partial \alpha} = 0, \frac{\partial S}{\partial \beta} = 0, \frac{\partial S}{\partial \gamma} = 0, \ldots$$

Эти равенства можно рассматривать как уравнения относительно α, β, γ…; которые в математической статистике называются **нормальными уравнениями**.

Доказано, что так как величина $S \geq 0$ при любых α, β, γ…, то у нее обязательно должен существовать хотя бы один минимум и, если система нормальных уравнений имеет единственное решение, то оно и является минимальным для величин *S*, и никаких дополнительных исследований проводить не нужно.

Используя правила дифференцирования, нормальным уравнениям можно придать следующий вид:

$$\begin{cases} \sum 2[y_i - f(x_i)]\dfrac{\partial f(x_i)}{\partial \alpha} = 0 \\ \sum 2[y_i - f(x_i)]\dfrac{\partial f(x_i)}{\partial \beta} = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases}$$

или, после небольших изменений

$$\begin{cases} \sum y_i \dfrac{\partial f(x_i)}{\partial \alpha} - \sum f(x_i)\dfrac{\partial f(x_i)}{\partial \alpha} = 0, \\ \sum y_i \dfrac{\partial f(x_i)}{\partial \beta} - \sum f(x_i)\dfrac{\partial f(x_i)}{\partial \beta} = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases}$$

Пример: $y = \alpha + \beta x + \gamma x^2$.

Тогда $\dfrac{\partial f(x)}{\partial \alpha} = 1, \dfrac{\partial f(x)}{\partial \beta} = x, \dfrac{\partial f(x)}{\partial \gamma} = x^2$. Поэтому нормальные уравнения имеют вид

$$\begin{cases} \sum y_i - \sum(\alpha + \beta x_i + \gamma x_i^2) = 0 \\ \sum y_i x_i - \sum(\alpha x_i + \beta x_i^2 + \gamma x_i^3) = 0 \\ \sum y_i x_i^2 - \sum(\alpha x_i^2 + \beta x_i^3 + \gamma x_i^4) = 0 \end{cases}.$$

Относительно неизвестных коэффициентов α, β и γ получилась линейная система уравнений третьего порядка; ее нетрудно решить, например, с помощью определителей.

После того, как уравнение найдено, его надо подвергнуть статистическому анализу:

1. Оценить ошибку от замены истинной регрессии приближенной. Проверить значимость всех слагаемых уравнения в сравнении со случайной ошибкой наблюдений.
2. Оценить силу связи (провести корреляционный анализ).

Линейная регрессия

Важный случай регрессии первого порядка $y = \alpha + \beta x$ называется **парной линейной регрессией**. Этот вид зависимости часто встречается в практических расчетах.

Так, если об изучаемом явлении нет никаких косвенных сведений, кроме наблюдений, проводимых в настоящий момент, то предварительный характер зависимости можно выяснить, нанося данные в виде точек на координатной плоскости, причем удобнее всего это сделать, предполагая зависимость линейной.

Наконец, во всех сложных случаях, когда регрессия заведомо будет нелинейной, изучение линейной регрессии можно использовать как первый этап исследования, с тем, чтобы в дальнейшем внести в нее необходимые поправки.

Пользуясь принципом наименьших квадратов, легко составить нормальные уравнения линейной регрессии:

$$\sum y_i - \sum(\alpha + \beta x_i) = 0$$

$$\sum y_i x_i - \sum(\alpha + \beta x_i) x_i = 0.$$

После простых преобразований приводим систему к виду (m – объем выборки)

$$\begin{cases} n\alpha + \beta \sum x_i = \sum y_i \\ \alpha \sum x_i + \beta \sum x_i^2 = \sum y_i x_i \end{cases}.$$

Число β называется **коэффициентом регрессии**; его легко найти с помощью определителей:

$$\beta = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}.$$

Число α называется **свободным членом регрессии**. Его тоже нетрудно найти с помощью определителей, но проще выразить его из первого уравнения через найденное уже β

$$\alpha = \frac{\sum y_i - \beta \sum x_i}{n}.$$

Полученные формулы полностью определяют линейную регрессию по заданной выборке. Это равенство можно переписать в виде:

$$\alpha = \frac{1}{n}\sum y_i - \beta \frac{1}{n}\sum x_i = \bar{y} - \beta \bar{x},$$

откуда

$$\bar{y} = \alpha + \beta \bar{x}.$$

Таким образом, средняя точка $(\bar{x}, \bar{y})$ совместного распределения изучаемых величин всегда лежит на линии регрессии. Отсюда вытекает, что для определения линии регрессии достаточно знать лишь ее угловой коэффициент β.

Перейдем к оценке силы найденной связи. Тот факт, что исследуемая зависимость предполагается линейной, позволяет использовать для оценки силы связи выборочный коэффициент корреляции *r*. Можно показать, что *r* и β связаны между собой

$$\beta = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\beta = r\frac{S_y}{S_x}, \text{ откуда } r = \frac{\beta S_x}{S_y}$$

или в развернутом виде

$$r = \beta\sqrt{\frac{n\sum x_i^2 - (\sum x_i)^2}{n\sum y_i^2 - (\sum y_i)^2}}.$$

Если коэффициент корреляции был вычислен ранее, то можно использовать обратную замену β на r.

Мы получим уравнение регрессии в виде

$$y = \alpha + r\frac{S_y}{S_x}x$$

или, заменяя α на $\bar{y} - \beta\bar{x}$, в виде

$$y - \bar{y} = r\frac{S_y}{S_x}(x - \bar{x}).$$

Нелинейная парная регрессия

Основным способом отыскания уравнения нелинейной регрессии (как и линейной) служит принцип наименьших квадратов. Это значит, что уравнение ищется в заданном классе функций и выборочные числовые данные используются лишь для определения неизвестных коэффициентов из системы нормальных уравнений. При этом различаются 2 случая: тип уравнения фиксируется сразу, так что принцип наименьших квадратов используется лишь один раз, или же уравнение регрессии в дальнейшем подвергается уточнениям, для чего принцип наименьших квадратов последовательно используется несколько раз.

Уравнение регрессии может быть известно заранее из соображений аналогии, из теоретических рассуждений или из сравнения эмпирических данных с известными формулами. Разумеется, никакая уверенность в типе регрессии не освобождает от регрессионного и корреляционного анализа найденного уравнения.

Известно, что любая непрерывная функция может быть со сколь угодно высокой точностью заменена многочленом, при этом повышение точности достигается за счет повышения степени многочлена. Поэтому на практике любую регрессию можно считать в виде многочлена, находя его степень путем последовательных подсчетов.

Но степень может оказаться очень высокой и в уравнении будет много неопределенных коэффициентов. А каждый коэффициент накладывает лишнюю связь на выборку и это увеличивает дисперсию.

Иногда используют регрессию показательного, логарифмического, дробно–степенного, тригонометрического и т.д. типов. Количество коэффициентов при этом сокращается, но подбор вида уравнения гораздо сложнее (нет соответствующего алгоритма).

При подборе формул можно руководствоваться следующими соображениями:

1) В тех случаях, когда с возрастанием одной величины замечается пропорциональное возрастание или убывание другой величины, прежде всего берется уравнение прямой

$$y = \alpha + \beta x.$$

2) Если с возрастанием одной величины наблюдается резкое возрастание другой, то может быть применимо уравнение показательной кривой

$$y = \alpha\beta^x.$$

3) Если, наоборот, с возрастанием одной величины имеется замедленное возрастание другой, то может быть пригодна логарифмическая кривая

$$y = \alpha + \beta x + \gamma \lg x.$$

4) В случае периодического изменения одной величины с возрастанием другой могут быть применимы различные тригонометрические функции.

5) Для дугообразных кривых, имеющих один изгиб и схематически изображенных на рисунке 12, сравнительно хорошее совпадение может дать парабола 2–го порядка:

$$y = \alpha + \beta x + \gamma x^2.$$

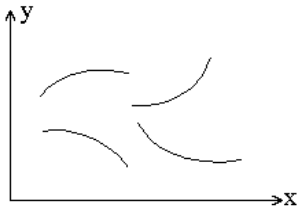


Рисунок 12 – График зависимости y от x

6) Для кривых *S*–образной формы, имеющих двойной изгиб, может подойти уравнение параболы 3–го порядка, в соответствии с рисунком 13.

$$y = \alpha + \beta x + \gamma x^2 + \eta x^3.$$

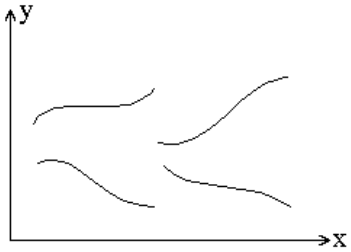


Рисунок 13 – График зависимости y от x

Вычисление трансцендентной регрессии упрощается, если провести замену переменных, превращающую регрессию в линейную. Например, зависимость показательного типа

$$y = \alpha \beta^x$$

превращается в линейную путем логарифмирования

$$\lg y = \lg \alpha + x \lg \beta x.$$

При использовании кривых, подобранных механически, нужно быть осторожным, во всяком случае, нельзя их применять за пределами крайних значений данных, на основе которых вычислены эти уравнения.

Множественная линейная регрессия

Обозначим через X_1 и X_2 независимые переменные, а через Y – зависимую случайную величину. Их реализации будут соответственно: x_{1i}, x_{2i}, y_i (i = 1, 2, …, m), если m – объем

выборки. Предположим, что случайная величина $У$ для любой фиксированной пары значений (x_1; x_2) распределена по нормальному закону.

Уравнение регрессии ищем в виде

$$y = \alpha + \beta x_1 + \gamma x_2 .$$

Оценки для коэффициентов α, β и γ получаем, используя метод наименьших квадратов, т.е. обеспечивается выполнение условий:

$$\sum_{i=1}^{m} [y_i - (\alpha + \beta x_{1i} + \gamma x_{2i})]^2 = \min .$$

Приравняв нулю частные производные по α, β и γ, получим систему линейных уравнений для определения α, β и γ:

$$\begin{cases} n\alpha + \beta \sum x_{1i} + \gamma \sum x_{2i} = \sum y_i \\ \alpha \sum x_{1i} + \beta \sum x_{1i}^2 + \gamma \sum x_{1i} x_{2i} = \sum x_{1i} y_i \\ \alpha \sum x_{2i} + \beta \sum x_{1i} x_{2i} + \gamma \sum x_{2i}^2 = \sum x_{2i} y_i \end{cases} .$$

Преобразование этой системы с учетом обозначений

$$S(x_\kappa^2) = \sum (x_{\kappa i} - \bar{x}_\kappa)^2, \quad \kappa = 1, 2$$
$$S(x_\kappa y) = \sum (x_{\kappa i} - \bar{x}_\kappa)(y_i - \bar{y})$$
$$S(x_1 x_2) = \sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$$
$$S(y^2) = \sum (y_i - \bar{y})^2$$

дает оценки коэффициентов α, β, γ:

$$\beta = \frac{S(x_2^2)S(x_1 y) - S(x_1 x_2)S(x_2 y)}{S(x_1^2)S(x_2^2) - [S(x_1 x_2)]^2},$$
$$\gamma = \frac{S(x_1^2)S(x_2 y) - S(x_1 x_2)S(x_1 y)}{S(x_1^2)S(x_2^2) - [S(x_1 x_2)]^2},$$
$$\alpha = \bar{y} - \beta \bar{x}_1 - \gamma \bar{x}_2 .$$

Коэффициенты β и γ являются коэффициентами регрессии(множественные коэффициенты регрессии).

Величина β показывает зависимость значений y от значений x_1 при постоянном x_2. Поэтому иногда используют обозначение $\beta = \beta_{yx_1 \cdot x_2}$. Соответственно ($\gamma = \gamma_{yx_2 \cdot x_1}$) выражает зависимость значений y от значений x_2 при постоянном x_1.

Уравнение плоскости регрессии получают в виде $y = \alpha + \beta x_1 + \gamma x_2$. Оно имеет смысл лишь в определенной области изменения значений x_1 и x_2.

Аналогично – при большем количестве x.

Определение параметров уравнения множественной регрессии методом активного эксперимента

Основа метода – активный спланированный эксперимент – однофакторный и многофакторный (фактор – независимая переменная).

Однофакторный эксперимент состоит в том, что по определенному плану изменяется значение только одного фактора x_1 и ведут наблюдение за зависящей от него переменной (функцией отклика $y = f(x_1, x_2\ldots)$. Остальные факторы x_2, $x_3,\ldots$ поддерживаются без изменений на одном базовом уровне. Затем та же схема повторяется со вторым фактором x_2 и т.д. После каждой серии опытов рассчитывают «свой» коэффициент регрессии, а затем окончательно находят общее выражение для функции $y = b_o + b_1x_1 + b_2x_2 + \ldots$ Недостатком такого подхода является то, что требуется проведение большого количества опытов, кроме того метод не позволяет учитывать эффекты взаимного влияния факторов.

Чаще применяют многофакторный активный эксперимент, при проведении которого в каждом опыте изменяют значения всех факторов. При этом сокращается число опытов, и задача решается при минимальных затратах времени и средств.

Суть метода

Ищется уравнение регрессии линейного вида

$$y = b_o + b_1x_1 + b_2x_2 + \ldots + b_\kappa x_\kappa, (*)$$

где x_1, $x_2,\ldots$, $x_к$ – варьируемые факторы, они поддерживаются на двух заранее выбранных фиксированных уровнях.

Верхний уровень $x_{i,\max}$ кодируется через +1, нижний – $x_{i,\min}$ – через –1. Соотношение между натуральными и кодированными переменными имеет вид

$$X_i = \frac{x_i - x_{io}}{\Delta x_i},$$

гдеx_i – натуральная переменная ($x_i = x_{i,\max}$или $x_i = x_{i,\min}$);

X_i– кодированная переменная (+1 или –1);

x_{io} – средний (нулевой) уровень, около которого осуществляется варьирование;

Δx_i– интервал (шаг) варьирования по отношению к x_{io};

$$\Delta x_i = x_{i\max} - x_{io} = x_{io} - x_{i\min} > 0.$$

Число опытов определяется из соотношения $N = 2^к$, где $к$ – число варьируемых факторов.

Следовательно, при двух факторах минимальное число опытов (без повторения) равно 4, при трех факторах – 8 и т.д.

Матрицы планирования для этих случаев приведены соответственно в таблицах2 и 3.

Таблица 2 – Два фактора

Номер опыта	X_1	X_2	y		Номер опыта	X_1	X_2	y
1	–1	–1	y_1		3	–1	+1	y_3
2	+1	–1	y_2		4	+1	+1	y_4

Таблица 3 – Три фактора

Номер опыта	X_1	X_2	X_3	y		Номер опыта	X_1	X_2	X_3	y
1	–1	–1	–1	y_1		5	–1	–1	+1	y_5
2	+1	–1	–1	y_2		6	+1	–1	+1	y_6
3	–1	+1	–1	y_3		7	–1	+1	+1	y_7
4	+1	+1	–1	y_4		8	+1	+1	+1	y_8

Как можно заметить, матрица так называемого полного факторного эксперимента (ПФЭ) строится по принципу – ни одной повторяющейся комбинации уровней факторов. Удобно использовать правило: матрица эксперимента 2^3 получается путем повторения матрицы эксперимента 2^2 при нижнем (–1) , а затем при верхнем (+1) значении нового фактора X_3.

При возрастании числа факторов число опытов может стать довольно большим. Так, для шести факторов число необходимых опытов равно 64.

Следует отметить, что при полном факторном эксперименте имеется возможность определить коэффициенты не только для уравнения регрессии линейного вида

$$y = b_o + b_1 x_1 + b_2 x_2 ,$$

но и для уравнения, отражающего взаимодействия факторов, например

$$y = b_o + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2 .$$

Для сокращения числа опытов часто используют дробный факторный эксперимент (ДФЭ). Идея состоит в том, что, если из каких–либо соображений можно пренебречь необходимостью определения коэффициентов при некоторых факторах или их взаимодействиях, то реализуется не вся матрица ПФЭ, а, например, половинная, четвертная, восьмая и т.д. часть полной матрицы.

Рассмотрим методику организации ПФЭ.

Прежде чем приступить к реализации матрицы планирования необходимо выбрать для каждого фактора опорный (нулевой) уровень x_{io} и интервал варьирования Δx_i, что позволит определить нижнее и верхнее значения уровня каждой их всех варьируемых переменных.

После того как составлена матрица планирования и выбраны уровни варьирования факторов можно перейти к постановке опытов, в каждом из которых должна быть реализована одна из строк матрицы. При этом кодированному значению переменной (–1) соответствует нижний уровень варьируемого фактора, а значению (+1) – верхний уровень.

Для устранения предвзятости или субъективизма исследователя, а также систематических ошибок, связанных, например, с разогревом или охлаждением агрегатов и приборов во время эксперимента, старением катализатора, опыты проводятся не в очередности, соответствующей их порядковому номеру в матрице планирования, а в случайном порядке, называемом порядком рандомизации. Порядок рандомизации может быть, например, разыгран путем вытаскивания номеров опытов из урны.

В результате реализации на объекте каждого из опытов заполняется последний столбец матрицы, т.е. записываются значения выходной величины у, полученные при проведении соответствующих вариантных опытов (строк матрицы).

Важным условием получения достоверных результатов является воспроизводимость опытов. Поэтому на каждом уровне могут проводить несколько опытов. Эта процедура называется проверкой воспроизводимости и осуществляется с использованием специального критерия. Необходимо подчеркнуть, что проверка воспроизводимости является важнейшей предпосылкой, лишь при выполнении которой результатам опытов и полученным на их основе закономерностям можно доверять.

При реализации опытов в соответствии с матрицей планирования и поверки воспроизводимости можно приступить к расчету коэффициентов уравнения регрессии. Благодаря переходу к кодированным переменным, которые принимают лишь два значения (–1) (+1), и специальному планированию экспериментов коэффициенты уравнения регрессии определяются раздельно, независимо друг от друга и по очень простой формуле.

Например, коэффициент b_i при i–ом факторе x_i:

$$b_i = \frac{\sum_{u=1}^{N} X_{iu}\bar{y}_u}{N},$$

гдеN – число вариантов опытов в матрице планирования;

X_{iu} – значение кодированной переменной вu–той строке, i–того столбца, равное либо (–1), либо (+1);

$\bar{y}_u$ – среднее значение выхода для u–того варианта опыта (строки).

Отсюда видно, что расчет коэффициентов сводится к простому алгебраическому суммированию построчных значений выходов со знаками столбца, соответствующего данному фактору, и делению на число вариантов опытов. Например, коэффициент при переменной X_1 для уравнения (*) с использованием таблицы 1 определяется следующим образом:

$$b_1 = \frac{-y_1 + y_2 - y_3 + y_4}{4}.$$

Коэффициент b_o по физическому смыслу соответствует опыту с поддержанием всех варьируемых факторов на средних (опорных) уровнях

$$b_o = \frac{\sum_{u=1}^{N} \bar{y}_u}{4}, \text{ т.е. } b_o = \frac{\bar{y}_1 + \bar{y}_2 + \bar{y}_3 + \bar{y}_4}{4}.$$

Недостатком активного метода является условие управляемости процессом по каждому из варьируемых факторов, т.е. возможность независимого изменения каждого из этих факторов и поддержания его на заданном уровне в период проведения опыта.

Проверка статистических гипотез

Доверительные интервалы и доверительные вероятности

Выборочные параметры могут служить приближенными оценками соответствующих генеральных параметров. Погрешность такой оценки, как об этом говорилось ранее, тем меньше, чем больше объем выборки. Как можно оценить эту погрешность более строго?

Все выборочные параметры ($\bar{x}$, S^2) являются случайными величинами и их отклонения от генеральных параметров (погрешности) также будут случайными. Поэтому вопрос об оценке этих отклонений носит вероятностный характер и можно лишь указать вероятность той или иной погрешности, т.е. найти вероятность того, что некоторая случайная величина $\Delta\nu$ (отклонение выборочного параметра ν от исследуемого генерального) не превосходит по абсолютной величине некоторого заданного числа ε. Задача легко решается, если известны функции распределения $F(x)$ или $\varphi(x)$ величины $\Delta\nu$.

Известно:

$$P\{|\Delta\nu| \le \varepsilon\} = F(\varepsilon) - F(-\varepsilon) = \int_{-\varepsilon}^{\varepsilon} \varphi(x)dx.$$

Но, к сожалению, обычно и $F(x)$, и $\varphi(x)$ неизвестны.

Задачу решают по другому. Находят с заданной вероятностью границы возможных значений изучаемого параметра. Эту заданную вероятность называют **доверительной вероятностью**. В зависимости от конкретных обстоятельств в качестве доверительной вероятности берут p = 0,95; 0,98; 0,99; реже p = 0,90 или p = 0,999.

Иногда вместо доверительной вероятности рассматривают связанную с ней величину – так называемый уровень значимости $\alpha=1-p$.

Соответствующие доверительной вероятности границы значений параметра называют **доверительными границами**, а образуемый ими интервал – **доверительным интервалом**.

Представляет интерес оценка генерального среднего и генеральной дисперсии с точки зрения определения их доверительных границ.

Оценка генерального среднего

Предполагается, что наблюдаемая случайная величина имеет нормальное распределение. Для оценки генерального среднего желательно знать генеральную дисперсию σ^2. Но ее нельзя найти из наблюдений и вместо нее обычно берут выборочную дисперсию S^2.

Вводят величину

$$t=\frac{\bar{x}-a}{S}\sqrt{n},$$

где $\bar{x}$ – выборочное среднее;

S^2 – выборочная дисперсия;

a – генеральное среднее;

n – объем выборки.

Таким образом, величина t – это не что иное, как нормированное отклонение $\bar{x}$ от a.

Функция распределения величины t называется ***t* –распределением**, или **распределением Стьюдента**. Она зависит только от числа fстепеней свободы, по которым подсчитана дисперсия S^2. Если дисперсия S^2 и среднее $\bar{x}$ подсчитывались по одним и тем же наблюдениям, то $f = n-1$. Плотность $\varphi(t)$ распределения Стьюдента имеет вид, рисунок 14:

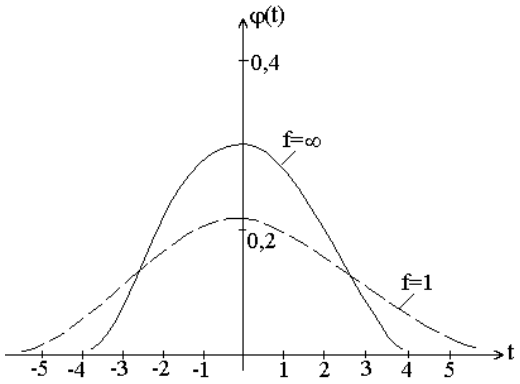


Рисунок 14 – Плотность распределения Стьюдента

По форме графики напоминают плотность нормального распределения, но при $t\to\pm\infty$ они значительно медленнее сближаются с осью абсцисс.

Представляет интерес рассмотрение одной из числовых характеристик случайной величины, которую называют **квантиль.**

Квантилем x_p случайной величины X с функцией распределения $F(x)$, в соотвествии с рисунком 15, называют решение уравнения

$$F(x_p) = P(x < x_p),$$

где вероятность P – заданная величина.

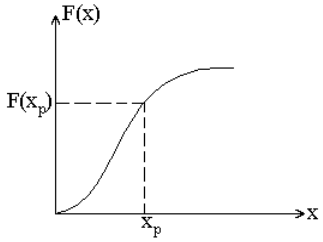


Рисунок 15 - Функциия распределения случайной величины

Величины квантилей представлены в специальных таблицах для соответствующих значений вероятности P.

Возвращаясь к распределению Стьюдента, обозначим t_p – квантиль t–распределения.

Если $t = \frac{\bar{x} - a}{S}\sqrt{n}$, то известно, что доверительные границы генерального среднего могут быть определены из двусторонних неравенств

$$-t_p \le t \le t_p \text{ или } -t_p \le \frac{\bar{x} - a}{S}\sqrt{n} \le t_p.$$

Преобразуем поэтапно двустороннее неравенство. Для этого:

1) Берем левую границу:

$$-t_p \le \frac{\bar{x} - a}{S}\sqrt{n}\,; -\frac{t_p S}{\sqrt{n}} \le \bar{x} - a\,;$$

$$-\bar{x} - \frac{t_p S}{\sqrt{n}} \le -a\,; \text{ откуда } \bar{x} + \frac{t_p S}{\sqrt{n}} > a.$$

2) Рассматриваем правую границу:

$$\frac{\bar{x} - a}{S}\sqrt{n} \le t_p\,;\ \bar{x} - a \le \frac{t_p S}{\sqrt{n}}\,;$$

$$-a \le -\bar{x} + \frac{t_p S}{\sqrt{n}}\,; \text{ или, в итоге } a \ge \bar{x} - \frac{t_p S}{\sqrt{n}}.$$

Окончательно получаем:

$$\bar{x} - \frac{S}{\sqrt{n}} t_p \le a \le \bar{x} + \frac{S}{\sqrt{n}} t_p,$$

что можно представить в виде:

$$\bar{x} - \varepsilon \le a \le \bar{x} + \varepsilon.$$

Таким образом, с вероятностью p – генеральное значение математического ожидания заключается в границах между $\bar{x}-\frac{S}{\sqrt{n}}t_p$ и $\bar{x}+\frac{S}{\sqrt{n}}t_p$, где величина t_p определяется по таблицам распределения Стьюдента для заданной доверительной вероятности p и определяемого числа степеней свободы $f = n-1$.

Распределение Стьюдента позволяет оценивать генеральное среднее (математическое ожидание), когда генеральная дисперсия неизвестна. При этом число наблюдений может быть очень малым, даже равным двум. Конечно, скудость информации сказывается на результатах – доверительные границы получаются довольно широкими. Поэтому везде, где это только возможно, нужно стараться увеличивать число степеней свободы у выборочной дисперсии.

Пример. Выборочное среднее $\bar{x}=18{,}6$ найдено по трем наблюдениям, а выборочная дисперсия определена как $S^2 = 0{,}25$, при этом$f = 3 - 1 = 2$. В качестве доверительной вероятности возьмем $p = 0{,}95$, Величину t_p найдем из таблицы Стьюдента, где для $f = 2$ и $p = 0{,}95$ находим $t_p = 4{,}30$.

Тогда учитывая, что $\varepsilon=\frac{S}{\sqrt{n}}t_p$, доверительная оценка определится как:

$$18{,}6-\frac{0{,}5}{\sqrt{3}}4{,}30\le a\le 18{,}6+\frac{0{,}5}{\sqrt{3}}4{,}30.$$

После вычислений получаем окончательные значения доверительной оценки генерального среднего

$$17{,}36\le a\le 19{,}84.$$

В некоторых задачах требуется найти одностороннюю доверительную оценку генерального среднего, т.е. оценку только сверху или только снизу. При этом задача решается аналогично.

Оценка генеральной дисперсии

Для оценки генеральной дисперсии σ^2 используется выборочная дисперсия S^2. Эта дисперсия в силу случайности выборки сама является случайной величиной. Но математическим ожиданием для S^2 служит генеральная дисперсия σ^2. Отсюда следует, что σ^2 можно оценить по S^2, если известно распределение величины S^2.

Распределение величины S^2 можно получить с помощью так называемого распределения Пирсона (или χ^2 – распределения). Для выборки с элементами $x_1, x_2,\ldots,x_n$ через χ^2 обозначается сумма

$$\chi^2=\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2.$$

В этой сумме есть связь $\bar{x}$, поэтому число степеней свободы $f = n–1$. Плотность χ^2 – распределения зависит только от f, соответствующие графики приведены на рисунке 16.

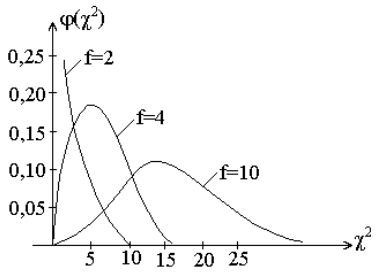


Рисунок 16– Плотность распределения

Нетрудно установить связь между величинами χ^2 и S^2:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \Rightarrow \quad \sum(x_i - \bar{x})^2 = S^2(n-1)$$

$$\chi^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ или } \chi^2 = \frac{S^2}{\sigma^2}(n-1) = \frac{S^2 \cdot f}{\sigma^2}.$$

Теоретически доказано и подтверждено практикой, что для уровня значимости $\alpha = 1 - p$ доверительная оценка величины χ^2 имеет вид

$$\chi^2_{\frac{\alpha}{2}} \le \chi^2 \le \chi^2_{1-\frac{\alpha}{2}},$$

где $\chi^2_{\frac{\alpha}{2}}$ и $\chi^2_{1-\frac{\alpha}{2}}$ – соответствующие квантили распределения Пирсона.

Отсюда

$$\chi^2_{\frac{\alpha}{2}} \le \frac{fS^2}{\sigma^2} \le \chi^2_{1-\frac{\alpha}{2}}$$

или после преобразований

$$\frac{fS^2}{\chi^2_{1-\frac{\alpha}{2}}} \le \sigma^2 \le \frac{fS^2}{\chi^2_{\frac{\alpha}{2}}}.$$

Таким образом получили двустороннюю доверительную оценку для генеральной дисперсии σ^2.

Проверка статистических гипотез применяется для того, чтобы использовать полученную по выборке информацию для суждения о законе распределения генеральной совокупности. При этом имеется определенное представление о неизвестном вероятностном законе $F(x)$ и его параметрах, которое формулируется в виде статистической гипотезы, обозначаемой символом H_0 (нулевая, или основная гипотеза).

Запись $H_o : F(x) = F_o(x)$ означает допущение («гипотезу») о том, что $F_0(x)$ есть функция распределения генеральной совокупности.

С помощью статистических методов или критериев для проверки гипотезы устанавливается, соответствуют ли взятые из выборки данные выдвинутой гипотезе или нет, т.е. нужно ли принять или отвергнуть гипотезу.

Если вид функции распределения $F(x)$ задан отдельными параметрами и если гипотеза строится именно по этим неизвестным параметрам, то говорят о параметрических гипотезах.

Для проверки гипотезы вводят критерий – правило, позволяющее принять или отвергнуть гипотезу в соответствии с данными эксперимента.

Особое положение занимает проверка адекватности модели регрессии.

Адекватная модель регрессии, как правило, неизвестна. Подбирая, например, параболическую модель мы не знаем заранее, какого она должна быть порядка, на какой степени следует остановиться. Не знаем мы также, сколько факторов надо учитывать. Поэтому обычно начинают с моделей первого порядка – линейных моделей и затем

повышают порядок модели (степень многочлена) до тех пор, пока при сравнительно небольшом числе параметров модель не станет адекватной, т.е. гипотеза об ее истинности не будет противоречить данным эксперимента (если это вообще возможно, если существует такая параболическая модель). Для проверки адекватности модели обычно используют критерий Фишера.

Гипотеза о равенстве дисперсий. Критерий Фишера

Сравнение двух или нескольких выборочных дисперсий является одной из важнейших задач статистической обработки наблюдений. Основной выясняемый вопрос при этом – можно ли считать сравниваемые выборочные дисперсии оценками одной и той же генеральной дисперсии.

Начнем со сравнения двух выборочных дисперсий S_1^2 и S_2^2, имеющих соответственно f_1 и f_2 степеней свободы. Будем считать, что первая выборка сделана из генеральной совокупности с дисперсией σ_1^2, вторая – из генеральной совокупности с дисперсией σ_2^2. Выдвигается нулевая гипотеза – гипотеза о равенстве генеральных дисперсий $H_o : \sigma_1^2 = \sigma_2^2$. Для того, чтобы отвергнуть эту гипотезу, нужно доказать значимость расхождения между S_1^2 и S_2^2 при выбранном уровне значимости $\alpha = 1 - p$. В качестве критерия значимости обычно используется так называемое **распределение Фишера**.

Распределением Фишера (или *F*–распределением) называется распределение случайной величины:

$$F = \frac{S_1^2}{\sigma_1^2} : \frac{S_2^2}{\sigma_2^2}.$$

Это распределение зависит только от f_1 и f_2 при этом

$$F(f_1, f_2) = \frac{1}{F(f_2, f_1)}.$$

На рисунке 17 приведены графики плотности *F*–распределения при сочетаниях (f_1, f_2) = (10,4) и (10,50).

