И. А. КАЦКО,
П. С. БОНДАРЕНКО,
Г. В. ГОРЕЛОВА

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебник

Издание третье, исправленное и дополненное

ЛАНЬ
САНКТ-ПЕТЕРБУРГ
МОСКВА
КРАСНОДАР
2023

УДК 519.2
ББК 22.17я73

К 30 **Кацко И. А.** Теория вероятностей и математическая статистика : учебник для вузов / И. А. Кацко, П. С. Бондаренко, Г. В. Горелова. — 3-е изд., испр. и доп. — Санкт-Петербург : Лань, 2023. — 436 с. : ил. — Текст : непосредственный.

ISBN 978-5-507-45492-1

Рассмотрены основные вопросы теории вероятностей и математической статистики в соответствии с программой подготовки студентов, обучающихся в вузах. Для углубленного изучения курса приведены приложения теории вероятностей в компьютерных науках (computer science). Содержание книги предполагает возможность заложить базу для изучения прикладной статистики (анализа данных) и науки о данных, поэтому особое внимание уделяется содержательному смыслу (геометрическому, физическому, логическому) вводимых понятий.

Учебник вместе со сборником задач авторов представляют собой учебно-методический комплекс, соответствующий требованиям ФГОС ВО последнего поколения, который предназначен для обучающихся в вузах по экономическим и связанным с IT-направлениям подготовки и может быть полезен студентам (бакалавриата, специалитета, магистратуры), аспирантам, преподавателям и специалистам.

УДК 519.2
ББК 22.17я73

Рецензенты:
Л. И. НИВОРОЖКИНА — доктор экономических наук, профессор, зав. кафедрой статистики, эконометрики и оценки рисков Ростовского государственного экономического университета, заслуженный деятель науки РФ;
А. И. ОРЛОВ — доктор технических наук, доктор экономических наук, кандидат физико-математических наук, профессор кафедры экономики и организации производства Московского государственного технического университета им. Н. Э. Баумана.

Обложка
П. И. ПОЛЯКОВА

Оглавление

А. А. Чупров сравнивает вероятность с центром тяжести тела, с Гринвичским меридианом, с линией экватора. Перечисленные понятия механики, географии не существуют в действительности, это воображаемые точки, линии, но они служат инструментом, методом познания окружающего нас мира, а связь частостей с вероятностями позволяет раскрыть причинные связи там, где принцип причинности в явной форме применить нельзя.

А. Л. Вайнштейн, Н. С. Четвериков.
Введение к книге Ог. Курно
Основы теории шансов и вероятностей, 1970

Одна из наиболее бросающихся в глаза общих тенденций современной математики и ее приложений состоит в резком повышении роли тех разделов науки, которые анализируют явления, имеющие «случайный» характер, и основываются на теории вероятностей. И всего лишь «небольшим преувеличением» прозвучала шутка известного американского математика Дж. Дуба, начавшего свой доклад в Московском математическом обществе словами: «Всем специалистам по теории вероятностей хорошо известно, что математика представляет собой часть теории вероятностей». ... в наше время основы теории вероятностей должны входить в научный багаж каждого образованного человека.

И. М. Яглом

Теорию вероятностей и статистику можно без преувеличения назвать основой всего машинного обучения вообще и значительной доли других исследований в рамках искусственного интеллекта. Даже если алгоритм на первый взгляд не использует вероятности или случайные процессы, при ближайшем рассмотрении наверняка окажется, что для его анализа придется привлекать вероятность.

С. И. Николенко, А. Л. Тулупьев.
Самообучающиеся системы

Математическая статистика — задача (раздел) теории вероятностей,
Машинное обучение — раздел теории вероятностей.
Есть две категории людей, говорящих на эту тему, но…
Программирующие на Python (точнее, подключающие библиотеки) говорят — «машинное обучение»,
Работающие с Power Point говорят — «искусственный интеллект».

Из серии „Relato Refero“

... истинной логикой этого мира является исчисление вероятностей, занимающееся нахождением величин вероятностей, которые учитывает или должен учитывать любой здравомыслящий человек.

Дж. Максвелл

Есть люди, полагающие, что математика — это нудное занятие, которое всегда уныло и скучно; мы же находим математику развлечением и не стыдимся признаться в этом.

Р. Грэхем, Д. Кнут, О. Паташник.
Конкретная математика. Основания информатики, 1998

На книжной полке рядом стоят два тома Пушкина: первый и второй. Страницы каждого тома имеют вместе толщину 2 см, а обложка каждая 2 мм. Червь прогрыз (перпендикулярно страницам) от первой страницы первого тома до последней страницы второго тома. Какой путь он прогрыз?

[Эта топологическая задача с невероятным ответом — 4 мм — совершенно недоступна академикам, но некоторые дошкольники легко справляются с ней.]

В. И. Арнольд. Задачи для детей от 5 до 15, 2004

Конечно, мы будем учиться доказывать, но будем также учиться догадываться.

Д. Пойа

Предисловие

Из предисловия к первому изданию

На современном этапе развития российского общества резко повысилась управленческая роль руководителя организации. В связи с этим в стране проводятся многочисленные исследования, перенимается и пропагандируется опыт в области менеджмента и маркетинга. Одним из важнейших моментов в деятельности руководителя, менеджера, экономиста является принятие решений в условиях неопределенности. При этом наиболее разработанным инструментарием является математическая статистика, позволяющая решать *задачи принятия решений*[1] *в условиях вероятностной неопределенности* и имеющая достаточно развитое программное обеспечение (например, в *Excel* (Пакет анализа) и одной из его доступных надстроек — *AtteStat,* созданной российским ученым И. Гайдышевым). Появляется большое количество свободно распространяемых пакетов: *JASP* (реализованы методы классической, непараметрической и байесовской статистики), *Gretl* (реализованы современные методы эконометрического моделирования), языки программирования с открытым кодом (*R, Python*) и др.

В процессе всей своей жизни человек часто сталкивается с событиями и явлениями, исход которых заранее не определен. Например, студент не знает, какие именно дополнительные вопросы задаст экзаменатор, служащий — сколько точно времени у него займет дорога на работу завтра (через неделю), инвестор — окупятся ли его инвестиции, страховщик — причину и размер выплаты страхового вознаграждения и т. д. Такие явления известный специалист по прикладной статистике В. В. Налимов называл диффузными, так как очень часто невозможно в чистом виде выявить факторы, однозначно влияющие на результат и позволяющие применить классические детерминированные методы. Тем не менее в подобных ситуациях, связанных с неопределенностью, человеку необходимо принимать решения.

Обычно принятию решений предшествует анализ известных данных (на основании предшествующего опыта, здравого смысла, интуиции и т. д.). Первые известные примеры обработки данных описаны несколько тысяч лет назад — в «Махабхарате» и Ветхом Завете (Книга Чисел). Стремясь увидеть и обосновать закономерности в неопределенных процессах, человечество выработало целый арсенал методов, совокупность которых называется математической статистикой (прикладной статистикой или анализом данных).

Кратко рассмотрим основные задачи, определяемые методами математической статистики:

– выборочное наблюдение решает задачу обобщения на всю совокупность результатов, полученных при изучении ее части, например рейтинг политиков, анкетирование, качество продукции и т. д.;

– проверка статистических гипотез позволяет ответить на вопрос о достоверности принимаемого решения (например, обоснованность рейтинга популярности);

[1] О теории принятия решений см., например, [70].

– дисперсионный анализ изучает влияние факторных признаков на результативный (например, зависит ли производительность труда рабочего от стажа, возраста, стажа и возраста);

– корреляционно-регрессионный анализ позволяет выявить связи и построить модели зависимости (например, какая зависимость существует между расходами населения и сбережениями);

– анализ временных рядов рассматривает последовательности чисел во времени и изучает их свойства (например, количество пятен на Солнце, характеризующее его активность по годам; курс доллара по дням и т. д.);

– последовательный анализ А. Вальда является одним из первых разработанных методов, позволивших осуществлять контроль качества выпускаемой продукции, давший начало развитию динамического программирования (детерминированного и стохастического) и управляемых цепей Маркова.

Перечисленные выше (и другие) методы основываются на теоретических положениях теории вероятностей и являются классическими.

Существует ряд направлений, которые не используют предпосылки классической теории вероятностей. Наиболее известные из них указаны ниже.

Непараметрическая статистика, в отличие от классической (параметрической), не предполагает, что наблюдения подчиняются определенному закону распределения.

Бутстреп — способ обработки выборочных данных с помощью метода статистических испытаний (метода Монте-Карло), при котором выборку «размножают» и изучают устойчивость получаемых выводов.

Реальные данные — это не числа, а интервалы (результат измерения плюс-минус погрешность), поэтому возникает необходимость в соответствующей статистике.

Статистика объектов нечисловой природы (А. И. Орлов). Например, качественные оценки признака (машина — очень плохая, плохая, хорошая, очень хорошая), ранжировка (распределение студентов по росту), классификация (группировка студентов курса не по одному, а по ряду признаков: интересы, достаток, оценки, симпатии, антипатии) и т. д.

Байесовская статистика приводит к удвоению всех методов математической статистики с учетом появления новых свидетельств (или реализации идеи размножения, генерации выборки, например, с использованием алгоритма *MCMC — Markov Chain Monte Carlo*).

Методы прикладной статистики быстро входят в нашу жизнь посредством пакетов прикладных математических ($Matcad, MatLab$), статистических ($Statistica, STADIA, SAS, IBM\ SPSS, AtteStat$) и других программ, в которых предусматриваются средства обработки данных. Здесь даже у серьезных людей возникает вопрос: «А зачем знать теоретические положения? Главное уметь кнопки нажимать! …». Ответом на этот вопрос может служить притча, приводимая Ф. Фишером.

В *N*-ской губернии разразилась эпидемия чумы. Крестьяне узнали, что можно «научно» выяснить причину болезни, для этого нужно посчитать коэффициент корреляции (глава 15). Таким образом, они выяснили, что между количеством заболеваний в деревнях и количеством врачей существует прямая корреляционная зависимость. Поэтому крестьяне решили избавиться от врачей, посчитав их причиной болезни … Притча, скорее всего, вызывает улыбку, однако в менее очевидном случае мы зачастую поступаем аналогично.

Поэтому, несомненно, необходимо знать основные идеи и методы прикладной статистики, условия их применения, а затем «нажимать кнопки». Методически более целесообразно изучать анализ данных на компьютере в *Excel*, а затем, по мере возникновения соответствующих вопросов, переходить к профессиональным программам.

Авторы выражают глубокую признательность:

– уважаемым рецензентам Л. И. Ниворожкиной, А. И. Орлову,

– нашим коллегам В. Н. Волковой, Ю. И. Бершицкому, Н. Б. Паклину, В. Н. Лаптеву, Е. В. Луценко, Ю. И. Лыпарю, Н. Н. Лябаху, А. М. Ляховецкому, С. Г. Чефранову, а также всем участникам ежегодной международной научно-практической конференции «Системный анализ в проектировании и управлении» (СПбГПУ, рук. В. Н. Волкова, В. Н. Козлов, В. Е. Ланкин), многолетнее общение с которыми способствовало формированию взглядов, отраженных в настоящей книге.

Из предисловия ко второму изданию

Во втором издании была существенно переработана вся книга. С учетом того, что студентам дается большая самостоятельность в изучении дисциплин, большее внимание уделено изложению полноценных в логическом отношении доказательств основных результатов теории вероятностей и математической статистики, опирающихся на известные факты из вводных курсов математического анализа, линейной алгебры и аналитической геометрии (часто необходимые пояснения даются по ходу изложения). При рассмотрении традиционных вопросов теории вероятностей и математической статистики, как правило, внимание уделялось содержательной стороне и возможности в дальнейшем облегчить понимание современных идей прикладной статистики (анализа данных) и науки о данных.

Авторы благодарят А. Е. Жминько, Д. и А. Кацко (выполнили все рисунки в *AutoCAD*), М. Ю. Киек (вопросы аудита — пример 10.12) за помощь в подготовке материалов к изданию.

Настоящая книга — один из итогов деятельности кафедры статистики и прикладной математики Кубанского ГАУ, созданной в 1961 г. под руководством доктора экономических наук, профессора, академика РАСХН Д. Н. Письменной. Авторы, осуществлявшие руководство кафедрой в последние десятилетия, благодарят сотрудников, работавших в разные годы существования коллектива за их профессиональные и личные качества, во многом благодаря которым наша работа завершена к 60-летнему юбилею кафедры и 100-летнему юбилею Кубанского ГАУ.

Предисловие к третьему изданию

В этом, третьем издании, исправлены замеченные опечатки и неточности, сделаны отдельные изменения и дополнения изложения. Содержание курса не претерпело существенных изменений, однако было решено выделить отдельно задачи и сформировать сборник задач и упражнений для проведения практических занятий и самостоятельной работы студентов, включающий как авторские, так и наиболее интересные задачи, принадлежащие авторам книг, отраженных в списке литературы, за что им приносится искренняя и глубокая благодарность. Кроме того, на базе третьей и четвертой частей второго издания было решено издать отдельный учебник и практикум по прикладной статистике, так как такая потребность в вузах уже назрела. Указанные изменения обусловлены необходимостью учитывать то, что сегодня уделяется все больше внимания вопросам анализа данных, оперирующих различными вероятностными (геометрическими, логическими) моделями, методами статистического обучения, понятиями зависимости и независимости и т. д. Кроме того, появляются курсы по аналитике данных, машинному обучению, статистическому обучению, $Big\ Data$ и др. Поэтому мы надеемся, что предлагаемый подход будет интересен и востребован как студентами, так и преподавателями вузов.

Краснодар, 2022 г. Авторы

Введение

Теория вероятностей — математическая дисциплина, изучающая закономерности, происходящие в массовых однородных случайных явлениях и процессах.

С возникновением теории вероятностей наука получила мощный аппарат исследования случайных явлений и процессов, до этого исследовались лишь детерминированные явления и опыты, в которых первоначальные условия однозначно позволяли определить их исход. Между тем случайные явления присутствуют во многих областях науки (биологии, генетике, агрономии, экономике, демографии, технике и т. д.), когда заранее невозможно предсказать результат опыта. Исторически зарождение и развитие теории вероятностей связано с азартными играми, в которых требовалось обосновать то или иное решение. Классическим примером является задача, рассматриваемая ниже.

Пример 1. Двое играют в безобидную игру (шансы выиграть у обоих одинаковы). Договариваются, что всю ставку забирает игрок, выигравший первым 6 партий. Как правильно разделить ставку, если игра остановилась при счете 5:3?

Решение. Для выигрыша первому игроку достаточно выиграть одну партию, второму игроку необходимо подряд выиграть три партии. Всего три партии предполагает 2·2·2=8 исходов (каждая партия имеет два исхода: выиграл, проиграл). В пользу второго игрока только один исход из 8 возможных, а в пользу первого 7 исходов. Поэтому справедливо разделить ставку пропорционально шансам выиграть, то есть 7:1.

Постановка и решение подобных задач Б. Паскалем, П. Ферма, Х. Гюйгенсом послужили исходной базой появления теории вероятностей. Дальнейшее ее развитие связано с именами Я. Бернулли, С. Пуассона, А. Муавра, П. Лапласа, К. Гаусса, П. Л. Чебышёва, А. А. Маркова, А. М. Ляпунова, А. Я. Хинчина, А. Н. Колмогорова, В. Феллера, Б. В. Гнеденко, В. С. Пугачева, Д. А. Вентцель, Е. С. Вентцель, А. А. Боровкова, А. Н. Ширяева и др. Как и всякая математическая теория, с точки зрения аксиоматического подхода теория вероятностей занимается изучением соотношений между неопределяемыми объектами (понятиями). В геометрии неопределяемые понятия — это точка, прямая и плоскость, аналогично, в теории вероятностей неопределяемые понятия — это элементарные события (исходы) (ω_i) и пространство элементарных событий ($\Omega = \{\omega_i\}$).