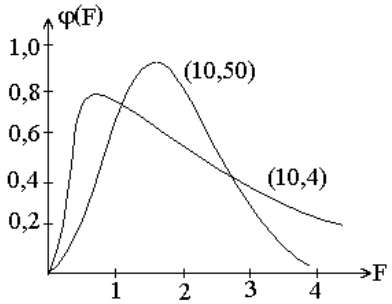


Рисунок 17 графики плотности F–распределения

Как и в случае χ^2–распределения, плотность рассматривается лишь на положительной полуоси, т.е. при $0 \le F \le \infty$. В литературе даются квантили F_{1-p} для некоторых наиболее употребительных уровней значимости $\alpha = 1 - p$ и различных комбинаций f_1 и f_2. При

нахождении квантилей F_p для значений p, не вошедших в таблицу, используется очевидное соотношение

$$F_p(f_1, f_2) = \frac{1}{F_{1-p}(f_2, f_1)}.$$

Например: $F_{0,95}(4,3) = 9,1$; $F_{0,05}(3,4) = \frac{1}{9,1} = 0,11$.

Вернемся к рассмотрению нулевой гипотезы, согласно которой $\frac{\sigma_1^2}{\sigma_2^2} = 1$. В этом случае $F = \frac{S_1^2}{S_2^2}$ и, следовательно, F–распределение может быть использовано непосредственно для оценки отношения $\frac{S_1^2}{S_2^2}$. При этом должно выполняться двустороннее неравенство

$$\frac{1}{F_{1-\alpha/2}(f_2, f_1)} \le \frac{S_1^2}{S_2^2} \le F_{1-\alpha/2}(f_1, f_2). \qquad (*)$$

При этом будем обозначать через S_1^2, большую из сравниваемых дисперсий. Если большей выборочной дисперсии заведомо не может соответствовать меньшая генеральная, т.е. если неравенство $\sigma_1^2 < \sigma_2^2$ заведомо невозможно, то нужно применять односторонний критерий, сравнивая отношение $\frac{S_1^2}{S_2^2}$ с односторонними доверительными оценками:

$$\frac{S_1^2}{S_2^2} \le F_{1-\alpha}(f_1, f_2).$$

Нулевая гипотеза отвергается, если $\frac{S_1^2}{S_2^2} \ge F_{1-\alpha}(f_1, f_2)$, где значение F берется из таблицы для квантилей распределения Фишера.

Пример. При изучении стабильности температуры в термостате получены данные 21,2; 21,8; 21,3; 21,0; 21,4; 21,3. К стабилизатору температуры применено некоторое усовершенствование, после чего (на другом режиме) получены данные: 37,7; 37,6; 37,6; 37,4. Можно ли при уровне значимости $\alpha = 0,05$ считать усовершенствование эффективным?

Эффективность стабилизаторов температуры, очевидно, зависит от даваемой ими дисперсии температур. Таким образом, задача состоит в том, чтобы сравнить генеральные дисперсии данных выборок температур. Вычисляем выборочные дисперсии, уменьшив для удобства вычислений все данные на 21 в первом случае и на 37,5 – во втором:

$$S_1^2 = \frac{1}{5}\left[0,2^2 + 0,8^2 + 0,3^2 + 0,4^2 + 0,3^2 - \right.$$
$$\left. - \frac{(0,2 + 0,8 + 0,3 + 0,4 + 0,3)^2}{6}\right] = \frac{1}{5}\left[1,02 - \frac{4}{6}\right] = 0,07,$$

$$S_2^2 = \frac{1}{3}\left[0{,}2^2 + 0{,}1^2 + 0{,}1^2 + 0{,}1^2 - \frac{(0{,}2+0{,}1+0{,}1-0{,}1)^2}{4}\right] =$$

$$= \frac{1}{3}\left(0{,}07 - \frac{0{,}09}{4}\right) = 0{,}016,$$

Отсюда $F = \frac{0{,}07}{0{,}016} = 4{,}4$.

Числа степеней свободы $f_1 = 5$, $f_2 = 3$.

Усовершенствование может лишь уменьшить дисперсию, поэтому применяем односторонний критерий значимости. По таблице квантилей распределения Фишера находим $F_{0{,}95}(5{,}3) = 9{,}0$. Мы видим, что $\frac{S_1^2}{S_2^2} = 4{,}4 < 9{,}0$. Следовательно, данные наблюдений не позволяют отвергнуть нулевую гипотезу и считать усовершенствование эффективным. Таким образом, различие дисперсий незначимо.

Проверка адекватности уравнения регрессии (математической модели)

После завершения вычислений, связанных с получением оценок коэффициентов регрессии, проверяется адекватность полученного уравнения.

Для проверки значимости (адекватности) уравнения регрессии в целом с использованием *F*–критерия Фишера общую дисперсию S_y^2 сравнивают с остаточной дисперсией $S_{y\,ост}^2$.

Общая дисперсия *У*:

$$S_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_в)^2}{n-1} = \frac{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}{n-1},$$

где *n* – объем выборки; y_i (i = 1, 2,…, *n*) – выборочные значения *y*.

Остаточная дисперсия (дисперсия неадекватности)

$$S_{y\,ост}^2 = S_{ад}^2 = \frac{\sum_{i=1}^{n}(y_i - y_{pi})^2}{n-2},$$

т.е. показатель ошибки предсказания уравнением регрессии результатов опытов, где y_{pi} – расчетное значение величины у, вычисленное по полученному уравнению регрессии (двойка в знаменателе) – количество переменных в уравнении регрессии.

Качество предсказания определяют, сравнивая $S_{y\,ост}^2$ с S_y^2. Находят $F = \frac{S_y^2}{S_{y\,ост}^2}$ (делят всегда большую величину на меньшую).

Для заданной величины уровня значимости $\alpha = 1 - p$ по таблице критерия Фишера (и для соответствующих значений степеней свободы f_1 и f_2) определяют табличное значение критерия $F_{таб}$.

Длятого,чтобыуравнение регрессии адекватно описывало результаты экспериментов с определенной доверительной вероятностью«*p*»требуетсявыполнение следующего условия: $F < F_{таб}$.

Пример. Данные эксперимента

№ п/п	x	y
1	1,5	5,0
2	4,0	4,5
3	5,0	7,0
4	7,0	6,5
5	8,5	9,5
6	10,0	9,0
7	11,0	11,0
8	12,5	9,0
∑	59,5	61,5

Расчеты по известной методике позволяют получить для уравнения регрессии $\alpha = 3{,}73$; $\beta = 0{,}53$, т.е. линейное уравнение регрессии имеет вид:

$$y = 3{,}73 + 0{,}53x\,.$$

Оценим значимость этого уравнения с использованием критерия Фишера. Для этого определяем общую дисперсию y:

$$S_y^2 = \frac{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}{n-1} = \frac{509{,}25 - \frac{61{,}5}{8}}{7} = \\ = \frac{509{,}25 - 472{,}78}{7} = \frac{36{,}47}{7} = 5{,}21$$

(при этом $\sum y_i^2$ и $\sum y_i$ взяты из таблицы, по которой определяют коэффициенты регрессии).

Остаточная дисперсия $S^2_{y\,ост} = S^2_{ад}$ и находится с помощью таблицы:

№ п/п	y_i	x_i	$y_{pi}=3{,}73+0{,}53x_i$	$y_i - y_{pi}$	$(y_i - y_{pi})^2$
1	5,0	1,5	$3{,}73+0{,}53\cdot 1{,}5 = 4{,}53$	+0,47	0,2209
2	4,5	4,0	$3{,}73+0{,}53\cdot 4{,}0 = 5{,}85$	–1,35	1,82225
3	7,0	5,0	$3{,}73+0{,}53\cdot 5{,}0 = 6{,}38$	+0,62	0,3844
4	6,5	7,0	$3{,}73+0{,}53\cdot 7{,}0 = 7{,}44$	–0,94	0,8836
5	9,5	8,5	$3{,}73+0{,}53\cdot 8{,}5 = 8{,}24$	+1,26	1,5876
6	9,0	10,0	$3{,}73+0{,}53\cdot 10{,}0 = 9{,}03$	–0,03	0,0009
7	11,0	11,0	$3{,}73+0{,}53\cdot 11{,}0 = 9{,}56$	+1,44	2,0736
8	9,0	12,5	$3{,}73+0{,}53\cdot 12{,}5 = 10{,}35$	–1,35	1,8225
∑	61,5	59,5		0,12	8,8

Откуда

$$S^2_{y\,ад} = \frac{\sum (y_i - y_{pi})^2}{n-2} = \frac{8{,}8}{6} = 1{,}46\,.$$

Определяем

$$F = \frac{S_y^2}{S^2_{y\,ад}} = \frac{5{,}21}{1{,}46} = 3{,}56\,.$$

Для 5 % уровня значимости ($\alpha = 0{,}05$) и для $f_1 = 7$, а$f_2 = 6$, $F_{таб} = 4{,}21$
$F < F_{таб}$, т.е. уравнение регрессии адекватно.

Математическое описание случайных сигналовв системах управления

Случайные процессы

Функция, значение которой при каждом данном значении независимой переменной является случайной величиной, называется **случайной функцией**. То есть, это бесконечная совокупность случайных величин, зависящая от непрерывно изменяющейся независимой переменной. Случайная функция, зарегистрированная по результатам опыта, называется реализацией случайной функции. Случайная функция, для которой независимой переменной является время *t*называется **случайным (стохастическим) процессом.**

Для характеристики случайной функции служат моменты случайной функции. Для их определения необходимо знать многомерные функции распределения случайной величины.

Предположим, что мы располагаем большим числом однотипных систем, работающих одновременно при одинаковых условиях. Будем наблюдать изменения величин на выходе этих устройств. Они будут характеризоваться некоторыми случайными функциями $x'(t), x''(t), x'''(t),...$, причем все эти функции будут отличаться друг от друга.

Рассмотрим какой–либо момент времени t и найдем, какая доля из общего числа функций x(t) имеет в этот момент времени значение, заключенное между x и $x + dx$.

Эта доля зависит от момента времени t и пропорциональна dx при малых dx. Обозначим ее через $w_1(x, t)dx$ – (при количестве реализаций, стремящихся к бесконечности, эта величина соответствует вероятности того, что в момент времени t величина x будет заключена в пределах x и x + dx) и назовем $w_1(x, t)$ – первой или одномерной функцией распределения вероятности.

Рассмотрим теперь все возможные пары значений x, наблюденные в два различных момента времени t_1 и t_2. Долю пар значенийx,для которойвеличина x заключена между (x_1, x_1+ dx_1) при t= t_1 и между (x_2, x_2 + dx_2) при t= t_2, отнесенную к общему числу наблюденных пар значений, обозначим через $w_2(x_1, t_1, x_2, t_2)$dx_1dx_2 и назовем второй или двумерной функцией распределения вероятности.

Этот процесс можно продолжить и определить третью, четвертую и все последующие функции распределения вероятности.

Итак, случайный процесс можно характеризовать некоторыми функциями распределения вероятности, полностью определяющими его в статистическом смысле.

Действительно, зная $w_1(x, t)$ можно определить математическое ожидание $m_{ox}(t)$ случайной величины $x(t)$:

$$m_{ox}(t) = \int_{-\infty}^{\infty} x(t) w_1(x,t) dx$$

и его дисперсию $\beta_{2x}(t)$:

$$\beta_{2x}(t) = M\left[x(t) - m_{ox}(t)\right]^2 = \int_{-\infty}^{\infty} \left[x(t) - m_{ox}(t)\right]^2 w_1(x,t) dx .$$

Зная вторую функцию распределения $w_2(x_1,t_1, x_2,t_2)$, можно определить как $m_{ox}(t)$, $\beta_{2x}(t)$, так и центральный момент 2–го порядка:

$$\beta_{2x}(t_1,t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[x_1 - m_{ox}(t_1)\right]\left[x_2 - m_{ox}(t_2)\right] w_2(x_1 t_1 x_2 t_2) dx_1 dx_2 ,$$

характеризующий связь между значениями случайной функции в различные моменты времени. Функция $\beta_{2x}(t_1,t_2)$ называется **корреляционной функцией**.

Зная *n*–мерную функцию распределения вероятности, можно определить все последующие моменты случайной функции, включая момент *n*–го порядка.

Если приходится иметь дело не с одной, а с несколькими взаимосвязанными функциями, то кроме их собственных моментов, приходится вводить еще и их взаимные моменты.

Так, например, если имеются две случайные функции $x(t)$ и $y(t)$, то простейшим взаимным моментом является момент 2–го порядка:

$$\beta_{2xy}(t_1,t_2) = M\{[x - m_{ox}(t_1)][y - m_{oy}(t_2)]\} =$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[x_1 - m_{ox}(t_1)][y - m_{oy}(t_2)]w_2(x_1t_1; yt_2)dxdy$$

называемый взаимной корреляционной функцией случайных процессов $x(t)$ и $y(t)$.

Стационарные случайные процессы

Общая теория случайных функций, требующая задания многомерных функций распределения вероятности, обычно оказывается очень сложной и громоздкой для практических применений.

Но, если случайные функции имеют нормальное распределение, то задание их первых двух моментов достаточно для определения всех последующих моментов, вследствие чего корреляционная теория (2–ой центральный момент – корреляционный момент) может рассматриваться как общая теория случайных функций с нормальным распределением.

Но и корреляционная теория достаточно сложна, поэтому обычно рассматривают те или иные виды случайных процессов, удовлетворяющих определенным допущениям.

Особое место занимают стационарные случайные процессы. Для них вид функции распределения вероятности не зависит от смещения начала отсчета вдоль оси времени. В случае стационарных случайных процессов определение функции распределения упрощается в том отношении, что она может быть определена в течение достаточно долгого промежутка времени из результатов наблюдения над одной единственной системой, а не над многими. Действительно, так как в этом случае функция распределения не зависит от начала отсчета времени, то можно предположить, что экспериментальную запись кривой $x(t)$, полученную из наблюдения над одной системой в течение достаточно долгого промежутка времени, можно разбить на ряд отрезков длиной T (где T велико по сравнению со всеми «периодами», которые имеются в исследуемом процессе) и считать, что функциями, входящими в совокупность, являются функции $x(t)$, представляющие собой части всей кривой $x(t)$ на протяжении каждого из отрезков T.

В основе этого предположения лежит так называемая **эргодическая гипотеза**, согласно которой большое число наблюдений над одной – единственной системой, движение которой представляет собой стационарный случайный процесс, в моменты времени, выбранные произвольным образом, имеет те же статистические свойства, то же число наблюдений над произвольно выбранными подобными ей системами в один и тот же момент времени.

Различают «среднее значение по совокупности (множеству)», т.е. средние значения, определенные на основании наблюдения над многими подобными системами в один и тот же момент времени, и «среднее по времени», т.е. среднее значение, определенное на основании наблюдения над одной из этих систем для достаточно большого числа последующих моментов времени.

Для стационарных процессов это одно и то же. Так, например, для функции $x(t)$ «среднее по совокупности» $\tilde{x}$:

$\tilde{x} = \int_{-\infty}^{\infty} x w_1(x,t)dx$ – зависит от момента времени t, а среднее по времени $\bar{x}$ для интервала времени $2T$:

$\bar{x} = \frac{1}{2T}\int_{-T}^{T} x(t)dt$ – не зависит от t и для стационарного процесса $\tilde{x} = \bar{x}$.

Аналогичное равенство имеет место и для моментов более высокого порядка.

Корреляционная функция

Важной вероятностной характеристикой случайного процесса является корреляционная функция, которая для стационарного случайного процесса $x(t)$ определится как:

$$R_x(\tau) = M\{[x(t) - m_{0x}][x(t+\tau) - m_{0x}]\} \approx$$
$$\approx \frac{1}{2T}\int_{-T}^{T}[x(t) - m_{0x}][x(t+\tau) - m_{0x}]dt$$

где $m_{ox} = \bar{x}$ – математическое ожидание («среднее по времени» на интервале T).

Для «центрированного» случайного процесса $m_{ox} = 0$ и

$$R_x(\tau) \approx \frac{1}{2T}\int_{-T}^{T} x(t)x(t+\tau)dt.$$

Корреляционную функцию случайного процесса $x(t)$ называют еще **автокорреляционной функцией** $R_x(\tau)$, так как если приходится иметь дело с двумя стационарными случайными процессами $x(t)$ и $y(t)$, то кроме их математических ожиданий m_{ox}, m_{oy} и автокорреляционных функций $R_x(\tau)$ и $R_y(\tau)$, обычно вводится в рассмотрение корреляционная функция связи, или взаимная корреляционная функция, которая для стационарного случайного процесса –

$$R_{xy}(\tau) \approx \frac{1}{2T}\int_{-T}^{T}[x(t) - m_{ox}][y(t+\tau) - m_{oy}]dt.$$

Если $m_{ox} = m_{oy} = 0$, то

$$R_{xy}(\tau) \approx \frac{1}{2T}\int_{-T}^{T} x(t)y(t+\tau)dt.$$

Она имеет непосредственную связь с понятием коэффициента корреляции. Так, если коэффициент корреляции характеризует меру зависимости между двумя системами чисел x_i и y_i (случайные величины), то взаимная корреляционная функция характеризует зависимость значений одной и той же или различных случайных функций в различные моменты времени.

Свойства корреляционной функции

Для простоты предположим, что $m_{ox} = 0$.

1. Начальное значение корреляционной функции $R(\tau)$ равно среднему значению квадрата случайной функции и поэтому существенно положительно, т.е.

$$R(0) = \frac{1}{2T}\int_{-T}^{T} x^2(t)dt = \bar{x^2} > 0.$$

2. Корреляционная функция есть четная функция τ, т.е.

$$R(\tau) = R(-\tau).$$

Действительно,

$$R(\tau) = \overline{x(t)x(t+\tau)} = \overline{x(t-\tau)x(t)} = R(-\tau).$$

3. Значение $R(\tau)$ при любом τ не может превышать ее начального значения $R(0)$, т.е.

$$R(0) \geq |R(\tau)|.$$

4. Корреляционная функция $R(\tau)$ для достаточно больших τ стремится к нулю, т.е.

$$\lim_{\tau\to\infty} R(\tau) = 0.$$

5. Если $x(t)$ представляет собой, например, стационарный случайный процесс с наложенной на него постоянной составляющей a_o, то $R(\tau)$ будет иметь вид, изображенный на рисунке 18:

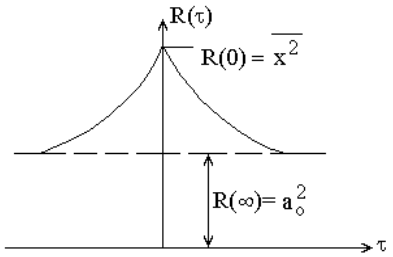


Рисунок18 – Стационарный случайный процесс с наложенной на него постоянной составляющей

При этом ее начальная ордината $R(0)$ равна среднему значению квадрата сигнала $x(t)$, а ее конечная ордината $R(\infty)$ – значению квадрата постоянной составляющей сигнала $x(t)$, т.е. величине a_o^2.

6. Если $x(t)$ представляет собой стационарный случайный процесс с наложенной на него периодической составляющей, рисунок 19, то $R(\tau)$ также будет содержать периодическую составляющую с тем же периодом и, следовательно, будет иметь вид:

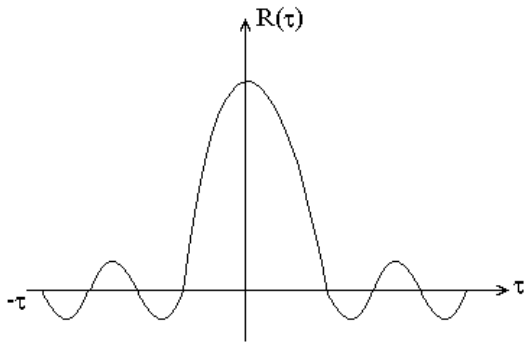


Рисунок 19 – Стационарный случайный процесс с наложенной на него периодической составляющей

7. «Белый шум». Случайный процесс $x(t)$, который характеризуется тем, что в нем отсутствует какая–либо взаимная связь между предыдущим и последующим значениями $x(t)$, называется **абсолютно случайным процессом** или **«белым шумом»**. Очевидно, что в этом случае корреляционная функция равна нулю при всех значениях τ, кроме $\tau = 0$, и ее можно представить в виде δ–функции или практически в виде импульса достаточно большой амплитуды, но малой ширины, площадь которого равна единице.

Взаимная корреляционная функция $R_{xy}(\tau)$ не является четной, в отличие от $R_x(\tau)$, но для нее справедливо равенство: $R_{xy}(\tau)=R_{yx}(-\tau)$. Кроме того: $\sqrt{R_x(0)}\sqrt{R_y(0)} \geq \left|R_{xy}(\tau)\right|$.

Спектральная плотность

Спектральная плотность случайного процесса $x(t)$ определяется как преобразование Фурье от корреляционной функции $R_x(\tau)$:

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

Это так называемая автоспектральная плотность. В свою очередь в соответствии с формулой обратного преобразования Фурье:

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

$S_x(\omega)$ играет большую роль при исследовании преобразования случайных сигналов линейными системами.

Автоспектральная плотность $S_x(\omega)$ – действительная и четная функция.

Так, используя формулу Эйлера, можно записать:

$$S_x(\omega) = \int_{-\infty}^{\infty} R_x(\tau)\cos\omega\tau d\tau - j\int_{-\infty}^{\infty} R_x(\tau)\sin\omega\tau d\tau .$$

Но $R_x(\tau)$ – четная функция, т.е. $R_x(\tau) = R_x(-\tau)$, а $\sin(-\omega\tau) = -\sin\omega\tau$, поэтому во втором слагаемом подынтегральное выражение представляет собой нечетную функцию переменного τ.

Поэтому $\int_{-\infty}^{\infty} R_x(\tau)\sin\omega\tau d\tau = 0$.

И тогда $S_x(\omega) = \int_{-\infty}^{\infty} R_x(\tau)\cos\omega\tau d\tau = 2\int_{o}^{\infty} R_x(\tau)\cos\omega\tau d\tau$, и следовательно

$S_x(\omega) = S_x(-\omega)$.

Аналогично

$$R_x(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_x(\omega)\cos\varpi\tau d\omega = \frac{1}{\pi}\int_{o}^{\infty} S_x(\omega)\cos\omega\tau d\omega.$$

Эти соотношения позволяют определить $S_x(\omega)$ по заданной аналитически или в виде графика $R_x(\tau)$, или наоборот. В любом случае для определения функции $S(\omega)$ по $R(\tau)$ и наоборот можно использовать таблицы преобразования Фурье.

Свойства спектральной плотности

1. Если $R(\tau)$ – монотонно убывающая функция τ, то $S(\omega)$ – также монотонно убывающая функция ω.

2. Чем уже функция $R(\tau)$, тем более пологой и широкой является функция $S(\omega)$.

3. Если $R(\tau)$ стремится к нулю в течение очень короткого времени Δ, то $S(\omega)$ сохраняет постоянное значение до частоты порядка $\frac{2\pi}{\Delta}$.

4. Спектральная плотность "белого шума" равна интегралу от δ–функции, т.е. равна единице. Таким образом, энергия "белого шума" распределена по спектру равномерно и его суммарная энергия равна бесконечности, что физически нереализуемо.

5. Если случайная функция содержит постоянную составляющую, то в кривой спектральной плотности в точке ω = 0 имеется δ–функция (практически – острый импульс).

6. Если случайная функция имеет периодическую составляющую частоты ω_o, то в составе кривой $S(\omega)$ имеется две δ–функции в точках $\pm\omega_o$.

Определение параметров функции распределения

Ошибки косвенных измерений

Различают прямые и косвенные измерения. В первом случае непосредственно измеряется определяемая величина, во втором она задается некоторой функцией от непосредственно измеряемых величин. Пусть случайная величина z зависит от наблюдений $x_1, x_2, \dots, x_n$ по известному закону:

$$z = f(x_1, x_2, \dots, x_n).$$

Истинное значение величины z может не совпадать с математическим ожиданием M_x, а определяться тем же законом:

$$a_z = f\left(m_{x_1}, m_{x_2}, \dots, m_{x_n}\right).$$

Величина a_z называется средним косвенного измерения.

Дисперсия косвенного измерения σ_z^2 определяется так же, как обычная дисперсия, только отклонения берутся от среднего косвенного измерения a_z. Ее можно найти, зная дисперсии отдельных наблюдений и вид функции f. На практике определяют выборочные дисперсии $s_{x_i}^2$ и по ним выборочную дисперсию косвенного измерения s_z^2, которая служит оценкой генеральной дисперсии σ_z^2.

Чтобы найти s_z^2, разложим функцию $z = f(x_1, x_2, \dots, x_n)$ в ряд Тейлора в точке $\left(m_{x_1}, m_{x_2}, \dots, m_{x_n}\right)$, ограничиваясь членами первого порядка:

$$z \approx f(m_{x_1}, m_{x_2}, \dots, m_{x_n}) + \frac{\partial f}{\partial x_1}(x_1 - m_{x_1}) +$$
$$+\frac{\partial f}{\partial x_2}(x_2 - m_{x_2}) + \frac{\partial f}{\partial x_n}(x_n - m_{x_n}), \quad (2.1)$$

и определим s_z^2 по закону сложения дисперсий:

$$s_z^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 s_{x_i}^2. \quad (2.2)$$

Выражение (2.2) называется законом накопления ошибок.

Определение дисперсии по текущим измерениям

Математическое ожидание и дисперсия генеральной совокупности оцениваются средним и дисперсией выборки тем точнее, чем больше объем выборки. При этом среднее характеризует результат измерений, а дисперсия – точность этого результата (дисперсия воспроизводимости). Если выполнено m параллельных опытов и получена выборка $y_1, y_2, \dots, y_m$ значений измеряемой величины, то дисперсия воспроизводимости равна:

$$s_{\text{воспр}}^2 = \frac{\sum_{u=1}^{m}(y_u - \bar{y})^2}{m-1}, \quad (2.3)$$

$$\bar{y} = \frac{\sum_{u=1}^{m} y_u}{m}.$$

и ошибка опыта (ошибка воспроизводимости):

$$s_{\text{воспр}} = \sqrt{s_{\text{воспр.}}^2}.$$

Часто для оценки точности применяемой методики многократно повторяют анализ одой и той же пробы. На проведение большой серии опытов требуется много времени, в течение которого может неконтролируемым образом измениться среднее значение результатов анализа. Значительно проще определять ошибку воспроизводимости по текущим измерениям.

Предположим, анализируется n проб. При анализе каждой пробы делается различное число параллельных опытов: $m_1, m_2, \dots, m_n$. Определим частные дисперсии $s_1^2, s_2^2, \dots, s_n^2$ для каждой такой выборки в отдельности. Число степеней свободы частных дисперсий соответственно равно: $f_1 = m_1 - 1$, $f_2 = m_2 - 1, \dots, f_n = m_n - 1$. Общая дисперсия воспроизводимости всех опытов будет равна средневзвешенному значению частных дисперсий:

$$s_{\text{воспр}}^2 = \frac{f_1 s_1^2 + f_2 s_2^2 + \dots + f_n s_n^2}{f_1 + f_2 + \dots + f_n} = \frac{(m_1-1)s_1^2 + (m_2-1)s_2^2 + \dots + (m_n-1)s_n^2}{m_1 + m_2 + \dots + m_n - n}. \quad 2.4)$$

Число степеней свободы общей дисперсии равно общему числу измерений минус число связей, использованных для определения n средних:

$$f_{\text{воспр}} = m_1 + m_2 + \dots + m_n - n = \sum_{i=0}^{n} m_i - n. (2.5)$$

Учитывая, что частные дисперсии определяются по результатам параллельных опытов по формуле:

$$s_i^2 = \frac{\sum_{u=1}^{m_i}(y_{iu} - \bar{y}_i)^2}{m_i - 1}, i = 1,2, \dots, n.$$

Из (2.4) следует:

$$s^2_{\text{воспр}} = \frac{(m_1-1)s_1^2 + (m_2-1)s_2^2 + \cdots + (m_n-1)s_n^2}{m_1 + m_2 + \cdots + m_n - n} =$$

$$= \frac{\sum_{i=1}^{n}(m_i-1)s_i^2}{\sum_{i=1}^{n}(m_i-n)} = \frac{\sum_{i=1}^{n}\sum_{u=1}^{m_i}(y_{iu}-\bar{y}_i)^2}{\sum_{i=1}^{n}(m_i-n)},$$

Если число параллельных опытов при анализе каждой пробы одинаково $m_1 = m_2 = \cdots = m_n = m$, формулы для расчета дисперсии воспроизводимости упрощаются. При этом:

$$s^2_{\text{воспр}} = \frac{(m-1)(s_1^2 + s_2^2 + \cdots + s_n^2)}{mn - n} =$$

$$= \frac{(m-1)\sum_{i=1}^{n}s_i^2}{n(m-1)} = \frac{\sum_{i=1}^{n}s_i^2}{n}. \quad (2.6)$$

Таким образом, при равном числе параллельных опытов общая дисперсия вопроизводимости равна среднеарифметическому значению частных дисперсий. Число степеней свободы общей дисперсии при этом равно:

$$f_{\text{воспр}} = n(m-1). \quad (2.7)$$

И окончательно

$$s^2_{\text{воспр}} = \frac{\sum_{i=1}^{n}\sum_{u=1}^{m_i}(y_{iu}-\bar{y}_i)^2}{n(m-1)}. \quad (2.8)$$

Общая дисперсия воспроизводимости намного точнее оценивает дисперсию генеральной совокупности $\sigma^2_{\text{воспр}}$. При вычислении дисперсии по текущим измерениям объединяют между собой только те пробы, которые можно рассматривать как выборки из генеральных совокупностей с равными дисперсиями. При этом каждое из значений $s_1^2, s_2^2, \ldots, s_n^2$ можно рассматривать как оценку одной и той же генеральной дисперсии.

Доверительные интервалы и доверительная вероятность

Выборочные параметры являются случайными величинами, их отклонения от генеральных (погрешности) также будут случайными. Оценка этих отклонений носит вероятностный характер: можно лишь указать вероятность той или иной погрешности. Для этого используют понятия «доверительный интервал» и «доверительная вероятность». Пусть для генерального параметра a получена несмещенная оценка a^*. Нужно оценить возможную при этом ошибку. Назначим достаточно большую вероятность β – такую, что событие с вероятностью β можно считать практически достоверным, и найдем такое значение $\varepsilon = f(\beta) = \varepsilon_\beta$, для которого

$$P\left(|a^* - a| \le \varepsilon_\beta\right) = \beta. \quad (2.9)$$

При этом диапазон практически возможных значений ошибки, возникающей при замене a наa^* будет $\pm\varepsilon_\beta$, большие по абсолютной величине ошибки будут появляться только с малой вероятностью:

$$p = 1 - \beta, \quad (2.10)$$

называемой уровнем значимости. Выражение (2.9) может быть интерпретировано как вероятность того, что истинное значение параметра a лежит в пределах:

$$a^* - \varepsilon_\beta \le a \le a^* + \varepsilon_\beta. \quad (2.11)$$

Вероятность β называется *доверительной вероятностью*. Она характеризует надежность полученной оценки. Интервал $a^* \pm \varepsilon_\beta$ называется *доверительным интервалом.* Границы интервала $a' = a^* - \varepsilon_\beta$ и $a'' = a^* + \varepsilon_\beta$ называются *доверительными границами*. Доверительный интервал при данной доверительной вероятности определяет точность оценки. Его величина зависит от доверительной вероятности, с которой гарантируется нахождение параметра a внутри доверительного интервала: чем больше величина β, тем больше и величина ε_β. На практике значения доверительной вероятности фиксируют на определенном уровне $(0{,}9;\ 0{,}95;\ 0{,}99)$. Исходя из этого определяют доверительный интервал результата.

Для выборки объемом n значений случайной величины X среднеезначение определяется по формуле:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}. \quad (2.12)$$

Для построения доверительного интервала необходимо знать распределение этой оценки. Для выборок из генеральной совокупности, распределенной нормально, $\bar{x}$ имеет нормальное распределение с математическим ожиданием m_x и средним квадратическим отклонением $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$. Тогда, используя функцию Лапласа, получим:

$$P\left(|\bar{x} - m_x| < \varepsilon_\beta\right) = \beta = 2\Phi\left(\frac{\varepsilon_\beta}{\sigma_{\bar{x}}}\right). \quad (2.13)$$

Задавшись доверительной вероятностью β, определим по таблицам функции Лапласа (Приложение 2), что $k_\beta = \frac{\varepsilon_\beta}{\sigma_{\bar{x}}}$. Тогда доверительный интервал для математического ожидания имеет вид:

$$\bar{x} - k_\beta \sigma_{\bar{x}} \le m_x \le \bar{x} + k_\beta \sigma_{\bar{x}}. \quad (2.14)$$

Знание генеральной дисперсии σ_x^2 позволяет оценивать математическое ожидание даже по одному наблюдению. Если для нормально распределенной случайной величины X получено значение x_1, то доверительный интервал для математического ожидания с доверительной вероятностью β составляет:

$$x_1 - \sigma u_{1-\frac{p}{2}} \le m_x \le x_1 + \sigma u_{1-\frac{p}{2}}, \quad (2.15)$$

где $u_{1-\frac{p}{2}}$ – квантиль стандартного нормального распределения.

Оценка математического ожидания нормально распределенной случайной величины

При отсутствии грубых и систематических ошибокматематическое ожидание случайной величины совпадает с истинным результатом наблюдений. Проще всего оценить математическое ожидание с помощью известной дисперсии генеральной совокупности σ^2. Генеральную дисперсию получить из наблюдений нельзя: ее можно оценить при помощи выборочной дисперсии s^2. Ошибка от замены генеральной дисперсии выборочной будет тем

меньше, чем больше объем выборки. При небольших объемах выборки n для построения доверительного интервала математического ожидания используют распределение Стьюдента или *t*–распределение. Распределение Стьюдента имеет случайная величина*t***:**

$$t = \frac{\bar{x}-m_x}{s_x}\sqrt{n}. \quad (2.16)$$

Если дисперсия s^2 и среднее $\bar{x}$ определяются по одной и той же выборке, то $f = n - 1$. Распределение Стьюдента симметрично, поэтому

$$t_{\frac{p}{2}} = -t_{1-\frac{p}{2}}. \quad (2.17)$$

Учитывая симметрию распределения, часто пользуются обозначением $t_{p,f}$ (f – число степеней свободы, p – вероятность того, что tнаходится за пределами интервала $\left(t_{\frac{p}{2}}, t_{1-\frac{p}{2}}\right)$). В соответствии с (2.16) и (2.17) справедливо:

$$\bar{x} - \frac{s_x}{\sqrt{n}}t_{1-\frac{p}{2}} \le m_x \le \bar{x} + \frac{s_x}{\sqrt{n}}t_{1-\frac{p}{2}}, \quad (2.18)$$

Значения квантилей $t_{1-\frac{p}{2}}$ для различных чисел степеней свободы f и уровней значимости p приведены в Приложении 3. В некоторых задачах требуется найти одностороннюю оценку математического ожидания, т. е. оценку только сверху или только снизу. При доверительной вероятности $\beta = 1 - p$ оценка для случайной величины *t*сверху имеет вид:

$$t \le t_{1-p},$$

если
$$\frac{\bar{x}-m_x}{s_x}\sqrt{n} \le t_{1-p}; \quad (2.19)$$

оценка для t снизу имеет вид:

$$\frac{\bar{x}-m_x}{s_x}\sqrt{n} \ge -t_{1-p}. \quad (2.20)$$

Из неравенств (2.19) и (2.20) получим односторонние доверительные оценки для математического ожидания сверху:

$$m_x \le \bar{x} + \frac{s_x}{\sqrt{n}}t_{1-p} \quad (2.21)$$

и снизу

$$m_x \le \bar{x} - \frac{s_x}{\sqrt{n}}t_{1-p}. \quad (2.22)$$

Оценка дисперсии нормально распределенной случайной величины

Дисперсию генеральной совокупности σ_x^2 нормально распределенной случайной величины можно оценить, если известно распределение ее оценки – выборочной дисперсии s_x^2. Распределение выборочной дисперсии можно получить при помощи распределения Пирсона или χ^2-распределения.

Если имеется выборка n независимых наблюдений x_1, x_2,…, x_nнад нормально распределенной случайной величиной, то можно показать, что сумма:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x}\right)^2 \quad (2.23)$$

имеет распределение χ^2 с $f = n - 1$.

Плотность χ^2 распределения зависит только от числа степеней свободы. При доверительной вероятности $\beta = 1 - p$ двухсторонняя доверительная оценка для χ^2имеет вид:

$$\chi^2_{\frac{p}{2}} \leq \chi^2 \leq \chi^2_{1-\frac{p}{2}}, \quad (2.24)$$

Односторонние оценки имеют вид:

$$\chi^2 \leq \chi^2_{1-p};\ \chi^2 \geq \chi^2_p. \quad (2.25)$$

Квантили χ^2_{1-p} при различных p и f приведены в Приложении 4. Так как выборочная дисперсия s_x^2 через элементы выборки определяется по формуле:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{f}. \quad (2.26)$$

Из (2.23) получим:

$$\chi^2 = \frac{f s_x^2}{\sigma_x^2}. \quad (2.27)$$

Подставляя (2.27) в (2.25), получим

$$\chi^2_{\frac{p}{2}} \leq \frac{f s_x^2}{\sigma_x^2} \leq \chi^2_{1-\frac{p}{2}}. \quad (2.28)$$

Решая неравенство относительно σ_x^2, получим доверительные двусторонние границы для генеральной дисперсии σ_x^2:

$$\frac{f s_x^2}{\chi^2_{1-\frac{p}{2}}} \leq \sigma_x^2 \leq \frac{f s_x^2}{\chi^2_{\frac{p}{2}}}. \quad (2.29)$$

Аналогично получаются односторонние доверительные оценки:

$$\sigma_x^2 \leq \frac{f s_x^2}{\chi_p^2},\ \sigma_x^2 \geq \frac{f s_x^2}{\chi^2_{1-p}}. \quad (2.30)$$

При $n \geq 30$ выборочное среднее распределено приближенно нормально с математическим ожиданием $m_x = \sigma$ и среднеквадратичной ошибкой.

$$\sigma_x = \frac{\sigma_x}{\sqrt{2f}}. \quad (2.31)$$

При $n \leq 30$ неизвестный генеральный стандарт в выражении (2.31) заменяют выборочным:

$$\sigma_x \approx \frac{s_x}{\sqrt{2f}}. \quad (2.32)$$

Доверительные границы для генерального стандарта определяются неравенством:

$$s_x - \frac{s_x}{\sqrt{2f}} u_{1-\frac{p}{2}} \leq \sigma_x \leq s_x + \frac{s_x}{\sqrt{2f}} u_{1-\frac{p}{2}}.$$

Сравнение двух дисперсий

При обработке наблюдений часто возникает необходимость сравнить две или несколько выборочных дисперсий. Основная гипотеза, которая при этом проверяется: можно ли считать сравниваемые выборочные дисперсии оценками одной и той же генеральной дисперсии.

Для двух выборок $x_1', x_2', \ldots, x_{n_1}'$ и $x_1'', x_2'', \ldots, x_{n_1}''$ средние значения соответственно равны $\overline{x_1}$ и $\overline{x_2}$. Выборочные дисперсии определяются по формулам:

$$s_1^2 = \frac{\sum_{i=1}^{n_1}(x_i' - \overline{x_1})^2}{n_1 - 1} \text{ и } s_2^2 = \frac{\sum_{i=1}^{n_2}(x_i'' - \overline{x_2})^2}{n_2 - 1}$$

со степенями свободы:

$$f_1 = n_1 - 1;\ f_2 = n_2 - 1.$$

Требуется определить, являются ли выборочные дисперсии s_1^2 и s_2^2 значимо различимыми или же полученные выборки можно рассматривать как взятые из генеральных совокупностей с равными дисперсиями. Предположим, что первая выборка сделана из генеральной совокупности с дисперсией σ_1^2, а вторая – из генеральной совокупности с дисперсией σ_2^2. Проверяется нулевая гипотеза о равенстве генеральных дисперсий H_0: $\sigma_1^2 = \sigma_2^2$. Чтобы отвергнуть эту гипотезу, нужно доказать значимость различия s_1^2 и s_2^2 при выбранном уровне значимости p. В качестве критерия значимости обычно используется критерий Фишера.

Распределением Фишера (F-распределением) называется распределение случайной величины:

$$F = \left(\frac{s_1^2}{\sigma_1^2}\right) : \left(\frac{s_2^2}{\sigma_2^2}\right). \quad (2.33)$$

Для определения квантилей F_p для значений p используется соотношение

$$F_p(f_1, f_2) = \frac{1}{F_{1-p}(f_2, f_1)}. \quad (2.34)$$

При доверительной вероятности $(1-p)$ двухсторонняя доверительная оценка величины F имеет вид:

$$F_{\frac{p}{2}}(f_1, f_2) \leq F \leq F_{1-\frac{p}{2}}(f_1, f_2).$$

С учетом (2.34)

$$\frac{1}{F_{1-\frac{p}{2}}(f_1, f_2)} \leq F \leq F_{1-\frac{p}{2}}(f_1, f_2). \quad (2.35)$$

В условиях нулевой гипотезы $F = \frac{s_1^2}{s_2^2}$, следовательно, с вероятностью $(1 - p)$ должно выполняться двухстороннее неравенство:

$$\frac{1}{F_{1-\frac{p}{2}}(f_1,f_2)} \leq \frac{s_1^2}{s_2^2} \leq F_{1-\frac{p}{2}}(f_1,f_2), \quad (2.36)$$

или одно из односторонних неравенств:

$$\frac{s_1^2}{s_2^2} \leq F_{1-\frac{p}{2}}(f_1,f_2) \text{ и } \frac{s_1^2}{s_2^2} \geq \frac{1}{F_{1-\frac{p}{2}}(f_1,f_2)}. \quad (2.37)$$

Вероятность неравенств, противоположных (2.36) и (2.37), равна уровню значимости p. Они образуют критическую область для нулевой гипотезы. Если полученное дисперсионное отношение попадает в критическую область, различие между дисперсиями считается значимым, т. е.

$$\frac{s_1^2}{s_2^2} > F_{1-p}(f_1,f_2). \quad (2.38)$$

Значения $F_{1-p}(f_1,f_2)$ определяются по таблице Приложения 2.

Критерий значимости (2.36) применяется, когда соотношение между генеральными дисперсиями неизвестно. При этом в неравенстве (2.36) нужно проверять только правую часть, так как левая часть всегда выполняется по условию:

$$\frac{s_1^2}{s_2^2} > 1, \text{а } \frac{1}{F_{1-\frac{p}{2}}(f_1,f_2)} < 1$$

для небольших p. При этом различие между дисперсиями следует считать значимым, если

$$\frac{s_1^2}{s_2^2} > F_{1-\frac{p}{2}}(f_1,f_2). \quad (2.39)$$

Сравнение нескольких дисперсий

При определении оценки дисперсии по текущим измерениям по формуле:

$$s_y^2 = \frac{f_1 s_1^2 + f_2 s_2^2 + \cdots + f_n s_n^2}{f_1 + f_2 + \cdots + f_n} \quad (2.40)$$

принимается нулевая гипотеза равенства соответствующих генеральных дисперсий. Проверить эту гипотезу для выборок разного объема можно по критерию Бартлета. В условиях нулевой гипотезы отношение $\frac{B}{C}$, где

$$B = 2{,}303\left(f \lg s_y^2 - \sum_{i=1}^{n} f_i \lg s_i^2\right); C = 1 + \frac{1}{3(n-1)} \sum_{i=1}^{n} \frac{1}{f_i} - \frac{1}{f}. \quad (2.41)$$

распределено приближенно как χ^2 с $(n-1)$ степенями свободы, если все $f_i > 2$. Гипотеза равенства генеральных дисперсий принимается, если

$$\frac{B}{C} \leq \chi^2_{1-p}. \quad (2.42)$$

при выбранном уровне значимости p. Различие между выборочными дисперсиями можно считать незначимыми, а сами выборочные дисперсии однородными. Так как всегда $C > 1$, если окажется $B \leq \chi^2_{1-p}$, нулевую гипотезу следует принять; если $B > \chi^2_{1-p}$, критерий Бартлета вычисляют полностью.

Сравнение двух средних

Для сравнения между собой двух средних, полученных по выборкам из нормально распределенных генеральных совокупностей, применяется критерий Стьюдента или t-критерий. Пусть имеются две случайные выборки: $x_1, x_2, \dots, x_{n_1}$ и $y_1, y_2, \dots, y_{n_2}$. Первая выборка взята из нормально распределенной генеральной совокупности с параметрами m_x и σ_x^2, вторая – из генеральной совокупности с параметрами m_y и σ_y^2. По выборкам получены оценки для параметров: $\bar{x}, s_x^2$ и $\bar{y}, s_y^2$. Требуется проверить нулевую гипотезу: $m_x = m_y$ при условии $\sigma_x^2 = \sigma_y^2 = \sigma^2$.

По свойству линейности нормального распределения случайная величина $z = \bar{x} - \bar{y}$ распределена нормально с параметрами: $m_z = m_x - m_y, \sigma_z^2 = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$. С учетом приведенных зависимостей нормированная функция имеет вид:

$$\frac{z - m_x}{\sigma_z} = \frac{(\bar{x} - \bar{y}) - (m_x - m_y)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \quad (2.43)$$

Если генеральный стандарт σ заменить выборочным, получится величина, имеющая распределение Стьюдента:

$$t = \frac{(\bar{x} - \bar{y}) - (m_x - m_y)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (2.44)$$

с числом степеней свободы $f = n_1 + n_2 - 2$.

Однородность выборочных дисперсий проверяется по критерию Фишера. При доверительной вероятности $\beta = 1 - p$ двусторонняя оценка для разности $m_x - m_y$имеет вид:

$$\bar{x} - \bar{y} - t_{1-\frac{p}{2}} s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq m_x - m_y \leq \bar{x} - \bar{y} + t_{1-\frac{p}{2}} s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (2.45)$$

или односторонние оценки:

$$m_x - m_y \leq \bar{x} - \bar{y} + t_{1-p} s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}};$$

$$m_x - m_y \geq \bar{x} - \bar{y} - t_{1-p} s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \quad (2.46)$$

В условиях нулевой гипотезы полученные неравенства дают критерий проверки этой гипотезы. Нулевая гипотеза отвергается при двухстороннем критерии, если

$$|\bar{x} - \bar{y}| \geq t_{1-\frac{p}{2}} s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (2.47)$$

и при одностороннем критерии, если:

$$|\bar{x} - \bar{y}| \geq t_{1-p} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \quad (2.48)$$

Приведенными критериями нельзя пользоваться, если генеральные дисперсии не равны между собой. Для этого случая существует несколько приближенных критериев для сравнения двух средних. При $n_1 = n_2 = n$ можно воспользоваться приближенным *t*-критерием:

$$t = \frac{(\bar{x}-\bar{y})\sqrt{n}}{\sqrt{s_x^2+s_y^2}} \quad (2.49)$$

с числом степеней свободы:

$$f = \frac{n-1}{c^2-(1-c)^2}, \; c = \frac{s_x^2}{s_x^2-s_y^2}.$$

Сравнение нескольких средних

При сравнении нескольких средних можно использовать множественный ранговый критерий Дункана. Пусть по *k* выборкам разного объема получено *k* средних значений:$\overline{x_1}$, $\overline{x_2}$, …, $\overline{x_j}$, … , $\overline{x_k}$:

$$\overline{x_j} = \frac{\sum_{i=1}^{n_j} x_{j_i}}{n_j}.$$

Генеральные дисперсии равны между собой. При применении критерия Дункана следует:

1) проранжировать *k* средних значений, расположив их в порядке возрастания;
2) определить ошибку воспроизводимости результатов s_x с соответствующим числом степеней свободы f_x;
3) определить ошибку для каждого среднего: $s_{\overline{x_j}} = \sqrt{\frac{s_x^2}{n_j}}$;
4) выписать из таблицы Дункана (приложение 6) $(k-1)$ значений рангов с выбранным уровнем значимости, числом $n_D = f_x$ и p = 2,3,…, k;
5) умножить эти значения рангов на $s_{\overline{x_j}}$ и определить $(k-1)$ наименьших значимых рангов;
6) проверить значимость различия между средними, начиная с крайних в ранжировочном ряду: разность максимального и минимального значений среднего сравнитьс наименьшим значимым рангом при $(p = k)$, затем найти разность максимального среднего и второго среднего в ранжировочном ряду и сравнить ее с наименьшим значимым рангом при $p = k - 1$ и т. д.;
7) сравнение продолжать для второго по величине среднего, которое сравнивается с наименьшим и т. д., пока не будут исследованы все значимые различия между всеми $\frac{k(k-1)}{2}$ парами.

Сравнение выборочного распределенияи распределения генеральной совокупности

Гипотеза о нормальности изучаемого распределения называется основной гипотезой. Проверку этой гипотезы по выборке проводят при помощи критериев согласия. Критерии согласия применяют для проверки гипотезы о предполагаемом виде закона распределения. Они позволяют определить вероятность того, что при гипотетическом законе распределения наблюдающееся в рассматриваемой выборке отклонение вызывается случайными причинами. Вероятностный характер критериев не позволяет однозначно принять или отвергнуть проверяемую гипотезу. Критерий позволяет утверждать, что гипотеза не противоречит опытным данным, если вероятность наблюдаемого отклонения от гипотетического закона велика, или что гипотеза не согласуется с опытными данными, если эта вероятность мала.

Чаще всего используется один из двух критериев согласия: *критерий Пирсона* (критерий χ^2) и *критерий Колмогорова*. Для применения критерия Пирсона диапазон изменения случайной величины в выборке объема n разбивается на k интервалов. Число интервалов берут в зависимости от объема выборки в пределах от 8 до 20. Число интервалов можно определить по формуле:

$$f = n - l,$$

где l – число связей, наложенных на выборку.

Число элементов выборки, попавших в i-й интервал, обозначим через n_i. Основанием для выбора закона распределения служит гистограмма выборочного распределения. Параметры этого закона могут быть определены или из теоретических соображений, или нахождением их оценок по выборке. На основании принятого закона распределения вычисляются вероятности p_i попадания случайной величины X в i-й интервал. Величина, характеризующая отклонение выборочного распределения от предполагаемого, определяется формулой:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \qquad (2.50)$$

где k – число интервалов, n – объем выборки.

Сумма в формуле (1.50) имеет приближенно χ^2 –распределение с $f = k - c - 1$ степенями свободы (c – число параметров гипотетического закона распределения распределения).

Вероятности попадания p_i значений случайной величины в интервал для нормального закона распределения определяются по формуле:

$$P(a \le X \le b) = \Phi\left(\frac{b-\bar{x}}{s}\right) - \Phi\left(\frac{a-\bar{x}}{s}\right). \qquad (2.51)$$

При определении вероятностей p_i нужно считать, что крайний левый интервал простирается до $-\infty$; крайний правый – до $+\infty$.

Для применении критерия согласия Колмогорова необходимо определить наибольшее абсолютное отклонение выборочной функции распределения $F_n(x)$ от генеральной $F(x)$:

$$D = \max|F_n(x) - F(x)|. \qquad (2.52)$$

Затем вычисляется величина λ:

$$\lambda = \sqrt{n}D. \qquad (2.53)$$

Если вычисленное значение λ меньше табличного λ_{1-p}, то гипотеза о совпадении теоретического закона распределения $F(x)$ с выборочным $F_n(x)$ не отвергается. При $\lambda \geq \lambda_{1-p}$гипотеза отклоняется. Уровень значимости при использовании критерия Колмогорова $p = 0{,}2 \div 0{,}3$.

Нормальное распределение определяется двумя параметрами: математическим ожиданием и дисперсией, через которые выражаются все остальные моменты нормального распределения. Выборочные коэффициенты эксцесса и асимметрии определяются по формулам:

$$\gamma_1^* = \frac{\mu_3^*}{s_x^3} = \frac{1}{ns_x^3}\sum_{i=1}^{n}(x_i - \bar{x})^3, \quad (2.54)$$

$$\gamma_2^* = \frac{\mu_4^*}{s_x^4} = \frac{1}{ns_x^4}\sum_{i=1}^{n}(x_i - \bar{x})^4 - 3. \quad (1.55)$$

Дисперсии этих величин рассчитывают по формулам:

$$D(\gamma_1^*) = \frac{6(n-1)}{(n+1)(n+3)}, \quad (2.56)$$

$$D(\gamma_2^*) = \frac{24(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}. \quad (2.57)$$

Зная дисперсии, $D(\gamma_1^*)$ и $D(\gamma_2^*)$, можно оценить, значимо ли выборочные коэффициенты асимметрии и эксцесса отличаются от нуля. Если:

$$|\gamma_1^*| \leq 3\sqrt{D(\gamma_1^*)}, \quad (2.58)$$

$$|\gamma_2^*| \leq 5\sqrt{D(\gamma_2^*)}, \quad (2.59)$$

то наблюдаемое распределение можно считать нормальным.

Примеры решения задач:

Задача 1 Оценить ошибку определения линейной скорости движения газа в трубопроводе v, пользуясь следующими результатами измерений: количество газа $G = 3000\ {}^{\text{м}^3}/_{\text{ч}}$; ошибка измерения $S_G = 10\ {}^{\text{м}^3}/_{\text{ч}}$; сечение трубопровода $F = 0{,}1\ \text{м}^2$; ошибка измерения $S_F = 1\text{см}^2$.

Решение. Рассмотрим линейную скорость как результат косвенного измерения, тогда:

$$v = \frac{G}{F} = \frac{3000}{0{,}1} = 30000\ {}^{\text{м}}/_{\text{ч}} = 8{,}82\ {}^{\text{м}}/_{\text{с}}.$$

Вычислим ошибку определения линейной скорости движения газа в трубопроводе:

$$s_v = \sqrt{\left(\frac{\partial v}{\partial G}\right)^2 s_G^2 + \left(\frac{\partial v}{\partial F}\right)^2 s_F^2} = \sqrt{\frac{1}{F^2}s_G^2 + \frac{G^2}{F^4}s_F^2} =$$

$$= \frac{\sqrt{1 \cdot 10^{-2} \cdot 10^2 + 9 \cdot 10^6 \cdot 10^{-8}}}{1 \cdot 10^{-2} \cdot 3600} = 0{,}03\ {}^{\text{м}}/_{\text{с}}.$$

Задача 2 Результаты определения концентрации (%) вещества А колориметрическим методом приведены в табл. 2.1. Определить ошибку метода измерения по текущим измерениям.

Таблица 2.1 - Исходные данные к задаче 2

Номер опыта	Номер пробы							
	1	2	3	4	5	6	7	8
1	27,9	19,3	4,5	22,3	10,8	16,3	8,8	12,6
2	27,2	19,7	5,2	23,5	8,9	15,8	8,7	13,5
3	26,8	-	4,8	21,7	-	17,2	9,2	13,3

Решение. Вычислим средние значения результатов эксперимента по столбцам:

$\overline{y_1} = \frac{27,9+27,2+26,8}{3} = \frac{81,9}{3} = 27,3;\ \overline{y_2} = \frac{39}{2} = 19,5;\ \overline{y_3} = \frac{14,5}{3} = 4,83;\overline{y_4} = 22,5;\ \overline{y_5} = 9,85;$ $\overline{y_6} = 16,43;\overline{y_7} = 8,9;\ \overline{y_8} = 13,1.$

Определим число параллельных опытов для каждой пробы:

$m_1 = 3, m_2 = 2, m_3 = 3, m_4 = 3, m_5 = 2, m_6 = m_7 = m_8 = 3.$

Рассчитаем частные дисперсии результатов измерений:

$s_1^2 = \frac{\sum_{u=1}^{m_i}(y_{u1}-\overline{y_1})^2}{m_1-1} = \frac{0,62}{3-1} = 0,31 \quad ; \quad s_2^2 = 0,08;\ s_3^2 = 0,123; s_4^2 = 0,84; s_5^2 = 1,805; s_6^2 =$ $0,503; s_7^2 = 0,07;\ s_8^2 = 0,22.$

Число степеней свободы общей дисперсии воспроизводимости равно:

$$f_{\text{воспр}} = \sum_{i=1}^{8} f_i = \sum_{i=1}^{8} m_i - 8 = 22 - 8 = 14.$$

Дисперсия воспроизводимости:

$$s^2_{\text{воспр}} = \frac{\sum_{i=1}^{8} f_i s_i^2}{f_{\text{воспр}}} = \frac{5,99}{14} = 0,4279.$$

Ошибка метода измерения, определяемая по текущим измерениям, составляет:

$$s_{\text{воспр}} = \sqrt{s^2_{\text{воспр}}} = \sqrt{0,4279} = 0,654.$$

Задача 3. Среднее значение температуры в печи, полученное по четырем независимым измерениям оптическим пирометром, 2250 °С. Ошибка при этом методе измерения 10 °С. Найти с надежностью 95% доверительные границы, внутри которых лежит истинное значение измеряемой температуры.

Решение. Полагая, что ошибка измерения известна $\sigma_x = 10$ °С и что случайная величина X (температура печи) распределена нормально, справедливо:

$$x_1 - \sigma u_{1-p/2} \ll m_x \ll x_1 + \sigma u_{1-p/2} = 2250 - k_\beta \frac{10}{\sqrt{4}} \ll m_x \ll 2250 + k_\beta \frac{10}{\sqrt{4}}.$$

При β = 95 % $k_\beta = 1{,}96$ и, следовательно, истинное значение измеряемой температуры находится с надежностью 95% в следующих доверительных границах:

$$2240{,}0 \ll m_x \ll 2259{,}8.$$

Задача 4. Основной примесью в фосфатах является фтор. Найти возможный предел содержания фтора в фосфатах по следующим результатам анализов в 100 кг готового продукта (*F*, %): 0,18; 0,12; 0,13; 0,15. Доверительная вероятность β = 0,95.

Решение. Обозначим через X результат анализа содержания фтора в 100 кг фосфатов. Среднее содержание фтора по четырем параллельным определениям равно $\bar{x} = 0{,}15\ \%$. Ошибка воспроизводимостиs_x = 0,03%. Число степеней свободы ошибки воспроизводимости равно 3. Для определения возможного верхнего предела содержания фтора в готовом продукте (m_x) воспользуемся формулой:

$$m_x \le \bar{x} + \frac{s_x}{\sqrt{n}} t_{1-p} = 0{,}15 + 0{,}03 \cdot \frac{2{,}35}{4} = 0{,}1676.$$

При β = 0,95 и *f*= 3 по Приложению 3 для 2*p* = 0,10 имеем $t_{0{,}95} = 2{,}35$.

Задача 5. Готовый продукт должен содержать не менее 99% основного вещества. Требуется определить гипотезу статистической значимости различия между данными технических условий и следующими результатами трех определений содержания основного вещества в готовом продукте: 98,3; 97,3; 97,8%.

Решение. Обозначим через X результат анализа. Среднее значение трех параллельных измерений равно $\bar{x} = 97{,}8\ \%$. Ошибка воспроизводимости s_x = 0,5%. Число степеней свободы ошибки воспроизводимости*f* = 2. В качестве нулевой гипотезы рассмотрим гипотезу H^0: m_x = 99%, т. е. исследуемый реактив доброкачественный. Альтернативная гипотеза $\overline{H}: m_x \neq 99\ \%$.Используя распределение Стьюдента, определимвначале критическую область при двухстороннем критерии. При β = 0,95 *p*= 0,05 и квантиль $t_{\frac{1-p}{2}} = 4{,}3$ при*f*= 2 (приложение 3). Критические значения нулевой гипотезы равны:

$$\bar{x} \le m_x - t_{1-\frac{p}{2}} s_x/\sqrt{n},\ \bar{x} \le m_x + t_{1-\frac{p}{2}} s_x/\sqrt{n}.$$

Физический смысл имеет только первое неравенство:

$$\bar{x} \le 99 - 4{,}3 \cdot \frac{0{,}5}{\sqrt{3}} = 97{,}76.$$

Значение $\bar{x} = 97{,}8$ не попадает в эту критическую область, следовательно, двухсторонний критерий не позволяет отвергнуть нулевую гипотезу и считать вещество недоброкачественным. По физическому смыслу задачи здесь можно применить односторонний критерий, поэтому выборочную оценку нужно сравнивать только с теми значениям, которые меньше 99%.

Приβ = 0,95и *f*= 2 по таблице. Приложения 4 для 2*p* = 0,10 имеем $t_{0{,}95} = 2{,}92$. Критическое значение нулевой гипотезы:

$$\bar{x} \le m_x - \frac{t_{1-\frac{p}{2}} s_x}{\sqrt{n}} = 99 - 2{,}92 \cdot \frac{0{,}5}{\sqrt{3}} = 98{,}16.$$

Значение $\bar{x} = 97{,}8$ меньше критического значения и, следовательно, попадает в критическую область. Таким образом, односторонний критерий, как правило, более точный сумел выявить при тех же исходных данных недоброкачественность вещества.

Задача6.Оценить ошибку воспроизводимости определения P_2O_5 в сложном соединении по результатам трех параллельных опытов: 17,2; 16,3; 15,5.

Решение. Выборочная оценка для дисперсии воспроизводимости равна:

$$s_x^2 = \frac{\sum_{i=1}^{3}(x_i - \bar{x})^2}{3-1} = 0{,}73.$$

Число степеней свободы дисперсии воспроизводимости $f_x = 2$. Задавшись доверительной вероятностью β = 0,9, по таблице Приложения 4при числе степеней свободы *f*= 2 находим $\chi^2_{0{,}05} = 6$ и $\chi^2_{0{,}95}$ =0,103. Определим двухстороннюю доверительную оценку для дисперсии воспроизводимости:

$$\frac{f s_x^2}{\chi^2_{1-\frac{p}{2}}} \le \sigma_x^2 \le \frac{f s_x^2}{\chi^2_{\frac{p}{2}}};$$

$$\frac{0{,}73 \cdot 2}{6} \le \sigma_x^2 \le \frac{0{,}73 \cdot 6}{0{,}103};$$

$$0{,}24 \le \sigma_x^2 \le 14{,}1.$$

Извлекая из всех частей неравенства квадратный корень, получим оценку для ошибки воспроизводимости $0{,}49 \le \sigma_x \le 3{,}61$. В связи с малым числом степеней свободы доверительные границы получились резко асимметричными.

Задача 7.Оценить ошибку воспроизводимости σ_x для выборки из 31 наблюдения с выборочным стандартом $s_x = 0{,}85$. Доверительную вероятность β принять равной0,9.

Решение. Построим доверительный интервал для ошибки воспроизводимости, используя χ^2 распределение. По таблице Приложения 4 при числе степеней свободы *f*= 30 и доверительной вероятности β = 0,9находим $\chi^2_{0{,}05} = 43{,}8$ и $\chi^2_{0{,}95} = 18{,}5$. Определим двухстороннюю доверительную оценку для дисперсии воспроизводимости

$$\frac{f s_x^2}{\chi^2_{1-\frac{p}{2}}} \le \sigma_x^2 \le \frac{f s_x^2}{\chi^2_{\frac{p}{2}}};$$

$$\frac{0{,}85^2 \cdot 30}{43{,}8} \le \sigma_x^2 \le \frac{0{,}85^2 \cdot 30}{18{,}5};$$

$$0{,}48 \le \sigma_x^2 \le 1{,}13.$$

Доверительные границы для ошибки воспроизводимости определяются неравенством $0{,}69 \leq \sigma_x \leq 1{,}05$. Определим доверительные границы для σ_x , воспользовавшись нормальным распределением.

$$\sigma_x \approx \frac{s_x}{\sqrt{2f}} = \frac{0{,}85}{\sqrt{2 \cdot 30}} = 0{,}11.$$

По таблице Приложения 2 для доверительной вероятностиβ = 0,9 находим $u_{0{,}95}$ =1,64. Доверительные границы для ошибки воспроизводимости определяются неравенством:

$$s_x - \frac{s_x}{\sqrt{2f}} u_{1-\frac{p}{2}} \leq \sigma_x \leq s_x + \frac{s_x}{\sqrt{2f}} u_{1-\frac{p}{2}}$$

$$0{,}85 - 0{,}11 \cdot 1{,}64 \leq \sigma_x^2 \leq 0{,}85 + 0{,}11 \cdot 1{,}64;$$

$0{,}67 \leq \sigma_x \leq 1{,}03.$

Полученные с использованием нормального распределения доверительные границы мало отличаются от приведенных выше.