Пример 2. Монета подбрасывается один раз. Возможные элементарные исходы: выпал герб — ω_1, выпала решка — ω_2. Пространство элементарных событий $\Omega = \{\omega_1, \omega_2,\}$.

Пример 3. Игральная кость подбрасывается один раз. Элементарные события: ω_1 — появление 1, ω_2 — 2, ω_3 — 3, ω_4 — 4, ω_5 — 5, ω_6 — 6. Пространство элементарных событий $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$.

Вероятность события — это число, имеющее ту же природу, что и расстояние в геометрии или масса в теоретической механике. Она всегда связана с каким-либо пространством элементарных событий, природа которого не имеет значения. Понятие вероятности обычно строится на интуитивных соображениях

(например, вероятность появления герба при подбрасывании симметричной монеты очевидно равна 1/2) и связано со статистической устойчивостью относительной частоты события при большом числе опытов ($\mu(A) = k/n$ — относительная частота события, где k — число появлений события А в серии из n опытов). При подбрасывании монеты достаточно большое число раз относительная частота появлений герба будет колебаться около 0,5, следовательно, можно говорить, что вероятность появления герба равна 0,5 (табл. 1.1). Устойчивость относительной частоты появления события позволяет судить о вероятности как об объективной характеристике события в данном опыте, имеющей вполне определенное значение, независимо от того, будут проводиться опыты или нет. (То есть имеет место вероятностная неопределенность.)

Предметом современной теории вероятностей является выявление общих закономерностей и зависимостей изучаемых явлений, а также описание изучаемых физических явлений с помощью абстрактных моделей.

Математическая статистика — это раздел математики, в котором изучаются методы сбора, систематизации, обработки и использования статистических данных для научных и практических выводов. В развитие математической статистики (помимо перечисленных ранее ученых) большой вклад внесли К. Гальтон, К. Пирсон, У. Госсет (Стьюдент), Р. Фишер, Ю. Нейман, Э. Пирсон, Ч. Спирмен, А. Вальд, Г. Харман, Г. Шеффе, Н. В. Смирнов, Л. Н. Большев и др.

Математическая статистика использует математический аппарат и выводы теории вероятностей, поэтому часто рассматривается как задача теории вероятностей. Связующим звеном между теорией вероятностей и математической статистикой являются закон больших чисел и так называемые предельные теоремы. В частности, закон больших чисел аргументирует применение средней арифметической в качестве оценки математического ожидания, относительной частоты появления события как оценки вероятности. Последнее обосновывает понятие статистической устойчивости.

Всю жизнь человек вынужден принимать решения: в личной сфере (в какой вуз поступать, с кем общаться, как учиться); в общественной (посещать вечера, театры, митинги, собрания, выборы); в производственной (определение факторов, существенно влияющих на урожайность, производительность труда, качество материалов и т. д.); научной (выдвижение и проверка научных гипотез).

Принятие решений обычно преследует одну из целей: прогнозирование будущего состояния процесса (объекта); управление (т. е. как следует изменять одни параметры объекта (процесса), чтобы другие параметры приняли желаемое значение); объяснение физики (природы) изучаемого объекта (процесса). Одним из основных подходов к обоснованию и последующему принятию решений является статистический.

Статистические методы обработки данных можно классифицировать по нижеследующим признакам.

По способу получения экспериментальных данных:

– активный эксперимент;

– пассивный эксперимент (выборочное или сплошное наблюдение).

По цели обработки данных:

– описательные (получение и сравнение числовых характеристик экспериментальных данных: анализ вариационных рядов, выборочный метод, проверка статистических гипотез и другие);

– аналитические (количественная оценка и анализ зависимостей, описывающих изучаемые объекты (процессы): дисперсионный анализ, регрессионный анализ, анализ временных рядов и другие).

В математической статистике предполагается, что результаты опытов и наблюдений являются реализацией различных случайных процессов, имеющих те или иные законы распределения (причем неизвестные заранее), а иногда и детерминированные составляющие (регрессионный анализ). Отсюда вытекают основные задачи математической статистики:

1) организация наблюдений;

2) нахождение по результатам выборочных наблюдений оценок числовых характеристик всей совокупности и исследование точности их приближения (выборочный метод);

3) решение вопроса согласования результатов оценивания с опытными данными (проверка статистических гипотез);

4) оценка существенности влияния факторных признаков на результативный (дисперсионный анализ);

5) выявление аналитической зависимости между признаками (корреляционно-регрессионный анализ).

А. Вальд говорил, что «математическая статистика — это теория принятия решений в условиях неопределенности».

По существу, математическая статистика дает единственный, математически обоснованный аппарат для решения задач управления и прогнозирования при отсутствии явных закономерностей (наличии случайностей) в изучаемых процессах.

Расширение первоначальных утверждений от вероятностной природы данных до геометрической, логической и т. д. приводит к анализу данных, развитие которого прошло путь от разведочного анализа данных (Дж. Тьюки, 1960-е годы), направленного на выявление природы данных и существенно расширяющий «конфирматорный» — проверочный анализ, позволяющий лишь проверку априорных гипотез (математическая статистика) и классических методов машинного обучения (эволюционное программирование — Л. Фогель; метод группового учета аргументов — А. Г. Ивахненко; нейронные сети — У. Маккалок, У. Питтс, Ф. Розенблатт; алгоритмы распознавания образов: «Форель» — Н. Г. Загоруйко, «Кора» — М. М. Бонгард; метод опорных векторов — В. Н. Вапник, А. Я. Червонискас), статистических методов контроля качества производимой продукции (Э. Деминг, Г. Тагути, У. Шухарт, Д. В. Свечарник, Г. В. Горелова, В. В. Здор), до современных информационных технологий анализа данных (*Data Mining, Knowledge discovery in Data Bases, Text Mining, Social Mining, …., Internet of Things, Internet of Everything, Big Data*), объединяемых сегодня терминами *Аналитика 1.0–N.0.* Увеличение объемов данных возвращает нас к теории вероятностей, что подтверждает ее актуальность.

«…не проворным достается успешный бег, не храбрым — победа, не мудрым — хлеб, и не у разумных богатство, и не искусным — благорасположение, но время и случай для всех их».

Экклезиаст, 9.11

Часть I

Теория вероятностей

Глава 1
Случайные события

1.1. Алгебра событий

Одним из основных понятий теории вероятностей является опыт. Под опытом понимается выполнение комплекса условий, в результате которого происходят или не происходят определенные события (факты).

Простейшие неразложимые результаты опыта называются элементарными событиями (ω), а совокупность элементарных событий называется пространством элементарных событий $\Omega = \{\omega\}$. С каждым опытом связано пространство элементарных событий Ω (введение, примеры 1, 2).

Любое конечное или счетное[2] подмножество Ω называется событием. Различают три типа событий:

– достоверные (Ω),
– случайные,
– невозможные ($\emptyset$ или $\overline{\Omega}$).

Достоверным называется событие, которое обязательно произойдет в данном опыте при выполнении комплекса условий. Например, при подбрасывании игральной кости выпадет не более шести очков.

Невозможным называется событие, которое в данном опыте не может произойти. Например, при подбрасывании игральной кости выпадет более шести очков.

Случайным называется событие, которое в данном опыте может либо произойти, либо не произойти. Например, выпадение шести очков при подбрасывании игральной кости.

События обычно обозначают первыми прописными буквами латинского алфавита: A, B, C, ... Например, в примере введения 3 событие $A = \{\omega_2, \omega_4, \omega_6\}$ — появление четного числа при подбрасывании игральной кости.

События A и B несовместны, если в результате одного опыта они не могут происходить вместе, в противном случае — совместны. Например, при одном подбрасывании монеты не могут одновременно появиться герб и решка. Элементы последовательности событий $A_1, A_2, \ldots, A_n$ попарно несовместны, если любые два из них несовместны. Например, при подбрасывании игральной кости никакие два элементарные исхода (появление цифр 1, 2, 3, 4, 5, 6) не могут произойти одновременно.

Несколько событий равновозможны, если ни одно из них не имеет объективного преимущества перед другими. Например, элементарные исходы при

---

[2] Счетным называется множество, элементам которого можно поставить в соответствие ряд натуральных чисел, например, последовательность $\left\{\frac{1}{n}\right\}$, где $n \in N$ — счетное множество, так как можно установить соответствие между элементами последовательности и множеством натуральных чисел:

$$\begin{array}{cccccccc} \frac{1}{1} & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \ldots & \frac{1}{n} & \ldots \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow & \ldots & \updownarrow & \ldots \\ 1 & 2 & 3 & 4 & \ldots & n & \ldots \end{array}$$

подбрасывании монеты, игральной кости. Несовместные события $A_1, A_2, \ldots, A_n$ образуют полную группу, если в результате опыта кроме этих событий ничего не может произойти. Два события, образующие полную группу, называются противоположными. Например, попадание и промах по мишени при выстреле.

Обычно Ω изображают на плоскости в виде некоторой области, а ω_i в виде точек этой области, устанавливая соответствие между событиями и точечными множествами. Над событиями вводятся операции, совпадающие с операциями над множествами: сумма, произведение, отрицание.

1. Суммой событий A и B называется такое третье событие $A + B$ (или $A \cup B$), которое заключается в наступлении хотя бы одного из этих событий, т. е. или A, или B, или в совместном их появлении AB. Суммой двух или нескольких несовместных событий называется событие, заключающееся в появлении только одного из них (рис. 1.1).

Рис. 1.1 — Сумма событий

2. Произведением двух событий A и B называется такое третье событие AB (или $A \cap B$), которое заключается в наступлении событий A и B одновременно. Если события A и B несовместны, то $AB = \emptyset$ (рис. 1.2). Произведением нескольких событий называется событие, заключающееся в совместном появлении этих событий.

Рис. 1.2 — Произведение событий

3. Отрицанием события A называется событие $\overline{A}$ (не A), заключающееся в ненаступлении события A ($A + \overline{A} = \Omega, A\overline{A} = \overline{\Omega}$). Причем если в результате опыта может произойти событие A, то может произойти и обратное ему событие $\overline{A}$ (рис. 1.3).

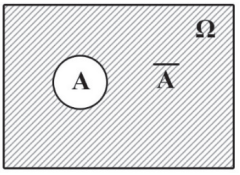


Рис. 1.3 — $\overline{A}$ — отрицание события A

Если наступление события A приводит к наступлению события B и наоборот (наступление B влечет наступление A), то события A и B равны (A = B).

Пусть S — множество всех подмножеств Ω, для которого выполняются следующие свойства:

1) если $A \in S$ и $B \in S$, то $A + B = A \cup B \in S$,
2) если $A \in S$ и $B \in S$, то $AB = A \cap B \in S$,
3) если $A \in S$, то $\overline{A} \in S$,

тогда множество S называется *алгеброй событий* (полем событий).

Замечание.

1. При более точном подходе достаточно одного из свойств 1) или 2), так как одно из них следует из другого.

2. При расширении операций сложения и умножения на случай счетного множества событий, алгебра событий S называется σ-алгеброй или борелевской алгеброй: если $A_i \in S, i \in N$, то $\sum_{i=1}^{\infty} A_i \epsilon S, \prod_{i=1}^{\infty} A_i \epsilon S$. ■

С каждым событием $A = \{\omega\} \subseteq \Omega$ можно связать *индикатор* события I_A — функцию, определенную на Ω:

$$I_A(\omega) = \begin{cases} 1, \text{при } \omega \in A, \\ 0, \text{при } \omega \notin A. \end{cases} \quad (1.1)$$

Иногда вводят обозначение $I_A(\omega) \coloneqq I(A)$[3]. Введение функции I_A позволяет рассматривать случайные события, а далее и случайные величины как частный случай общего понятия функции. В частности, известны свойства функции I_A:

1) если $A \subset B$, то $I_A \leq I_B$,
2) если $A = B$, то $I_A = I_B$,
3) $I_{\overline{A}} = 1 - I_A$,
4) $I_{A \cup B} = I_A + I_B - I_{AB}$,
5) $I_{A \cap B} = I_{AB}$.

Пример 1.1. Стрелок произвел 3 выстрела по мишени, элементарные события:

A_1 — попал при 1-м выстреле; $\overline{A}_1$ — не попал при 1-м выстреле;

A_2 — попал при 2-м выстреле; $\overline{A}_2$ — не попал при 2-м выстреле;

A_3 — попал при 3-м выстреле; $\overline{A}_3$ — не попал при 3-м выстреле.

Выразим через A_1, A_2, A_3 и их отрицания следующие события:

а) одно попадание: $A = A_1\overline{A}_2\overline{A}_3 + \overline{A}_1A_2\overline{A}_3 + \overline{A}_1\overline{A}_2A_3$,

б) три промаха: $B = \overline{A}_1\overline{A}_2\overline{A}_3$,

в) три попадания: $C = A_1A_2A_3$,

г) хотя бы один промах: $D = A + B + A_1A_2\overline{A}_3 + A_1\overline{A}_2A_3 + \overline{A}_1A_2A_3$,

или $D = \overline{A}_1 + \overline{A}_2 + \overline{A}_3$,

д) не менее двух попаданий: $E = A_1A_2\overline{A}_3 + A_1\overline{A}_2A_3 + \overline{A}_1A_2A_3 + C$,

е) не более одного попадания: $F = A + B$,

ж) попадание в мишень после промаха при первом выстреле:

$G = \overline{A}_1(\overline{A}_2A_3 + A_2\overline{A}_3 + A_2A_3)$,

з) хотя бы одно попадание: $H = A_1 + A_2 + A_3 = A + E$.

[3] $a \coloneqq b$ означает «a по определению равно b», $a =: b$ означает «a обозначим как b».

1.2. Вероятность события

Существует несколько подходов к определению вероятности события. Рассмотрим основные определения.

Аксиоматическое определение вероятности.

Вероятность события — это численная мера объективной возможности его появления. Вероятностной мерой называется числовая функция, определенная на поле событий S и удовлетворяющая следующим трем аксиомам.

Аксиомы вероятности.

1. Каждому событию A ставится в соответствие неотрицательное число p, которое называется вероятностью события A:

$$P(A) = p \geq 0, \text{ где } A \in S, S \subseteq \Omega.$$

2. Если события $A_1, A_2, \dots, A_n$ несовместны, то верно равенство

$$P(A_1 + A_2 + \cdots + A_n) = P(A_1) + P(A_2) + \cdots + P(A_n),$$

где $A_i \in S$ $(i = 1, 2, \dots, n), S \subseteq \Omega$.

3. $P(\Omega) = 1$, где Ω — истинное (достоверное) событие.

Пространство элементарных событий Ω с заданной в нем алгеброй S (или σ— алгеброй) и определенной на S вероятностью — неотрицательной мерой $P(A), A \in S$ называется *вероятностным пространством* и обозначается (Ω, S, P). Вероятностное пространство служит математической моделью любого случайного явления в теории вероятностей.

Аксиоматический подход не указывает, как конкретно находить вероятность, поэтому для решения задач целесообразно использовать подходы к определению вероятности, которые перечислены ниже.

Классическое определение вероятности.

Пусть события $A_1, A_2, \dots, A_n \in$ S образуют множество элементарных событий. Тогда события, которые приводят к наступлению события A $(A_i \in A)$, называются благоприятными исходами для события A, $m(A)$ — число благоприятных исходов.

Вероятностью события A называется отношение числа исходов, благоприятствующих наступлению события A, к числу всех возможных элементарных исходов:

$$P(A) = \frac{m(A)}{n}. \qquad (1.2)$$

В примере введения 2 вероятность появления герба $P(A) = 0{,}5$; в примере введения 3 вероятность появления числа больше четырех

$$P(A) = \frac{2}{6} = \frac{1}{3}.$$

Пример 1.2. Брошены две игральные кости. Найти вероятность того, что сумма очков, выпавших на гранях, равна 7.