Задача 8.При оценке точности определения содержания вещества *A*первым методом дисперсия воспроизводимости составила $s_1^2 = 0{,}73$; $f = 2$. Требуется сравнить этот метод с более точным вторым методом по результатам четырех параллельных определений содержания вещества *A*: 16,5; 15,9; 16,6; 15,8.

Решение. Дисперсия воспроизводимости второго метода:

$$s_2^2 = \frac{\sum_{i=1}^{4}(x_i - \bar{x})^2}{4 - 1} = 0{,}16$$

при числе степеней свободы $f_2 = 3$. По условиям задачи для оценки значимости различия между дисперсиями s_1^2 и s_2^2 можно использовать односторонний критерий значимости. Дисперсионное отношение $F = \frac{0{,}73}{0{,}16} = 4{,}5$ надо сравнить с табличным для уровня значимости $p = 0{,}05$ и чисел степеней свободы $f_1 = 2$ и $f_2 = 3F_{1-p}(f_1, f_2) = 9{,}6$. Таким образом, выборочное дисперсионное отношение меньше табличного и данные опытов не позволяют считать точность методов значимо различной.

Задача 9. При получении вещества *B* измерялась его степень восстановления при четырех различных температурах. Определить, меняется ли точность анализа с температурой? В табл. 2.2. приведены результаты статистического анализа однородности дисперсий воспроизводимости результатов при разных температурах.

Таблица 2.2 - Результаты статистического анализа однородности дисперсий вопроизводимости

Значение температуры	s_i^2	f_i^2	s_i^2	$\lg s_i^2$	$f_i \lg s_i^2$	$\frac{1}{f_i}$
T_1	1,72	5	8,60	0,2355	1,177	0,200
T_2	1,60	4	6,40	0,2041	0,816	0,250
T_3	1,97	6	11,82	0,2945	1,767	0,167
T_4	2,37	8	18,96	0,3747	2,995	0,125
Σ		23	45,78		6,755	0,735

Решение. Определяем дисперсию воспроизводимости:

$$s_y^2 = \frac{\sum_{i=1}^{4} f_i s_i^2}{f_1 + f_2 + f_3 + f_4} = \frac{45{,}78}{23} = 1{,}99.$$

В соответствии с критерием Бартлета:

$$B = 2{,}303\left(f \lg s_y^2 - \sum_{i=1}^{n} f_i \lg s_i^2\right) = 2{,}303(23 \cdot 0{,}2988 - 6{,}755) = 27;$$

$$C = 1 + \frac{1}{3(n-1)}\sum_{i=1}^{n}\frac{1}{f_i} - \frac{1}{f} = 1 + \frac{1}{3(4-1)}(0{,}735 - 0{,}043) = 1{,}077.$$

По таблице приложения 4 находим при трех степенях свободы и уровне значимости $p = 0{,}05$, $\chi^2_{0,95} = 7{,}8$, чтовеличина $B < \chi^2_{0,95}$ и, значит на уровне значимости $p = 0{,}05$ можно принять гипотезу о равенстве генеральных дисперсий. Величину C можно было не вычислять. Критерий Бартлета позволяет считать, что точность анализа не зависит от температуры. Выборочные дисперсии однородны, поэтому в качестве оценки для дисперсии воспроизводимости можно взять средневзвешенную дисперсию s_y^2 с числом степеней свободы f, равным 23.

Контрольные вопросы:

1. Классификация методов анализа данных. Этапы анализа данных: выявление закономерностей, прогнозирование, анализ исключений.
2. Равномерное распределение случайных величин, его свойства. Числовые характеристики. Область применения.
3. Нормальное (гауссово) распределение распределения случайных величин, его свойства. Числовые характеристики. Область применения. Функция Лапласа, ее свойства.
4. Распределение Стьюдента, его свойства. Числовые характеристики. Область применения.
5. Распределение «хи-квадрат», его свойства. Числовые характеристики. Область применения.
6. Экспоненциальное распределение, его свойства. Числовые характеристики. Область применения.
7. Распределение Фишера. Его свойства. Числовые характеристики. Область применения.

Контрольные вопросы

8. Что называют генеральной совокупностью?
9. Что такое выборка (выборочная совокупность)? Что называют объёмом выборки?
10. Напишите формулы для вычисления основных выборочных характеристик: среднего, дисперсии, ковариации, коэффициента корреляции.
11. Напишите формулы точечных оценок ковариации и коэффициента корреляции.
12. Что называют доверительной вероятностью и доверительным интервалом для неизвестного параметра θ?
13. Что такое статистическая гипотеза? Какие статистические гипотезы называют: основными или альтернативными, сложными или простыми?
14. Что называют статистическим критерием и его уровнем значимости при проверке статистической гипотезы?

15. Какие статистические критерии называют критериями значимости различий?

16. Для проверки каких гипотез используются одно- и двухвыборочные t-критерии Стьюдента? Какую статистику используют эти критерии?

17. Для проверки каких гипотез используется Фишера -критерий значимости различий? Какую статистику использует этот критерий?

18. Какие задачи являются объектом исследования в дисперсионном анализе?

25. В каком случае дисперсионный анализ называют одно- и многофакторным? 26. Каковы предпосылки однофакторного дисперсионного анализа? 27. Как формулируются основная и альтернативная гипотезы однофакторного дисперсионного анализа?

26. Какую зависимость называют регрессионной? В чем отличие регрессионной зависимости от функциональной?

27. Как формулируется задача регрессионного анализа? Из каких соображений выбирается форма регрессионной зависимости?

28. Какой вид имеет линейная регрессионная модель? Как называются переменные, представленные в модели?

3. Элементы дисперсионного анализа

Однофакторный дисперсионный анализ

Рассмотрим действие единичного фактора *A* (количественного или качественного), который принимает *k* различных значений (уровней фактора). На *i*-м уровне производится n_i наблюдений, результаты которых можно записать следующим образом:

$$\begin{bmatrix} y_{11} & y_{21} & \cdots & y_{k1} \\ y_{12} & y_{22} & \cdots & y_{k2} \\ \cdots & \cdots & \cdots & \cdots \\ y_{1n_1} & y_{1n_2} & \cdots & y_{1n_k} \end{bmatrix}.$$

Будем полагать, что результат любого наблюдения можно представить в виде модели:

$y_{ij} = \mu + d_i + \varepsilon_{ij}$,

где μ – суммарный эффект во всех опытах, d_i - эффект фактора *A* на *i*-м уровне,ε_{ij} – ошибка измерения на *i*-м уровне.

Положим также, что наблюдения на фиксированном уровне фактора нормально распределены относительно среднего значения $\mu + d_i$ с общей дисперсией σ^2. Общее число опытовопределяется по формуле:

$N = n_1 + n_2 + \cdots + n_k$.

Проверяется нулевая гипотеза равенства средних значений на различных уровнях фактора *A*:

$$m_1 = m_2 = \cdots = m_k = m.$$

Наиболее простые расчеты получаются при равном числе опытов на каждом уровне фактора *A*: $n_1 = n_2 = \cdots = n_k = n$. При этом число наблюдений *N* равно *kn*.

Обозначим через $\overline{y_i}$ среднее значение наблюдений на *i*-м уровне:

$\overline{y_i} = \frac{\sum_{j=1}^{n} y_{ij}}{n}$,

а общее среднее значение для всей выборки из *N* наблюдений:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} = \frac{1}{k}\sum_{i=1}^{k} \overline{y_i}.$$

Для проведения дисперсионного анализа необходимо определить общую выборочную дисперсию:

$s^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y})^2}{N-1}$,

которую затем следует разложить на составляющие, которые характеризовали бы вклад фактора *A* и фактора случайности. Фактор случайности можно оценить благодаря наличию повторных опытов на каждом уровне. Определим выборочную дисперсию на каждом уровне:

$s_i^2 = \frac{\sum_{j=1}^{n}(y_{ij}-\overline{y_i})^2}{n-1}$,*i*=1,2,…,*k* .

Если между дисперсиями нет значимых различий, для оценки генеральной дисперсии σ^2, характеризующей факторслучайности, используем выборочную дисперсию:

$$s_{\text{ош}}^2 = \frac{1}{k}\sum_{i=1}^{k} {s_i}^2.$$

Число степеней свободы дисперсии$s_{\text{ош}}^2$ равно$k(n-1) = N-k$. Приближенную оценку для дисперсии фактора *A* можно получить так:

$$\sigma_{\text{A}}^2 \approx s^2 - s_{\text{ош}}^2.$$

Более точную оценку для σ_{A}^2 можно получить, рассматривая отклонения средних $\bar{y}_i$на отдельных уровнях от общего среднего всей выборки $\bar{y}$:

$$\sigma_{\text{A}}^2 \approx \frac{1}{k-1}\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2 - \frac{s_{\text{ош}}^2}{n}.$$

Дисперсия фактора *A* для модели с фиксированными уровнями σ_{A}^2 не связана ни с какой случайной величиной, это условное название для математического ожидания среднего квадрата отклонений, обусловленного влиянием фактора *A*.

Введем следующее обозначение:

$$s_{\text{A}}^2 = \frac{n}{k-1}\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2 \approx n\sigma_{\text{A}}^2 + s_{\text{ош}}^2.$$

Эта дисперсия имеет *k*-1 степеней свободы. Если дисперсия s_{A}^2значимо отличается от$s_{\text{ош}}^2$, нулевая гипотеза$m_1 = m_2 = \cdots = m_k = m$отвергается и влияние фактора *A* считается существенным. Проверяется нулевая гипотеза по критерию Фишера. Так как альтернативой $\sigma_{\text{A}}^2 = s_{\text{ош}}^2$ является неравенство $\sigma_{\text{A}}^2 > s_{\text{ош}}^2$, для проверки гипотезы применяется односторонний критерий Фишера. Влияние фактора *A* считается значимым, если

$$\frac{s_{\text{A}}^2}{s_{\text{ош}}^2} > F_{1-p}(f_1, f_2), f_1 = k-1, f_2 = k(n-1) = N-k.$$

Дисперсионный анализ можно провести, рассчитывая последовательно:

1) суммы по столбцам:

$$A_i = \sum_{j=1}^{n} y_{ji},$$

2) сумму квадратов всех наблюдений:

$$SS_1 = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij})^2,$$

3) сумму квадратов по столбцам, деленную на число наблюдений в столбце:

$$SS_2 = \frac{1}{n}\sum_{i=1}^{k} {A_i}^2,$$

4) квадрат общего итога, деленный на число всех наблюдений:

$$SS_3 = \frac{1}{N}\left(\sum_{i=1}^{k} A_i\right)^2,$$

5) сумму квадратов каждого столбца:

$$SS_{\text{A}} = SS_2 - SS_3,$$

6) общую сумму квадратов:

$SS_{\text{общ}} = SS_1 - SS_3$,

7) остаточную сумму квадратов

$SS_{\text{ост}} = SS_1 - SS_2$,

8) дисперсию s_{A}^2:
$s_{\text{A}}^2 = \frac{SS_{\text{A}}}{k-1}$,

9) дисперсию $s_{\text{ош}}^2$:
$s_{\text{ош}}^2 = \frac{SS_{\text{ост}}}{k(n-1)}$.

Если отношение $\frac{s_{\text{A}}^2}{s_{\text{ош}}^2} \leq F_{1-p}$, влияние фактора *A* следует считать незначимым. При этом общая дисперсияs^2 связана только с фактором случайности и может служить оценкой для дисперсии воспроизводимости. Такая оценка лучше, так как имеет большее число степеней свободы. При интерпретации результатов дисперсионного анализа надо учитывать, что низкое значение дисперсионного отношения может быть связано с тем, что влияние какого-либо важного неконтролируемого фактора не было рандомизировано. Это может увеличить дисперсию внутри уровней, а дисперсию между уровнями оставить неизменной, что уменьшает дисперсионное отношение.

Если отношение$\frac{s_{\text{A}}^2}{s_{\text{ош}}^2} > F_{1-p}$, различие между дисперсиями значимо и, следовательно, значимо влияние фактора *A*.

Определим оценку влияния фактора *A*:

$\sigma_{\text{A}}^2 = \frac{s_{\text{A}}^2 - s_{\text{ош}}^2}{n}$.

При этом нулевая гипотеза $m_1 = m_2 = \cdots = m_k = m$ отвергается и различие между средними $m_1, m_2, \ldots m_k$ следует считать значимым.

В отличие от модели с фиксированными уровнями выводы по случайной модели распространяются на всю генеральную совокупность уровней. Рассмотрим схему вычислений для разного числа параллельных опытов. Пусть на уровне a_i проведено n_i число параллельных наблюдений. Общее число наблюдений равно:

$N = \sum_{i=1}^{k} n_i$.

Рассчитываем:

1) итоги по столбцам:
$A_i = \sum_{j=1}^{n} y_{ji}$,

2) сумму квадратов всех наблюдений:

$SS_1 = \sum_{i=1}^{k} \sum_{j=1}^{n_j} (y_{ij})^2$,

3) сумму итогов по столбцам, деленных на число наблюдений в соответствующем столбце:

$$SS_2 = \sum_{i=1}^{k} \frac{{A_i}^2}{n_i},$$

4) квадрат общего итога, деленный на число всех наблюдений:

$$SS_3 = \frac{1}{N}\left(\sum_{i=1}^{k} A_i\right)^2.$$

Дальнейшие расчеты будут проводиться по известным зависимостям. Если дисперсии $s_{\text{А}}^2$и $s_{\text{ош}}^2$ значимо отличаются друг от друга, дисперсию фактора вычисляют по формуле:

$$\sigma_{\text{А}}^2 \approx \frac{(k-1)N}{N^2-\sum_{i=1}^{k} n_i}(s_A^2 - s_{\text{ош}}^2).$$

Двухфакторный дисперсионный анализ

Изучается влияние на процесс одновременно двух факторов *A* и *B*. Фактор *A* исследуется на уровнях a_1, a_2, …, a_{k}, фактор *B* – на уровнях b_1, b_2, …, b_{m}.Допустим, что при каждом сочетании уровней факторов *A* и *B* проводится*n*параллельных наблюдений. Общее число наблюдений равно $N = nkm$. Результат наблюдения можно представить в виде следующей модели:

$$y_{ijq} = \mu + a_i + b_j + a_i b_j + \varepsilon_{ijq},$$

где μ – общее среднее, a_i – эффект фактора А на i-м уровне, i=1, 2, …, k; b_j – эффект фактора Вна j- м уровне, j = 1, 2,…, m;$a_i b_j$ – эффект взаимодействия факторов.

Эффект взаимодействия представляет собой отклонение среднего по наблюдениям в (ij) –й серии от суммы первых трех членов в модели, а ε_{ijq} (q = 1,2,…,n) и учитывает вариацию внутри серии наблюдений (ошибка воспроизводимости). Будем полагать, чтоε_{ijq} распределена нормально с нулевым математическим ожиданием и дисперсией $\sigma_{\text{ош}}^2$.

Если предположить,что между факторами нет взаимодействия, то можно принять линейную модель:

$$y_{ij} = \mu + a_i + b_j + a_i b_j + \varepsilon_{ij}.$$

Эта модель применяется при отсутствии параллельных опытов. Рассмотрим линейную модель. Через $\bar{y}_i$ и $\bar{y}'_i$обозначим средние, соответственно, по столбцам и строкам:

$$\bar{y}_i = \frac{A_i}{m}, \bar{y}'_j = \frac{B_j}{k},$$

через $\bar{\bar{y}}$ – среднее всех результатов:

$$\bar{\bar{y}} = \frac{1}{km}\sum_{i=1}^{k}\sum_{j=1}^{m} y_{ij}.$$

Рассеяние в средних по столбцам $\overline{y_1}, \overline{y_2}, \ldots, \overline{y_k}$ относительно общего среднего. Это рассеяние связано с влиянием фактора *A* и случайного фактора. Так какдисперсия среднего в *m* раз меньше дисперсии единичного измерения, справедливо:

$$\frac{1}{k-1}\sum_{i=1}^{k}(\overline{y_i}-\overline{\overline{y}})^2 \approx \sigma_A^2+\frac{\sigma^2}{m}.$$

В средних по строчкам не зависит от фактора *A* и связано с влиянием фактора *B*:

$$\frac{1}{m-1}\sum_{j=0}^{m}\left(\overline{y_j'}-\overline{\overline{y}}\right)^2 \approx \sigma_B^2+\frac{\sigma^2}{k}.$$

Для оценки фактора случайности (при отсутствии параллельных опытов) найдем дисперсию наблюдений по *i*-му столбцу:

$$s_i^2=\frac{1}{m-1}\sum_{j=1}^{m}\left(y_{ij}-\overline{y_i}\right)^2.$$

Эта дисперсия обусловлена влиянием фактора B и фактора случайности $s_i^2 \approx \sigma_B^2+\sigma^2$. Равенство станет точным, если вместо s_i^2 использоватьсредневзвешенную дисперсию по всем столбцам:

$$\sigma^2 =\frac{1}{(k-1)(m-1)}\left[\sum_{i=1}^{k}\sum_{j=1}^{m}\left(y_{ij}-\overline{y_i}\right)^2-k\sum_{j=1}^{m}\left(\overline{y_j'}-\overline{\overline{y}}\right)^2.\right]$$

Обозначим полученную оценку для дисперсии σ^2 через$s_{\text{ош}}^2$. Число степеней свободы $s_{\text{ош}}^2$ равно $(k-1)(m-1)$. Введем обозначения:

$s_A^2=\frac{m}{k-1}\sum_{i=1}^{k}(\overline{y_i}-\overline{\overline{y}})^2 \approx m\sigma_A^2+s_{\text{ош}}^2$,

$s_A^2=\frac{m}{k-1}\sum_{i=1}^{k}(\overline{y_i}-\overline{\overline{y}})^2 \approx m\sigma_A^2+s_{\text{ош}}^2$,

$$s_B^2=\frac{m}{k-1}\sum_{j=1}^{km}(\overline{y_i}'-\overline{\overline{y}})^2 \approx k\sigma_B^2+s_{\text{ош}}^2.$$

Величины s_A^2 и s_B^2 можно считать свободными дисперсиями с $(k-1)$и $(m-1)$ степенями свободы соответственно. Проверяют нулевые гипотезы о незначимости влияния A и B по критерию Фишера. Если дисперсионное отношение

$$F=\frac{s_{\text{A}}^2}{s_{\text{ош}}^2}<F_{1-p}(f_1,f_2), f_1=k-1, f_2=(k-1)(m-1),$$

то принимается гипотеза H^0: $\alpha_i=0$. Если

$$F=\frac{s_{\text{A}}^2}{s_{\text{ош}}^2}>F_{1-p}(f_1,f_2),$$

то нулевая гипотеза отвергается и влияние фактора *A* считается значимым. Аналогично, если

$$F = \frac{s_B^2}{s_{ош}^2} < F_{1-p}(f_1, f_2), f_1 = m - 1, f_2 = (k-1)(m-1),$$

то принимается гипотеза H^0: $\beta_i = 0$. При справедливости

$$F = \frac{s_B^2}{s_{ош}^2} > F_{1-p}(f_1, f_2)$$

влияние фактора *B* считается значимым. При проверке нулевых гипотез применяется односторонний критерий Фишера. При проведении дисперсионного анализа в условиях линейной модели можно использовать следующий алгоритм расчета:

Находят:

1) итоги по столбцам:

$$A_i = \sum_{j=1}^{m} y_{ij}, i = 1,2,\dots,k;$$

2) итоги по строкам:

$$B_i = \sum_{i=1}^{k} y_{ij}, j = 1,2,\dots,m;$$

3) сумму квадратов всех наблюдений:

$$SS_1 = \sum_{i=1}^{k}\sum_{j=1}^{m}(y_{ij})^2,$$

4) сумму квадратовитогов по столбцам, деленную на число наблюдений в столбце:

$$SS_2 = \frac{1}{m}\sum_{i=1}^{k} {A_i}^2,$$

5) сумму квадратов общего итога по строкам, деленную на число наблюдений в строке:

$$SS_3 = \frac{1}{k}\sum_{j=1}^{m} B_j^2,$$

6) квадрат общего итога, деленный на число всех наблюдений:

$$SS_4 = \frac{1}{mk}\left(\sum_{j=1}^{m} B_j^2\right)^2 = \frac{1}{mk}\left(\sum_{i=1}^{k} A_j^2\right)^2;$$

7) сумму квадратов для столбца:

$$SS_A = SS_2 - SS_4,$$

8) сумму квадратов для строки:

$$SS_B = SS_3 - SS_4,$$

9) общую сумму квадратов:

$SS_{\text{общ}} = SS_1 + SS_4$,

10) остаточную сумму квадратов

$SS_{\text{ост}} = SS_{\text{общ}} - SS_{\text{A}} - SS_{\text{B}}$,

11) дисперсию s_A^2

$s_A^2 = \frac{SS_{\text{A}}}{k-1}$,

12) дисперсию s_B^2

$$s_B^2 = \frac{SS_B}{m-1},$$

13) дисперсию $s_{\text{ош}}^2$

$s_{\text{ош}}^2 = \frac{SS_{\text{ост}}}{(k-1)(m-1)}$.

Установив при помощи дисперсионного анализа значимость влияния данного фактора, выясняют затем при помощи критерия Стьюдента или рангового критерия Дункана, какие именно средние значения у различны. Линейная модель справедлива, если между факторами *A* и *B* нет взаимодействия. В противном случае этому взаимодействию как фактору присуща своя дисперсия σ_{AB}^2. Взаимодействие *AB*, σ_{AB}^2 служит мерой того, насколько влияние фактора *A* зависит от уровня фактора *B* и наоборот.

Пусть при каждом сочетании уровней факторов *A* и *B* проводится *n* параллельных опытов. Тогда выборочная дисперсия результатов определяется по формуле:

$s_{ij} = \frac{1}{n-1}\sum_{u=1}^{n}\left(y_{iju} - \overline{y_{ij}}\right)^2$.

Если выборочные дисперсии однородны, то их можно усреднить:

$$s_{\text{ош}}^2 = \frac{1}{mk}\sum_{i=1}^{k}\sum_{j=1}^{m} s_{ij}^2$$

и использовать в качестве оценки для дисперсии воспроизводимости σ^2. Число степеней свободы $s_{\text{ош}}^2$ равно $mk(n-1)$.Более удобная формула для расчета дисперсии воспроизводимости имеет вид:

$$s_{\text{ош}}^2 = \frac{\sum_{j=1}^{k}\sum_{i=1}^{m}\sum_{u=1}^{n} y_{iju}^2 - \frac{\sum_{j=1}^{k}\sum_{i=1}^{m} y_{ij}^2}{n}}{mk(n-1)}.$$

При проведении дисперсионного анализа для нелинейной модели целесообразно использовать следующий алгоритм расчета.

1. Исходные данные представить в виде табл. 3.1.

Таблица3.1 – Данные для двухфакторного дисперсионного анализа с повторениями

							Итоги
	a_1	a_2	…	a_i	…	a_k	
b_1	$y_{111}, y_{112}, \dots, y_{11n}$	$y_{211}, y_{212}, \dots, y_{21n}$	…	$y_{i11}, y_{i12}, \dots, y_{i1n}$	…	$y_{k11}, y_{k12}, \dots, y_{k1n}$	B_1
b_2	$y_{121}, y_{122}, \dots, y_{12n}$	$y_{221}, y_{222}, \dots, y_{22n}$	…	$y_{i21}, y_{i22}, \dots, y_{i2n}$	…	$y_{k21}, y_{k22}, \dots, y_{k2n}$	B_2
b_j	$y_{1j1}, y_{1j2}, \dots, y_{1jn}$	$y_{2j1}, y_{2j2}, \dots, y_{2jn}$	…	$y_{ij1}, y_{ij2}, \dots, y_{ijn}$	…	$y_{kj1}, y_{kj2}, \dots, y_{kjn}$	B_j
…	…	…	…	…	…	…	…
b_m	$y_{1m1}, y_{1m2}, \dots, y_{1mn}$	$y_{2m1}, y_{2m2}, \dots, y_{2mn}$	…	$y_{im1}, y_{im2}, \dots, y_{imn}$	…	$y_{km1}, y_{km2}, \dots, y_{kmn}$	B_m
Итоги	A_1	A_2	…	A_i	…	A_k	

По таблице 3.1. находим:

1. Суммы наблюдений в каждой ячейке:

$$y_{ij} = \sum_{u=1}^{n} y_{iju}, j = 1,2,\dots,m; i = 1,2,\dots,k;$$

2. Квадрат суммы наблюдений в каждой ячейке:

$$y_{ij}^2 = \left(\sum_{u=1}^{n} y_{iju}\right)^2;$$

3. Итоги по столбцам:

$$A_i = \sum_{j=1}^{m}\sum_{u=1}^{n} y_{iju};$$

4. Итоги по строкам:

$$B_j = \sum_{i=1}^{k}\sum_{u=1}^{n} y_{iju};$$

5. Сумму всех наблюдений:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{u=1}^{n} y_{iju} = \sum_{i=1}^{k} A_i = \sum_{j=1}^{m} B_j;$$

6. Сумму квадратов всех наблюдений:

$$SS_1 = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{u=1}^{n} {y_{iju}}^2;$$

7. Сумму квадратов итогов по столбцам, деленную на число наблюдений в столбце:

$$SS_2 = \frac{1}{mn}\sum_{i=1}^{k} \mathrm{A}_i^2.$$

8. Сумму квадратов итогов по строкам, деленную на число наблюдений в строке:

$SS_3 = \frac{1}{kn}\sum_{j=1}^{m} B_j^2$;

9. Квадрат общего итога, деленный на число всех наблюдений:

$$SS_4 = \frac{1}{mkn}\left(\sum_{i=1}^{k} A_i^2\right)^2 = \frac{1}{mkn}\left(\sum_{j=1}^{m} B_j^2\right)^2$$

10. Сумму квадратов для столбца:

$SS_A = SS_2 - SS_4$.

Примеры решения задач

Модель с фиксированным уровнем факторов

Пусть ставится задача исследования влияния на выход продукта B нескольких видов реагента A. Виды реагента A естественно назвать уровнями фактора $A_i (i = 1, \dots, m)$. Здесьm полное число применяемых видов реагентов, например, m = 5.

Обозначим y_{ij} – выход продукта, получаемый в j-м наблюдении при использовании i-го вида реагента.Сведем данные в табл. 3.2.

Таблица 3.2 – Исходные данные к задаче

Номер наблюдения	Уровень фактора А				
	a_1	a_2	a_3	a_4	a_5
1	79,80	87,30	42,45	76,0	70,70
2	86,30	69,60	64,30	83,5	64,65
3	86,50	81,75	78,9	72,80	38,50
4	92,30	77,95	61,00	89,00	77,00
5	76,50	83,65	31,30	76,50	91,50
6	87,05	64,80	72,85	87,45	68,00
7	82,50	67,30	58,65	74,50	38,05
8	90,00	75,45	52,50	93,15	79,95

Определяем суммы по столбцам:

$A1_1 = 680{,}95; A1_2 = 607{,}8; A1_3 = 461{,}7; A1_4 = 652{,}9; A1_5 = 528{,}35.$

Найдем средние значения выхода для каждого вида реагента:

$$\overline{y_1} = \frac{A1_1}{8} = \frac{680{,}95}{8} = 85{,}1; \overline{y_2} = \frac{A1_2}{8} = \frac{607{,}8}{8} = 75{,}97;$$
$$\overline{y_3} = \frac{A1_3}{8} = \frac{461{,}7}{8} = 57{,}74;$$
$$\overline{y_4} = \frac{A1_4}{8} = \frac{652{,}9}{8} = 81{,}61; \overline{y_2} = \frac{A1_5}{8} = \frac{528{,}35}{8} = 66{,}04.$$

Рассчитываем общее среднее для всех результатов:

$$\bar{\bar{y}} = \frac{1}{5}\sum_{i=1}^{5} \overline{y_i} = \frac{366{,}46}{5} = 73{,}2 \approx 73.$$

Для упрощения вычислений будем рассматривать вместо значений *y* отклонения этих значений от величины, близкой к общему среднему значению (табл. 3.3.):

Таблица 3.3

Номер наблюдения	Уровень фактора *A*				
	a_1	a_2	a_3	a_4	a_5
1	6,8	14,3	-30,55	3,0	-2,3
2	13,3	-3,4	-8,7	10,5	-8,35
3	13,5	8,75	5,9	-0,2	-34,5
4	19,3	4,95	-12,0	16,0	4,0
5	3,5	10,65	-41,7	3,5	18.5
6	14,05	-8,2	-0,15	14,45	-5,0
7	9,5	-5,7	-14,35	1,5	-34,95
8	17,0	2,45	-20,5	20,15	6,95

Определяем суммы по столбцам:

$A_1 = 96{,}95; A_2 = 23{,}8; A_3 = -122{,}05; A_4 = 68{,}9; A_5 = -55{,}65.$

Найдем средние значения выхода для каждого вида реагента:

$$\overline{y_1} = \frac{A_1}{8} = \frac{96{,}95}{8} = 12{,}11; \overline{y_2} = \frac{A_2}{8} = \frac{23{,}8}{8} = 2{,}97;$$
$$\overline{y_3} = \frac{A_3}{8} = \frac{-122.05}{8} = -15{,}25;$$
$$\overline{y_4} = \frac{A_4}{8} = \frac{68{,}9}{8} = 8{,}61; \overline{y_5} = \frac{A_5}{8} = \frac{-55{,}65}{8} = -6{,}95.$$

Рассчитываем общее среднее для всех результатов:

$$\bar{\bar{y}} = \frac{1}{5}\sum_{i=1}^{5} \overline{y_i} = \frac{1{,}494}{5} = 0{,}299.$$

Найдем сумму квадратов всех наблюдений:

$$SS_1 = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij})^2 = 9379.$$

Определим сумму итогов по столбцам, деленных на число наблюдений в соответствующем столбце:

$$SS_2 = \frac{1}{n}\sum_{i=1}^{k} A_i{}^2 = \frac{1}{8}\cdot(96{,}95^2 + 23{,}8^2 + \cdots + (-55{,}65)^2) = 4088.$$

Рассчитаем квадрат общего итога, деленный на число всех наблюдений:

$$SS_3 = \frac{1}{N}\left(\sum_{i=1}^{k} A_i\right)^2 = \frac{1}{5\cdot 8}\cdot 142{,}802 = 3{,}57.$$

Определим сумму квадратов для столбца:

$$SS_A = SS_2 - SS_3 = 4088 - 3{,}57 = 4085.$$

Найдем общую сумму квадратов:

$$SS_{общ} = SS_1 - SS_3 = 9379 - 3{,}57 = 9375$$

и остаточную сумму квадратов:

$$SS_{\text{ост}} = SS_1 - SS_2 = 9379 - 4088 = 5291.$$

Дисперсия $s_{\text{А}}^2$ составляет:

$s_{\text{А}}^2 = \frac{SS_{\text{А}}}{k-1} = \frac{4085}{5-1} = 1021,$
дисперсия $s_{\text{ош}}^2$:
$s_{\text{ош}}^2 = \frac{SS_{\text{ост}}}{k(n-1)} = \frac{5291}{5(8-1)} = 151{,}16.$

Полученные в результате расчета дисперсии сравним по критерию Фишера:

$$F = \frac{s_{\text{А}}^2}{s_{\text{ош}}^2} = \frac{1021}{151{,}16} = 6{,}765.$$

По Приложению 1 и Приложению5 находим: $F_{0,05}(4{,}35) = 2{,}6$. Так как $F > 2,6$, различие видов реагентов следует принять значимым. Перейдем к сравнению влияния отдельных видов реагентов на основе множественного рангового критерия Дункана.