Решение. Согласно классическому определению вероятности (1.2) найдем $m(A)$ и n. 1+6 = 6+1 = 2+5 = 5+2 = 3+4 = 4+3 — все возможные варианты получения в сумме 7 очков при подбрасывании двух игральных костей, следовательно, $m(A) = 6$. Общее число возможных случаев $n = 6 \cdot 6 = 36$, поэтому

$$P(A) = \frac{6}{36} = \frac{1}{6}.$$

Из классического определения следуют свойства вероятности:

1) $0 \le p(A) \le 1$,

так как $0 \le m(A) \le n$;

2) $P(\Omega) = 1$,

так как $m(\Omega) = n$;

3) $P(\overline{\Omega}) = 0$,

так как $m(\overline{\Omega}) = 0$.

$A + \bar{A} = \Omega$ — достоверное событие, поэтому

$$P(A) + P(\overline{A}) = 1 \text{ или } P(\overline{A}) = 1 - P(A).$$

Статистическое определение вероятности.

Пусть проводится серия, состоящая из *n* испытаний, в результате которой некоторое событие А наступает *k* раз, тогда число, к которому стремится отношение $\frac{k}{n}$ при $n \to \infty$, называется статистической вероятностью события А (относительной частотой события):

$$P(A) = \lim_{n\to\infty} \frac{k}{n}. \quad (1.3)$$

Таблица 1.1

Опыты по подбрасыванию монеты

Опыт	Число опытов, n	Появление герба, k	$\frac{k}{n}$
Опыт Керриха	10 000	5087	0,5087
Опыт Бюффона	4040	2048	0,5069
1-й опыт Пирсона	12 000	6019	0,5016
2-й опыт Пирсона	24 000	12 012	0,5005

Из таблицы 1.1, описывающей опыт подбрасывания монеты, следует, что $\frac{k}{n} \to 0{,}5$ ($\frac{k}{n}$ — относительная частота или частость события *A*).

Геометрическое определение вероятности удобно рассматривать при изучении бесконечных множеств.

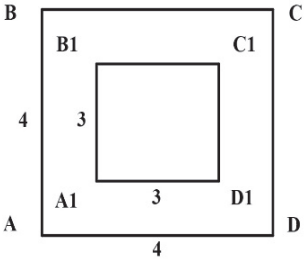
Геометрической вероятностью события А называется отношение меры области, благоприятствующей появлению события *A*, к мере всей области:

$$P(A) = \frac{mes(A)}{mes(\Omega)}, \quad (1.4)$$

где $mes(A)$[4] — мера области *A* (длина, площадь, объем).

Пример 1.3. Найти вероятность того, что точка, случайным образом брошенная в квадрат *ABCD* со стороной 4, попадет в квадрат $A_1B_1C_1D_1$ со стороной 3, находящийся внутри *ABCD*.

[4] mesure (*фр.*) — мера.



Решение. Вероятность события определяется как отношение меры части области (в данном случае площади), благоприятствующей событию A — $S_{A_1B_1C_1D_1}$, к мере всей области — S_{ABCD}:

$$P(A) = \frac{S_{A_1B_1C_1D_1}}{S_{ABCD}} = \frac{3 \cdot 3}{4 \cdot 4} = \frac{9}{16}.$$

Рис. 1.4 — Иллюстрация события A

Пример 1.4. Два лица A и B договорились встретиться в определенном месте в промежутке времени от 9^{00} до 10^{00}. Каждый из них приходит наудачу, независимо от другого лица, и ожидает 10 минут. Какова вероятность того, что они встретятся?

Решение. Рассмотрим прямоугольную систему координат XOY, в качестве единиц масштаба выберем часы. Пусть x и y — моменты прихода A и B соответственно. Необходимым и достаточным условием встречи является выполнение неравенства $|y - x| \le 1/6$ (или $x - 1/6 \le y \le x + 1/6$).

Тогда все возможные исходы будут являться точками квадрата 1×1. Заштрихованной области квадрата, ограниченной сторонами квадрата, а также прямыми $y = x - 1/6$ и $y = x + 1/6$ — событию D, соответствуют исходы, благоприятствующие встрече (рис. 1.5).

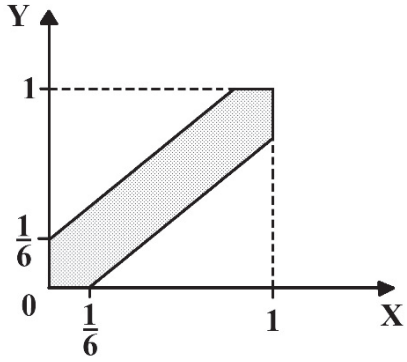


Рис. 1.5 — Иллюстрация вероятности встречи

Искомая вероятность равна отношению площади заштрихованной фигуры к площади всего квадрата.

$$P(D) = \frac{1^2 - \left(\frac{5}{6}\right)^2}{1^2} = 1 - \frac{25}{36} = \frac{11}{36}.$$

Замечание. Известны и другие подходы к определению вероятности, нашедшие свое применение в экспертных оценках, теории искусственного интеллекта, байесовской статистике. Например, *субъективная вероятность* — степень уверенности субъекта в наступлении события [36, 53]. ■

1.3. Комбинаторика

Для вычисления вероятности классическим способом нас в комбинаторике будет интересовать возможность определения количественно различных подмножеств конечных множеств.

Комбинаторика (комбинаторный анализ) — раздел дискретной математики, посвященный решению задач выбора и расположения элементов некоторого, обычно конечного, множества в соответствии с заданными правилами. Рождение комбинаторики связано с работами Б. Паскаля и П. Ферма по поводу азартных игр, большой вклад внесли Г. Лейбниц, Я. Бернулли, Л. Эйлер. В настоящее время интерес к комбинаторике связан с развитием компьютеров.

Правило произведения. Пусть из некоторого конечного множества:

1-й объект можно выбрать k_1 способами,

2-й объект — k_2 способами,

……...

n-й объект — k_n способами.

Тогда произвольный набор перечисленных n объектов из данного множества можно выбрать $k_1 k_2 ... k_n$ способами.

Пример 1.5. Сколько существует трехзначных чисел с разными цифрами?

Решение. В десятичной системе исчисления десять цифр: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. На первом месте может стоять любая из девяти цифр (кроме нуля). На втором месте — любая из оставшихся 9 цифр, кроме выбранной. На последнем месте любая из оставшихся 8 цифр. По правилу произведения, 9·9·8 = 648 трехзначных чисел имеют разные цифры.

Правило суммы. При выполнении условий правила произведения, любой из объектов можно выбрать $k_1 + k_2 + k_3 + ... + k_n$ способами.

Пример 1.6. Сколько существует способов выбора одного карандаша из коробки, содержащей 5 красных, 7 синих, 3 зеленых карандаша?

Решение. Один карандаш, по правилу суммы, можно выбрать 5+7+3=15 способами.

Обычно в комбинаторике рассматривается идеализированный эксперимент по выбору наудачу k элементов из n элементов. При этом элементы: а) не возвращаются обратно (схема выбора без возвращений); б) возвращаются обратно (схема выбора с возвращением).

I. Схема выбора без возвращений.

Размещением из n элементов по k называют любой упорядоченный набор из k элементов, принадлежащих n элементному множеству. Различные размещения отличны друг от друга или порядком элементов, или составом.

Число *размещений из n элементов по k* обозначается A_n^k и вычисляется по формуле

$$A_n^k = n(n-1)(n-2)\ldots(n-k+1) = \frac{n!}{(n-k)!}, \qquad (1.5)$$

где $n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$, $1! = 1$, $0! = 1$.

Доказательство формулы числа размещений следует из правила произведения.

Пример 1.7. В соревнованиях участвует 10 человек, трое из них займут 1, 2, 3-е места. Сколько существует различных вариантов?

Решение. Число различных вариантов равно

$$A_{10}^3 = \frac{10!}{(10-3)!} = \frac{10!}{7!} = 8 \cdot 9 \cdot 10 = 720.$$

Перестановкой из n элементов называют размещение из n элементов по n элементам. Различные перестановки различаются порядком своих элементов. Число перестановок из n элементов обозначают P_n и вычисляют по формуле

$$P_n = A_n^n = \frac{n!}{0!} = \frac{n!}{1} = n! \qquad (1.6)$$

Пример 1.8. Сколько существует способов расстановки 10 книг на полке?

Решение. Общее число способов расстановки определяется как число перестановок из 10 элементов и равно $P_{10} = 10! = 3\ 628\ 800$.

Сочетанием из n элементов по k называется любой набор из k элементов, принадлежащих n элементному множеству. Различные сочетания отличаются друг от друга только составом своих элементов.

Число сочетаний из n элементов по k обозначается C_n^k и вычисляется по формуле

$$C_n^k = \frac{A_n^k}{P_k} = \frac{n(n-1)(n-2)\ldots(n-k+1)}{1 \cdot 2 \cdot 3 \cdot \ldots \cdot k} = \frac{n!}{k!(n-k)!}. \qquad (1.7)$$

Формула C_n^k получается путем деления формулы числа размещений A_n^k на число перестановок из k элементов P_k, так как рассматриваются неупорядоченные наборы из n элементного множества по k. Числа C_n^k называют биномиальными коэффициентами, так как они являются коэффициентами в разложении бинома

$$(a+b)^n = \sum_{k=0}^{n} C_n^k a^k b^{n-k}, \qquad (1.8)$$

где $0 \le k \le n$.

Справедливы тождества:

$$C_n^k = C_n^{n-k}, \quad C_n^0 = 1, \quad C_n^1 = n,$$
$$(1+1)^n = C_n^0 + C_n^1 + \cdots + C_n^n = 2^n,$$
$$\sum_{k=0}^{n} (C_n^k)^2 = C_{2n}^n.$$

Известно также *правило Паскаля*:

$$C_{n+1}^k = C_n^k + C_n^{k-1}, \quad (1 \le k \le n). \qquad (1.9)$$

Тождества можно доказать, опираясь на определение числа сочетаний через факториалы и проделав необходимые упрощения, или воспользоваться методом математической индукции. Докажем *правило Паскаля*. Число сочетаний без повторений из $(n+1)$ объектов по k равно C_{n+1}^k. Рассмотрим один из $(n+1)$ объектов. Если он входит в сочетание из k объектов, то остальные $(k-1)$ объектов можно выбрать из n оставшихся C_n^{k-1}способами. Если объект не входит в сочетание, то C_n^k — число способов выбора k объектов из n. Два рассмотренных случая несовместны и единственно возможны, поэтому, *по правилу суммы*, общее число сочетаний — сумма числа сочетаний, где объект включен в сочетание из k объектов, с числом сочетаний, когда объект не содержится среди k объектов. Следовательно, общее число сочетаний из $(n+1)$ по k определяется по формуле (1.9).

Правило Паскаля позволяет представить разложение бинома Ньютона по C_n^k, $(0 \le k \le n)$ в виде арифметического треугольника Паскаля.

n	C_n^k														
0							1								
1						1		1							
2					1		2		1						
3				1		3		3		1					
4			1		4		6		4		1				
5		1		5		10		10		5		1			
6	1		6		15		20		15		6		1		
7	1	7		21		35		35		21		7		1	
	.....														
n	C_n^0	C_n^1		C_n^2		C_n^3	...	C_n^{k-1}		C_n^k		C_n^{k+1}	...	C_n^{n-1}	C_n^n

Заметим, что если провести под цифрами треугольника Паскаля прямые, параллельные его сторонам, и соединить соседние точки пересечения прямых векторами сверху вниз, то получится граф-решетка (бесконечный ориентированный граф[5]), а цифры будут соответствовать числу траекторий, позволяющих достичь соответствующей вершины графа (рис. 1.6).

Физически эта идея иллюстрируется с использованием доски Гальтона (рис. 5.5). Считая, что все пути равновозможны для достижения уровня (яруса) n, можно определить вероятность достижения вершины k $(0 \le k \le n)$ на уровне n как

$$\frac{C_n^k}{2^n}.$$

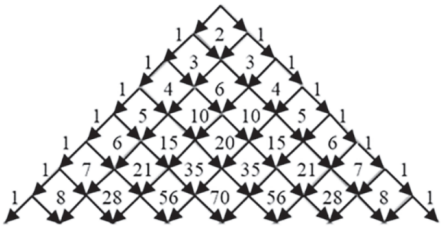


Рис. 1.6 — Граф-решетка, иллюстрирующий число траекторий симметричного случайного блуждания

Еще одна интерпретация связана с симметричным случайным блужданием точки (то есть переходящей из точки в точку с одинаковой вероятностью) по «ближайшим соседям» в пространстве целочисленных точек Z^d ($d-$мерном пространстве). При $d = 1$ первоначально точка находится в начале координат, а затем в каждый дискретный момент времени сдвигается влево или вправо на одну единицу по оси абсцисс с вероятностями $p_1 = p_2 = \frac{1}{2}$ (простейшая модель броуновского движения и диффузии частицы). Аналогично можно рассмотреть симметричное случайное блуждание точки на плоскости ($d = 2$) — движение

[5] О графах см. 1.5.

вверх-вниз или влево-направо с вероятностями $p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$; в трехмерном пространстве ($d = 3$) — $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}$ и так далее. Причем при $d = 1$ и $d = 2$ можно доказать, что точка с вероятностью 1 бесконечное число раз достигает любую точку пространства с целочисленными координатами, но среднее время ожидания этих событий бесконечно. В пространствах большей размерности свойство «возвратности» точки не сохраняется [65, 131].

Пример 1.9. Сколько существует способов выбора трех человек из десяти человек.

Решение. В данном случае для нас важен только состав наборов по три человека, порядок выбора роли не играет, поэтому, в отличие от примера 1.7, число способов выбора подсчитаем по формуле сочетаний

$$C_{10}^3 = \frac{10!}{3!(10-3)!} = \frac{7!\cdot 8\cdot 9\cdot 10}{7!\cdot 1\cdot 2\cdot 3} = 120.$$

II. Схема выбора с возвращениями. Если при выборе k элементов из n — элементы возвращаются обратно и упорядочиваются, то говорят, что это *размещения с повторениями*. Формула числа размещений с повторениями следует из правила произведения

$$\overline{A_n^k} = n^k. \qquad (1.10)$$

Пример 1.10. В гостинице 10 комнат, в каждой из которых можно разместить четырех человек. Сколько существует упорядоченных вариантов размещения прибывших четырех гостей?

Решение. Каждый следующий гость из 4 может быть помещен в любую из 10 комнат, поэтому общее число размещений по формуле размещений с повторениями равно $\overline{A_{10}^4} = 10^4 = 10000$.

Если при выборе k элементов из n элементы возвращаются обратно без последующего упорядочивания, то говорят, что это *сочетания с повторениями*. Число сочетаний с повторениями из n элементов по k

$$\overline{C_n^k} = C_{n+k-1}^k = \frac{(n+k-1)!}{(n-1)!k!}. \qquad (1.11)$$

Доказательство. Рассмотрим схему размещения k шаров по n ящикам и изобразим n ящиков в виде промежутков между $(n+1)$ черточками, а шары точками, например, $|\ldots|.||\ldots..||\ldots|$ — изображено 6 ящиков, а в них 12 шаров (3, 1, 0, 5, 0, 3). Рассмотрим общее число перестановок из $(n+k-1)$, так как крайние черточки остаются неподвижными, то по формуле числа перестановок

$$P_{n+k-1} = (n+k-1)! —$$

общее число перестановок внутри крайних черточек из $(n-1)$ черточек и k точек.

Число перестановок из $(n-1)$ черточек $P_{n-1} = (n-1)!$, из k точек $P_k = k!$. Следовательно, общее число сочетаний с повторениями равно

$$\overline{C_n^k} = C_{n+k-1}^k = \frac{(n+k-1)!}{(n-1)!k!}.$$

Пример 1.11. В магазине продается 10 видов тортов. Очередной покупатель выбил чек на три торта. Считая, что любой набор товаров равновозможен, определить число возможных заказов.

Решение. Число равновозможных заказов по формуле (1.11) равно

$$\overline{C_{10}^3} = C_{10+3-1}^3 = C_{12}^3 = \frac{12!}{9!\,3!} = \frac{9!\cdot 10\cdot 11\cdot 12}{9!\cdot 3!} = 220.$$

III. Схема упорядоченных разбиений. Пусть $k_1, k_2, \ldots, k_r$ — целые числа, такие, что $k_1 + k_2 + \ldots + k_r = n$, $k_i \geq 0$ $(i = 1, 2, \ldots, r)$. Число способов, которыми множество из n элементов можно разделить на r упорядоченных частей (r подмножеств или r групп), из которых первая содержит k_1 элементов, вторая — k_2 элементов и r-ая — k_r элементов, обозначается $C_n(k_1, k_2, \ldots, k_r)$ и вычисляется по формуле

$$C_n(k_1, k_2, \ldots, k_r) = \frac{n!}{k_1!k_2!\ldots k_r!}. \qquad (1.12)$$

Формулу (1.12) можно получить по правилу произведения. Числа, которые определяются по формуле (1.12), называются *полиномиальными коэффициентами*, так как они являются коэффициентами в разложении полинома $(a_1 + a_2 + \ldots + a_r)^n$.

Пример 1.12. Девять человек размещается в гостинице в четырехместный, трехместный и двухместный номера. Сколько существует способов их размещения?

Решение. Число способов размещения по формуле (1.12) равно

$$C_9(4, 3, 2) = \frac{9!}{4!\,3!\,2!} = 1260.$$

Пример 1.13. Набирая номер телефона, абонент забыл последние 3 цифры и набрал их наудачу, помня, что они различны. Найти вероятность того, что набраны нужные цифры.

Решение. Событие A — номер набран верно:

$$P(A) = \frac{m(A)}{n} = \frac{1}{720},$$

где $m(A) = 1$, так как только один набор из 3 цифр является нужным, всего таких наборов $n = A_{10}^3 = \frac{10!}{(10-3)!} = \frac{7!\cdot 8\cdot 9\cdot 10}{7!} = 720$.

Пример 1.14. В ящике 15 деталей, среди которых 10 окрашены. Сборщик наудачу выбрал три детали. Найти вероятность того, что все три детали окрашены.

Решение. Событие A — взятые наугад 3 детали окрашены.

$m(A) = C_{10}^3 = \frac{10!}{(10-3)!3!} = \frac{7!\cdot 8\cdot 9\cdot 10}{7!\cdot 1\cdot 2\cdot 3} = 120$ — число благоприятствующих исходов. $n = C_{15}^3 = \frac{15!}{(15-3)!3!} = \frac{12!\cdot 13\cdot 14\cdot 15}{12!\cdot 1\cdot 2\cdot 3} = 455$ — общее число возможных исходов. Имеем $P(A) = \frac{m(A)}{n} = \frac{120}{455} = \frac{24}{91}$.

Замечание. 1. Формулы размещений и сочетаний обычно используются в двух классических задачах: 1) размещения k различимых и не различимых шаров по n урнам с запретом (в каждой урне может находиться не более одного элемента) или без запрета (в урне может находиться любое число элементов); 2) выбора k упорядоченных и неупорядоченных шаров из урны с n шарами с возвращениями или без возвращений. Указанные способы выбора соответствуют физическим статистикам, рассмотренным в заданиях сборника задач 1.1 (№ 3). Эти задачи сведены в таблицу 1.2.

2. Область определения биномиальных коэффициентов можно расширить следующим образом:

$$C_x^k = \binom{x}{k} = \frac{(x)_k}{k!} = \frac{x(x-1)\dots(x-k+1)}{k!}, \quad (1.13)$$

$$A_n^k = (x)_k = x(x-1)\dots(x-k+1), \quad (1.14)$$

где $k = 0, 1, 2, \dots$, а $x \in R$. В частности

$$\binom{x}{0} = 1, \binom{x}{1} = x, \binom{x}{2} = \frac{x(x-1)}{2}.$$

В этом случае формула бинома Ньютона

$$(a+b)^x = \sum_k \binom{x}{k} a^k b^{x-k} \quad (1.15)$$

может использоваться для дробных и отрицательных степеней (x), что, собственно, и открыл Ньютон, так как для натуральной степени формула использовалась и ранее.

Таблица 1.2

Задача размещения k шаров по n урнам			
Тип элементов / Размещение	Элементы различимы	Элементы не различимы	
Без запрета	$\overline{A_n^k}$ (статистика Больцмана — Максвелла)	$\overline{C_n^k}$ (статистика Бозе — Эйнштейна)	С возвращением
С запретом	A_n^k	C_n^k (статистика Ферми — Дирака)	Без возвращения
	Упорядоченный	Неупорядоченный	Выбор / Набор
	Задача выбора k шаров из урны с n шарами		

3. Пусть $x = m \in Z_+$, тогда

$$\binom{-m}{k} = \frac{(-m)_k}{k!} = \frac{(-m)(-m-1)\dots(-m-k+1)}{k!} = (-1)^k \binom{m+k-1}{k}, \quad (1.16)$$

в частности, при $m = 1$:

$$\binom{-1}{k} = \frac{(-1)_k}{k!} = \frac{(-1)(-1-1)\dots(-1-k+1)}{k!} = (-1)^k. \quad (1.17)$$

Например, из формулы (1.13) можно получить, что при $|t| \le 1$:

$$\frac{1}{1+t} = (1+t)^{-1} = 1 - t + t^2 - t^3 + t^4 + \dots + (-1)^k t^k + \dots \quad \blacksquare \quad (1.18)$$

Пример 1.15. В группе 12 студентов, среди которых 8 отличников. По списку отбирают 9. Найти вероятность того, что отберут 5 отличников.

Решение. Событие A — отобрали 5 отличников.

$$P(A) = \frac{m(A)}{n} = \frac{56}{220} = \frac{14}{55},$$

так как по правилу произведения $m(A) = C_8^5 C_4^4 = 56,$ где C_8^5 — число возможных наборов из 8 отличников по 5, C_4^4 — число возможных наборов по 4 из остальных четырех студентов; n — общее число способов выбора из 12 студентов 9 равно:

$$n = C_{12}^9 = \frac{12!}{(12-9)!9!} = \frac{9!\cdot 10\cdot 11\cdot 12}{3!9!} = 220.$$

Пример 1.16. В коробке 5 красных, 3 зеленых и 2 синих карандаша. Наудачу, без возвращения, извлекают 3 карандаша. Найти вероятности следующих событий:

1) A — все извлеченные карандаши разного цвета,

2) B — все извлеченные карандаши одного цвета,

3) C — среди извлеченных карандашей один синий,

4) D — среди извлеченных карандашей в точности два одного цвета.

Решение. Всего в коробке 5+3+2=10 карандашей.

По правилу произведения $m(A) = 5 \cdot 3 \cdot 2 = 30$ — число исходов, благоприятствующих наступлению события A. Общее число способов выбора из 10 карандашей 3 вычисляется как число сочетаний из 10 по 3. $n = C_{10}^3 = \frac{10!}{(10-3)!3!} =$ $= \frac{7!\cdot 8\cdot 9\cdot 10}{7!\cdot 1\cdot 2\cdot 3} = 120,$ отсюда $P(A) = \frac{m(A)}{n} = \frac{30}{120} = \frac{1}{4}.$

1) Если все извлеченные карандаши одного цвета, то это либо 3 красных, либо 3 зеленых (3 синих не может быть — их в коробке всего 2). Поэтому, по правилу суммы,

$$m(B) = C_5^3 + C_3^3 = \frac{5!}{(5-3)!3!} + 1 = \frac{3!\cdot 4\cdot 5}{2!3!} = 10 + 1 = 11.$$

$n = 120,$ следовательно, $P(B) = \frac{m(B)}{n} = \frac{11}{120} \approx 0{,}0917.$

2) Из 10 карандашей (так как по условию: 5 + 3 + 2 = 10 карандашей) один синий можно выбрать $C_2^1 = 2$ — способами, 2 из оставшихся 8 карандашей не синего цвета можно выбрать $C_8^2 = 28$ — способами. Отсюда, по правилу произведения, $m(C) = C_2^1 C_8^2 = 2 \cdot 28 = 56.$

$$P(C) = \frac{m(C)}{n} = \frac{56}{120} = \frac{7}{15}.$$

3) Событие D — два карандаша одного цвета — произойдет, если из трех карандашей вытащили: 2 красных + (1 зеленый или 1 синий) или 2 зеленых + (1 красный или 1 синий), или 2 синих + (1 красный или 1 зеленый). По правилу произведения: $C_5^2 C_5^1$ — число способов выбора 2 красных карандашей и 1 другого цвета; $C_3^2 C_7^1$ — число способов выбора 2 зеленых карандашей и 1 другого цвета; $C_2^2 C_8^1$ — число способов выбора 2 синих карандашей и 1 другого цвета. Общее число исходов, благоприятствующих наступлению события D, по правилу суммы равно

$$m(D) = C_5^2 C_5^1 + C_3^2 C_7^1 + C_2^2 C_8^1 = 79.$$

Следовательно, $P(D) = \frac{m(D)}{n} = \frac{79}{120}.$

Пример 1.17. Лифт начинает движение с четырьмя пассажирами и останавливается на 10 этажах. Какова вероятность, что никакие два пассажира не выйдут на одном этаже.

Решение. Пусть все возможные случаи выхода пассажиров равновероятны, тогда первый пассажир имеет 10 возможностей выхода на 10 этажах, второй — 9 на 9 оставшихся этажах, третий — 8 на 8 оставшихся этажах, четвертый — 7. По правилу произведения, общее число исходов, благоприятствующих событию A (никакие два пассажира не выйдут на одном этаже), $m(A) = 10 \cdot 9 \cdot 8 \cdot 7 = A_{10}^4$. Общее число вариантов выхода четырех пассажиров на 10 этажах равно числу размещений с возвращениями из 10 элементов по 4, $\overline{A_{10}^4}$.

Отсюда $P(A) = \frac{m(A)}{n} = \frac{A_{10}^4}{\overline{A_{10}^4}} = \frac{10 \cdot 9 \cdot 8 \cdot 7}{10^4} = \frac{504}{1000} = 0{,}504.$

Пример 1.18. На полке стоит n книг, какова вероятность, что k из них находятся рядом ($k < n$)?

Решение. В качестве элементарных событий условно можно рассматривать различные подмножества пространства элементарных событий, полагая, что они имеют соответствующие свойства (неразложимости, равновозможности).

I способ. Элементарные события — книги, которые могут стоять на n местах. В одном ряду на соседних местах k книг можно поставить $(n-k+1)$ способами. Причем по правилу произведения,

$m(A) = (n-k+1)\,k!\,(n-k)!$, где $k!$ — количество перестановок из k книг, стоящих рядом, $(n-k)!$ — количество перестановок из $(n-k)$ оставшихся книг.

Общее число перестановок из n книг равно $n!$.

Имеем

$$P(A) = \frac{(n-k+1)\,k!\,(n-k)!}{n!} = \frac{(n-k+1)!\,k!}{n!}.$$

II способ. Рассмотрим в качестве элементарных событий все размещения из n книг по k, получим

$$P(A) = \frac{(n-k+1)P_k}{A_n^k} = \frac{(n-k+1)k!}{\frac{n!}{(n-k)!}} = \frac{(n-k+1)!\,k!}{n!},$$

где P_k — количество всех перестановок из k книг, стоящих рядом, A_n^k — количество всех размещений из n книг по k.

III способ. Рассмотрим в качестве пространства элементарных событий множество всех сочетаний из n книг по k, имеем

$$P(A) = \frac{(n-k+1)}{C_n^k} = \frac{(n-k+1)}{\frac{n!}{k!(n-k)!}} = \frac{(n-k+1)!\,k!}{n!},$$

где C_n^k — количество всех сочетаний из n книг по k.

Пример 1.19. Какова вероятность того, что никакие два человека из n находящихся в одной комнате не имеют день рождения в один день года.

Решение. Будем считать, что год невисокосный, то есть в году 365 дней. Тогда число благоприятствующих исходов $m(A) = A_{365}^n$. Общее число размещений с возвращениями n человек по 365 дням равно $\overline{A_{365}^n}$.

Имеем

$$P(A) = \frac{A_{365}^n}{A_{365}^n} = \frac{\frac{365!}{(365-n)!}}{365^n}.$$

Используя компьютер (например, *MS Excel*), оцените, при каком значении n можно заключить пари на равных условиях, что никакие два человека из n не имеют день рождения в один день года.

1.4. Основные теоремы теории вероятностей

Теорема 1. Вероятность суммы двух несовместных событий A и B равна сумме вероятностей этих событий:

$$P(A + B) = P(A) + P(B). \quad (1.19)$$

Доказательство. Обозначим через n — общее число элементарных исходов событий A и B, а через m_1 и m_2 число исходов, благоприятствующих появлению событий A и B соответственно, тогда $P(A) = \frac{m_2}{n}$, $P(B) = \frac{m_2}{n}$, $P(A + B) = \frac{m_1+m_2}{n}$.

Отсюда

$$P(A + B) = \frac{m_1}{n} + \frac{m_2}{n} = P(A) + P(B).$$

Следствие 1. Если $A_1, A_2, \dots, A_n$ — попарно несовместные события, то вероятность их суммы равна сумме вероятностей этих событий:

$$P(A_1 + A_2 + \cdots + A_n) = P(A_1) + P(A_2) + \cdots + P(A_n). \quad (1.20)$$

Следствие 2. Вероятность суммы попарно несовместных событий $A_1, A_2, \dots, A_n$, образующих полную группу, равна 1:

$$P(A_1 + A_2 + \cdots + A_n) = P(A_1) + P(A_2) + \cdots + P(A_n) = 1. \quad (1.21)$$

Следствие 3. События A и $\overline{A}$ несовместны и образуют полную группу событий, поэтому

$$P\left(A + \overline{A}\right) = P(A) + P\left(\overline{A}\right) = 1. \quad (1.22)$$

Отсюда

$$P\left(\overline{A}\right) = 1 - P(A). \quad (1.23)$$

Теорема 2. Вероятность суммы двух совместных событий A и B равна сумме вероятностей этих событий без вероятности их произведения:

$$P(A + B) = P(A) + P(B) - P(AB). \quad (1.24)$$

Доказательство. Обозначим через n — общее число элементарных исходов событий A и B, а через m_1 и m_2 число исходов, благоприятствующих появлению событий A и B соответственно, а l — число исходов, в результате которых события A и B наступают одновременно. Тогда $P(A) = \frac{m_1}{n}$, $P(B) = \frac{m_2}{n}$, $P(AB) = \frac{l}{n}$. Из рисунка 1.2 для случая совместных событий следует, что в случае одновременного наступления событий A и B, результат их произведения AB принадлежит и событию A, и событию B одновременно. Следовательно,

$$P(A + B) = \frac{m_1}{n} + \frac{m_2}{n} - \frac{l}{n} = P(A) + P(B) - P(AB).$$

Введем понятие зависимых и независимых событий.

Два события A и B называются *независимыми*, если появление одного из них не влияет на вероятность появления другого (в противном случае события *зависимы*).