Нормированная ошибка среднего равна:

$$s_{\bar{y}} = \sqrt{\frac{s_{\text{ош}}^2}{n}} = \sqrt{\frac{151{,}16}{8}} = 4{,}35.$$

Расположим средние значения в порядке возрастания их величин:

Реагент А3	Реагент А5	Реагент А2	Реагент А4	Реагент А5
57,74	66,04	75,97	81,61	85,1

Выпишем из Приложения 6 значимые ранги для$f = 35; p = 0{,}05$:

p	2	3	4	5
Ранги,r	2,875	3,025	3,11	3,185
$rs_{\bar{y}}$	12,5	13,2	13,52	13,9

$\overline{y_1} - \overline{y_3} = 27{,}3 > 13{,}9$– различие значимо
$\overline{y_1} - \overline{y_5} = 19{,}05 > 13{,}52$– различие значимо
$\overline{y_1} - \overline{y_2} = 9{,}13 < 13{,}2$– различие незначимо
$\overline{y_1} - \overline{y_4} = 3{,}49 < 13{,}9$– различие незначимо
$\overline{y_4} - \overline{y_3} = 23{,}81 > 13{,}52$– различие значимо
$\overline{y_4} - \overline{y_5} = 15{,}56 > 13{,}2$– различие значимо
$\overline{y_4} - \overline{y_2} = 5{,}64 < 12{,}5$– различие незначимо
$\overline{y_2} - \overline{y_3} = 18{,}17 > 13{,}9$– различие значимо
$\overline{y_2} - \overline{y_5} = 9{,}92 < 12{,}5$– различие незначимо
$\overline{y_5} - \overline{y_3} = 8{,}25 < 12{,}5$– различие незначимо

Модель со случайным уровнем факторов

Решается задача, аналогичная предыдущей. Исходные данные представлены в табл. 3.4.

Таблица 3.4 – Исходные данные к задаче

Номер наблюдения	Уровень фактора *A*				
	a_1	a_2	a_3	a_4	a_5
1	6,8	14,3	-30,55	3,0	-2,3
2	13,3	-3,4	-8,7		
3	13,5	8,75	5,9	-0,2	
4	19,3	4,95	-12,0	16,0	4,0
5	3,5	10,65	-41,7		
6	14,05	-8,2	-0,15	14,45	
7	9,5	-5,7	-14,35	1,5	
8	17,0	2,45	-20,5		

В данном случае число параллельных опытов на уровнях разное. Определим общее число всех наблюдений:

$$N = \sum_{i=1}^{k} n_i = 8 + 8 + 8 + 5 + 2 = 31.$$

Рассчитываем:
итоги по столбцам:

$A_1 = 96{,}95$, $A_2 = 23{,}8$, $A_3 = -122{,}05$, $A_4 = 34{,}75$, $A_5 = 1{,}7$;

сумму квадратов всех наблюдений:

$$SS_1 = \sum_{i=1}^{k} \sum_{j=1}^{n_j} (y_{ij})^2 = 5953;$$

сумму итогов по столбцам, деленных на число наблюдений в соответствующем столбце:

$$SS_2 = \sum_{i=1}^{k} \frac{A_i^2}{n_i} = \frac{96{,}95^2}{8} + \frac{23{,}8^2}{8} + \frac{(-122{,}05)^2}{8} + \frac{34{,}75^2}{5} + \frac{1{,}7^2}{2} = 3351;$$

квадрат общего итога, деленный на число всех наблюдений:

$$SS_3 = \frac{1}{N}\left(\sum_{i=1}^{k} A_i\right)^2 = \frac{1236}{31} = 39{,}87.$$

Дальнейшие расчеты проводятся по известным зависимостям.
Определим сумму квадратов для столбца:

$$SS_A = SS_2 - SS_3 = 3351 - 39{,}87 = 3311.$$

Найдем общую сумму квадратов:

$$SS_{общ} = SS_1 - SS_3 = 5953 - 39{,}87 = 5913$$

и остаточную сумму квадратов:

$$SS_{\text{ост}} = SS_1 - SS_2 = 5953 - 3351 = 2602.$$

Дисперсия $s_{\text{А}}^2$ составляет:

$s_{\text{А}}^2 = \frac{SS_{\text{А}}}{k-1} = \frac{3311}{5-1} = 827{,}75,$
дисперсия $s_{\text{ош}}^2$:
$s_{\text{ош}}^2 = \frac{SS_{\text{ост}}}{k(n-1)} = \frac{2602}{5(8-1)} = 74{,}35.$
Полученные в результате расчета дисперсии сравним по критерию Фишера:

$$F = \frac{s_{\text{А}}^2}{s_{\text{ош}}^2} = \frac{827{,}75}{74{,}35} = 11{,}133.$$

В соответствии с Приложением $5F_{0{,}05}(4{,}35) = 2{,}6$. Так как $F > 1{,}77$, различие видов реагентов следует принять значимым.

Задания для самостоятельного решения:

1. Провести исследование исследования влияния на выход продукта *B* нескольких видов реагента *A*. Виды реагента А назвать уровнями фактора $A_i(i = 1, \dots, m)$при*m* полном числе применяемых видов реагентов. Исходные данные для решения задачи приведены в табл. 3.5.

Таблица 3.5 – Исходные данные к задаче

Вариант 1					
Номер наблюдения	Уровень фактора А				
	a_1	a_2	a_3	a_4	a_5
1	2	3	4	5	6
1	69,80	77,30	42,45	46,0	70,70
2	76,30	59,60	54,30	53,5	64,65
3	76,50	61,75	68,9	62,80	38,50
4	72,30	57,95	51,00	49,00	77,00
5	76,50	63,65	61,30	56,50	91,50
6	77,05	64,80	52,85	57,45	68,00
7	72,50	77,30	58,65	64,50	38,05
Вариант 2					
Номер наблюдения	Уровень фактора А				
	a_1	a_2	a_3	a_4	a_5
1	69,80	77,30	42,45	46,0	70,70
2	76,30	59,60	54,30	53,5	64,65
3	76,50	61,75	68,9	62,80	38,50
4	72,30	57,95	51,00	49,00	77,00
5	76,50	63,65	61,30	56,50	91,50
6	77,05	64,80	52,85	57,45	68,00
Вариант 3					
Номер наблюдения	Уровень фактора А				
	a_1	a_2	a_3	a_4	a_5
1	42,45	67,30	46,0	69,80	60,70
2	54,30	59,60	53,5	76,30	64,65
3	78,9	51,75	62,80	76,50	38,50
4	51,00	57,95	49,00	62,30	77,00
5	61,30	63,65	56,50	76,50	91,50

6	52,85	64,80	57,45	67,05	68,00
7	58,65	77,30	64,50	62,50	38,05
8	72,50	7,45	73,15	60,00	79,95
9	51,00	57,95	49,00	62,30	77,00
10	61,30	63,65	56,50	76,50	91,50
Вариант 4					
Номер наблюдения	Уровень фактора А				
	a_1	a_2	a_3	a_4	a_5
1	82,45	77,30	76,0	79,80	40,70
2	54,30	79,60	73,5	56,30	44,65
3	88,9	71,75	62,80	56,50	48,50
4	51,00	77,95	49,00	52,30	77,00
5	61,30	63,65	56,50	76,50	51,50
6	852,85	64,80	57,45	67,05	48,00

Контрольные вопросы:

1. Как проводится дисперсионный анализ для определения значимости уравнения множественной регрессии?
2. Однофакторный дисперсионный анализ с одинаковым числом испытаний на различных уровнях.
3. Однофакторный дисперсионный анализ с различным числом испытаний на различных уровнях.
4. Двух- и многофакторный дисперсионный анализ. Критерий адекватности Фишера.
5. Дайте определение понятию «регрессия».
6. На основе какого метода вычисляются коэффициенты уравнения регрессии?
7. Что такое величина достоверности аппроксимации?
8. Из чего состоит проверка достоверности модели регрессии?
9. Что такое дисперсионный анализ?
10. Назовите назначение переменных - индикаторов?
11. Какая связь между регрессионным и дисперсионным анализом?
12. Основные компоненты дисперсионного анализа?
13. Охарактеризуйтерезультат дисперсионного анализа?

4. Анализ данных с помощью R

Основные компоненты статистической среды R

Статистическая среда R выполняет любой набор осмысленных инструкций языка R, содержащихся в файле скрипта или представленных последовательностью команд, задаваемых с консоли. Работа с консолью может показаться трудной для современных пользователей, привыкших к кнопочным меню, поскольку надо запоминать синтаксис отдельных команд. Однако, после приобретения некоторых навыков, оказывается, что многие процедуры обработки данных можно выполнять быстрее и с меньшим трудом, чем, предположим, в том же пакете Statistica.

Консоль R представляет собой диалоговое окно, в котором пользователь вводит команды и где видит результаты их выполнения. Это окно возникает сразу при запуске среды (например, после клика мышью на ярлыке R на рабочем столе). Кроме того, стандартный графический пользовательский интерфейс R (RGui) включает окно редактирования скриптов и всплывающие окна с графической информацией (рисунками, диаграммами и проч.)

В командном режиме R может работать, например, как обычный калькулятор:

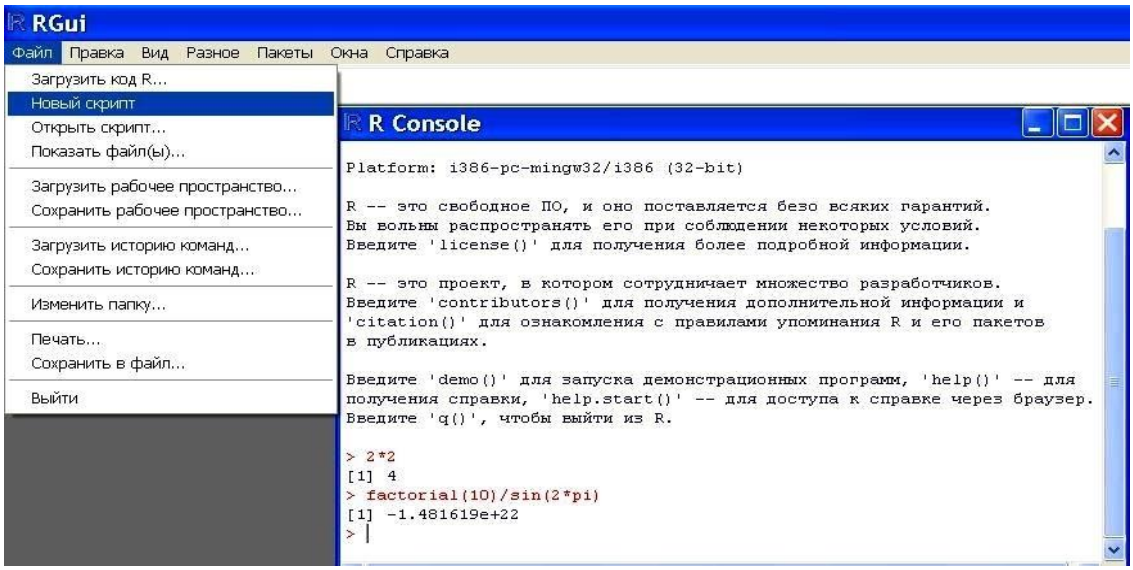


Рисунок 22 Внешний вид программы

Справа от символа приглашения пользователь может ввести произвольное арифметическое выражение, нажать клавишу Enter и тут же получить результат. Например, во второй команде на приведенном выше рисунке мы использовали функции факториала и синуса, а также встроенное число p. Результаты, полученные в текстовой форме можно выделить мышью и скопировать через буфер обмена в любой текстовый файл операционной системы (например, документ Word).

При работе с использованием RGui мы рекомендуем во всех случаях создавать файл со скриптом (т.е. последовательностью команд языка R, выполняющей определенные действия). Как правило, это обычный текстовый файл с любым именем (но, для определенности, лучше с расширением *.r), который можно создавать и редактировать обычным редактором типа "Блокнот". Если этот файл существует, его лучше всего поместить в рабочий каталог, и тогда после запуска R и выбора пункта меню "*Файл > Открыть скрипт*" содержимое этого файла появится в окне "*Редактор R*". Выполнить последовательность команд скрипта можно из пункта меню "*Правка > Запустить все*". Можно также выделить мышью осмысленный фрагмент из любого места подготовленного

скрипта (от имени одной переменной до всего содержимого) и осуществить запуск этого блока на выполнение.

Справку по отдельным функциям можно получить с использованием следующих команд:

° **help**("foo") или ? foo–справка по функции foo (кавычки необязательны);

° **help.search**("foo") или ?? foo – поиск всех справочных файлов, содержащих foo;

° **example**("foo")– примеры использования функции foo;

° **RSiteSearch**("foo")– поиск ссылок в онлайн-руководствах и архивах рассылок; ° **apropos**("foo", mode="function")– список всех функций с комбинацией foo; ° **vignette**("foo")– список руководств по теме foo.

Объекты, пакеты, функции, устройства

Язык R принадлежит к семейству так называемых высокоуровневых объектноориентированных языков программирования. Для неспециалиста строгое определение понятия "объект" является достаточно абстрактным. Однако для простоты можно называть объектами все, что было создано в процессе работы с R. Выделяют два основных типа объектов:

1. **Объекты, предназначенные для хранения данных** ("*data objects*") – это отдельные переменные, векторы, матрицы и массивы, списки, факторы, таблицы данных; 2. **Функции** ("*function objects*") – это поименованные программы, предназначенные для создания новых объектов или выполнения определенных действий над ними.

Объекты среды R, предназначенные для коллективного и свободного использования, комплектуются в пакеты, объединяемые сходной тематикой или методами обработки данных. Есть некоторое отличие между терминами *пакет* ("*package*") и *библиотека* ("*library*"). Термин "library" определяет директорию, которая может содержать один или несколько пакетов. Термин "package" обозначает совокупность функций, HTMLстраниц руководств и примеров объектов данных, предназначенных для тестирования или обучения.

Пакеты инсталлируются в определенной директории операционной системы или, в неустановленном виде, могут храниться и распространяться в архивных *.zip файлах Windows (версия пакета должна корреспондироваться с конкретной версией вашей R). Полная информация о пакете (версия, основное тематическое направление, авторы, даты изменений, лицензии, другие функционально связанные пакеты, полный список функций с указанием на их назначение и проч.) может быть получена командой library(help=<имя_пакета>), например:

```
library(help=Matrix)
```

Все пакеты R относятся к одной из трех категорий: базовые ("base"), рекомендуемые ("recommended") и прочие, установленные пользователем. Получить их список на конкретном компьютере можно, подав команду library() или:

```
installed.packages(priority = "base") installed.packages(priority = "recommended")
# Получениеполногоспискапакетовpacklist <- rownames(installed.packages())
# Выводинформациивбуферобменавформатедля Excel
write.table(packlist,"clipboard",sep="\t", col.names=NA)
```

Базовые и рекомендуемые пакеты обычно включаются в инсталляционный файл R. Разумеется, нет необходимости сразу устанавливать "про запас" много разных пакетов. Для установки пакета достаточно в командном окне R Console выбрать пункт меню "*Пакеты > Установить пакет(ы)*" или ввести, например, команду:

```
install.packages(c("vegan", "xlsReadWrite", "car"))
```

При запуске консоли RGui загружаются только некоторые базовые пакеты. Для инициализации любого другого пакета перед непосредственным использованием его функций нужно ввести команду **library** (<*имя_пакета*>).

Установить, какие пакеты загружены в каждый момент проводимой сессии, можно, подав команду:

```
sessionInfo()
R version 2.13.2 (2011-09-30)
Platform: i386-pc-mingw32/i386 (32-bit)
attached base packages:
[1] statsgraphicsgrDevices utilsdatasetsmethodsbase
other attached packages:
[1] vegan_2.0-2permute_0.6-3
loaded via a namespace (and not attached):
[1] grid_2.13.2lattice_0.19-33 tools_2.13.2
```

Наконец, рассмотрим некоторые простейшие приемы сохранения результатов работы, полученных во время сессии R:

° sink(file=<*имя файла*>) – выводит результаты выполнения последующих команд в режиме реального времени в файл с заданным именем; для прекращения действия этой команды необходимо выполнить команду sink() без параметров;

° save(file=<*имя файла*>, <*список сохраняемых объектов*>) – сохраняет указанные объекты в двоичном файле XDR-формата, с которым можно работать в любой операционной системе;

° load(file=<*имя файла*>) – восстанавливает сохраненные объекты в текущей среде;

° save.image(file=<*имя файла*>) – сохраняет все объекты, созданные в ходе работы, в виде специфичного для R rda-файла.

Пример передачи сформированной таблицы с данными в буфер обмена в формате, совместимом со структурой листа Excel, был приведен выше в настоящем разделе. В главе 6 будет приведен пример передачи данных из объекта линейной модели в файл Word.

Среда R может генерировать пиксельное изображение необходимого качества почти для любого разрешения дисплея или устройства печати, а также сохранить полученные графические окна в файлах разного формата. Для каждого устройства графического вывода существует функция драйвера: для получения полного списка драйверов можно ввести команду help(Devices). Среди графических устройств наиболее употребительными являются:

° windows()– графическое окноWindows (экран, принтер или метафайл).

° png(),jpeg(), bmp(),tiff()– вывод в растровый файл соответствующего формата;

° pdf(),postscript() – вывод графической информации в файл PDF или PostScript.

По завершению работы с устройством вывода следует отключить его драйвер командой dev.off().

Типы данных языка R

Все объекты данных (а, следовательно, и переменные) в R можно разделить на следующие классы (т.е. типы объектов):

°**numeric** – объекты, к которым относятся целочисленные (integer) и действительные числа (double);

°**logical** – логические объекты, которые принимают только два значения: FALSE (сокращенно F) и TRUE (T);

° **character** – символьные объекты (значения переменных задаются в двойных, либо одинарных кавычках).

В R можно создавать имена для различных объектов (функций или переменных) как на латинице, так и на кириллице, но следует учесть, что **а** (кириллица) и **a** (латиница) – это два разных объекта. Кроме того, среда R чувствительна к регистру, т.е. строчные и заглавные буквы в ней различаются. Имена переменных (идентификаторы) в R должны начинаться с буквы (или точки **.**) и состоять из букв, цифр, знаков точки и подчёркивания. При помощи команды ? <*имя*> можно проверить, существует липеременная или функция суказанными именем.

Проверка на принадлежность переменной к определенному классу проверяется функциями is.numeric(<*имя_объекта*>), is.integer(<*имя*>), is.logical(<*имя*>), is.character(<*имя*>), а для преобразования объекта в другой тип можно использовать функции as.numeric(<*имя*>), as.integer(<*имя*>), as.logical(<*имя*>), as.character(<*имя*>).

В R существует ряд специальных объектов:

° Inf – положительная или отрицательная бесконечность (обычно результат деления вещественного числа на 0);

° NA– "отсутствующее значение" (Not Available);

° NaN– "нечисло" (Not a Number).

Проверить, относится ли переменная к какому-либо из этих специальных типов, можно, соответственно, функциями is.□nite(<*имя*>), is.na(<*имя*>) и is.nan(<*имя*>).

Выражение (*expression*) языка R представляет собой сочетание таких элементов, как оператор присваивания, арифметические или логические операторы,имена объектов и имена функций. Результат выполнения выражения, как правило, сразу отображается в командном или графическом окне. Однако при выполнении операции присваивания результат сохраняется в соответствующем объекте и на экран не выводится.

В качестве оператора присваивания в R можно использовать либо символ “=“, либо пару символов “<-“ (присваивание определенного значения объекту слева) или “->“ (присваивание значения объекту справа). Хорошим стилем программирования считается использование “<-“.

Выражения языка R организуются в скрипте по строкам.В одной строке можно ввести несколько команд, разделяя их символом “;“. Одну команду можно также расположить на двух (и более) строках.

Объекты типа **numeric** могут составлять выражения с использованием традиционных арифметических операций + (сложение), – (вычитание), * (умножение), / (деление), ^ (возведение в степень), %/% (целочисленное деление), %% (остаток от деления). Операции имеют обычный приоритет, т.е. сначала выполняется возведение в степень, затем умножение или деление, потом уже сложение или вычитание. В выражениях могут использоваться круглые скобки и операции в них имеют наибольший приоритет.

Логическиевыражения могутсоставлятьсяс использованием следующих логических операторов:

° "Равно" ==

° "Не равно" !=

° "Меньше" <° "Больше" >

° "Меньше либо равно" <=

° "Больше либо равно" >=

° "Логическое И" &

° "Логическое ИЛИ" | ° "Логическое НЕ" !

Векторы и матрицы

Вектор представляет собой поименованный одномерный объект, содержащий набор однотипных элементов (числовые, логические, либо текстовые значения – никакие их сочетанияне допускаются). Для создания векторов небольшой длины в R используется

функция конкатенации c() (от "*concatenate*" – объединять, связывать). В качестве аргументов этой функции через запятую перечисляют объединяемые в вектор значения, например:

```
my.vector <- c(1, 2, 3, 4, 5) my.vector
[1] 1 2 3 4 5
```

Для создания векторов, содержащих последовательную совокупность чисел, удобна функция seq() (от "*sequence*" – последовательность). Так, вектор с именем S, содержащий совокупность целых чисел от 1 до 7, можно создать следующим образом:

```
S <- seq(1,7)
S
[1] 1 2 3 4 5 6 7
```

Идентичный результат будет получен при помощи команды

```
S <- 1:7
S
[1] 1 2 3 4 5 6 7
```

Векторы, содержащие одинаковые значения, создают при помощи функции rep() (от "*repeat*" – повторять). Например, для формирования текстового вектора Text, содержащего пять значений "test", следует выполнить команду

```
Text <- rep("test", 5)
Text
[1] "test" "test" "test" "test" "test"
```

Система R способна выполнять самые разнообразные операции над векторами. Так, несколько векторов можно объединить в один, используя уже рассмотренную выше функцию конкатенации:

```
v1 <- c(1, 2, 3) v2 <- c(4, 5, 6)
V <- c(v1, v2)
V
[1] 1 2 3 4 5 6
```

Используяиндексныеномера,можновыполнятьразличныеопераciис избранными элементами разных векторов:

```
# создадим еще один числовой вектор z, содержащий 3 значения: z <- c(0.5, 0.1, 0.6)
# умножим первый элемент вектора y на третий элемент вектора z # (т.е. 5*0.6):
y[1]*z[3]
[1] 3
```

Для упорядочения значений вектора по возрастанию или убыванию используют функцию sort() в сочетании с аргументом decreasing = FALSE или decreasing = TRUE соответственно ("*decreasing*" значит "убывающий"):

```
sort(z)# поумолчанию decreasing = FALSE
[1] 0.3 0.5 0.6
sort(z, decreasing = TRUE)
[1] 0.6 0.5 0.3
```

Матрица представляет собой двумерный вектор. В R для создания матриц служит одноименная функция:

```
my.mat <- matrix(seq(1, 16), nrow = 4, ncol = 4) my.mat
[,1] [,2] [,3] [,4]
[1,]15913
[2,]261014
[3,]371115
[4,]481216
```

В качестве заголовков строк и столбцов создаваемой матрицы автоматически выводятся соответствующие индексные номера (строки: [1,], [2,], и т.д.; столбцы: [,1], [,2], и т.д.). Для придания пользовательских заголовков строкам и столбцам матриц используют функции rownames() и colnames() соответственно. Например, для обозначения строк матрицы my.mat буквами A, B, C и D необходимо выполнитьследующее:

```
rownames(my.mat) <- c("A", "B", "C", "D") my.mat
[,1] [,2] [,3] [,4]
A       1234
B       5678
C       9101112
D       13141516
E
```

Таблица данных (*data frame*) представляет собой объект R, по структуре напоминающий лист электронной таблицы Microsoft Excel. Каждый столбец таблицы является вектором, содержащим данные определенного типа. При этом действует правило, согласно которому все столбцы должны иметь одинаковую длину (собственно, с "точки зрения" R таблица данных является частным случаем списка, в котором все компонентывекторы имеют одинаковый размер).

Таблицы данных – это основной класс объектов R, используемых для хранения данных. Обычно такие таблицы подготавливаются при помощи внешних приложений (особенно популярна и удобна программа Microsoft Excel) и затем загружаются в среду R.

Подробнее об импортировании данных в R будет рассказано ниже. Тем не менее, небольшую таблицу можно собрать из нескольких векторов средствами самой системы R. Для этого используют функцию data.frame().

При необходимости внесения исправлений в таблицу можно воспользоваться встроенным в R редактором данных. Внешне этот редактор напоминает обычный лист Excel, однако имеет весьма ограниченные функциональные возможности. Все, что он позволяет делать – это добавлять новые или исправлять уже введенные значения переменных, изменять заголовки столбцов, а также добавлять новые строки и столбцы. Работая в стандартной версии R, редактор данных можно запустить из меню "*Файлы* > *Редактор данных*", либо выполнив команду fix() (*fix* – исправлять, чинить) из командной строки консоли R (например, fix(CITY)). После внесения исправлений редактор просто закрывают – все изменения будут сохранены автоматически.

Заполнение пустых значений

Часто на практике некоторые значения в таблице отсутствуют, что может быть обусловлено множеством причин: на момент измерения прибор вышел из строя, по невнимательности персонала измерение не было занесено в протокол исследования, испытуемый отказался отвечать на определенный вопрос(ы) в анкете, была утеряна проба, и т.п. Ячейки с такими отсутствующими значениями (*missing values*) в таблицах данных R не

могут быть просто пустыми – иначе столбцы таблицы окажутся разной длины. Для обозначения отсутствующих наблюдений в языке R, как указывалось ранее, имеется специальное значение – NA (*not available* – не доступно). В разделе 4.4 мы остановимся на решении проблемы заполнения пропусков подробнее. Здесь же отметим, что если значение NA имеет смысл нуля (например, экземпляров некого вида обнаружено не было), то легко произвести эту замену в таблице DF командой

```
DF[is.na(DF)] <- 0
```

Сортировка таблиц

Сортировка строк таблицы по различным ключам не представляет труда. Для этого используется функция order():

```
DF<- data.frame(X1=c(1,15,1,3), X2=c(1,0,7,0), X3=c(1,0,1,2),
                                X4=c(7,4,41,0), X5=c(1,0,5,3))
row.names(DF) <- c("A","B","C","D")
#DF1 – таблица, столбцы которой отсортированы
#по убыванию суммы значений
DF1 <-DF[ , rev(order(colSums(DF)))]
# DF2 – таблица, строки которой отсортированы в восходящем
#порядке по 1 столбцу, затем в нисходящем по второму DF2 <- DF[order(DF$X1, -
DF$X2), ]
```

Импортирование данных в R

В предыдущих разделах было рассмотрено, как, работая непосредственно в системе R, можно создать небольшие по объему объекты для хранения данных (векторы, матрицы, списки, таблицы данных). Следует отметить, однако, что возможности системы R по вводу и редактированию данных умышленно ограничены ее создателями, которые предполагали, что для этого будут использоваться другие средства (например, программа Microsoft Excel или базы данных). Поэтому подлежащие анализу объемные таблицы данных обычно подготавливаются при помощи сторонних приложений, и только потом загружаются в рабочую среду R из внешних файлов. Хотя предпочтение при этом отдается текстовым файлам, выше был упомянут специальный пакет foreign, функции которого позволяют импортировать таблицы, сохраненные во множестве других распространенных форматов (Excel, SPSS, SAS, STATA, Access, Matlab, SQL, Oracle, и т.п).

Импортирование данных в систему R часто вызывает проблемы у тех, кто только начинает работать с этой программой. Тем не менее, ничего сложного в этом нет. Ниже будут подробно рассмотрены наиболее распространенные способы импорта таблиц данных в рабочую среду R, однако сначала стоит ознакомиться с правилами подготовки загружаемых файлов:

- В импортируемой таблице с данными не должно быть пустых ячеек. Если некоторые значения по тем или иным причинам отсутствуют, вместо них следует ввести NA.
- Импортируемую таблицу с данными рекомендуется преобразовать в простой текстовый файл с одним из допустимых расширений. На практике обычно используются файлы с расширением .txt, в которых значения переменных разделены знаками табуляции (*tab-delimited files*), а также файлы с расширением .csv (*comma separated values*), в которых значения переменных разделены запятыми или другим разделяющим символом.
- В качестве первой строки в импортируемой таблице рекомендуется ввести заголовки столбцов-переменных. Такая строка – удобный, но не обязательный элемент загружаемого файла. Если она отсутствует, то об этом необходимо сообщить в описании команды, которая будет управлять загрузкой файла (например, read.table() – см. ниже). Все последующие строки файла в качестве первого элемента могут содержать

заголовки строк (если таковые предусмотрены), после которых следуют значения каждой из имеющихся в таблице переменных.

Имена столбцов таблицы лучше присвоить с соблюдением правил идентификации переменных R, т.е. исключить пробелы и другие специальные символы, кроме точки и подчеркивания. Во избежание проблем, связанных с кодировкой, все текстовые величины в импортируемых файлах стоит создавать с использованием букв латинского алфавита.

Подлежащий импортированию файл рекомендуется поместить в рабочую папку программы, т.е. папку, в которой R по умолчанию будет "пытаться найти" этот файл (см. раздел 1.1). Чтобы выяснить путь к рабочей папке R на своем компьютере используйте команду getwd() (*getworking directory* – узнать рабочую директорию); например:

```
getwd() [1] "C:/Temp/"
```

Изменить рабочую директорию можно при помощи команды setwd() (*setworking directory*– создать рабочую директорию):

```
setwd("C:/My Documents")
# при выполнении этой команды внешне ничего не произойдет,
# однако последующее применение команды getwd() покажет,
# что путь к рабочей папке изменился: getwd()
[1] "C:/My Documents/"
```

Основной функцией для импортирования данных в рабочую среду R является read.table(). Эта мощная функция позволяет достаточно тонко настроить процесс загрузки внешних файлов, в связи с чем, она имеет большое количество управляющих аргументов. Наиболее важные из этих аргументов перечислены ниже в таблице 3.1 (подробнее см. файл помощи, доступный по команде ?read.table).

Таблица 4.1 – Функциии для импортирования данных в рабочую среду R

Аргумент	**Назначение**
file	Служит для указания пути к импортируемому файлу. Путь приводят либо в абсолютном виде (например, file = "C:/Temp/MyData.dat"), либо указывают только имя импортируемого файла (например, file = "MyData.txt"), но при условии, что последний хранится в рабочей папке программы (см. выше). В качестве имени можно также указывать полную URL-ссылку на файл, который предполагается загрузить из Сети (например: file = "http://somesite.net/YourData.csv"). Начиная с версии R 2.10, появилась возможность импортировать архивированные файлы в zip-формате.
header	Служит для сообщения программе о наличии в загружаемом файле строки с заголовками столбцов. По умолчанию принимает значение FALSE. Если строка с заголовками столбцов имеется, этому аргументу следует присвоить значение TRUE.
row.names	Служит для указания номера столбца, в котором содержатся имена строк (например, в рассмотренном выше примере это был первый столбец, поэтому row.names = 1). Важно помнить, что все имена строк должны быть уникальными, т.е. одинаковые имена для двух или более строк не допускаются.

sep	Служит для указания разделителя значений переменных, используемого в файле (*separator* – разделитель). По умолчанию предполагается, что значения переменных разделены "пустым пространством", например, в виде пробела или знака табуляции (sep = ""). В файлах формата csv значения переменных разделены запятыми, и поэтому для них sep = ",".
dec	Служит для указания знака, используемого в файле для отделения целой части числа от дроби. По умолчанию dec = ".". Однако во многих странах в качестве десятичного знака применяют запятую, о чем важно вспомнить перед загрузкой файла и, при необходимости, использовать dec = ",". Следите, чтобы dec и sep не были бы одинаковыми.
nrows	Выражается целым числом, указывающим количество строк, которое должно быть считано из загружаемой таблицы. Отрицательные и иные значения игнорируются. Пример: nrows = 100.
skip	Выражается целым числом, указывающим количество строк в файле, которое должно быть пропущено перед началом импортирования. Пример: skip = 5

Графическое представление данных

Функция plot() – главная "рабочая лошадка", используемая для построения графиков в R. Поведение этой *функции высокого уровня* определяется классом объектов, указываемых в качестве ее аргументов. Соответственно, с помощью plot() можно создать очень большой набор разнотипных графиков.

В качестве примера используем данные по скорости выведения из организма человека *индометацина*– одного из наиболее активных противовоспалительных препаратов. В эксперименте приняли участие шесть испытуемых. Результаты этого исследования входят в базовый набор данных R и доступны по команде

```
data(Indometh)
Применивкоманду names(Indometh)
[1] "Subject" "time" "conc"
```

видим, что в состав таблицы Indometh входят переменные Subject (испытуемый), time (время с момента введения препарата) и conc (концентрация препарата в крови). Чтобы облегчить дальнейшую работу, прикрепим таблицу Indometh к *поисковому пути* R:

```
attach(Indometh)
```

Благодаря этой команде, теперь мыможем напрямую обращаться к переменным таблицы Indometh (т.е. использовать их имена непосредственно, например, time вместо Indometh$time).

Зависимость концентрации индометацина в крови от времени можно легко изобразить при помощи следующей команды:

Предположим, что перед нами стоит задача отобразить на графике не все исходные данные, а только средние значения концентрации индометацина для каждой временной точки. Рассчитать средние значения (или любые другие количественные величины) для отдельных групп данных позволяет функция tapply():

(means<- **tapply**(conc, time, **mean**))
0.250.50.7511.2523
2.07666667 1.32166667 0.91833333 0.68333333 0.55666667 0.33166667 0.19833333
4568
0.13666667 0.12500000 0.09000000 0.07166667

Управляющие параметры функции plot()

Функция plot() имеет большое количество управляющих параметров, которые позволяют осуществить тонкую настройку внешнего вида графика. Ниже рассмотрены лишь некоторые из них.

1. Параметры xlab и ylab

Параметры xlab и ylab служат для изменения названий осей X и Y соответственно:
plot(indo.times, means, xlab = "Время", ylab = "Концентрация")

2. Параметр type

Параметр type позволяет изменять внешний вид точек на графике. Он принимает одно из следующихзначений:

° "p"– точки (*points*; используется по умолчанию)
° "l"– линии (*lines*)
° "b"– изображаются и точки, и линии (*both points and lines*)
° "o"– точки изображаются поверх линий (*points over lines*)
° "h"– гистограмма (*histogram*)
° "s"– ступенчатая кривая (*steps*)
° "n"– данные не отображаются (*no points*)

3. Параметры xlim и ylim

Эти два параметра контролируют размах значений на каждой из осей графика. По умолчанию они оба принимают значение NULL – в этом случае размах выбирается программой автоматически. Для отмены автоматических настроек соответствующему параметру необходимо присвоить значение в виде числового вектора, содержащего минимальное и максимальное значения, которые должны отображаться на оси. Например:
plot(indo.times, means, xlab="Время", ylab="Концентрация", xlim=**c**(0, 15))

plot(indo.times, means, xlab="Время", ylab="Концентрация", ylim=**c**(0, 5))

4. Параметры axes и ann

Эти два параметра контролируют отображение осей и их названий соответственно. Каждый из них может принимать одно из двух возможных значений – TRUEили FALSE:

plot(indo.times, means, xlab = "Время", ylab = "Концентрация",
axes = TRUE, ann = TRUE)
plot(indo.times, means, xlab = "Время", ylab = "Концентрация",
axes = FALSE, ann = TRUE)
plot(indo.times, means, xlab = "Время", ylab = "Концентрация",axes = TRUE, ann = FALSE)

5. Параметр log

При помощи аргумента log можно перевести одну или обе оси графика на логарифмическую шкалу, например:

plot(indo.times, means, xlab = "Время", ylab = "Концентрация", log = "x")
plot(indo.times, means, xlab = "Время", ylab = "Концентрация", log = "y")
plot(indo.times, means, xlab = "Время", ylab = "Концентрация", log = "xy")

6. Параметр main

Аргумент main служит для создания заголовка графика. По умолчанию название размещается в верхней части рисунка:

plot(indo.times, means, xlab = "Время", ylab = "Концентрация",main = "Скорость выведения индометацина", type = "o")

Далее будут рассмотрены графические параметры, контролирующие внешний вид графиков, например, тип, размер и цвет символов и линий, тип и размер шрифта в названиях графика и его осей, использование математических символов в названиях, размещение легенды, и т.п. Они применяются в качестве аргументов не только при вызове plot(), но и многих других функций.

Управление общими параметрами – аргументами графических функций

1. Тип символа

Как видно из приведенного выше рисунка, отдельные измерения по умолчанию изображаются в виде кружков. Изменить тип символов, используемых для отображения наблюдений, позволяет аргумент pch (*plotting character* – символ изображения). В стандартных случаях этот аргумент принимает численные значения от 1 до 25. Например, при pch = 2 символы превратятся из кружков в незакрашенные треугольники:

plot(indo.times, means, xlab = "Время", ylab = "Концентрация",main = "Скорость выведения индометацина", type = "o", pch = 2)

Таблица 25-ти стандартных маркеров и соответствующие им численные коды представлена на рисунке 23

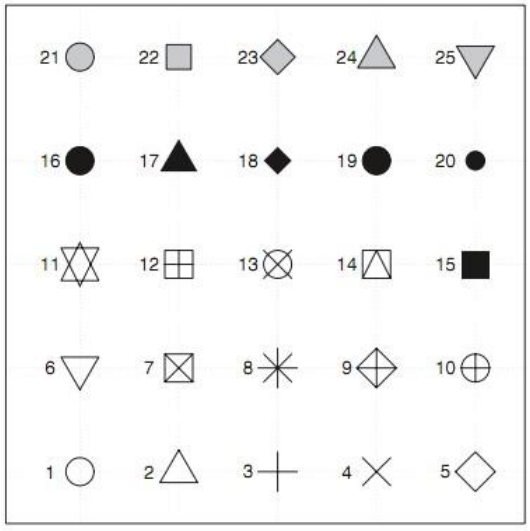


Рисунок 23 – Стандартные маркеры

Набор стандартных маркеров может быть значительно расширен в случае, когда аргумент pch используется в комбинации с другим аргументом – font, задающим шрифт символов. Параметр pch может при этом принимать любое целое число от 1 до 128 и от 160 до 254. Например, при font = 5 маркеру в виде "сердечка" соответствует код 169:

plot(indo.times, means, xlab = "Время", ylab = "Концентрация",main = "Скорость выведения индометацина", type = "o",pch = 169, font = 5)

В качестве маркеров можно также использовать обычные печатные символы, например, буквы:

plot(indo.times, means, xlab = "Время", ylab = "Концентрация",main = "Скорость выведения индометацина", type = "o",pch = "A")

2. Размер маркера

Размер маркеров задается при помощи аргумента cex (*character extension* – размер символа), который по умолчанию равен 1. Уменьшение или увеличение этого параметра приводит к соответствующим пропорциональным изменениям размера символов.

При необходимости мы можем также изменить ширину линии обводки символа. Для этого служит параметр lwd (*line width* – ширина линии). Далее будет приведен пример для символа с кодом 21 ("заполненный кружок").

3. Цвет маркера

Цвет любого графического объекта может быть задан несколькими способами:

° по названию цвета: например, col = "red" (красный), col = "green" (зеленый), или col = "black" (черный). Всего в R имеется 675 стандартных цветов. Их названия доступны по команде colors();

° путем непосредственного указания красного, зеленого и синего компонентов RGB спектра, например: "#RRGGBB;

° по численному коду, например: col = 2 (красный), col = 3 (зеленый), или col = 1 (черный).

Цвет маркеров задается при помощи аргумента col (*color* – цвет). Подобрать его можно следующим образом:

```
n <- 20
plot(1:n, pch=CIRCLE<-16, cex=1:n, col=1:n) text(1:n)
```

Имеются также отдельные параметры для настройки цвета других элементов графика (например, заголовка col.main, названий осей col.lab, меток осей col.axes и др.).

4. Ширина линии

Ширина линии задается при помощи аргумента lwd (от *line width*) функции plot(). Аргумент принимает положительные числовые значения, показывающие, во сколько *раз* ширина линии должна быть больше относительно ширины, заданной по умолчанию. Ширина линии (по умолчанию равна 1) является безразмерной величиной, поскольку на разных графических устройствах линии с одинаковыми параметрами могут выглядеть по-разному. Ниже приведены примеры трех графиков с разными значениями параметра lwd:

```
plot(indo.times, means, xlab = "Время", ylab = "Концентрация",
main = "lwd = 2", type = "l", lwd = 2)
plot(indo.times, means, xlab = "Время", ylab = "Концентрация",
main = "lwd = 5", type = "l", lwd = 5)
plot(indo.times, means, xlab = "Время", ylab = "Концентрация",
main = "lwd = 10", type = "l", lwd = 10)
```

5. Концы и места соединения линий

Аргумент lend (от *line end* – окончание линии) функции plot() позволяет настроить внешний вид концов линии. Этот аргумент принимает значения 0 (по умолчанию), 1 или 2, что соответствует округлым, усеченным квадратным и квадратным концам соответственно. Места соединения линий также могут выглядеть по-разному, что определяется аргументом ljoin (от *line* – линия, и *join* – место соединения). Аргумент ljoin принимает значения 0 (по умолчанию), 1 или 2, что соответствует округлому, остроугольному и усеченному соединениям соответственно.

6. Тип линии

Тип линии настраивается при помощи аргумента lty (от *line* – линия, и *type* – тип) функции plot(). Существует шесть предустановленных типов линий, которые задаются числами от 1 до 6 соответственно.

При необходимости можно создать пользовательские типы линий. В таких случаях в качестве значения аргумента lty выступает текстовая последовательность из четырех цифр.

Эти числа (от 1 до 9) определяют размер четырех элементов, составляющих повторяющийся паттерн "штрих - пробел - штрих - пробел". Например, при lty = "4241" линия будет состоять из повторяющегося паттерна, в котором имеется штрих длиной 4 единицы, пробел длиной 2 единицы, опять штрих длиной 4 единицы, и пробел в 1 единицу.

Цвет линий задается при помощи аргумента col (от *color* – цвет). Использование параметра col в отношении линий ничем не отличается от его использования в отношении графических символов.

7. Рамка графика

Для настройки внешнего вида рамки графика служит аргумент bty (от *box* – коробка, и *type* – тип) функции plot(). Этот аргумент принимает одно из следущих шести текстовых значений:

"O" "L" "7" "C" "U" "["

Рамка будет принимать вид в соответствии с формой указанного символа (допускается использование также строчных букв o, l, c, и u).

Гистограммы, функции ядерной плотности и функция cdplot()

Гистограмма является важным инструментом статистики, позволяющим наглядно представить распределение значений анализируемой переменной. В системе R для построения гистограмм служит функция hist(). Ее основным аргументом выступает имя анализируемой переменной.

В качестве примера создадим нормально распределенную совокупность X из 100 наблюдений со средним значением 15 и стандартным отклонением 5:

X <- **rnorm**(n = 100, mean = 15, sd= 5)

Для создания переменной X использована функция rnorm() (от *random* – случайный, и *norm* – нормальный). Используя генератор случайных чисел, эта функция формирует нормально распределенные совокупности с заданными размером (n), средним значением (mean) и стандартным отклонением (sd).

Изобразить значения переменной X в виде гистограммы очень просто:

```
hist(X)
```

Функция hist() автоматически выбирает количество столбцов для отображения на графике, а также создает названия осей и заголовок графика. Такого рисунка, получаемого с использованием автоматических настроек, может оказаться вполне достаточно (например, при проведении быстрого разведочного анализа данных). Однако часто требуется его дополнительная доработка.

Прежде всего, важно обратить внимание на размер шага, используемого для разбиения данных на классы при построении гистограммы. В приведенном выше примере программа автоматически разбила значения переменной X на 6 классов. Однако такое грубое разбиение может недостаточно точно отражать свойства анализируемой совокупности. Для более детального изучения этих свойств можно увеличить дробность деления данных на классы (т.е. использовать меньший классовый промежуток). Сделать это позволяет аргумент breaks (*разломы*) функции hist(). При необходимости столбцы гистограммы можно залить желаемым цветом. Для этого следует воспользоваться аргументом col – это тот же аргумент, который мы использовали при рассмотрении настроек функции plot()

Диаграммы размахов

Диаграммы размахов, или "ящики с усами" (англ. *box-whisker plots*), получили свое название за характерный вид: точку или линию, соответствующую среднему положению совокупности данных, окружает прямоугольник ("ящик"), длина которого соответствует одному из показателей разброса или точности оценки генерального параметра.

Дополнительно от этого прямоугольника отходят "усы", также соответствующие по длине одному из показателей разброса или точности. Графики этого типа очень популярны, поскольку позволяют дать очень полную статистическую характеристику анализируемой совокупности. Кроме того, диаграммы размахов можно использовать для визуальной экспресс-оценки разницы между двумя и более группами (например, между датами отбора проб, экспериментальными группами, участками пространства и т.п.).

В R для построения диаграмм размахов служит функция boxplot(). Здесь, в отличие от других статистических программ, при построении диаграмм размахов используются устойчивые (робастные) оценки центральной тенденции (медиана) и разброса (интерквартильный размах – ИКР). Верхний "ус" простирается от верхней границы "ящика" до наибольшего выборочного значения, находящегося в пределах расстояния 1.5´ИКР от этой границы. Аналогично, нижний "ус" простирается от нижней границы "ящика" до наименьшего выборочного значения, находящегося в пределах расстояния 1.5´ИКР от этой границы. Длину данного интервала (т.е. 1.5´ИКР) можно изменить при помощи аргумента range функции boxplot().

Наблюдения, находящиеся за пределами "усов", потенциально могут быть выбросами. Однако всегда следует внимательно относиться к такого рода нестандартным наблюдениям – они вполне могут оказаться "нормальными" для исследуемой совокупности, и поэтому не должны удаляться из анализа без дополнительного расследования причин их появления.

Круговые и столбиковые диаграммы

Круговые диаграммы (англ. *pie charts*), мягко говоря, не в почете у профессиональных статистиков. Информация, представляемая при помощи круговой диаграммы, плохо воспринимается визуально и практически всегда лучшей альтернативой этому способу визуализации данных будет рассмотренная ниже точечная диаграмма Кливленда или, в крайнем случае, столбиковая диаграмма. Не удивительно поэтому, что в первых версиях R даже не было отдельной функции для построения круговых диаграмм. Позднее такая функция появилась, поскольку в ряде случаев этот вид диаграмм все же может оказаться полезным. Несложно догадаться, что соответствующая функция называется pie().

Функция pie() имеет несколько аргументов (подробнее см. ? pie). Основными из них являются следующие:

° x – вектор из положительных чисел, на основе которых строится диаграмма;

° labels – текстовый вектор, содержащий подписи секторов диаграммы; если значения x уже имеют атрибут names (имена), то аргумент labels указывать не обязательно (см. ниже);

° radius – изменяет размер квадрата, внутри которого строится диаграмма; в случаях, когда подписи секторов диаграммы слишком длинные, размер этого

квадрата можно уменьшить (возможные значения: от -1 до 1; см. ниже);

° init.angle – угол поворота диаграммы;

° col – вектор (числовой или текстовый), содержащий коды цветов для заливки секторов диаграммы;

° main – текстовый вектор, содержащий заголовок диаграммы;

° ... – другие графические параметры (например, параметры, определяющие размер подписей секторов диаграммы, цвет линий, и т.п.).

Столбиковые диаграммы

Для создания столбиковых (или "*столбчатых*", реже "*линейчатых*"; англ. *bar plots* или *bar charts*) диаграмм в системе R служит функция barplot(). У этой функции имеется большое количество аргументов (подробнее см. ? barplot), к основным из которых относятся:

° hight ("высота") – числовой вектор или матрица со значениями, используемыми для построения диаграммы. Если аргумент hight указан в виде вектора, то строится график из последовательно расположенных столбцов, высоты которых соответствуют значениям этого вектора. Если hight указан в виде матрицы и аргумент beside = FALSE, то будет построена

столбчатая *диаграмма с накоплением*. Если же hight указан в виде матрицы и аргумент beside = TRUE, то столбцы диаграммы будут *сгруппированы* в соответствии со столбцами матрицы.

°width ("ширина") – необязательный параметр, позволяющий регулировать ширину столбцов на диаграмме. Указывается в виде числового вектора, значения которого соответствуют ширине столбцов.

° space ("пространство") – величина зазора между столбцами (пропорционально их средней ширине). Может быть указана, либо в виде одного числа, либо в виде вектора из чисел, соответствующих каждому столбцу диаграммы.

° names.arg – текстовый вектор, содержащий подписи (вдоль оси X) для каждого столбца или группы столбцов. Если этот аргумент не указан, то в качестве подписей автоматически будут использованы имена элементов вектора height (если таковые имеются), либо заголовки столбцов, если height представляет собой матрицу.

° legend.text – вектор, содержащий текстовые элементы легенды графика. Этот аргумент полезен, если только height является матрицей. В этом случае метки легенды должны соответствовать строкам матрицы. Аргументу legend.text можно также присвоить значение TRUE, и тогда имена строк матрицы (если таковые имеются) будут использованы в качестве меток легенды автоматически.

° beside – принимает логическое значение и имеет смысл, если только height является матрицей. Значение FALSE приведет к построению диаграммы с накоплением. При значении TRUE столбцы будут сгруппированы.

° horiz – принимает логическое значение: TRUE для горизонтального расположения столбцов и FALSE – для вертикального.

° density – числовой вектор, задающий плотность заштриховки столбцов.

° angle – угол наклона штрихов (в градусах).

° col – вектор цветовых кодов для столбцов или их элементов. По умолчанию столбцы закрашиваются серым цветом, если height – вектор, и разными градациями серого, если height – матрица.

° border – код цвета для обводки столбцов. Если границу столбцов обводить не предполагается, можно указать border = NA.

° ... – другие графические параметры (см., например, ? plot и ? par).

В качестве первого примера используем рассмотренные ранее данные по эффективности шести инсектицидных средств (A – F), входящие в базовую комплектацию R. Загрузим таблицу с этими данными в рабочую среду R:

data(InsectSprays) InsectSprays

Стоит задача отобразить в виде столбиковой диаграммы средние значения количества насекомых, учтенных на экспериментальных растениях после обработки каждым инсектицидом. Эти средние значения можно быстро рассчитать для каждой группы при помощи функции tapply().

Для построения столбчатой диаграммы в ее простейшем виде достаточно выполнить следующую команду:

barplot(height = Means) *# илипросто barplot(Means)*

Добавим подписи осей и закрасим столбцы диаграммы голубым цветом:

barplot(Means, col = "steelblue",xlab = "Инсектицид",ylab = "Количество выживших насекомых")

Заполнение пропущенных значений в таблицах данных

К сожалению, большинство статистических методов предполагает, что в ходе наблюдений были получены полностью укомплектованные матрицы, векторы и другие информационные структуры эксперимента. Поскольку на практике пропуски в данных все же являются повсеместным явлением, прежде чем начать аналитические изыскания, необходимо привести обрабатываемые таблицы к "каноническому" виду, т.е. либо удалить

фрагменты объектов с недостающими элементами, либо заменить имеющиеся пропуски на некоторые разумные значения.

Несмотря на то, что книги по статистике скупо разбирают проблему исследования недостающих данных, в этой области существует впечатляющее множество подходов, методологий и их критических анализов. На практике процедура "борьбы с пропусками" обычно включает следующие шаги:

1. Идентификация недостающих данных.
2. Исследование закономерностей появления отсутствующих значений.
3. Формирование наборов данных, не содержащих пропуски (в результате удаления или замены соответствующих фрагментов).

Необходимо признать, что идентификация недостающих данных является здесь единственным однозначным шагом. Анализ, почему данные отсутствуют, зависит от вашего понимания процессов, которые воспроизводят экспериментальную информацию. Решение о способе устранения пропущенных значений также будет зависеть от вашей оценки того, какие процедуры приведут к самым надежным и точным результатам. Мы лишь бегло остановимся на некоторых функциях пакета mice, который является хорошим средством реализации всех шагов упомянутой процедуры (Kabacoff, 2011).

Для конкретизации наших дальнейших рассуждений рассмотрим набор данных sleep из пакета VIM, составленный по результатам наблюдений за процессом сна 62 млекопитающих разных видов. Авторы заинтересовались тем, почему потребность в сне у животных меняется от вида к виду и от каких экологических и таксономических переменных она зависит. Зависимые переменные включали продолжительность сна со сновидениями (Dream), сна без сновидений (NonD), и их сумма (sleep). Таксономические переменные включали массу тела (BodyWgt), вес мозга (BrainWgt), продолжительность жизни (Span) и время беременности (Gest). Экологические переменные состояли из 5-бальных оценок степени хищничества животных (Pred), меры защищенности их места для сна (Exp), изменяющегося от глубокой норы до полностью открытого пространства, и показателя риска (Danger), который основывался на логической комбинации остальных двух (Pred и Exp).

Идентификация пропущенных значений обычно легко делается с использованием функций **is.na**() и **complete.cases**():

```
data(sleep, package="VIM") head(sleep)
BodyWgt BrainWgt NonD Dream Sleep Span Gest Pred Exp Danger 1
6654.0005712.0NANA3.3 38.6645353 21.0006.66.32.08.34.542313 33.38544.5NANA12.5
14.060111 40.9205.7NANA16.5NA25523 5 2547.0004603.02.11.83.9 69.0624354
# списокстрок, вкоторыхнетпропущенныхзначений sleep[complete.cases(sleep),]
# список строк, в которых хотя бы одно пропущенное значение
sleep[!complete.cases(sleep),] sum(is.na(sleep$Dream))
[1] 12
```

Таким образом, мы имеем 42 полностью укомплектованные строки и 20 строк, имеющих хотя бы одно пропущенное значение. Из них 12 недостающих значений принадлежит переменной Dream.

Дополнительную статистическую информацию о пропусках предоставляет функция **md.pattern**(), а функции **aggr**(), **matrixplot**() и **scattMiss**() позволяют отобразить на графиках закономерности во встречаемости отсутствующих наблюдений:

```
library(mice) md.pattern(sleep)
BodyWgt BrainWgt Pred Exp Danger Sleep Span Gest Dream NonD
4211111111110
2       11111101111
3       11111110111
 91111111100
 21111101110
```

11111110011
21111101100
11111110100

000004441214 38

Значения 0 в теле приведенной таблицы соответствуют недостающим значениям, причем в первой строке пропусков нет, а последующие строки упорядочены по числу их появления. Первый столбец указывает число случаев в каждом образце данных, а последний столбец – число переменных с отсутствующими значениями в каждой строке. Например, в четвертой строке в исходных данных содержится 9 случаев с одновременно отсутствующими NonD и Dream. Набор данных содержит в общей сложности (42´0) +

(2´1) + … + (1´3) = 38 недостающих значений, а в последней строке приведена частота пропусков для каждой переменной.

Законы распределения вероятностей, реализованные в R

В базовой версии R имеются функции для работы с целым рядом распространенных законов распределения вероятностей. В зависимости от назначения, имена этих функций начинаются с одной из следующих четырех букв:

° d(от"*density*",плотность):функциииплотностивероятности("функция распределения масс" для дискретных величин);

° p(от"*probability*", *вероятность*):кумулятивныефункциираспределения вероятностей;

° q (от "*quantile*", квантиль): функции для нахождения квантилей того или иного распределения;

° r (от "*random*", случайный): функции для генерации случайных чисел в соответствии с параметрами того или иного закона распределения вероятностей.

В частности, в базовой версии R реализованы следующие законы распределения вероятностей:

° Бета-распределение (см. dbeta)
° Биномиальное распределение (включая распределение Бернулли) (dbinom)
° Распределение Коши (dcauchy)
° Распределение хи-квадрат (dchisq)
° Экспоненциальное распределение (dexp)
° Распределение Фишера (df)
° Гамма-распределение (dgamma)
° Геометрическоераспределение(какчастныйслучайотрицательного биномиального распределения) (dgeom) ° Гипергеометрическое распределение (dhyper)
° Логнормальное распределение (dlnorm)
° Полиномиальное (или мультиномиальное) распределение (dmultinom)
° Отрицательное биномиальное распределение (dnbinom)
° Нормальное распределение (dnorm)
° Распределение Пуассона (dpois)
° Распределение Стьюдента (dt)
° Равномерное распределение (dunif)
° Распределение Вейбулла (dweibull)≤

В качестве примера рассмотрим d-, p-, q- и r-функции, предназначенные для работы с нормальным распределением. Предположим, что мы имеем дело с непрерывной количественной величиной X , значения которой распределены в соответствии со стандартным нормальным распределением (среднее значение = 0, стандартное отклонение = 1).

R-функция drnorm() позволяет рассчитать значение функции плотности вероятности для заданного значения (или сразу для вектора из нескольких значений) *X*. Так, для *x* = -1 в случае со стандартным нормальным распределением получаем

dnorm(-1) [1] 0.2419707

Кумулятивная функция распределения (или просто "функция распределения") описывает вероятность того, что вещественнозначная случайная переменная *X* примет *какое-либо* значение, не превышающее, либо равное *x*. Для нашего примера получаем:

pnorm(-1) [1] 0.1586553

Подбор закона и параметров распределения в R

Подбор распределения заключается в нахождении математической функции, которая бы наилучшим образом представляла наши экспериментальные данные. Исследователь часто сталкивается с такой задачей: у него есть некоторые наблюдения количественного показателя x_1, x_2, … x_n и он желает проверить, принадлежит ли полученная им выборка из неизвестной популяции некой теоретической генеральной совокупности с функцией плотности вероятности $f(\mathbf{x}, q)$, где q является вектором параметров, оцениваемых по имеющимся данным.

Можно выделить 4 шага при подборе распределений (Ricci, 2005):

1) Выбор модели: выдвигается гипотеза о принадлежности выборки некоторому семейству распределений;
2) Оценка параметров теоретического распределения;
3) Оценка качества приближения;
4) Проверка согласия между наблюдаемыми и ожидаемыми значениями с использованием статистических тестов.

Имея одну конкретную выборку и задав форму некоторого теоретического распределения, можно рассчитать точечные оценки его параметров $\hat{\boldsymbol{\theta}}(\mathbf{x})$ (например, среднее и стандартное отклонение нормального распределения). Для этого используются следующие процедуры оценивания: *а*) по аналогии, *б*) метод моментов и *в*) метод максимального правдоподобия.

Метод моментов – это техника конструирования оценок неизвестных параметров, основанная на предполагаемом соотношении выборочных моментов для заданных распределений. В общем случае, для использования метода нужно составить нелинейное уравнение (или систему уравнений, в зависимости от числа неизвестных параметров), где выборочные моменты приравниваются к соответствующим теоретическим. Нахождение корней уравнений выполняется аналитическими методами или с использованием функций, таких, например, как **multiroot**().

При оценке параметров известного семейства вероятностных распределений этот метод в значительной мере вытесняется методом максимального правдоподобия Фишера (MLE – Maximum Likelihood Estimation), так как максимально правдоподобная оценка имеет большую вероятность оказаться ближе к истинному значению оцениваемой величины.

В большинстве функций R, специально предназначенных для оценки параметров, по выбору пользователя реализуется и метод моментов, и метод максимального правдоподобия (более того, последний используется по умолчанию). Для подробного рассмотрения практических коллизий при подборе распределений создадим выборку значений из смеси распределений и постараемся выбрать наиболее подходящую теоретическую модель. При оценке параметров будем использовать функции fitdistr() из пакета MASS и fitdist()из пакета fitdistrplus.

Для выбора *наилучшего* закона распределения из трех возможных можно воспользоваться целым набором мер, таких как средняя абсолютная разность между фактическими и прогнозируемыми значениями, сумма квадратов этих разностей, относительные средние разности, критерий c^2 или та же *D*-статистика

КолмогороваСмирнова. Ориентируясь, например, на значения последнего критерия, можно полагать нашу выборку распределенной по закону Вейбулла (однако, возможно, другие используемые меры дадут иную предупорядоченность моделей).

В общем случае вопрос о том, являются ли полученные оценки параметров наилучшими, обычно остается открытым. Во-первых, соображения, из которых задается вид теоретического распределения, часто оказываются субъективными. Это нередко приводит к тому, что, выполнив вычисления, мы полагаем, что найдены оценки истинных характеристик генерального распределения, тогда как на самом деле получены оценки параметров некого теоретического распределения, которое может быть всего лишь "похожим" на распределение генеральной совокупности, из которого взята выборка. Вовторых, обычно мы ничего не можем сказать об устойчивости модели и ошибке ее воспроизводимости: нет никакой гарантии, что при взятии повторной выборки расхождение искомых величин параметров не окажется слишком большим.

Проблема подбора наилучшего распределения еще более усложняется при обработке счетных данных (*count data*), выборочные элементы которых можно рассматривать как подмножество натуральных чисел. К таким выборкам относятся, например, количество отловленных мышей, результаты стрельбы по мишеням, число избирателей, проголосовавших за каждого кандидата и т.д. Счетные данные занимают некоторое промежуточное положение между числами, представленными на непрерывных шкалах, и измерениями в порядковой шкале.

Традиционно считается, что счетные данные следует аппроксимировать дискретными распределениями типа Пуассона, биномиального или геометрического. Однако известно, что при значении l> 9 и достаточно большом объеме выборки распределение Пуассона хорошо аппроксимируется нормальным распределением. Поэтому "спектр" возможных распределений-претендентов для выборок, состоящих из счетных данных, может включать как дискретные, так и непрерывные теоретические распределения.

Гипотеза о равенстве средних двух генеральных совокупностей

Критерий Стьюдента*t* относится к одним из наиболее давно разработанных и широко используемых статистических критериев. Чаще всего он применяется для проверки нулевой гипотезы о равенстве средних значений двух совокупностей, хотя существует также и одновыборочная модификация этого метода. В этом разделе мы продемонстрируем, как можно реализовать в R статистические тесты, основанные на этом критерии.

Начать, пожалуй, следует с математических допущений, на которых основан критерий Стьюдента. Основных таких допущений, как известно, два:

°сравниваемые выборки должны происходить из нормальнораспределенных совокупностей;

°дисперсии сравниваемых генеральных совокупностей должны быть равны.

Кроме того, в своей исходной форме, *t*-критерий предполагает независимость сравниваемых выборок.

Проверка указанных требований применительно к анализируемым данным должна всегда предшествовать формальному статистическому анализу, в котором задействован критерий Стьюдента (к сожалению, многие исследователи забывают об этом). Методы проверки одного из этих предположений мы рассмотрели в разделе 4.8. Отметим, однако, что условие нормальности распределения данных становится не таким жестким при достаточно больших объемах выборок, а для выборок с разными дисперсиями существует особая модификация *t*-критерия (критерий Уэлча; см. ниже). Кроме того, если сам по себе *t*-критерий корректно отражает степень различия выборок, но неверно оценивается его статистическая значимость с использованием таблиц стандартного распределения Стьюдента, то для решения этой проблемы и расчета *p*-значения, не зависящего от законов

распределения, с успехом могут использоваться методы рандомизации и бутстрепа (Шитиков, Розенберг, 2014).

Одновыборочный t-критерий

Этот вариант критерия Стьюдента служит для проверки нулевой гипотезы о равенстве среднего значения (m_1) генеральной совокупности, из которой была взята выборка, некоторому априори известному значению (m_0).

В общем виде проверка (или, иными словами, "тестирование") этой гипотезы выполняется при помощи *t*-критерия, который рассчитывается как отношение разницы между выборочным средним и известным значением к стандартной ошибке выборочного среднего.

Рассмотрим, как это можно сделать в системе R.