Теорема 3. Вероятность произведения двух независимых событий A и B равна произведению их вероятностей:

$$P(AB) = P(A)P(B). \tag{1.25}$$

Доказательство. Так как события A и B являются независимыми, то любой исход из m_1 благоприятствующих событию A из n_1 возможных может совпасть с любым из m_2 исходов, благоприятствующих событию B из n_2 возможных, т. е.

$$P(A) = \frac{m_1}{n_1},\ P(B) = \frac{m_2}{n_2},\ P(AB) = \frac{m_1 m_2}{n_1 n_2}.$$

Отсюда следует, что

$$P(AB) = \frac{m_1}{n_1}\frac{m_2}{n_2} = P(A)P(B).$$

Следствие. Вероятность произведения n независимых событий $A_1, A_2, \dots, A_n$ равна произведению их вероятностей:

$$P(A_1 A_2 \dots A_n) = P(A_1)P(A_2) \dots P(A_n). \tag{1.26}$$

Вообще, говорят, что *события $A_1, A_2, \dots, A_n$ статистически или стохастически взаимно независимы, если для любых их подмножеств выполняется формула* (1.26).

Условной вероятностью события B при условии, что событие A уже произошло, называется число $P(AB)/P(A)$, которое обозначается

$$\frac{P(AB)}{P(A)} = P(B/A) = P_A(B).$$

Аналогично,

$$\frac{P(AB)}{P(B)} = P(A/B) = P_B(A) —$$

условная вероятность события A при условии, что событие B уже произошло.

Теорема 4. Вероятность произведения двух зависимых событий A и B равна произведению вероятности наступления события A на условную вероятность события B при условии, что событие A уже произошло:

$$P(AB) = P(A)P(B/A). \tag{1.27}$$

Доказательство. Пусть m — число исходов, благоприятствующих появлению события A, k — число исходов, благоприятствующих появлению события AB, а n — общее число элементарных исходов. Следовательно, $P(B/A) = \frac{k}{m}$. Разделим числитель и знаменатель дроби на n, тогда $P(B/A) = \frac{k/n}{m/n}$. Но учитывая, что

$$P(AB) = \frac{k}{n} \text{ и } P(A) = \frac{m}{n},$$

получаем

$$P(B/A) = \frac{P(AB)}{P(A)} \text{ или } P(AB) = P(A)P(B/A).$$

Следствие. Если события A и B независимы, то из теоремы 4 следует теорема 3.

Событие B не зависит от события A, если $P(B/A) = P(B)$.

Теорему 4 можно обобщить на случай n событий.

Теорема 5. Вероятность произведения n зависимых событий $A_1, A_2, \dots, A_n$ равна произведению последовательных условных вероятностей:

$$P(A_1A_2 \dots A_n) = P(A_1)P(A_2/A_1) \dots P(A_n/A_1A_2 \dots A_{n-1}). \quad (1.28)$$

Следствие. Если события $A_1, A_2, \dots, A_n$ независимы, то верна формула (1.26).

Теорема 6. Вероятность наступления хотя бы одного из событий $A_1, A_2, \dots, A_n$ равна разности между единицей и произведением условных вероятностей противоположных событий $\overline{A}_1, \overline{A}_2, \overline{A}_3, \dots, \overline{A}_n$:

$$P(A) = 1 - P(\overline{A}_1)P(\overline{A}_2/\overline{A}_1) \dots P(\overline{A}_n/\overline{A}_1\overline{A}_2 \dots \overline{A}_{n-1}). \quad (1.29)$$

Следствие 1. Вероятность наступления хотя бы одного из событий $A_1, A_2, \dots, A_n$, независимых в совокупности, равна разности между единицей и произведением вероятностей противоположных событий:

$$P(A) = 1 - P(\overline{A}_1)P(\overline{A}_2) \dots P(\overline{A}_n). \quad (1.30)$$

Следствие 2. Если события имеют одинаковую вероятность появиться ($P(A_i) = p$, $P(\overline{A}_i) = 1 - p = q$, где $i = 1, 2, \dots, n$), то вероятность появления хотя бы одного из них равна

$$P(A) = 1 - q^n. \quad (1.31)$$

Замечание. 1. В теоремах 1–6 неявно предполагается, что все события, в рамках каждой теоремы, принадлежат одному пространству элементарных событий.

2. Если события взаимно независимы, то они и попарно независимы, обратное неверно. То есть если события попарно независимы, то они могут и не быть взаимно независимы, как следует из нижеследующего примера, придуманного С. Н. Бернштейном. ■

Пример 1.20. Три грани тетраэдра окрашены в красный (событие A), синий (событие B) и зеленый (событие С) цвет, а на четвертую грань нанесены все три цвета (событие ABC). Найти вероятности наступления событий AB, BC, AC, ABC и сделать вывод.

Решение. Из условия следует, что

$$P(A) = P(B) = P(C) = \frac{2}{4} = \frac{1}{2}.$$

Рассмотрим вероятности указанных событий:

$$P(AB) = P(BC) = P(AC) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2},$$

следовательно, события A, B, C попарно независимы; если события A, B, C независимы в совокупности, то

$$P(A)P(B)P(C) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8},$$

что неверно, так как

$$P(ABC) = \frac{1}{4}.$$

Таким образом, попарно независимые события A, B, C зависимы в совокупности.

Пример 1.21. В урне 10 шаров, из которых два белые, а остальные черные. Наудачу взято два шара. Найдем вероятность того, что оба шара черные.

Решение. Пусть события: A_1 — первый шар черный; A_2 — второй шар черный. Тогда событие $A = A_1A_2$ — оба шара черные. Вероятность того, что второй шар черный, будет зависеть от того какого цвета первый шар. Если первый шар черный, то вероятность того, что второй шар также черного цвета, равна условной вероятности $P(A_2/A_1) = 7/9$, так как после наступления события A_1 — всего шаров останется 9, из них 7 черных. Отсюда

$$P(A) = P(A_1A_2) = P(A_1)P(A_2/A_1) = \frac{8}{10}\cdot\frac{7}{9} = \frac{28}{45}.$$

Пример 1.22. Два стрелка сделали по одному выстрелу в мишень, вероятность попадания первого 0,8, а второго 0,6. Найти вероятность следующих событий: 1) событие A — оба попали; 2) событие B — попал один; 3) событие C — попал хотя бы один стрелок.

Решение. Пусть A_1, A_2 — события, обозначающие соответственно, что первый и второй стрелок попали в цель. По условию:

$$P(A_1) = 0{,}8;\ P(A_2) = 0{,}6.$$

1) Событие A — оба стрелка попали в цель, наступит при одновременном попадании, поэтому $A = A_1\,A_2$. Отсюда, в силу независимости событий A_1, A_2, по теореме 3 имеем

$$\mathrm{P}(A)=\mathrm{P}(A_1)\mathrm{P}(A_2)= 0{,}8\cdot 0{,}6 = 0{,}48.$$

2) Событие B — попал один стрелок $B = \bar{A}_1A_2+A_1\bar{A}_2$. Применим последовательно теоремы 1 и 3:

$$\mathrm{P}(B)=\mathrm{P}(\bar{A}_1A_2+A_1\bar{A}_2)=\mathrm{P}(\bar{A}_1\,A_2\,)+\mathrm{P}(A_1\bar{A}_2)=$$
$$=\mathrm{P}(\bar{A}_1)\mathrm{P}(A_2)+\mathrm{P}(A_1)\mathrm{P}(\bar{A}_2)=(1-0{,}8)\cdot 0{,}6+0{,}8\cdot(1-0{,}6)=0{,}12+0{,}32=0{,}44.$$

3) Событие C — хотя бы один стрелок попал в мишень, $C = A_1+A_2$,

$$\mathrm{P}(C)=\mathrm{P}(A_1+A_2)=\mathrm{P}(A_1)+\mathrm{P}(A_2)-\mathrm{P}(A_1A_2)=\mathrm{P}(A_1)+\mathrm{P}(A_2)-\mathrm{P}(A_1)P(A_2)=0{,}8+0{,}6-$$
$$-0{,}8\cdot 0{,}6=0{,}92$$

или

$$\mathrm{P}(\bar{C})=\mathrm{P}(\bar{A}_1\bar{A}_2)=\mathrm{P}(\bar{A}_1)\mathrm{P}(\bar{A}_2)=(1-\mathrm{P}(A_1))(1-\mathrm{P}(A_2))=$$
$$=(1-0{,}8)(1-0{,}6)=0{,}08,$$

отсюда (по следствию 3 из теоремы 1),

$$\mathrm{P}(C) = 1-0{,}08=0{,}92.$$

Пример 1.23. В урне два белых и три черных шара. Из урны вынимают подряд два шара. Найти вероятность того, что оба шара белые.

Решение. Событие A — оба шара белые, событие A_1 — первым вытащили белый шар, событие A_2 — вторым вытащили белый шар. Событие A наступит, если наступят одновременно и A_1 и A_2, $A = A_1A_2$, отсюда

$$\mathrm{P}(A)=\mathrm{P}(A_1A_2) = \mathrm{P}(A_1)\mathrm{P}(A_2/A_1)=\frac{2}{5}\cdot\frac{1}{4}=\frac{1}{10}.$$

Пример 1.24. Условие примера 1.23, но после извлечения первого шара этот шар возвращается в урну.

Решение. В этом случае события A_1 и A_2 — независимые.

$$A=A_1A_2,\ \ \mathrm{P}(A)=\mathrm{P}(A_1A_2)=\mathrm{P}(A_1)\mathrm{P}(A_2)=\frac{2}{5}\cdot\frac{2}{5}=0{,}16.$$

Пример 1.25. Вероятность одного попадания в цель при одном залпе из двух орудий равна 0,38. Найти вероятность поражения цели при одном выстреле из первого орудия, если известно, что для второго орудия эта вероятность равна 0,8.

Решение. Событие A — попадание одного орудия при одновременном залпе из двух орудий. Обозначим события:

A_1 — первое орудие попало в цель,

A_2 — второе орудие попало в цель.

По условию: $\mathrm{P}(A) = 0{,}38$, $\mathrm{P}(A_2) = 0{,}8$. Событие A наступит, если наступит A_1, но не наступит A_2 или наступит событие A_2, но не наступит событие A_1. Имеем:

$$A=\bar{A}_1A_2+A_1\bar{A}_2, \quad P(A)= P(\bar{A}_1A_2+A_1\bar{A}_2) = \mathrm{P}(\bar{A}_1A_2) + \mathrm{P}(A_1\bar{A}_2).$$

Так как события A_1 и A_2 независимы, то

$$\mathrm{P}(\bar{A}_1A_2)+\mathrm{P}(A_1\bar{A}_2)=$$
$$= \mathrm{P}(A_1)\mathrm{P}(\bar{A}_2)+\mathrm{P}(\bar{A}_1)\mathrm{P}(A_2)=\mathrm{P}(A_1)(1-\mathrm{P}(A_2))+(1-\mathrm{P}(A_1))\cdot\mathrm{P}(A_2)=$$
$$=0{,}38 \Rightarrow P(A_1)(1-0{,}8)+(1-P(A_1))0{,}8=0{,}38;$$
$$0{,}2\mathrm{P}(A_1)+0{,}8-0{,}8\mathrm{P}(A_1)=0{,}38;\ \mathrm{P}(A_1)=0{,}7.$$

Пример 1.26. Студент разыскивает нужную ему формулу в трех справочниках. Вероятность того, что формула содержится в первом, втором и третьем справочнике соответственно равна 0,6; 0,7; 0,8.

Найти вероятность того, что формула содержится:

а) только в одном справочнике (событие A);

б) только в двух справочниках (событие B);

в) во всех трех справочниках (событие C);

г) хотя бы в одном справочнике (событие D);

д) не содержится ни в одном справочнике (событие E).

Решение. Рассмотрим следующие события и их вероятности:

A_1 — формула находится в 1-м справочнике, $\mathrm{P}(A_1)= 0{,}6$; $\mathrm{P}\left(\overline{A}_1\right) = 0{,}4$;

A_2 — формула находится во 2-м справочнике, $\mathrm{P}(A_2) = 0{,}7$; $\mathrm{P}\left(\overline{A}_2\right) = 0{,}3$;

A_3 — формула находится в 3-м справочнике, $\mathrm{P}(A_3) = 0{,}8$, $\mathrm{P}\left(\overline{A}_3\right) = 0{,}2$.

Выразим через исходные события и их отрицания все события A–E, применим теоремы параграфа 1.6:

а) $A=A_1\overline{A}_2\overline{A}_3+\overline{A}_1A_2\overline{A}_3+\overline{A}_1\overline{A}_2A_3$,

$P(A)=P(A_1\bar{A}_2\bar{A}_3)+P(\bar{A}_1A_2\bar{A}_3)+P(\bar{A}_1\bar{A}_2A_3)=P(A_1)P(\bar{A}_2)P(\bar{A}_3)+$
$+P(\bar{A}_1)P(A_2)P(\bar{A}_3)+P(\bar{A}_1)P(\bar{A}_2)P(A_3)=0{,}6\cdot0{,}3\cdot0{,}2+0{,}4\cdot0{,}7\cdot0{,}2+0{,}4\cdot0{,}3\cdot0{,}8=0{,}188;$

б) $B=A_1A_2\overline{A}_3+A_1\overline{A}_2A_3+\overline{A}_1A_2A_3$,

далее, аналогично пункту а), получим, что $\mathrm{P}(\mathrm{B})=0{,}452$;

в) $\mathrm{C} = A_1A_2A_3$,

$\mathrm{P}(\mathrm{C}) = \mathrm{P}(A_1\,A_2\,A_3) = \mathrm{P}(A_1)\mathrm{P}(A_2)\mathrm{P}(A_3) = 0{,}6\cdot0{,}7\cdot0{,}8 = 0{,}336$;

г) $D = A_1+A_2+A_3$,

вероятность события D можно найти, обобщив теорему 2 для трех событий:

$$\mathrm{P}(\mathrm{D}) = \mathrm{P}(A_1+A_2+A_3) =$$
$$= \mathrm{P}(A_1) + \mathrm{P}(\mathrm{A}_2) + \mathrm{P}(A_3) - \mathrm{P}(A_1A_2) - \mathrm{P}(A_1A_3) - \mathrm{P}(A_2A_3) + \mathrm{P}(A_1A_2A_3).$$

Но проще воспользоваться следствием к теореме 6:

$$\mathrm{P}(D) = 1 - \mathrm{P}\left(\overline{A}_1\right)\mathrm{P}\left(\overline{A}_2\right)\mathrm{P}\left(\overline{A}_3\right) = 1 - 0{,}4\cdot0{,}3\cdot0{,}2 = 0{,}976;$$

д) $E=\overline{A}_1\overline{A}_2\overline{A}_3$, $\mathrm{P}(\mathrm{E}) =0{,}4\cdot0{,}3\cdot0{,}2 = 0{,}024$.

Пример 1.27. Два игрока поочередно бросают игральную кость. Выигрывает первый, у которого появится шесть очков. Найти вероятность выигрыша каждого игрока.

Решение. Пусть событие A_i — выиграл первый игрок, событие B_i — выиграл второй при i −ом подбрасывании (i=1, 2, …):

$$\mathrm{P}(A_i) = P(B_i) = \frac{1}{6}, \mathrm{P}(\overline{A}_i) = P(\overline{B}_i) = \frac{5}{6}.$$

Выигрыш первого игрока до $(k+1)$-го подбрасывания — событие А:

$$A = A_1 + \overline{A}_1\overline{B}_1A_2 + \overline{A}_1\overline{B}_1\overline{A}_2\overline{B}_2A_3 + \cdots + \overline{A}_1\overline{B}_1\overline{A}_2\overline{B}_2 \ldots \overline{A}_{k-1}\overline{B}_{k-1}A_k.$$

Событие A — сумма несовместных событий, каждый член которой, начиная со второго, является произведением независимых событий, поэтому

$$\mathrm{P}(A) = \frac{1}{6} + \frac{5}{6}\cdot\frac{5}{6}\cdot\frac{1}{6} + \frac{5}{6}\cdot\frac{5}{6}\cdot\frac{5}{6}\cdot\frac{5}{6}\cdot\frac{1}{6} + \cdots + \left(\frac{5}{6}\right)^{k-1}\cdot\frac{1}{6} = \frac{1}{6}\cdot\frac{1-\left(\frac{25}{36}\right)^k}{1-\frac{25}{36}} =$$

$$= \frac{1}{6}\cdot\frac{1-\left(\frac{25}{36}\right)^k}{\frac{11}{36}},$$

как сумма k членов геометрической прогрессии с первым членом $\frac{1}{6}$ и знаменателем $\frac{25}{36}$. При $k\to\infty$, по формуле суммы членов бесконечной убывающей геометрической прогрессии, имеем

$$P(A) = \frac{\frac{1}{6}}{1-\frac{25}{36}} = \frac{6}{11}.$$

Аналогично, если выигрыш второго игрока до $(k+1)$-го подбрасывания — событие В, то

$$B = \overline{A}_1B_1 + \overline{A}_1\overline{B}_1\overline{A}_2B_2 + \cdots + \overline{A}_1\overline{B}_1\overline{A}_2\overline{B}_2 \ldots \overline{A}_{k-1}\overline{B}_{k-1}\overline{A}_kB_k,$$

$$\mathrm{P}(B) = \frac{5}{6}\cdot\frac{1}{6} + \frac{5}{6}\cdot\frac{5}{6}\cdot\frac{5}{6}\cdot\frac{1}{6} + \cdots + \left(\frac{5}{6}\right)^{2(k-1)}\frac{5}{6}\cdot\frac{1}{6} = \frac{5}{36}\cdot\frac{1-\left(\frac{25}{36}\right)^k}{1-\frac{25}{36}}.$$

При $k\to\infty$, $\mathrm{P}(B) = \frac{\frac{5}{36}}{1-\frac{25}{36}} = \frac{5}{11}$.

Пример 1.28. Гардеробщица выдала сразу номерки владельцам n шляп, повесив их наугад. Какова вероятность того, что хотя бы одно из n лиц, сдавших в гардероб шляпы, получит свою шляпу?

Решение. Пусть событие A_i $(i = 1, 2, \ldots, n)$ означает, что i-ый владелец шляпы получит ее обратно. Тогда, если A — событие, означающее, что хотя бы одно лицо получит свою шляпу, то по определению операции суммы событий, получим, что

$$A = A_1 + A_2 + A_3 + \cdots + A_i + \cdots + A_{n-1} + A_n.$$

Пусть $n = 3$, тогда, используя теорему сложения для совместных событий, получим вероятность наступления события A:

$$\mathrm{P}\big(A_1+(A_2+A_3)\big)=P(A_1)+P(A_2+A_3)-\ \mathrm{P}\big(A_1(A_2+A_3)\big)=\ P(A_1)+$$
$$+P(A_2)+P(A_3)-P(A_2A_3)-\ \mathrm{P}(A_1)(P(A_2)+P(A_3)-P(A_2A_3))=$$
$$=P(A_1)+P(A_2)+P(A_3)-P(A_2A_3)-\ \mathrm{P}(A_1A_2)-P(A_1A_3)+P(A_1A_2A_3).$$

Обобщая полученный результат, опираясь на метод математической индукции, можно получить вероятность наступления события A:

$$P(A)=P(\textstyle\sum_{i=1}^{n}A_i)=\sum_{i=1}^{n}\mathrm{P}(A_i)-\sum_{1\le i<j\le n}\mathrm{P}\big(A_iA_j\big)+$$
$$+\textstyle\sum_{1\le i<j<k\le n}\mathrm{P}\big(A_iA_jA_k\big)+\ldots+(-1)^{n+1}P(A_1A_2\ldots A_n).$$

Рассмотрим вероятности сумм:

$P(A_i)=\frac{1}{n}$, количество таких слагаемых равно $\mathrm{C}_n^1=\frac{n}{1!}$;

$\mathrm{P}\big(A_iA_j\big)=\frac{1}{n(n-1)}$, количество слагаемых равно $\mathrm{C}_n^2=\ \frac{n(n-1)}{2!}$;

$\mathrm{P}\big(A_iA_jA_k\big)=\frac{1}{n(n-1)(n-2)}$, количество слагаемых равно $\mathrm{C}_n^3=\ \frac{n(n-1)(n-2)}{3!}$ и т. д.

Перемножая, получим

$$P(A)=n\frac{1}{n}-\frac{n(n-1)}{2!}\frac{1}{n(n-1)}+\frac{n(n-1)(n-2)}{3!}\frac{1}{n(n-1)(n-2)}+\cdots+(-1)^{n+1}\frac{1}{n!}$$

или

$$P(A)=1-\frac{1}{2!}+\frac{1}{3!}+\cdots+(-1)^{n+1}\frac{1}{n!}.$$

Как известно из курса математического анализа, разложение функции e^x в ряд по формуле Маклорена имеет вид

$$e^x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\cdots+\frac{x^n}{n!}+\cdots,$$

при $x=-1$ имеем

$$e^{-1}=\frac{1}{2!}-\frac{1}{3!}+\cdots+(-1)^n\frac{1}{n!}+\cdots$$

Следовательно, при $n\to+\infty$

$$P(A)=1-e^{-1}\approx 0{,}6321.$$

Тогда вероятность того, что ни один владелец не получит своей шляпы будет стремиться к $e^{-1}\approx 0{,}3679$.

Самое интересное, что вероятности почти не зависят от числа шляп, сданных в гардероб. (С использованием *MS Excel* найдите соответствующие вероятности при $n=2,\ldots,10$ и сделайте вывод.)

1.5. Формулы полной вероятности и вероятности гипотез

Пусть событие A может наступать вместе с одним из несовместных событий $H_1, H_2, \ldots, H_i, \ldots, H_n$, образующих полную группу (рис. 1.7).

Тогда вероятность события A определяется по *формуле полной вероятности*:

$$P(A)=P(H_1)P(\mathrm{A}/H_1)+\ldots+P(H_i)P(A/H_i)+\cdots+\mathrm{P}(H_\mathrm{n})\mathrm{P}(\mathrm{A}/H_\mathrm{n}).$$

Или

$$P(A)=\textstyle\sum_{i=1}^{n}\mathrm{P}(H_i)\mathrm{P}(A/H_i),\qquad(1.32)$$

где события $H_1, H_2, \ldots, H_i, \ldots, H_n$ — гипотезы, а $P(A/H_i)$ — условная вероятность наступления события A при наступлении i-ой гипотезы ($i=1,2,\ldots,n$).

$$\textstyle\sum_{i=1}^{n}P(H_i)=1.$$

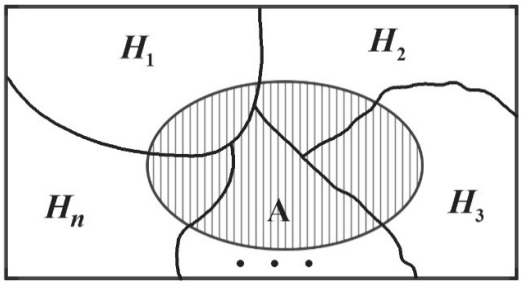


Рис. 1.7 — Иллюстрация наступления события A и гипотез Hi

Доказательство. Событие A может наступить вместе с одним из событий $H_1, H_2, \ldots, H_i, \ldots, H_n$. Значит,

$$A = AH_1 + AH_2 + \ldots + AH_i + \cdots + AH_n.$$

Так как события $AH_i (i = 1, 2, \ldots, n)$ несовместны, то по теореме сложения вероятностей

$$P(A) = P(AH_1) + P(AH_2) + \ldots + P(AH_i) + \ldots + \mathrm{P}(AH_\mathrm{n}).$$

Применяя к событиям AH_i теорему умножения вероятностей для зависимых событий, находим

$$P(A) = P(H_1)P(\mathrm{A}/H_1) + \ldots + P(H_i)P(\mathrm{A}/H_\mathrm{i}) + \cdots + \mathrm{P}(H_\mathrm{n})\mathrm{P}(\mathrm{A}/H_\mathrm{n}). \quad (1.33)$$

Заметим, что из теоремы 4 о вероятности произведения зависимых событий следует, что

$$P(AH_i) = P(H_i)P(A/H_i) = P(A)P(H_i/A). \quad (1.34)$$

Пусть событие A, вероятность появления которого определена по формуле полной вероятности, уже произошло. Тогда из (1.34) условная вероятность осуществления гипотезы H_i при условии того, что событие A произошло, определяется по формуле вероятности гипотез или *формуле Байеса* (она позволяет пересмотреть вероятности гипотез после наступления события A):

$$\mathrm{P}(H_i/A) = \frac{\mathrm{P}(H_i)\mathrm{P}(A/H_i)}{P(A)}, \quad (1.35)$$

где $P(H_i)$ — *априорные вероятности* гипотез H_i, $\mathrm{P}(H_i/A)$ — *апостериорные вероятности* гипотез H_i ($i = 1, 2, \ldots, n$); $\sum_{i=1}^{n} P(H_i) = 1$, $\sum_{i=1}^{n} P(H_i/A) = 1$.

Замечание. Формула (1.35) была получена Т. Байесом в XVIII в., предложившим использовать новые данные для корректировки первоначальных убеждений. Сегодня формула Байеса — основа байесовской статистики и имеет большое значение в теории искусственного интеллекта и современных информационных технологиях обработки данных, в том числе в сети Интернет. ■

Пример 1.29. Команда стрелков состоит из 5 человек, трое из них попадают в цель с вероятностью 0,8, а двое — с вероятностью 0,6. Наудачу из команды берется стрелок и производит выстрел.

а) Какова вероятность того, что стрелок попадет в цель?

б) Если стрелок попал в цель, то какова вероятность, что это один из трех (один из двух) стрелков?

Решение. а) Обозначим через A событие, что случайно взятый стрелок попадет в цель. Событие A может произойти, если произойдет одно из несовместных событий: H_1 — наудачу взятый стрелок один из трех, H_2 — наудачу взятый стрелок один из двух стрелков. Для определения вероятности события A воспользуемся формулой (1.32):

$$\mathrm{P}(A) = \mathrm{P}(H_1)\mathrm{P}(A/H_1) + \mathrm{P}(H_2)\mathrm{P}(A/H_2) = 0{,}6 \cdot 0{,}8 + 0{,}4 \cdot 0{,}6 = 0{,}72,$$

так как

$$\mathrm{P}(H_1) = \frac{3}{5} = 0{,}6,\ \mathrm{P}(A/H_1) = 0{,}8, \quad \mathrm{P}(H_2) = \frac{2}{5} = 0{,}4,\ \mathrm{P}(A/H_2) = 0{,}6.$$

б) По формуле (1.35):

$$\mathrm{P}(H_1/A) = \frac{\mathrm{P}(H_1)\mathrm{P}(A/H_1)}{\mathrm{P}(A)} = \frac{0{,}6 \cdot 0{,}8}{0{,}72} = \frac{2}{3},$$

$$\mathrm{P}(H_2/A) = \frac{\mathrm{P}(H_2)\mathrm{P}(A/H_2)}{\mathrm{P}(A)} = \frac{0{,}4 \cdot 0{,}6}{0{,}72} = \frac{1}{3},$$

или

$$P(H_2/A) = 1 - P(H_1/A) = 1 - \frac{2}{3} = \frac{1}{3}.$$

Пример 1.30. По предмету теория вероятностей и математическая статистика имеется 30 экзаменационных билетов. Студент Павлов выучил только 20. Каким выгоднее ему зайти на экзамен: первым или вторым?

Решение. Событие A — студент Павлов заходит первым на экзамен,

$$\mathrm{P}(A) = \frac{20}{30} = \frac{2}{3}.$$

Событие B — студент Павлов заходит вторым на экзамен, которое может произойти только с одним из попарно несовместных событий A_1, A_2, где: событие A_1 — 1-й студент вытащит 1 из 20 билетов, которые Павлов знает; событие A_2 — 1-й студент вытащит 1 из 10 остальных билетов.

$$B = A_1B + A_2B,$$

$$\mathrm{P}(B) = \mathrm{P}(A_1)\mathrm{P}(B/A_1) + \mathrm{P}(A_2)\mathrm{P}(B/A_2) =$$

$$= \frac{20}{30} \cdot \frac{19}{29} + \frac{10}{30} \cdot \frac{20}{29} = \frac{2}{3}.$$

То есть все равно, зайдет студент первым или вторым.

Графическое представление. Анализ логических возможностей при проведении опыта или исследовании какого-то процесса обычно проводится с использованием графов.

Граф (G) *— это множество вершин (V), соединенных ребрами (E)*, что обозначается как $G =< V, E >$. Граф G называется взвешенным, если каждому ребру (i, j) ставится в соответствие вес w_{ij}. Если веса представляют собой вероятности p_{ij}, причем $\sum_i p_{ij} = 1$, то граф называется вероятностным.

С помощью графов можно иллюстрировать правила сложения и умножения вероятностей событий (рис. 1.8, 1.9).

1. $P(AB + CD) = P(A)\,P(B/A) + P(C)\,P(D/C)$,

где AB и CD — несовместные события.

Если выделяется одна вершина (корень) графа, которая обозначает начало изучаемого процесса, тогда (при отсутствии циклов) такой граф называют деревом. Дерево представляет графический вариант классификации объектов, когда одна группа детализируется по ярусам (уровням).

Серия испытаний, где каждый эксперимент зависит от исходов предыдущих, изображается в виде дерева, *последовательность ветвей* которого — *путь,* определяет множество всех логических возможностей. Дерево начинается из исходной точки, а ветви, исходящие из этой точки, образуют первый ряд дерева. Конец каждой ветви отвечает возможному результату первого эксперимента. Из конца каждой ветви начинается новое множество ветвей, отвечающих результатам второго опыта. Так, ряд за рядом строится дерево, пока не будут исчерпаны все эксперименты. Каждый ярус (ряд) дерева отвечает одному из возможных исходов предыдущего эксперимента. Каждая точка ветвления соответствует единственной логической возможности предшествующих опытов.

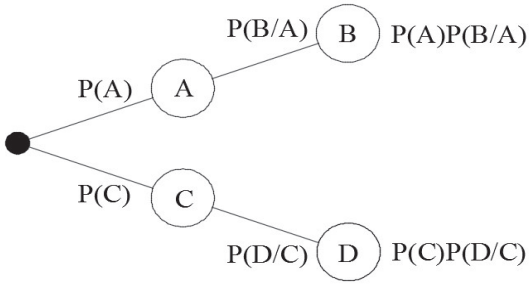


Рис. 1.8 — Граф правила сложения и умножения вероятностей событий

2. $P(A_1 A_2 \ldots A_{n-1} A_n) = P(A_1) P(A_2/A_1) \ldots P(A_n/A_1 A_2 \ldots A_{n-1})$.

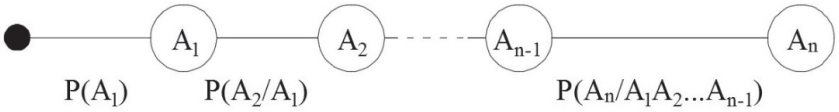


Рис. 1.9 — Граф правила умножения вероятностей событий

Для прогнозирования возможных исходов последовательности экспериментов, которая называется *случайным процессом,* строится вероятностная мера (весовая функция, определенная на пространстве элементарных событий Ω, удовлетворяющая аксиомам вероятности) путей дерева, определенная на множестве всех путей конкретного дерева. Сумма вероятностей всех ветвей, выходящих из одной точки (*пучок ветвей*), равна единице (то есть образует полную группу событий).