Предположим, у нас имеются данные по суточному потреблению энергии, поступающей с пищей (кДж/сутки), для 11 женщин:

d.intake <- **c**(5260, 5470, 5640, 6180, 6390, 6515,6805, 7515, 7515, 8230, 8770)

Среднее значение для этих 11 наблюдений составляет:

mean(d.intake) [1] 6753.6

Зададимся вопросом: отличается ли это выборочное среднее значение от установленной нормы в 7725 кДж/сутки? Разница между нашим выборочным значением и этим нормативом довольно прилична: 7725 - 6753.6 = 971.4. Но насколько эта разница статистически значима с учетом уровня вариации приведенных выше 11 значений? Ответить на этот вопрос поможет одновыборочный *t*-тест. Как и другие варианты *t*-теста, одновыборочный тест Стьюдента выполняется в R при помощи функции t.test():

t.test(d.intake, mu = 7725)One Sample t-test data:d.intake
t = -2.8208, df = 10, p-value = 0.01814
alternative hypothesis: true mean is not equal to 7725
95 percent confidence interval:
5986.348 7520.925 sample estimates: mean of x6753.636

Видим, что для имеющихся выборочных данных *t*-критерий составляет -2.821 при 10 степенях свободы (df). Вероятность получить такое (либо большее) значение *t* при условии, что проверяемая нулевая гипотеза верна, оказалась весьма мала: p-value = 0.01814(во всяком случае, это меньше 5%). Следовательно, по правилу, представленному выше, мы можем отклонить проверяемую нулевую гипотезу о равенстве выборочного среднего значения нормативу и принять альтернативную гипотезу (alternative hypothesis: true mean is not equal to 7725). Принимая это предположение, мы рискуем ошибиться с вероятностью менее 5%.

Помимо *t*-критерия, числа степеней свободы, *p*-значения и выборочного среднего (sample estimates: mean of x), программа рассчитала также 95%-ный доверительный интервал (95 percent confidence interval) для истинной разницы между выборочным средним значением суточного потребления энергии и нормативом. Если бы мы повторили аналогичный тест много раз для разных групп из 11 женщин, то в 95% случаев эта разница оказалась бы в диапазоне от 5986.3 до 7520.9 кДж/сутки.

Сравнение двух независимых выборок

При сравнении двух выборок проверяемая нулевая гипотеза состоит в том, что обе эти выборки происходят из нормально распределенных генеральных совокупностей с одинаковыми средними значениями.

Рассмотрим пример о суточном расходе энергии (expend) у худощавых женщин (lean) и женщин с избыточным весом (obese), приведенный в книге П. Дальгаарда (Dalgaard, 2008). Данные из этого примера (подробнее см. ?energy) входят в состав пакета ISwR, сопровождающего эту книгу:

library(ISwR) **data**(energy) **attach**(energy) **head**(energy)expend stature 19.21obese
2 7.53lean
3 7.48lean

4 8.08lean
5 8.09lean
6 10.15lean

Соответствующие средние значения потребления энергии в рассматриваемых группах пациенток можно найти с использованием знакомой нам функции tapply(): **tapply**(expend, stature, **mean**) lean obese 8.07 10.30

Вопрос заключается в том, различаются ли эти средние значения статистически?

Проверим гипотезу об отсутствии разницы при помощи *t*-теста:

t.test(expend ~ stature)

Welch Two Sample t-test data:expend by stature

t = -3.8555, df = 15.919, p-value = 0.001411 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval:

-3.459167 -1.004081

sample estimates:

mean in group lean mean in group obese8.06615410.297778

Обратите внимание на использование знака ~ в вызове функции t.test(). Это стандартный для R способ записи *формул*, описывающих связь между переменными. В нашем случае выражение expend ~ stature можно расшифровать как "зависимость суточного потребления энергии (expend) от статуса пациентки (stature)".

Согласно величине полученного *p*-значения (p-value = 0.001411), средний уровень потребления энергии у женщин из рассматриваемых весовых групп статистически значимо различается. Отвергая нулевую гипотезу о равенстве этих средних значений, мы рискуем ошибиться с вероятностью лишь около 0.15%. При этом истинная разница между средними значениями с вероятностью 95% находится в диапазоне от -3.5 до -1.0 (см. 95 percent confidence interval).

Следует подчеркнуть, что при выполнении двухвыборочного *t*-теста функция R по умолчанию принимает, что дисперсии сравниваемых совокупностей *не равны*, и, как следствие, выполняет *t*-тест в модификации Уэлча. Мы можем изменить такое поведение программы, воспользовавшись аргументом var.equal = TRUE: (от *variance* – дисперсия, и *equal* – равный):

t.test(expend ~ stature, var.equal = TRUE)

Two Sample t-test data:expend by stature

t = -3.9456, df = 20, p-value = 0.000799

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-3.411451 -1.051796 sample estimates:

mean in group lean mean in group obese8.06615410.297778

P-значение стало еще меньше, и мы так же, как и после теста в модификации Уэлча, можем сделать вывод о наличии существенной разницы групповых средних. Однако такое совпадение выводов будет иметь место не всегда и, следовательно, на разницу между групповыми дисперсиями (или ее отсутствие) следует обращать серьезное внимание при выборе и интерпретации того или иного варианта *t*-теста.

Сравнение двух зависимых выборок

Зависимыми, или *парными*, являются две выборки, содержащие результаты измерений какого-либо количественного признака, выполненных на одних и тех же объектах. Во многих исследованиях определенный отклик измеряется у одних и тех же объектов *до и после* экспериментального воздействия. При такой схеме эксперимента исследователь более точно оценивает эффект воздействия именно потому, что прослеживает его фактически у каждого уникального объекта.

Нас интересуют свойства выборки, составленной из разностей значений признака у одних и тех же объектов, а точнее – "истинная средняя разность" как результат

экспериментального воздействия (обозначим его δ). Если верна нулевая гипотеза H_0: δ = 0, утверждающая, что средняя разность δ между парами реализаций случайных величин статистически значимо не отличается от нуля, то нет оснований предполагать, что эффект воздействия имеет место.

В книге П. Дальгаарда (Dalgaard, 2008) имеется другой пример о суточном потреблении энергии, измеренном уже у *одних и тех же* 11 женщин до и после периода менструаций:

data(intake) *# изпакета ISwR* **attach**(intake) **head**(intake)pre post 15260 3910
2 5470 4220
3 5640 3885
4 6180 5160
5 6390 5645

Индивидуальные разностипотребления энергии у этих женщин составляют:

post - pre
[1] -1350 -1250 -1755 -1020-745 -1835 -1540 -1540
[9]-725 -1330 -1435

Усреднив эти индивидуальные разницы, получим

mean(post - pre) [1] -1320.5

Задача заключается в том, чтобы оценить, насколько статистически значимо эта средняя разность отличается от нуля. Применим парный критерий Стьюдента (обратите внимание на использование аргумента paired = TRUE):

t.test(pre, post, paired = TRUE)
Paired t-test data:pre and post
t = 11.9414, df = 10, p-value = 3.059e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
1074.072 1566.838 sample estimates:
mean of the differences1320.455

Как видим, рассчитанное программой *p*-значение оказалось намного меньше 0.05, что позволяет нам сделать заключение о наличии существенной разницы в потреблении энергии у исследованных женщин до и после менструации. Истинная величина эффекта (в абсолютном выражении) с вероятностью 95% находится в интервале от 1074.1 до 1566.8 кДж/сутки.

Приведенные выше примеры охватывают наиболее типичные случаи применения критерия Стьюдента. За рамками этого раздела остаются так называемые *односторонние* варианты *t*-теста, когда проверяемая нулевая гипотеза заключается в том, что одно из сравниваемых средних значений больше (или меньше) другого. Однако можно отметить, что односторонний вариант *t*-теста легко реализуется при помощи функции t.test() в сочетании с аргументом alternative, который может принимать одно из трех значений - "two.sided" ("двухсторонний"; выбирается программой по умолчанию), "greater" ("больше") или "less" ("меньше").

Использование рангового критерия Уилкоксона-Манна-Уитни

Как описано выше, одно из важных условий корректного применения критерия Стьюдента состоит в том, что анализируемые выборки должны принадлежать нормально распределенным генеральным совокупностям. В случаях, когда это условие не выполняется, вместо критерия Стьюдента следует использовать его непараметрический аналог – *критерий Уилкоксона* (*Wilcoxon rank test*). Здесь необходимо сразу пояснить, что создатели системы R под названием "критерий Уилкоксона" (или "тест Уилкоксона") объединяют как метод, предложенный собственно Ф. Уилкоксоном (*Wilcoxon*) в 1945 г., так и опубликованный несколько позднее (1947 г.) метод Манна-Уитни. Первый из этих методов обычно

используется для сравнения двух парных выборок, тогда как второй предназначен для сравнения двух независимых выборок. Ниже мы не будем разграничивать эти методы, придерживаясь подхода, принятого в системе R.

Одновыборочный критерий Уилкоксона

Этот вариант критерия (*Wilcoxon signed rank test*) служит для проверки нулевой гипотезы о том, что анализируемая выборка происходит из симметрично распределенной генеральной совокупности с центром в точке μ_0. Алгоритм метода заключается в следующем:

°μ_0 отнимают от каждого выборочного значения;

°получившиеся величины ранжируют по возрастанию, игнорируя знак;

°ранговые номера со знаком + суммируют, получая величину V;

° критерий V сравнивают с критическим значением для заданного уровня значимости и числа степеней свободы.

Альтернативный вариант интерпретации результатов теста заключается в нахождении условной вероятности случайным образом получить значение критерия, равное или превышающее эмпирическую величину V, при условии истинности нулевой гипотезы.

Ниже при описании теста Уилкоксона мы будем использовать примеры, рассмотренные в предыдущем разделе для иллюстрации *t*-теста Стьюдента и представленные в пакете ISwR. Это позволит нам сравнить результаты, получаемые при помощи обоих методов. В частности, обратимся к данным о суточном потреблении энергии у 11 женщин из книги и выясним, имеются ли отличия от нормативного значения 7725 кДж/сутки:

d.intake <- **c**(5260, 5470, 5640, 6180, 6390, 6515,6805, 7515, 7515, 8230, 8770)

Для выполнения теста Уилкоксона в системе R используется функция wilcox.test():

wilcox.test(d.intake, mu = 7725)

Wilcoxon signed rank test with continuity correction

data:d.intake

V = 8, p-value = 0.0293

alternative hypothesis: true location is not equal to 7725 Warning message:

In wilcox.test.default(d.intake, mu = 7725) :cannot compute exact p-value with ties

Видим, что рассчитанное значение критерия V = 8. Вероятность случайно получить такое (или превышающее его) значение при условии, что нулевая гипотеза верна, не превышает 0.05 (p-value = 0.0293). Это позволяет нам отклонить нулевую гипотезу о том, что суточное потребление энергии у обследованных 11 женщин не отличается от принятой нормы. Обратите внимание на выданное программой предупреждение о том, что полученное значение вероятности *p* не является точным из-за наличия в данных значений с одинаковыми рангами (Warning message... cannot compute exact pvalue with ties). Проблема расчета точных *p*-значений при наличии повторяющихся значений в данных характерна для статистических методов, основанных на рангах, и критерий Уилкоксона здесь, увы, не исключение (хотя можно упомянуть функцию wilcox_test() из пакета coin). При наличии повторяющихся наблюдений *p*-значение рассчитывается путем аппроксимации распределения критерия Уилкоксона нормальным распределением (см. справочный файл по функции ?wilcox.test).

Сравнение двух независимых выборок

Если сравниваемые выборки являются независимыми (аргумент paired = FALSE), то мы имеем дело с критерием Уилкоксона, который в англоязычной литературе называют *Wilcoxon rank sum test* (как было отмечено выше, этот метод называют также методом Манна-Уитни). Проверяемая с его помощью нулевая гипотеза состоит в том, что центры распределений, из которых происходят сравниваемые выборки, смещены относительно друг друга на величину μ (например, $\mu = 0$). Алгоритм метода состоит в следующем:

°все имеющиеся значения ранжируют, игнорируя их знак;

°ранги значений, принадлежащих к первой группе, суммируют, получая величину W;

° W сравнивают со значением, которое можно было бы ожидать при верной нулевой гипотезе и имеющемся числе степеней свободы.

Альтернативный подход – находят вероятность случайным образом получить значение W, равное или превышающее наблюдаемое значение (при условии истинности нулевой гипотезы).

Используем рассмотренный ранее пример о суточном расходе энергии (expend) у худощавых женщин (lean) и женщин с избыточным весом (obese):

data(energy) *# изпакета ISwR* **attach**(energy) **str**(energy) 'data.frame': 22 obs. of2 variables:

$ expend : num9.21 7.53 7.48 8.08 8.09 ...

$ stature: Factor w/ 2 levels "lean","obese": 2 1 1 1 1 1 1 1 1 1 ...

Проверим гипотезу об отсутствии разницы в потреблении энергии у женщин из этих двух групп при помощи критерия Уилкоксона для независимых выборок:

wilcox.test(expend ~ stature, paired = FALSE)

Wilcoxon rank sum test with continuity correction

data:expend by stature

W = 12, p-value = 0.002122

alternative hypothesis: true location shift is not equal to 0 Warning message:

In wilcox.test.default(x = c(7.53, 7.48, 8.08, 8.09, 10.15, 8.4,:cannot compute exact p-value with ties

Согласно полученному *p*-значению (p-value = 0.002122), потребление энергии у женщин из рассматриваемых весовых групп статистически значимо различается. Отвергая нулевую гипотезу о равенстве потребляемой энергии, мы рискуем ошибиться с вероятностью лишь около 0.2%.

Сравнение двух зависимых выборок

Понятие "зависимые", или "связные выборки" обсуждалось ранее в разделе, посвященном критерию Стьюдента. Сейчас для нас более важен тот факт, что обе сравниваемые выборки происходят из ненормально распределенных генеральных совокупностей. Это дает нам весомые основания выполнить сравнение при помощи парного рангового критерия Уилкоксона.

Как и в парном тесте Стьюдента, находят разницу между всеми имеющимися парными выборочными наблюдениями с целью проверить нулевую гипотезу о том, что медиана полученных разностей равна нулю (либо какому-либо другому, отличному от нуля значению). Здесь (псевдо)-медианой распределения F называют медиану распределения $(u + v) / 2$, где u и v являются независимыми переменными, каждая из которых имеет распределение F. Если распределение F симметрично, псевдомедиана и медиана совпадают (подробнее см. ?wilcox.test).

Используем рассмотренный ранее пример о суточном потреблении энергии, измеренном у *одних и тех же* 11 женщин до и после периода менструаций:

data(intake) *# изпакета ISwR* **attach**(intake)

Сравнить два периода по потреблению энергии при помощи критерия Уилкоксона можно следующим образом (обратите внимание на использование аргумента paired = TRUE):

wilcox.test(pre, post, paired = TRUE)

Wilcoxon signed rank test with continuity correction data:pre and post

V = 66, p-value = 0.00384

alternative hypothesis: true location shift is not equal to 0 Warning message:

In wilcox.test.default(pre, post, paired = T) :cannot compute exact p-value with ties

Как видим, рассчитанное программой *p*-значение оказалось меньше 0.05, что позволяет нам сделать заключение о наличии статистически значимой разницы в потреблении энергии у исследованных женщин до и после менструации. (Для сравнения: *p*-значение, полученное при помощи критерия Стьюдента было << 0.001). Мы можем оценить

доверительный интервал, в котором с определенной вероятностью находится истинная величина эффекта, воспользовавшись аргументом conf.int (вероятность задается при помощи аргумента conf.level; по умолчанию рассчитывается 95%-ный доверительный интервал):

wilcox.test(pre, post, paired = TRUE, conf.int = TRUE) Wilcoxon signed rank test with continuity correction

data:pre and post

V = 66, p-value = 0.00384

alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:

1037.5 1582.5 sample estimates:

(pseudo)median

1341.332 Warning messages:

1: In wilcox.test.default(pre, post, paired = TRUE, conf.int = TRUE):

cannot compute exact p-value with ties

2: In wilcox.test.default(pre, post, paired = TRUE, conf.int = TRUE):cannot compute exact confidence interval with ties

Видим, что истинная разность уровней потребленной энергии с вероятностью 95% находится в интервале от 1037.5 до 1581.5 кДж/сутки. Опять-таки, из-за наличия повторяющихся наблюдений, расчет точных доверительных пределов оказался невозможным. Псевдомедиана ((pseudo)median) индивидуальных разностей между парными значениями потребления энергии была оценена в 1341.3 кДж/сутки.

Важно отметить одно из ограничений критерия Уилкоксона для двух выборок (зависимых или независимых): если общее количество наблюдений не превышает 6, то обнаружить разницу между выборками с уровнем ошибки в 5% просто невозможно.

Гипотеза об однородности дисперсий

В общем виде ***F-критерий Фишера***, или *F*-тест, используется для сравнения дисперсий двух нормально распределенных генеральных совокупностей. Генеральные дисперсии оцениваются на основе выборок, и сам критерий непосредственно рассчитывается как отношение одной выборочной дисперсии к другой: $F = s_1^2 / s_2^2$. На практике в числитель приведенной формулы обычно помещают бóльшую дисперсию, а в знаменатель – меньшую.

При выполнении *F*-теста и интерпретации получаемых с его помощью результатов важно помнить о следующих ограничениях:

°сравниваемые совокупности должны быть нормально распределены;

°сравниваемые совокупности должны быть статистически независимыми.

Для выполнения теста Фишера в R имеется функция var.test() (от *variance* – дисперсия, и *test* – тест). Используем рассмотренный ранее пример о суточном потреблении энергии у худощавых женщин (lean) и женщин с избыточным весом (obese):

data(energy, package="ISwR")

attach(energy) **colnames**(energy)expend stature

Дисперсии в этих двух весовых группах женщин можно легко сравнить следующим образом:

var.test(expend ~ stature)

F test to compare two variances

data:expend by stature

F = 0.7844, num df = 12, denom df = 8, p-value = 0.6797 alternative hypothesis: true ratio of variances is not equal to 1 95 percent confidence interval:

0.1867876 2.7547991 sample estimates: ratio of variances0.784446

Как видим, полученное *p*-значение значительно превышает 5%-ный уровень значимости, на основании чего мы *не можем* отклонить нулевую гипотезу о равенстве

дисперсий в исследованных совокупностях. Истинное отношение сравниваемых дисперсий с вероятностью 95% находится в интервале от 0.19 до 2.75 (см. 95 percent confidence interval).

Исходя из данного результата, мы, например, вправе были бы использовать вариант *t*-критерия Стьюдента для совокупностей с одинаковыми дисперсиями при сравнении среднего потребления энергии у женщин из рассматриваемых весовых групп.

Проверка однородности дисперсии в нескольких группах

Решение более общей задачи проверки однородности дисперсии в двух или более группах осуществляется с использованием различных классических и непараметрических тестов: Левене, Бартлетта, Кохрана, Хартли, Флигнера, Ансари-Бредли, Сижела-Тьюки, Муда и др.Подробное исследование распределения используемых при этом статистик и сравнительный анализ мощности критериев выполнили.

Критерий Флигнера-Килина (*median-centering Fligner-Killeen test*) не требует предположений о нормальности сравниваемых выборок и оценивает совокупное распеределение *рангов* абсолютных разностей $z_{ij} = |x_{ij} - c_i|$,

где c_i – выборочная медиана для i-й группы.

Тестовая статистика Флигнера-Килина, как и статистика Бартлетта, имеет стандартное c^2-распределение с $(k - 1)$ степенями свободы.

Рассмотрим реализацию в среде R перечисленных тестов на примере данных, полученных в ходе эксперимента по изучению эффективности шести видов инсектицидных средств:

data(InsectSprays)

Тест Левене можно выполнить при помощи функции leveneTest() из пакета car. **library**(car)

leveneTest(count ~ spray, data = InsectSprays)

Levene's Test for Homogeneity of Variance (center = median)

Df F valuePr(>F) group53.8214 0.004223 **

66

Signif. codes:0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

leveneTest(count ~ spray, data = InsectSprays, center = mean)

Levene's Test for Homogeneity of Variance (center = mean)

Df F valuePr(>F) group56.4554 6.104e-05 ***

66

Signif. codes:0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Аналогично тесты Бартлетта и Флигнера-Килина выполняются при помощи функций bartlett.test() и fligner.test() соответственно:

bartlett.test(count ~ spray, data = InsectSprays)

Bartlett test of homogeneity of variances

data:count by spray Bartlett's K-squared = 25.9598, df = 5, p-value = 9.085e-05

fligner.test(count ~ spray, data = InsectSprays) Fligner-Killeen test of homogeneity of variances data:count by spray

Fligner-Killeen: med chi-squared = 14.4828, df = 5, p-value = 0.01282

Все тесты показывают, что нулевая гипотеза о равенстве дисперсий в группах должна быть отклонена, однако оценки значимости p существенно разнятся в зависимости от того, используется для анализа среднее или медиана (что косвенно свидетельствует о слабом предположении в отношении нормальности распределения исследуемой совокупности).

Дисперсионный анализ

Рассмотренный ранее *t*-критерий Стьюдента (равно как и его непараметрические аналоги) предназначен для сравнения исключительно двух совокупностей. Однако часто он

неверно используется для попарного сравнения большего количества групп, что приводит к так называемому *эффекту множественных сравнений* (англ. *multiple comparisons;* Гланц, 1999, с. 101-104). Об этом эффекте и о том, как с ним бороться, мы поговорим позднее в граве 6. Здесь же мы опишем принципы *однофакторного дисперсионного анализа*, как раз предназначенного для *одновременного* сравнения средних значений двух и более групп. Принципы дисперсионного анализа (англ. *analysis of variance*, ANOVA) были разработаны в 1920-х гг. сэром Рональдом Эйлмером Фишером (англ. Ronald Aylmer Fisher) – «*гением, едва ли не в одиночку заложившим основы современной статистики*».

Может возникнуть вопрос: почему метод, используемый для сравнения *средних* значений, называется *дисперсионным* анализом? Все дело в том, что при установлении разницы между средними значениями мы в действительности сравниваем дисперсии анализируемых совокупностей. Однако обо всем по порядку...

Постановка задачи

Рассмотренный ниже пример заимствован из книги (Maindonald, Braun, 2010). Имеются данные о массе кустов томатов (всё растение целиком; weight, в кг), которые выращивали в течение 2 месяцев при трех разных экспериментальных условиях (trt, от *treatment*)– при поливе водой (Water), в среде с добавлением удобрения (Nutrient), а

также в среде с добавлением удобрения и гербицида 2,4-D (Nutrient+24D):

```
# Создадим таблицу с данными: tomato <-
data.frame(weight= c(1.5, 1.9, 1.3, 1.5, 2.4, 1.5, # water
1.5, 1.2, 1.2, 2.1, 2.9, 1.6, # nutrient
1.9, 1.6, 0.8, 1.15, 0.9, 1.6), # nutrient+24Dtrt = rep(c("Water", "Nutrient",
"Nutrient+24D"), c(6, 6, 6)))
```

Переменная trt представляет собой фактор с тремя уровнями. Для более наглядного сравнения экспериментальных условий в последующем, сделаем уровень "Water" базовым (англ. *reference*), т.е. уровнем, с которым R будет сравнивать все остальные уровни. Этоможносделатьприпомощифункцииrelevel(): tomato$trt<- **relevel**(tomato$trt, ref = "Water")

Подлежащую проверке нулевую гипотезу можно сформулировать так:

H_0: $\mu_1 = \mu_2 = \mu_3$,

т.е. всеполученные измерения веса растений происходят из одной нормально распределенной генеральной совокупности. Другими словами, нулевая гипотеза утверждает, что в действительности исследованные условия выращивания растений не оказывают никакого влияния на вес последних, а *наблюдаемые различия между групповыми средними несущественны и вызваны влиянием случайных факторов*.

Подчеркнем еще раз, что рассматриваемый пример соответствует случаю *однофакторного* дисперсионного анализа: изучается действие одного фактора – условий выращивания (с тремя уровнями Water, Nutrient и Nutrient+24D) на интересующую нас переменную-отклик (вес растений).

К сожалению, исследователь почти никогда не имеет возможности изучить всю генеральную совокупность. Как же нам тогда узнать, верна ли приведенная выше нулевая гипотеза, располагая только выборочными данными? Мы можем сформулировать этот вопрос иначе: *какова вероятностьполучить наблюдаемые различия между групповыми средними, извлекая случайные выборки из одной нормально распределенной генеральной совокупности*? Для ответа на этот вопрос нам потребуется статистический критерий, который количественно характеризовалбы величину различий между сравниваемыми группами.

Перед тем, как сконструировать такой критерий, зададимся еще одним вопросом:

что заставляет насдумать, что различия между ними неслучайны? Чтобы лучше понять свойства имеющихся данных, визуализируем их при помощи одномерной диаграммы рассеяния:

```
attach(tomato) stripchart(weight ~ trt, xlab = "Вес, кг", ylab = "Условия")
```

Для подтверждения своего визуального впечатления рассчитаем групповые средние (Means<- **tapply**(weight, trt, **mean**))

WaterNutrient Nutrient+24D

1.6833331.7500001.325000

Таким образом, для оценки различий между группами следует каким-то образом сравнить разброс групповых средних с разбросом значений внутри групп. Это – ключевая идея дисперсионного анализа.

Две оценки дисперсии при дисперсионном анализе

Чем больше разброс выборочных средних и чем меньше разброс значений внутри групп, тем меньше вероятность того, что наши группы являются случайными выборками из одной совокупности. Дисперсию генеральной совокупности можно оценить двумя способами. С одной стороны, оценкой дисперсии генеральной совокупностью будет дисперсия, вычисленная *для каждой группы*. Такая оценка не будет зависеть от различий групповых средних. С другой стороны, при верной нулевой гипотезе (см. выше) *разбросгрупповых средних* тоже позволит оценить дисперсию генеральной совокупности. Очевидно, что такая оценка уже будет зависеть от различий между группами.

Если экспериментальные группы – это случайные выборки из одной и той же нормально распределенной генеральной совокупности, то оба способа оценки генеральной дисперсии должны давать примерно одинаковые результаты. Соответственно, если эти оценки действительно оказываются близки, то мы не можем отвергнуть нулевую гипотезу. И наоборот: если разница между этими оценками оказывается существенной, мы можем принять альтернативную гипотезу: *маловероятно, что мы получили бы наблюдаемые различия между группами, если бы они были просто случайными выборками из одной нормально распределенной генеральной совокупности*.

Выполнение дисперсионного анализа в R

Дисперсионный анализ в R можно выполнить при помощи базовых функций aov() и lm(). Для нашего примера получаем:

summary(**aov**(weight ~ trt, data = tomato))

Df Sum Sq Mean Sq F value Pr(>F) trt20.6270.31351.2020.328 Residuals153.9120.2608

В приведенных результатах строка, обозначенная как trt, соответствует источнику дисперсии в данных, связанному с действием изучаемого экспериментального фактора − условий выращивания растений. Строка, обозначенная как Residuals, характеризует внутригрупповую дисперсию (ее еще называют *флуктуирующей* или *остаточной* дисперсией − в том смысле, что она не может быть объяснена влиянием экспериментального фактора и является следствием случайной флуктуации данных).

Столбец Sum Sq содержит *SSB* и *SSW*, а столбец Mean Sq −меж- и внутригрупповую дисперсии *MSB* и *MSW*. В столбце F value представлено рассчитанное по имеющимся данным значение *F*-критерия. Наконец, в столбце Pr(>F) представлена вероятность получить *F*-значение, равное или превышающее выборочное наблюдаемое значение, при условии, что нулевая гипотеза верна. Как видим, эта вероятность достаточно высока. Во всяком случае, она превышает 5%-ный уровень значимости и, на этом основании, мы заключаем, что нулевая гипотеза верна. Таким образом, с достаточно высокой степенью уверенности мы можем утверждать, что экспериментальные условия не оказали существенного влияния на вес растений.

Двухфакторный дисперсионный анализ

Описанный однофакторный дисперсионный анализ заключается в выяснении влияния какого-то одного фактора на интересующую нас количественную переменную. Однако очень редко тот или иной процесс определяется только одним фактором. Напротив − обычнонаблюдается одновременное влияние многих факторов. Задача исследователя − выявить, какие факторы оказывают существенное влияние на изучаемое явление, а какие

можно исключить из рассмотрения. Как будет показано ниже, двухфакторный дисперсионный анализ (англ. *two-way analysis of variance,* или *two-way ANOVA*) позволяет установить одновременное влияние двух факторов, а также взаимодействие между этими факторами. При наличии более двух факторов говорят о *многофакторном дисперсионном анализе* (англ. *multifactor ANOVA*, котрый не следует путать с многомерным анализом *MANOVA − multivariate ANOVA*!).

Критерий хи-квадрат

Критерий χ^2 ("хи-квадрат", также "критерий согласия Пирсона") имеет чрезвычайно широкое применение в статистике. В общем виде можно сказать, что он используется для проверки нулевой гипотезы о подчинении наблюдаемой случайной величины определенному теоретическому закону распределения. Однако конкретная формулировка проверяемой гипотезы от случая к случаю будет варьировать.

В этом разделе мы опишем принцип работы критерия χ^2 на гипотетическом примере из иммунологии. Представим, что мы выполнили эксперимент по установлению эффективности подавления развития микробного заболевания при введении в организм соответствующих антител. Всего в эксперименте было задействовано 111 мышей, которых мы разделили на две группы, включающие 57 и 54 животных соответственно. Первой группе мышей сделали инъекции патогенных бактерий с последующим введением сыворотки крови, содержащей антитела против этих бактерий. Животные из второй группы служили контролем – им сделали только бактериальные инъекции. После некоторого времени инкубации оказалось, что 38 мышей погибли, а 73 выжили. Из погибших 13 принадлежали первой группе, а 25 – ко второй (контрольной).

Проверяемую в этом эксперименте нулевую гипотезу можно сформулировать так: введение сыворотки с антителами не оказывает никакого влияния на выживаемость мышей. Иными словами, мы утверждаем, что наблюдаемые различия в выживаемости мышей (77.2% в первой группе против 53.7% во второй группе) совершенно случайны и не связаны с действием антител.

Для проверки сформулированной выше нулевой гипотезы нам необходимо знать, какова была бы ситуация, если бы антитела действительно не оказывали никакого действия на выживаемость мышей. Другими словами, нужно рассчитать *ожидаемые частоты* для соответствующих ячеек таблицы сопряженности. Как это сделать?

В эксперименте всего погибло 38 мышей, что составляет 34.2% от общего числа задействованных животных. Если введение антител не влияет на выживаемость мышей, в обеих экспериментальных группах должен наблюдаться одинаковый процент смертности, а именно 34.2%. Рассчитав, сколько составляет 34.2% от 57 и 54, получим 19.5 и 18.5. Это и есть ожидаемые величины смертности в наших экспериментальных группах. Аналогичным образом рассчитываются и ожидаемые величины выживаемости: поскольку всего выжили 73 мыши, или 65.8% от общего их числа, то ожидаемые частоты выживаемости составят 37.5 и 35.5. Ожидаемые частоты довольно сильно отличаются от наблюдаемых, т.е. введение антител, похоже, все-таки оказывает влияние на выживаемость мышей, зараженных патогенным микроорганизмом. Это впечатление мы можем выразить количественно при помощи критерия согласия Пирсона χ^2:

$$\chi^2 = \sum(f_o - f_e)^2 / f_e,$$

где f_o и f_e − наблюдаемые и ожидаемые частоты соответственно.

Суммирование производится по всем ячейкам таблицы. Так, для рассматриваемого примера имеем

$$\chi^2 = (13 - 19.5)^2/19.5 + (44 - 37.5)^2/37.5 + (25 - 18.5)^2/18.5 + (29 - 35.5)^2/35.5 = 2.16 + \\ +1.12 + 2.31 + 1.20 = 6.79$$

Достаточно ли велико полученное значение χ^2, чтобы отклонить нулевую гипотезу? Для ответа на этот вопрос необходимо найти соответствующее критическое значение критерия. Число степеней свободы для χ^2 рассчитывается как $df = (R - 1)(C - 1)$, где R и C– количество строк и столбцов в таблице сопряженности.

В нашем случае $df = (2 - 1)(2 - 1) = 1$. Зная число степеней свободы, мы легко можем узнать критическое значение χ^2 при помощи стандартной R-функции qchisq():

qchisq(p = 0.95, **df** = 1)

[1] 3.841459

Таким образом, при одной степени свободы только в 5% случаев величина критерия χ^2 превышает 3.841. Полученное нами значение 6.79 значительно превышает это критическое значение, что дает нам право отвергнуть нулевую гипотезу об отсутствии связи между введением антител и выживаемостью зараженных мышей. Отвергая эту гипотезу, мы рискуем ошибиться с вероятностью менее 5%.