Замечание. Конечный стохастический процесс полностью описывается деревом логических возможностей и заданием на его ветвях вероятностных мер. При этом можно выделить три наиболее часто встречающихся типа (конечных) случайных процессов, отличающихся требованиями к пучкам ветвей из каждой точки ветвления:

– с независимыми значениями — пучки ветвей, принадлежащие одному ярусу (ряду), эквивалентны, например последовательность подбрасывания монеты (Г, Р) и игральной кости (1 очко, больше 1 очка);

– *независимых испытаний* (все пучки ветвей эквивалентны, гл. 2);

– *марковских цепей* (веса ветвей одного яруса (ряда) зависят от исхода предыдущего эксперимента, гл. 9). ■

Пример 1.31. Имеется две урны, первая содержит два черных и один белый шар, а вторая один черный и два белых шара. Наудачу выбирается урна и из нее последовательно выбирается два шара. Какова вероятность того, что второй шар белый? Если шар белый, то какова вероятность, что выбрали вторую урну?

Таблица 1.3

Логические возможности

Случай	Урна	Первый шар	Второй шар
1	1	Черный	Черный
2	1	Черный	Белый
3	1	Белый	Черный
4	2	Черный	Белый
5	2	Белый	Черный
6	2	Белый	Белый

Решение. Событие A — выбранный шар белый; гипотеза H_1 — выбрана первая урна, H_2 – выбрана вторая урна. Логические возможности выбора представлены в таблице 1.3 и на рисунке 1.10, соответствующее вероятностное дерево на рисунке 1.11.

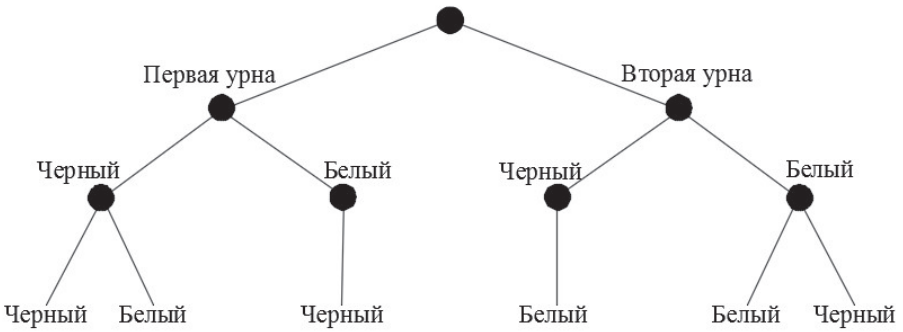


Рис. 1.10 — Дерево логических возможностей

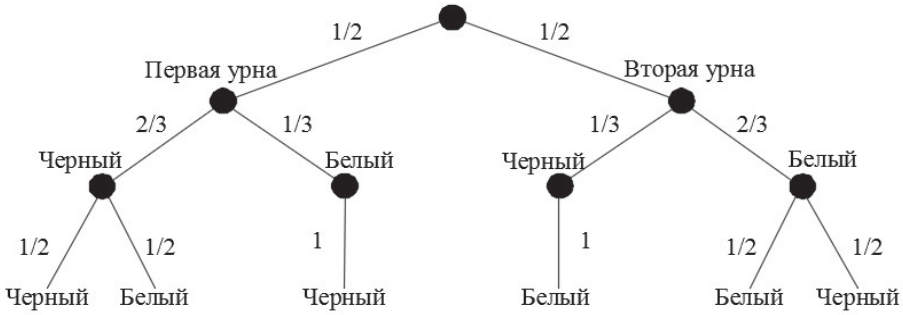


Рис. 1.11 — Вероятностное дерево

Согласно формуле полной вероятности:

$$P(A) = P(H_1)P(A/H_1) + P(H_2)P(A/H_2),$$

$$P(A) = \frac{1}{2}\cdot\frac{2}{3}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{3}\cdot 1 + \frac{1}{2}\cdot\frac{2}{3}\cdot\frac{1}{2} = \frac{6}{12} = \frac{1}{2}.$$

Значит, вероятность того, что второй шар белый, равна 0,5.

Если второй выбранный шар белый, то согласно формуле Байеса вероятность того, что была выбрана вторая урна, составит

$$P(H_2/A) = \frac{P(H_2)P(A/H_2)}{P(A)} = \frac{4/12}{1/2} = \frac{2}{3}.$$

Современная экономика базируется на том, что два действующих лица — производитель и покупатель решают, что и по какой цене выгодно производить и покупать соответственно. На протяжении долгого времени считалось, что человек, действуя в условиях неопределенности и имея знания о вероятностях тех или иных событий, поступает рационально, то есть в явной или неявной форме пользуется так называемыми аксиомами рационального поведения. Предполагается, что существует некоторая функция полезности, которую рационально действующий экономический субъект в процессе выбора стратегии поведения старается максимизировать.

Пример 1.32. На рынке кофемашин покупатель хочет купить одну кофемашину, продавец продать кофемашину. Есть два типа кофемашин: высокого и низкого качества. Вероятность отказа машин высокого качества за определенный период времени составляет 0,2, а низкого — 0,75. Если кофемашина работает без сбоев, то полезность для покупателя 400 у. е., сбой уменьшает полезность до 200 у. е. Известно, что 25% кофемашин реализуется высокого качества. Какую кофемашину следует выбрать покупателю?

Решение. Пусть H_1, H_2 — выбор кофемашин высокого и низкого качества соответственно. Условия задачи можно представить в виде таблицы 1.4 и графически в виде дерева решений (рис. 1.12).

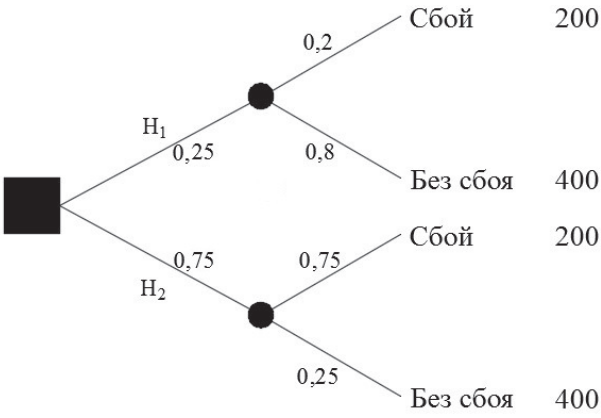


Рис. 1.12 — Дерево решений

Таблица 1.4

Тип машины	Доля на рынке	Вероятность отказа машины	Выигрыш при работе без сбоя, у. е.	Выигрыш при сбое, у. е.
Высокого качества, H_1	0,25	0,20	400	200
Низкого качества, H_2	0,75	0,75	400	200

Оценим среднюю ожидаемую полезность U (*utility*).

$U(H_1) = 0{,}25(0{,}2 \cdot 200 + 0{,}8 \cdot 400) = 90$ у. е.

$U(H_2) = 0{,}75(0{,}75 \cdot 200 + 0{,}25 \cdot 400) = 187{,}5$ у. е.

Если выбирать действие с максимальной ожидаемой полезностью, то рациональный человек должен выбрать действие H_2, а не H_1.

Приведенный ранее рисунок 1.12 называют деревом решений. Квадратик — место, где человек принимает решение, а кружочек — место, где все решает случай. На ветвях дерева написаны значения вероятностей, а справа у конечных ветвей значения исходов (результаты). Деревья решений используются для представления возможных действий.

Темы (вопросы) для самоконтроля

1. Понятие случайного события и его свойства.
2. Алгебраические операции над событиями.
3. Аксиоматическое, классическое и статистическое определения вероятности события.
4. Геометрическое определение вероятности события.
5. Основные понятия комбинаторики и их применение к вычислению вероятности классическим способом.
6. Условная вероятность. Независимость событий.
7. Теоремы сложения и умножения.
8. Вероятности сложных событий.
9. Формула полной вероятности.
10. Формула Байеса.

Глава 2
Повторные независимые испытания

2.1. Схемы повторных независимых испытаний и формула Бернулли

1. *Постоянные условия опыта.*

а) Пусть некоторый опыт повторяется в неизменных условиях n раз, причем событие A в каждом испытании может либо наступить (успех), либо не наступить (неудача).

Обозначим $P(A) = p$ — вероятность успеха, $P(\bar{A}) = 1 - p = q$ — вероятность неудачи. Тогда вероятность того, что событие A произойдет в n испытаниях k раз вычисляется по *формуле Бернулли*:

$$P_n(k) = C_n^k \, p^k \, q^{n-k}. \tag{2.1}$$

Доказательство. Последовательности наступления и ненаступления событий A и $\overline{A}$ могут чередоваться различным образом. Каждая комбинация, в которую A входит k раз и $\overline{A}$ входит $(n-k)$ раз, называется благоприятной. Число благоприятных ситуаций равно числу способов выбора k элементов из n, то есть $m = C_n^k$. Рассмотрим событие B_1 — благоприятную ситуацию, кода первые k раз событие A произошло, а $(n-k)$ не произошло:

$$B_1 = A\,A\,\ldots A\,\overline{A}\,\overline{A}\,\ldots\overline{A}.$$

Вероятность события B_1 равна $p^k q^{n-k}$. В любой благоприятной ситуации, когда событие A произошло k раз и $(n-k)$ не произошло, вероятность будет равна $p^k q^{n-k}$:

$$P(B_1) = P(B_2) = \cdots = P(B_m) = p^k q^{n-k}.$$

Все благоприятные комбинации несовместны, поэтому

$$P_n(k) = P(B_1 + B_2 + \cdots + B_m) = P(B_1) + P(B_2) + \cdots + P(B_m) = \\ = mp^k q^{n-k} = C_n^k \, p^k \, q^{n-k}.$$

Рассмотрение и построение соответствующего вероятностного дерева позволяет визуализировать доказательство формулы (рис. 2.1).

Условия, приводящие к формуле Бернулли, называются *частной схемой повторных независимых испытаний*, или *схемой Бернулли*. Так как вероятности $P_n(k)$ для различных значений k представляют собой слагаемые в разложении бинома Ньютона:

$$(p+q)^n = C_n^0 p^0 q^n + C_n^1 p^1 q^{n-1} + \cdots + C_n^k p^k q^{n-k} + \cdots + C_n^n p^n q^0 = 1,$$

то распределение вероятностей $P_n(k)$, где $0 \le k \le n$, называется *биномиальным распределением*.

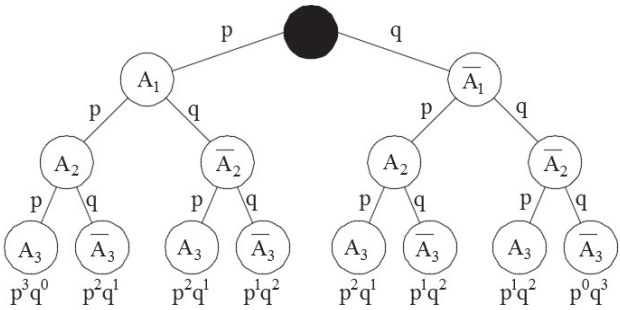


Рис. 2.1 — Вероятностное дерево независимых испытаний
с постоянными условиями опыта

б) Пусть проводятся испытания по схеме Бернулли. Вычислим по формуле Бернулли вероятность наступления $(k-1)$ успехов $(k \geq 1)$, m неудач и обозначим как

$$P_{m+k-1}(k-1,m) = C_{m+k-1}^{m} p^{k-1} q^{m}, \qquad (2.2)$$

$(m = 0, 1, 2, \ldots)$.

Тогда вероятность появления m неудач до получения k успехов будет равна

$$P_{m+k}(k,m) = pP_{m+k-1}(k-1,m) = C_{m+k-1}^{m} p^{k} q^{m}, \qquad (2.3)$$

соответствующее распределение вероятностей называется *отрицательным биномиальным* (причем множество возможных случаев (неудач) может быть бесконечно).

2. *Переменные условия опыта*.

Если в каждом из независимых испытаний вероятности наступления события A разные (*общая схема повторения опытов*), то вероятность наступления события A k раз в n испытаниях определяется как коэффициент при k – ой степени полинома

$$\varphi(Z) = \prod_{i=1}^{n}(q_i + p_i Z) = a_n Z^n + a_{n-1} Z^{n-1} + \cdots + a_1 Z^1 + a_0, \qquad (2.4)$$

где $\varphi(Z)$ — *производящая функция.*

Соответствующее вероятностное дерево (рис. 2.2) позволяет визуализировать смысл формулы (2.4).

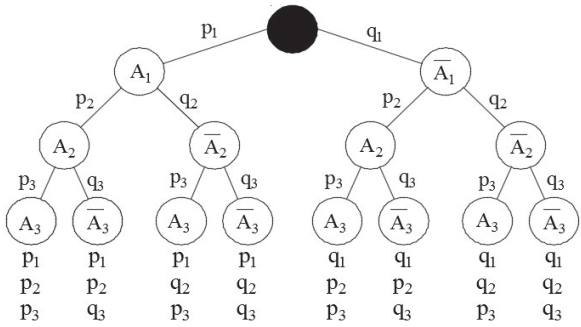


Рис. 2.2 — Вероятностное дерево независимых испытаний
с переменными условиями опыта

3. *Опыт с несколькими событиями*.

Если в результате опыта может появиться одно из несовместных событий $A_1, A_2, \ldots, A_r$, образующих полную группу, где

$$\mathrm{P}(A_1) = p_1, \mathrm{P}(A_2) = p_2, \ldots, \mathrm{P}(A_r) = p_r \text{ и } \sum_i p_i = 1,$$

то вероятность того, что в n опытах появится событие A_1 — k_1 раз, A_2 — k_2 раз, …, A_r — k_r раз ($\sum_i k_i = n$), определяется по формуле

$$P_n(k_1, k_2, \ldots, k_r) = \frac{n!}{k_1! k_2! \ldots k_r!} p_1^{k_1} p_2^{k_2} \ldots p_r^{k_r}. \quad (2.5)$$

Соответствующее распределение вероятностей называется *полиномиальным (мультиномиальным)*.

Часто применяемые формулы в схеме Бернулли.

Вероятность наступления события A в n независимых испытаниях:

а) менее k раз: $P_n(0) + P_n(1) + \cdots + P_n(k-1)$;

б) более k раз: $P_n(k+1) + P_n(k+2) + \cdots + P_n(n)$;

в) не менее k раз: $P_n(k) + P_n(k+1) + \ldots + P_n(n)$;

г) не более k раз: $P_n(0) + P_n(1) + \cdots + P_n(k)$;

д) хотя бы один раз: $1 - P_n(0)$.

Пример 2.1. Монета, подбрасывается 6 раз в неизменных условиях. Успехом считается выпадение герба (событие A). Найти вероятность того, что герб появится 4 раза.

Решение. Условия проведения опыта соответствуют схеме Бернулли.

$P(A) = p = \frac{1}{2}, P(\bar{A}) = 1 - p = q = 1 - \frac{1}{2} = \frac{1}{2}$. Согласно условию $n = 6$, $k = 4$.

По формуле Бернулли имеем

$$P_6(4) = C_6^4 p^4 q^{6-4} = \frac{6!}{4!(6-4)!}\left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^2 = \frac{15}{64}.$$

Пример 2.2. Стрелок попадает в цель с вероятностью 0,6, и он собирается произвести 10 выстрелов. Найти вероятность того, что стрелок попадет в цель: а) три раза, б) хотя бы один раз.

Решение. $p = 0{,}6$; $q = 1 - p = 1 - 0{,}6 = 0{,}4$; $n = 10$.

а) $P_{10}(3) = C_{10}^3 p^3 q^{10-3} = \frac{10!}{(10-3)!3!} 0{,}6^3\, 0{,}4^7 = 0{,}255$;

б) $P_{10}(k \geq 1) = 1 - P_{10}(0) = 1 - C_{10}^0 p^0 q^{10-0} = 1 - 0{,}4^{10} = 0{,}9999$.

Если в независимых испытаниях найти вероятности всех возможных исходов, то их значения будут вначале возрастать, а затем, достигнув наибольшего значения, станут убывать. Число наступления события A в n независимых испытаниях называется наивероятнейшим, если вероятность наступить событию A в независимых испытаниях является наибольшей по сравнению с вероятностями других исходов и обозначается k_0.