Следует отметить, что приведенная выше формула для критерия χ^2 дает несколько завышенные значения при работе с таблицами сопряженности размером 2´2. Причина заключается в том, что распределение самого критерия χ^2 является непрерывным, тогда как частоты бинарных признаков ("погибло" / "выжило") по определению дискретны. В связи с этим при расчете критерия принято вводить так называемую *поправку на непрерывность*, или *поправку Йетса*.

В нашем случае критерий χ^2 с поправкой на непрерывность составил бы 5.792, и нулевая гипотеза об отсутствии эффекта антител все равно была бы отклонена. Возможно, однако, что в других ситуациях это сделать так легко не удалось бы.

Безусловно, нет необходимости выполнять приведенные выше вычисления вручную. В R для этого имеется стандартная функция chisq.test(). При работе с этой функцией данные оформляются в виде матрицы, напоминающей приведенную выше таблицу сопряженности:

mice <- **matrix**(**c**(13, 44, 25, 29), nrow = 2, byrow = TRUE) mice *# просмотр содержимого матрицы*

[,1] [,2]

[1,]1344 [2,]2529

chisq.test(mice) *# тест хи-квадрат*

Pearson's Chi-squared test with Yates' continuity correction

data:mice X-squared = 5.7923, df = 1, p-value = 0.0161

Каквидим, R автоматическиприменяетпоправкуЙетсананепрерывность (Pearson's Chi-squared test with Yates' continuity correction).

Рассчитанное программой значение χ^2 составило 5.79213. Мы можем отклонить нулевую гипотезу об отсутствии эффекта антител, рискуя ошибиться с вероятностью чуть более 1% (p-value = 0.0161).

Критерий Мак-Немара

Рассмотренный выше критерий χ^2 для анализа таблиц сопряженности размером 2´2 применим только в отношении независимых наблюдений. Если же учет какого-либо дихотомического признака выполняется, например, на одних и тех же испытуемых, то вместо критерия χ^2 следует использовать *критерий Мак-Немара*, названный по имени автора – американского психолога и статистика, описавшего этот критерий.

Представим такую ситуацию: 15-ти испытуемым дали попробовать новый освежительный напиток, который скоро должен будет поступить в продажу. Испытуемых попросили ответить на вопрос, понравился ли им этот напиток - с вариантами ответа "да" или "нет". После этого *тем же* 15 участникам эксперимента показали рекламу, содержащую информацию о пользе и вкусовых качествах напитка. По завершении просмотра испытуемым снова дали выпить напиток и попросили ответить на ранее поставленный

вопрос. Производитель напитка желает выяснить эффективность рекламы, изучив изменения в предпочтениях участников эксперимента после ее просмотра. Данные, полученые в ходе эксперимента, могли бы выглядеть так:

data<- **read.table**(header=TRUE, con<- textConnection("subjecttimeresult1pre0
1 post1
2 pre1….
15pre0
15 post1
"))

В таблице data столбец subject содержит идентификационные номера испытуемых, столбец time – метки, отражающие время проведения опроса (pre – до показа рекламы, post – после), а столбец result – ответы на вопрос (закодированы как 0 – "нет" и 1 – "да"). Данные умышленно приведены сначала в так называемом "длинном формате", чтобы продемонстрировать еще раз, что *одни и те же* 15 испытуемых были опрошены дважды и что их ответы могли измениться после просмотра рекламы.

Подобно критерию χ^2, критерий Мак-Немара работает с таблицей сопряженности размером 2´2. Для преобразования объекта data в такую таблицу необходимо выполнить несколько дополнительных операций. Сначала преобразуем data в таблицу "широкого формата", для чего используем возможности пакета reshape2:

 # Если пакет еще не установлен на Вашем компьютере: **install.packages**("reshape2")
 # Преобразуемданныев "широкийформат": data.wide <- **dcast**(data, subject ~ time, value.var="result") data.widesubject post pre 1110
2 211
3 310 ….
14 1400
15 1510

Теперь эти данные можно легко свести в матрицу, соответствующую таблице сопряженности (использованная ниже функция table() входит в базовый набор функций R):

ct <- **table**(data.wide[,**c**("pre","post")]) ctpost pre 0 10 2 8
1 1 4

Полученная таблица сопряженности отражает изменения предпочтений участников эксперимента до и после показа рекламы (сравните с критерием Стьюдента для парных выборок). Так, например, мы видим, что двоим испытуемым напиток не нравился как до, так и после показа рекламы (пересечение строки 0 и столбца 0); восьми испытуемым напиток не нравился до показа рекламы, но понравился после (пересечение строки 0 и столбца 1), и т.д. В общем виде таблицу сопряженности для критерия Мак-Немара мы можем представить в виде таблицы 3.2:

Таблица 3.2 – Критерий Мак-Немара

	Тест 2 (-)	Тест 2 (+)	Всего
Тест 1 (-)	*a*	*b*	*a* + *b*
Тест 1 (+)	*c*	*d*	*c* + *d*
Всего	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d*

Критерий Мак-Немара предназначен для проверки нулевой гипотезы о том, что маргинальные частоты строк и столбцов таблицы сопряженности не различаются, т.е.

$p(a) + p(b) = p(a) + p(c)$ и$p(c) + p(d) = p(b) + p(d)$.

После сокращения получаем:

H_0: $p(b) = p(c)$иH_1: $p(b) \neq p(c)$.

Критерий рассчитывается следующим образом:

$$Q = (b - c)^2 / (b + c)$$

При больших объемах выборок (примерно 50 и выше) Q имеет распределение, близкое к распределению χ^2 с одной степенью свободы. Более точное приближение достигается при помощи *поправки Эдвардса*):$Q = (|b - c| - 1)^2 / (b + c)$

В R приведенные формулы реализованы в базовой функции mcnemar.test():

```
mcnemar.test(ct)
McNemar's Chi-squared test with continuity correction
data:ct McNemar's chi-squared = 4, df = 1, p-value = 0.0455
```

По умолчанию применяется поправка Эдвардса. При необходимости ее можно отключить, воспользовавшись аргументом correct:

```
mcnemar.test(ct, correct = FALSE)
McNemar's Chi-squared test data:ct McNemar's chi-squared = 5.4444, df = 1, p-value = 0.01963
```

Как видим, в обоих случаях p-value< 0.05, что дает нам основания отклонить нулевую гипотезу об отсутствии эффекта рекламы - число испытуемых, которым напиток понравился после ее просмотра, увеличилось.

Обратите внимание: хотя критерий Мак-Немара имеет дело с таблицей сопряженности 2´2, структура этой таблицы в корне отличается от таковой для критерия χ^2. В случае с χ^2 факт повторного учета дихотомического признака на тех же объектах не отражен:

```
table(data[,c("time","result")])
result time01post3 12pre105
```

Список использованной литературы:

1. Crawley, M.J. The R Book. 2nd ed. – Wiley, 2013. – 1076 p.
2. Højsgaard, S., Edwards, D., Lauritzen, S. Graphical Models with R.– Springer, 2012.– 182 p.
3. Jockers, M.L. Text Analysis with R for Students of Literature. – Springer, 2014. – 199 p.
4. Kabacoff, R. R in Action: Data Analysis and Graphics With R. Manning Publications, 2011. 447p. (Рус. пер.: Кабаков Р.И. R в действии: Анализ и визуализация данных в программе R / пер. с англ. П.А. Волкова. Москва: ДМКПресс, 2014. 580 с.)
5. Karp, N. R Commander: – An Introduction. 2014. –52 p.
6. Kateri, M. Contingency Table Analysis: Methods and Implementation Using R. – Springer, 2014. –315 p.
7. Klemelä, J. Multivariate Nonparametric Regression and Visualization: With R and Applications to Finance.– Wiley, 2014. – 396 p.
8. Анализ данных : учебник для академического бакалавриата / В. С. Мхитарян [и др.] ; под редакцией В. С. Мхитаряна. — Москва : Издательство Юрайт, 2019. — 490 с.— ISBN 978-5-534-00616-2. — Текст : электронный // ЭБС Юрайт [сайт]. — URL: https://biblio-online.ru/bcode/432178 (дата обращения: 10.04.2019).
9. Соловьев, В. И.. Анализ данных в экономике. Теория вероятностей и прикладная статистика в Microsoft Excel: [Текст]: учебник – Москва: КНОРУС, 2018. – 387 с.
10. Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. [Электронный ресурс] – Режим доступа:http://r-analytics.blogspot.com (дата обращения 21.01.2019).
11. Зададаев,. С.А. Математика на языке R: учебник [Текст]:/ Финансовый университет при Правительстве РФ. – Москва: ПРОМЕТЕЙ, 2018. – 324 с.
12. Савельев, А.А., Мухарамова С.С., Пилюгин А.Г. Использование языка R для статистической обработки данных. [Текст]: учебно-методическое пособие. – Казань: Казанский университет, 2007. – 28 с.
13. Статистический анализ данных в системе R. [Текст] :Учебное пособие / А. Г. Буховец [и др.]; под ред. А. Г. Буховец. – Воронеж: ВГАУ, 2010. – 124 с.
14. Яу, Н. Искусство визуализации в бизнесе: Как представить сложную информацию простыми образами [Текст]:/ пер. с англ. А. Кирова. – Москва: Манн, Иванов и Фербер, 2013. – 352 с.
15. Потемкин А. В. Анализ данных = Data analysis. Tutorial: учебное пособие / А.В. Потемкин, И.М. Эйсымонт ; Финуниверситет, Каф. "Теория вероятностей и математич. статистика". – М.: Финуниверситет, 2014 .– 160 с.

Приложение 1

Таблица значений функции $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-z^2/2}\, dz$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,0000	0,32	0,1255	0,64	0,2389	0,96	0,3315
0,01	0,0040	0,33	0,1293	0,65	0,2422	0,97	0,3340
0,02	0,0080	0,34	0,1331	0,66	0,2454	0,98	0,3365
0,03	0,0120	0,35	0,1368	0,67	0,2486	0,99	0,3389
0,04	0,0160	0,36	0,1406	0,68	0,2517	1,00	0,3413
0,05	0,0199	0,37	0,1443	0,69	0,2549	1,01	0,3438
0,06	0,0239	0,38	0,1480	0,70	0,2580	1,02	0,3461
0,07	0,0279	0,39	0,1517	0,71	0,2611	1,03	0,3485
0,08	0,0319	0,40	0,1554	0,72	0,2642	1,04	0,3508
0,09	0,0359	0,41	0,1591	0,73	0,2673	1,05	0,3531
0,10	0,0398	0,42	0,1628	0,74	0,2703	1,06	0,3554
0,11	0,0438	0,43	0,1664	0,75	0,2734	1,07	0,3577
0,12	0,0478	0,44	0,1700	0,76	0,2764	1,08	0,3599
0,13	0,0517	0,45	0,1736	0,77	0,2794	1,09	0,3621
0,14	0,0557	0,46	0,1772	0,78	0,2823	1,10	0,3643
0,15	0,0596	0,47	0,1808	0,79	0,2852	1,11	0,3665
0,16	0,0636	0,48	0,1844	0,80	0,2881	1,12	0,3686
0,17	0,0675	0,49	0,1879	0,81	0,2910	1,13	0,3708
0,18	0,0714	0,50	0,1915	0,82	0,2939	1,14	0,3729
0,19	0,0753	0,51	0,1950	0,83	0,2967	1,15	0,3749
0,20	0,0793	0,52	0,1985	0,84	0,2995	1,16	0,3770
0,21	0,0832	0,53	0,2019	0,85	0,3023	1,17	0,3790
0,22	0,0871	0,54	0,2054	0,86	0,3051	1,18	0,3810
0,23	0,0910	0,55	0,2088	0,87	0,3078	1,19	0,3830
0,24	0,0948	0,56	0,2123	0,88	0,3106	1,20	0,3849
0,25	0,0987	0,57	0,2157	0,89	0,3133	1,21	0,3869
0,26	0,1026	0,58	0,2190	0,90	0,3159	1,22	0,3883
0,27	0,1064	0,59	0,2224	0,91	0,3186	1,23	0,3907
0,28	0,1103	0,60	0,2257	0,92	0,3212	1,24	0,3925
0,29	0,1141	0,61	0,2291	0,93	0,3238	1,25	0,3944
0,30	0,1179	0,62	0,2324	0,94	0,3264		
0,31	0,1217	0,63	0,2357	0,95	0,3289		

Продолжение Приложения 1

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
1,26	0,3962	1,59	0,4441	1,92	0,4726	2,50	0,4938
1,27	0,3980	1,60	0,4452	1,93	0,4732	2,52	0,4941
1,28	0,3997	1,61	0,4463	1,94	0,4738	2,54	0,4945
1,29	0,4015	1,62	0,4474	1,95	0,4744	2,56	0,4948
1,30	0,4032	1,63	0,4484	1,96	0,4750	2,58	0,4951
1,31	0,4049	1,64	0,4495	1,97	0,4756	2,60	0,4953
1,32	0,4066	1,65	0,4505	1,98	0,4761	2,62	0,4956
1,33	0,4082	1,66	0,4515	1,99	0,4767	2,64	0,4959
1,34	0,4099	1,67	0,4525	2,00	0,4772	2,66	0,4961
1,35	0,4115	1,68	0,4535	2,02	0,4783	2,68	0,4963
1,36	0,4131	1,69	0,4545	2,04	0,4793	2,70	0,4965
1,37	0,4147	1,70	0,4554	2,06	0,4803	2,72	0,4967
1,38	0,4162	1,71	0,4564	2,08	0,4812	2,74	0,4969
1,39	0,4177	1,72	0,4573	2,10	0,4821	2,76	0,4971
1,40	0,4192	1,73	0,4582	2,12	0,4830	2,78	0,4973
1,41	0,4207	1,74	0,4591	2,14	0,4838	2,80	0,4974
1,42	0,4222	1,75	0,4599	2,16	0,4846	2,82	0,4976
1,43	0,4236	1,76	0,4608	2,18	0,4854	2,84	0,4977
1,44	0,4251	1,77	0,4616	2,20	0,4861	2,86	0,4979
1,45	0,4265	1,78	0,4625	2,22	0,4868	2,88	0,4980
1,46	0,4279	1,79	0,4633	2,24	0,4875	2,90	0,4981
1,47	0,4292	1,80	0,4641	2,26	0,4881	2,92	0,4982
1,48	0,4306	1,81	0,4649	2,28	0,4887	2,94	0,4984
1,49	0,4319	1,82	0,4656	2,30	0,4893	2,96	0,4985
1,50	0,4332	1,83	0,4664	2,32	0,4898	2,98	0,4986
1,51	0,4345	1,84	0,4671	2,34	0,4904	3,00	0,49865
1,52	0,4357	1,85	0,4678	2,36	0,4909	3,20	0,49931
1,53	0,4370	1,86	0,4686	2,38	0,4913	3,40	0,49966
1,54	0,4382	1,87	0,4693	2,40	0,4918	3,60	0,499841
1,55	0,4394	1,88	0,4699	2,42	0,4922	3,80	0,499928
1,56	0,4406	1,89	0,4706	2,44	0,4927	4,00	0,499968
1,57	0,4418	1,90	0,4713	2,46	0,4931	4,50	0,499997
1,58	0,4429	1,91	0,4719	2,48	0,4934	5,00	0,499997

Приложение 2

Квантили нормального распределения

Таблица 2. **Квантили нормального распределения**

p	$1-\frac{p}{2}$	$u_{1-\frac{p}{2}}$	p	$1-\frac{p}{2}$	$u_{1-\frac{p}{2}}$
0,80	0,60	0,25	0,05	0,975	1,96
0,50	0,75	0,67	0,04	0,980	2,05
0,40	0,80	0,84	0,02	0,990	2,33
0,30	0,85	1,04	0,01	0,995	2,58
0,25	0,875	1,15	0,005	0,9975	2,81
0,20	0,90	1,28	0,002	0,999	3,09
0,15	0,925	1,44	0,001	0,9995	3,29
0,10	0,95	1,64	0,0001	0,99995	3,89

Приложение 3

Квантили распределения Стьюдента

Число степеней свободы f	Уровни значимости p						
	0,20	0,10	0,05	0,02	0,01	0,005	0,001
1	3,08	6,31	12,71	31,82	63,66	127,32	636,62
2	1,89	2,92	4,30	6,97	9,93	14,09	31,60
3	1,64	2,35	3,18	4,54	5,84	7,45	12,94
4	1,53	2,13	2,78	3,75	4,60	5,60	8,61
5	1,48	2,02	2,57	3,37	4,03	4,77	6,86
6	1,44	1,94	2,45	3,14	3,71	4,32	5,96
7	1,42	1,90	2,37	3,00	3,50	4,03	5,41
8	1,40	1,86	2,31	2,90	3,36	3,83	5,04
9	1,38	1,83	2,26	2,82	3,25	3,69	4,78
10	1,37	1,81	2,23	2,76	3,17	3,58	4,59
11	1,36	1,80	2,20	2,72	3,11	3,50	4,44
12	1,36	1,78	2,18	2,68	3,06	3,43	4,32
13	1,35	1,77	2,16	2,65	3,01	3,37	4,22
14	1,34	1,76	2,15	2,62	2,98	3,33	4,14
15	1,34	1,75	2,13	2,60	2,95	3,29	4,07
16	1,34	1,75	2,12	2,58	2,92	3,25	4,02
17	1,33	1,74	2,11	2,57	2,90	3,22	3,97
18	1,33	1,73	2,10	2,55	2,88	3,20	3,92
19	1,33	1,73	2,09	2,54	2,86	3,17	3,88
20	1,33	1,73	2,09	2,53	2,85	3,15	3,85
21	1,32	1,72	2,08	2,52	2,83	3,14	3,82
22	1,32	1,72	2,07	2,51	2,82	3,12	3,79
23	1,32	1,71	2,07	2,50	2,81	3,10	3,77
24	1,32	1,71	2,06	2,49	2,80	3,09	3,75
25	1,32	1,71	2,06	2,48	2,79	3,08	3,73
26	1,32	1,71	2,06	2,48	2,78	3,07	3,71
27	1,31	1,70	2,05	2,47	2,77	3,06	3,69
28	1,31	1,70	2,05	2,47	2,76	3,05	3,67
29	1,31	1,70	2,04	2,46	2,76	3,04	3,66
30	1,31	1,70	2,04	2,46	2,75	3,03	3,65
40	1,30	1,68	2,02	2,42	2,70	2,97	3,55
60	1,30	1,67	2,00	2,39	2,66	2,91	3,46
120	1,29	1,66	1,98	2,36	2,62	2,86	3,37
∞	1,28	1,64	1,96	2,33	2,58	2,81	3,29

Приложение 4

Квантили распределения Пирсона χ^2_{1-p}

Число степеней свободы f	Уровни значимости p							
	0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30
1	0,00016	0,0006	0,0039	0,016	0,064	0,148	0,455	1,07
2	0,020	0,040	0,103	0,211	0,446	0,713	1,386	2,41
3	0,115	0,185	0,352	0,584	1,005	1,424	2,336	3,66
4	0,30	0,43	0,71	1,06	1,65	2,19	3,36	4,9
5	0,55	0,75	1,14	1,61	2,34	3,00	4,35	6,1
6	0,87	1,13	1,63	2,20	3,07	3,83	5,35	7,2
7	1,24	1,56	2,17	2,83	3,82	4,67	6,35	8,4
8	1,65	2,03	2,73	3,49	4,59	5,53	7,34	9,5
9	2, 09	2,53	3,32	4,17	5,38	6,39	8,34	10,7
10	2,56	3,06	3,94	4,86	6,18	7,27	9,34	11,8
11	3,1	3,6	4,6	5,6	7,0	8,1	10,3	12,9
12	3,6	4,2	5,2	6,3	7,8	9,0	11,3	14,0
13	4,1	4,8	5,9	7,0	8,6	9,9	12,3	15,1
14	4,7	5,4	6,6	7,8	9,5	10,8	13,3	16,2
15	5,2	6,0	7,3	8,5	10,3	11,7	14,3	17,3
16	5,8	6,6	8,0	9,3	11,2	12,6	15,3	18,4
17	6,4	7,3	8,7	10,1	12,0	13,5	16,3	19,5
18	7,0	7,9	9,4	10,9	12,9	14,4	17,3	20,6
19	7,6	8,6	10,1	11,7	13,7	15,4	18,3	21,7
20	8,3	9,2	10,9	12,4	14,6	16,3	19,3	22,8
21	8,9	9,9	11,6	13,2	15,4	17,2	20,3	23,9
22	9,5	10,6	12,3	14,0	16,3	18,1	21,3	24,9
23	10,2	11,3	13,1	14,8	17,2	19,0	22,3	26,0
24	10,9	12,0	13,8	15,7	18,1	19,9	23,3	27,1
25	11,5	12,7	14,6	16,5	18,9	20,9	24,3	28,2
26	12,2	13,4	15,4	17,3	19,8	21,8	25,3	29,3
27	12,9	14,1	16,2	18,1	20,7	22,7	26,3	30,3
28	13,6	14,8	16,9	18,9	21,6	23,6	27,3	31,4
29	14,3	15,6	17,7	19,8	22,4	24,6	28,3	32,5
30	15.0	16,3	18,5	20,6	23,4	25,5	29,3	33,5

Число степеней свободы f	Уровни значимости p							
	0,20	0,10	0,05	0,02	0,01	0,005	0,002	0,001
1	1,64	2,7	3,8	5,4	6,6	7,9	9,5	10,8
2	3,22	4,6	6,0	7,8	9,2	10,6	12,4	13,8
3	4,64	6,3	7,8	9,8	11,3	12,8	14,8	16,3
4	6,0	7,8	9,5	11,7	13,3	14,9	16,9	18,5
5	7,3	9,2	11,1	13,4	15,1	16,3	18,9	20,5
6	8,6	10,6	12,6	15,0	16,8	18,6	20,7	22,5
7	9,8	12,0	14,1	16,6	18,5	20,3	22,6	24,3
8	11,0	13,4	15,5	18,2	20,1	21,9	24,3	26,1
9	12,2	14,7	16,9	19,7	21,7	23,6	26,1	27,9
10	13,4	16,0	18,3	21,2	23,2	25,2	27,7	29,6
11	14,6	17,3	19,7	22,6	24,7	26,8	29,4	31,3
12	15,8	18,5	21,0	24,1	26,2	28,3	31	32,9
13	17,0	19,8	22,4	25,5	27,7	29,8	32,5	34,5
14	18,2	21,1	23,7	26,9	29,1	31,3	34	36,1
15	19,3	22,3	25,0	28,3	30,6	32,8	35,5	37,7
16	20,5	23,5,	26,3	29,6	32,0	34,3	37	39,2
17	21,6	24,8	27,6	31,0	33,4	35,7	38,5	40,8
18	22,8	26,0	28,9	32.3	34,8	37,2	40	42,3
19	23,9	27,2	30,1	33,7	36,2	38,6	41,5	43,8
20	25,0	28,4	31,4	35,0	37,6	40,0	43	45,3
21	26,2	29,6	32,7	36,3	38,9	41,4	44,5	46,8
22	27,3	30,8	33,9	37,7	40,3	42,8	46	48,3
23	28,4	32,0	35,2	39,0	41,6	44,2	47,5	49,7
24	29,6	33,2	36,4	40,3	43,0	45,6	48,5	51,2
25	30,7	34,4	37,7	41,6	44,3	46,9	50	52,6
26	31,8	35,6	38,9	42,9	45,6	48,3	51,5	54,1
27	32,9	36,7	40,1	44,1	47,0	49,6	53	55,5
28	34,0	37,9	41,3	45,4	48,3	51,0	54,5	56,9
29	35,1	39,1	42,6	46,7	49,6	52,3	56	58,3
30	36,3	40,3	43,8	48,0	50,9	53,7	57,5	59,7

Приложение 5

Квантили распределения Фишера F_{1-p} для $p = 0{,}05$

f_2	f_1								
	1	2	3	4	5	6	12	24	∞
1	164,4	199,5	215,7	224,6	230,2	234,0	244,9	249,0	254,3
2	18,5	19,2	19,2	19,3	19,3	19,3	19,4	19,5	19,5
3	10,1	9,6	9,3	9,1	9,0	8,9	8,7	8,6	8,5
4	7,7	6,9	6,6	6,4	6,3	6,2	5,9	5,8	5,6
5	6,6	5,8	5,4	5,2	5,1	5,0	4,7	4,5	4,4
6	6,0	5,1	4,8	4,5	4,4	4,3	4,0	3,8	3,7
7	5,6	4,7	4,4	4,1	4,0	3,9	3,6	3,4	3,2
8	5,3	4,5	4,1	3,8	3,7	3,6	3,3	3,1	2,9
9	5,1	4,3	3,9	3,6	3,5	3,4	3,1	2,9	2,7
10	5,0	4,1	3,7	3,5	3,3	3,2	2,9	2,7	2,5
11	4,8	4,0	3,6	3,4	3,2	3,1	2,8	2,6	2,4
12	4,8	3,9	3,5	3,3	3,1	3,0	2,7	2,5	2,3
13	4,7	3,8	3,4	3,2	3,0	2,9	2,6	2,4	2,2
14	4,6	3,7	3,3	3,1	3,0	2,9	2,5	2,3	2,1
15	4,5	3,7	3,3	3,1	2,9	2,8	2,5	2,3	2,1
16	4,5	3,6	3,2	3,0	2,9	2,7	2,4	2,2	2,0
17	4,5	3,6	3,2	3,0	2,8	2,7	2,4	2,2	2,0
18	4,4	3,6	3,2	2,9	2,8	2,7	2,3	2,1	1,9
19	4,4	3,5	3,1	2,9	2,7	2,6	2,3	2,1	1,8
20	4,4	3,5	3,1	2,9	2,7	2,6	2,3	2,1	1,8
22	4,3	3,4	3,1	2,8	2,7	2,6	2,2	2,0	1,8
24	4,3	3,4	3,0	2,8	2,6	2,5	2,2	2,0	1,7
26	4,2	3,4	3,0	2,7	2,6	2,4	2,1	1,9	1,7
28	4,2	3,3	2,9	2,7	2,6	2,4	2,1	1,9	1,6
30	4,2	3,3	2,9	2,7	2,5	2,4	2,1	1,9	1,6
40	4,1	3,2	2,9	2,6	2,5	2,3	2,0	1,8	1,5
60	4,0	3,2	2,8	2,5	2,4	2,3	1,9	1,7	1,4
120	3,9	3,1	2,7	2,5	2,3	2,2	1,8	1,6	1,3
∞	3,8	3,0	2,6	2,4	2,2	2,1	1,8	1,5	1,0

Приложение 6

Значимые ранги множественного рангового критерия Дункана при $p = 0{,}05$

n_D	p															
	2	3	4	5	6	7	8	9	10	12	14	16	18	20	50	100
1	18,00	18,00	18,00	18,00	18,00	18,00	18,00	18,00	18,00	18,00	18,00	18,00	18,00	18,00	18,00	18,00
2	6,09	6,09	6,09	6,09	6,09	6,09	6,09	6,09	6,09	6,09	6,09	6,09	6,09	6,09	6,09	6,09
3	4,50	4,50	4,50	4,50	4,50	4,50	4,50	4,50	4,50	4,50	4,50	4,50	4,50	4,50	4,50	4,50
4	3,98	4,01	4,02	4,02	4,02	4,02	4,02	4,02	4,02	4,02	4,02	4,02	4,02	4,02	402	4,02
5	3,64	3,74	3,79	3,83	3,83	3,83	3,83	3,83	3,83	3,83	3,83	3,83	3,83	3,83	3,83	3,83
6	3,46	3,58	3,64	3,68	3,68	3,68	3,68	3,68	3,68	368	3,68	3,68	3,68	3,68	3,68	3,68
7	3,35	3,47	3,54	3,58	3,60	3,61	3,61	3,61	3,61	3,61	3,61	3,61	3,61	3,61	3,61	3,61
8	3,26	3,39	3,47	3,52	3,55	3,56	3,56	3,56	3,56	3,56	3,56	3,56	3,56	3,56	3,56	3,56
9	3,20	3,34	3,41	3,47	3,50	3,52	3,52	3,52	3,52	3,52	3,52	3,52	3,52	3,52	3,52	3,52
10	3,15	3,30	3,37	3,43	3,46	3,47	3,47	3,47	3,47	3,47	3,47	3,47	3,47	3,48	3,48	3,48
11	3,11	3,27	3,35	3,39	3,43	3,44	3,45	3,46	3,46	3,46	3,46	3,46	3,47	3,48	3,48	3,48
12	3,08	3,23	3,33	3,36	3,40	3,42	3,44	3,44	3,46	3,46	3,46	3,46	3,47	3,48	3,48	3,48
13	3,06	3,21	3,30	3,35	3,38	3,41	3,42	3,44	3,45	3,45	3,46	3,46	3,47	3,47	3,47	3,47
14	3,03	3,18	3,27	3,33	3,37	3,39	3,41	3,42	3,44	3,45	3,46	3,46	3,47	3,47	3,47	3,47
15	3,01	3,16	3,25	3,31	3,36	3,38	3,40	3,42	3,43	3,44	3,45	3,46	3,47	3,47	3,47	3,47
16	3,00	3,15	3,23	3,30	3,34	3,37	3,39	3,41	3,43	3,44	3,45	3,46	3,47	3,47	3,47	3,47
17	2,98	3,13	3,22	3,28	3,33	3,36	3,38	3,42	3,42	3,44	3,45	3,46	3,47	3,47	3,47	3,47
18	2,97	3,12	3,21	3,27	3,32	3,35	3,37	3,39	3,41	3,43	3,45	3,46	3,47	3,47	3,47	3,47
19	2,96	3,11	3,19	3,26	3,31	3,35	3,37	3,39	3,41	3,43	3,44	3,46	3,47	3,47	3,47	3,47
20	2,95	3,10	3,18	3,25	3,30	3,34	3,36	3,38	3,40	3,43	3,44	3,46	3,47	3,47	3,47	3,47
22	2,93	3,08	3,17	3,24	3,29	3,32	3,35	3,37	3,39	3,42	3,44	3,45	3,46	3,47	3,47	3,47
24	2,92	3,07	3,15	3,22	3,28	3,31	3,34	3,37	3,38	3,41	3,44	3,45	3,46	3,47	3,47	3,47
26	2,91	3,06	3,14	3,21	3,27	3,30	3,34	3,36	3,38	3,41	3,43	3,45	3,46	3,47	3,47	3,47
28	2,90	3,04	3,13	3,20	3,26	3,30	3,33	3,35	3,37	3,40	3,43	3,46	3,47	3,47	3,47	3,47
30	2,89	3,04	3,12	3,20	3,25	3,29	3,32	3,35	3,37	3,40	3,43	3,44	3,46	3,47	3,47	3,47
40	2,86	3,01	3,10	3,17	3,22	3,27	3,30	3,33	3,35	3,39	3,42	3,44	3,46	3,47	3,47	3,47
60	2,83	2,98	3,08	3,14	3,20	3,24	3,28	3,31	3,33	3,37	3,40	3,43	3,45	3,47	3,48	3,48
100	2,80	2,95	3,05	3,12	3,18	3,22	3,26	3,29	3,32	3,36	3,40	3,42	3,45	3,47	3,53	3,53
∞	2,77	2,92	3,02	3,09	3,15	3,19	3,23	3,26	3,29	3,34	3,38	3,41	3,44	3,47	3,61	3,67

Ковалева Мария Александровна

Волошин Сергей Борисович

Анализ данных

Учебное пособие издано в авторской редакции

Сетевое издание

Главный редактор – Кирсанов К.А.

Вёрстка – Павлов А.А.

Ответственный за выпуск - Алимова Н.К.

Учебное издание

Системные требования:
операционная система Windows XP или новее, macOS 10.12 или новее, Linux.
Программное обеспечение для чтения файлов PDF.

Объем данных 3,49 Мб

Принято к публикации «09» сентября 2019 года

Режим доступа: https://izd-mn.com/PDF/32MNNPU19.pdfсвободный. – Загл. с экрана. - Яз. рус., англ.

ООО «Издательство «Мирнауки»

«Publishing company «World of science», LLC

Адрес:

Юридический адрес — 127055, г. Москва, пер. Порядковый, д. 21, офис 401.

Почтовый адрес — 127055, г. Москва, пер. Порядковый, д. 21, офис 401.

https://izd-mn.com/