Наивероятнейшее число наступивших событий в схеме Бернулли — k_0 ($k_0 \in N$), определяется из следующего неравенства:

$$np - q \leq k_0 \leq np + p. \quad (2.6)$$

Доказательство. Так как k_0 — наивероятнейшее число, то

$$P_n(k_0) \geq P_n(k_0 + 1) \text{ и } P_n(k_0) \geq P_n(k_0 - 1).$$

Рассмотрим первое неравенство. По формуле Бернулли находим

$$P_n(k_0) = C_n^{k_0} p^{k_0} q^{n-k_0} \text{ и } P_n(k_0+1) = C_n^{k_0+1} p^{k_0+1} q^{n-k_0-1}.$$

Следовательно,

$$\frac{n(n-1)\ldots(n-k_0+1)}{1\cdot 2\cdot\ldots\cdot k_0} p^{k_0} q^{n-k_0} \geq \frac{n(n-1)\ldots(n-k_0)}{1\cdot 2\cdot\ldots\cdot(k_0+1)} p^{k_0+1} q^{n-k_0-1}.$$

Так как $p > 0$ и $q > 0$, то $q \geq \frac{n-k_0}{k_0+1} p$ или $qk_0 + q \geq np - pk_0$.

$qk_0 + pk_0 + q \geq np$ или $k_0(q+p) \geq np - q$, но $p + q = 1$, значит, $k_0 \geq np - q$.

Аналогично из неравенства $P_n(k_0) \geq P_n(k_0 - 1)$ следует, что $k_0 \leq np + p$. Таким образом, $np - q \leq k_0 \leq np + p$.

Наивероятнейшее число k_0 — целое. В случае, если $(np - q)$ — целое, то имеется два наивероятнейших числа.

В *примере* 2.1: $6 \cdot \frac{1}{2} - \frac{1}{2} \leq k_0 \leq 6 \cdot \frac{1}{2} + \frac{1}{2}$,

$$2{,}5 \leq k_0 \leq 3{,}5,$$

$$k_0 = 3.$$

В *примере* 2.2: $10 \cdot 0{,}6 - 0{,}4 \leq k_0 \leq 10 \cdot 0{,}6 + 0{,}6$,

$$5{,}6 \leq k_0 \leq 6{,}6,$$

$$k_0 = 6.$$

Пример 2.3. Подбрасывается 12 игральных костей. Какова вероятность, что выпадут 7 одинаковых цифр, а остальные цифры выпадут по одному разу?

Решение. Для каждой игральной кости элементарные события $A_1, A_2, A_3, A_4, A_5, A_6$ соответствуют шести граням, вероятность каждого события $P(A_i) = \frac{1}{6}$. Одинаковые цифры могут быть: 1, 2, 3, 4, 5, 6. Поэтому сначала найдем вероятность того, что появилось семь единиц. По формуле полиномиального распределения (2.4) имеем

$$\frac{12!}{7!1!1!1!1!1!} \left(\frac{1}{6}\right)^7 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^1 = \frac{12!}{7!} \left(\frac{1}{6}\right)^{12}.$$

Вероятность появления любых семи одинаковых цифр очевидно будет в 6 раз больше:

$$P(A) = 6 \cdot \frac{12!}{7!} \cdot \left(\frac{1}{6}\right)^{12} = \frac{12!}{7!} \cdot \left(\frac{1}{6}\right)^{11} = 0{,}000262.$$

Пример 2.4. Производится три независимых выстрела по некоторой цели. Вероятности попадания при разных выстрелах различны и равны: p_1=0,7; p_2=0,8; p_3=0,9. Найти вероятность трех промахов, одного, двух и всех трех попаданий.

Решение. Производящая функция:

$$\varphi_3(Z) = (0{,}3 + 0{,}7Z)(0{,}2 + 0{,}8Z)(0{,}1 + 0{,}9Z) =$$

$$= 0{,}504Z^3 + 0{,}398Z^2 + 0{,}092Z + 0{,}006.$$

Отсюда, вероятность:

− трех промахов $P_3(0) = 0{,}006$;

− одного попадания $P_3(1) = 0{,}092$;

− двух попаданий $P_3(2) = 0{,}398$;

− трех попаданий $P_3(3) = 0{,}504$.

2.2. Приближенные формулы в схеме Бернулли

При большом числе опытов по схеме Бернулли удобнее пользоваться приближенными формулами.

1. Локальная формула Муавра — Лапласа. Если вероятность p наступления события A в n независимых испытаниях постоянна, отлична от нуля и единицы, а число испытаний достаточно велико ($npq \geq 10$), то вероятность того, что событие A появится k раз и не появится m раз ($k + m = n$), приближенно равна:

$$P_n(k) \approx \frac{1}{\sqrt{npq}}\varphi(x), \tag{2.7}$$

где

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}, \tag{2.8}$$

$$x = \frac{k - np}{\sqrt{npq}}, \tag{2.9}$$

x — находится в конечном интервале.

Доказательство. Из формулы (2.9) следует, что

$$k = np + x\sqrt{npq}\,, m = nq - x\sqrt{npq}. \tag{2.10}$$

Для вычисления факториалов при больших значениях n пользуются приближенной *формулой Стирлинга* (1730 г.):

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{O\left(\frac{\theta(n)}{12n}\right)}. \tag{2.11}$$

Замечание. 1. Символ «O» (о-большое) ввел Пауль Бахман (1894) для использования в приближенных формулах. Этот символ позволяет заменить знак « ≈ » знаком « = » и количественно выразить степень точности, например, в формуле Стирлинга (2.11) он показывает, что разница между логарифмом левой и правой части точно не определена, но какой бы она ни была, обозначение $\theta_n = O\left(\frac{\theta(n)}{12n}\right)$ позволяет утверждать, что она не превосходит константу, $\theta(n)$ $(0 < \theta(n) < 1)$, умноженную на $\frac{1}{12n}$:

$$\theta_n = O\left(\frac{\theta(n)}{12n}\right) < \frac{\theta(n)}{12n}.$$

2. Наряду с «о-большое» в асимптотическом анализе используется «о-малое». Пусть $f(x) = h(x)g(x)$ при $x \to a$ (соответственно, «при $x \to \pm\infty$»), тогда: если $\lim\limits_{x\to a}\frac{f(x)}{g(x)} = 1$, то $f(x)$ эквивалентна $g(x)$ и пишут

$$f(x) \sim g(x),$$

если $\lim\limits_{x\to a}\left|\frac{f(x)}{g(x)}\right| < C$ — предел отношения по модулю ограничен константой $C \in$ $\in R$, то пишут

$$f(x) = O(g(x)),$$

если $\lim\limits_{x\to a}\frac{f(x)}{g(x)} = 0$, то пишут

$$f(x) = o(g(x)).$$

Формулы, использующие «о-большое» и «о-малое», эффективно используются при вычислении пределов элементарных функций. Среди свойств введенных соотношений выделим следующее. При $x \to 0$:

$$O(x^{m+1}) = o(x^m).$$

Далее по тексту, чтобы избежать неоднозначности понимания, везде используется символ о-большое. ■

Относительная погрешность формулы (2.11) убывает с ростом n (от 8% при $n = 1$, до 0,08% при $n = 100$). При больших значениях n можно использовать приближенную формулу:

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

В формуле Бернулли, заменив факториалы по формуле Стирлинга (2.11), получим

$$P_n(k) = \frac{n!}{k!\,m!}p^k q^m = \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^{m+k} e^{\theta}}{\sqrt{2\pi k}\left(\frac{k}{e}\right)^k \sqrt{2\pi m}\left(\frac{m}{e}\right)^m} p^k q^m =$$

$$= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{km}}\; e^{\theta}\left(\frac{n}{k}p\right)^k \left(\frac{n}{m}q\right)^m, \qquad (2.12)$$

где $e^{\theta} \to 1$, так как при большом числе опытов

$$\theta = \theta_n - \theta_k - \theta_m < \frac{1}{12}\left(\frac{1}{n} + \frac{1}{k} + \frac{1}{m}\right) \to 0.$$

Рассмотрим величину $\ln A_n = \ln\left[\left(\frac{np}{k}\right)^k \left(\frac{nq}{m}\right)^m\right]$ и преобразуем ее по правилам логарифмирования:

$$\ln A_n = \ln\left[\left(\frac{np}{k}\right)^k \left(\frac{nq}{m}\right)^m\right] = k\,ln\left(\frac{np}{k}\right) + m\,ln\left(\frac{nq}{m}\right) =$$

$$= -\,k\,ln\left(\frac{k}{np}\right) - m\,ln\left(\frac{m}{nq}\right).$$

Из формул (2.10) следует, что

$$\ln A_n = -\left(np + x\sqrt{npq}\right)ln\left(1 + x\sqrt{\frac{q}{np}}\right) - \left(nq - x\sqrt{npq}\right)ln\left(1 - x\sqrt{\frac{p}{nq}}\right). \qquad (2.13)$$

Как известно из курса математического анализа, при $\alpha \to 0$ имеют место приближенные формулы

$$\ln(1 + \alpha) = \alpha - \frac{\alpha^2}{2} + O(\alpha^3),\ \ln(1 - \alpha) = -\alpha - \frac{\alpha^2}{2} + O(\alpha^3), \qquad (2.14)$$

где $O(\alpha^3)$ — функция о-большое, означающая, что разность между левой и правой частью в формулах (2.14) мала, но она не превосходит величину $c\alpha^3$, при $\alpha \to 0$, где $c = const$ $(0 < c < 1)$. Так как по условию считается, что n большое число, а переменная x конечна, то можно считать, что величины $x\sqrt{\frac{q}{np}}$ и $x\sqrt{\frac{p}{nq}}$ уменьшаются с ростом n, поэтому заменим их в равенстве (2.13) по формулам (2.14):

$$\ln A_n \approx \left(np + x\sqrt{npq}\right)\left(x\sqrt{\frac{q}{np}} - \frac{1}{2}x^2\frac{q}{np} + O(x^3\sqrt{\left(\frac{q}{np}\right)^3}\right) -$$

$$-(q - x\sqrt{npq})\left(-x\sqrt{\frac{p}{nq}} - \frac{1}{2}x^2\frac{p}{nq} + O(x^3\sqrt{\left(\frac{p}{nq}\right)^3}\right).$$

После элементарных преобразований получим

$$ln\, A_n = -\frac{x^2}{2}, \tag{2.15}$$

следовательно,

$$A_n = e^{-\frac{x^2}{2}}. \tag{2.16}$$

По формуле (2.11) имеем

$$P_n(k) \approx \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{km}}\; e^{-\frac{x^2}{2}}, \tag{2.17}$$

так как по условию n — большое число, то из формул (2.9) следует, что с ростом n ($n \to \infty$)

$$\frac{k}{n} \to p, \frac{m}{n} \to q. \tag{2.18}$$

Поэтому

$$P_n(k) \approx \frac{1}{\sqrt{npq}}\,\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} = \frac{1}{\sqrt{npq}}\varphi(x). \tag{2.19}$$

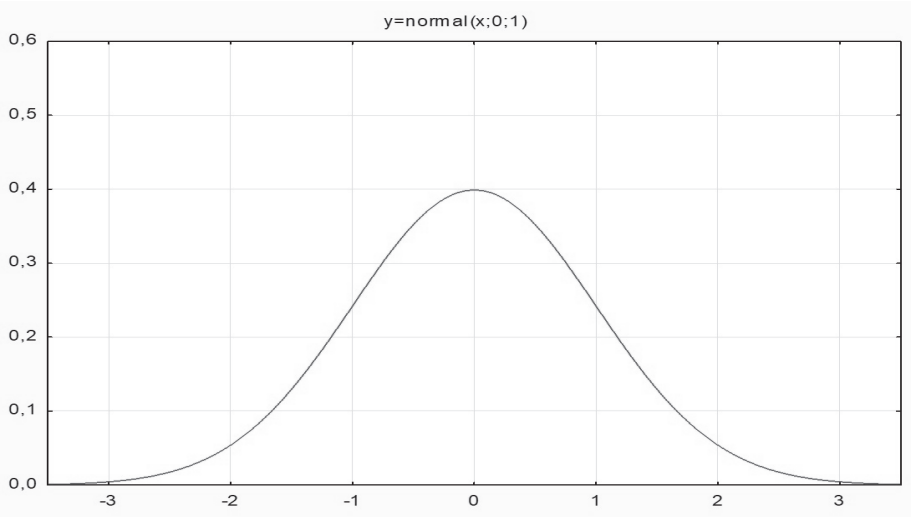


Рис. 2.3 — Функция плотности распределения вероятностей стандартного нормального распределения

Для облегчения вычислений значения функции $\varphi(x)$ представлена в виде таблицы (приложение 1). $\varphi(x)$ — функция плотности вероятности нормального распределения (рис. 2.3) имеет следующие свойства:

1) функция четная: $\varphi(-x) = \varphi(x)$;

2) точки перегиба $x = \pm 1$;

3) при $x \geq 5$, $\varphi(x) \to 0$, поэтому функция $\varphi(x)$ представлена в виде таблицы (затабулирована) для $0 \leq x \leq 5$.

2. *Формула Пуассона*. Если вероятность наступления события A в каждом из n независимых испытаний постоянна, причем $npq < 10$ и $p < 0{,}1$, то вероятность того, что событие A появится в n испытаниях k раз приближенно равна:

$$P_n(k) \approx \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \tag{2.20}$$

где $\lambda = np$.

Доказательство. Так как число опытов большое, то будем считать, что $n \to \infty$, $\lambda = np$ — не мало и не велико, $p = \frac{\lambda}{n}$.

Пусть в формуле Бернулли $k = 0$, тогда

$$P_n(0) = C_n^0 p^0 q^n = (1-p)^n = \left[\left(1 - \frac{\lambda}{n}\right)^{-\frac{n}{\lambda}}\right]^{-\lambda} = e^{-\lambda}, \tag{2.21}$$

так как по формуле второго замечательного предела

$$\lim_{n\to\infty} (1 + \theta(n))^{\frac{1}{\theta(n)}} = e,$$

где $\theta(n) = -\frac{\lambda}{n}$ бесконечно малая функция ($\theta(n) \to 0$ при $n \to \infty$).

Рассмотрим отношение $P_n(k)/P_n(k-1)$:

$$\frac{\frac{n!}{(n-k)!k!} p^k q^{n-k}}{\frac{n!}{(n-k+1)!(k-1)!} p^{k-1} q^{n-k+1}} = \frac{(n-k+1)p}{kq} = \frac{np - (k-1)p}{kq} = \frac{\lambda}{k}, \tag{2.22}$$

так как по условию $p \to 0$, а $q \to 1$.

Рассмотрим *доказательство* формулы Пуассона с использованием метода математической индукции. (База индукции.) Пусть $k = 0$, согласно (2.21):

$$P_n(0) = \frac{\lambda^0}{0!} e^{-\lambda} = e^{-\lambda} \text{ — верно,}$$

при $k = 1$, согласно формуле (2.22):

$$P_n(1) = \frac{\lambda^1}{1!} P_n(0) = \frac{\lambda^1}{1!} e^{-\lambda} \text{ — верно.}$$

Аналогично при $k = 2$:

$$P_n(2) = \frac{\lambda^2}{2!} P_n(2) = \frac{\lambda^2}{2!} e^{-\lambda} \text{ — верно.}$$

(Шаг индукции.) Пусть формула (2.20) верна при числе успехов в схеме Бернулли равном $(k-1)$:

$$P_n(k-1) = \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda}.$$

Тогда согласно формуле (2.22):

$$P_n(k) = \frac{\lambda}{k} \cdot P_n(k-1) = \frac{\lambda^k}{k!} e^{-\lambda},$$

что и требовалось доказать.

Пример 2.5. Стрелок произвел 400 выстрелов по мишени. Найти вероятность того, что число попаданий составит 325, если вероятность попадания при каждом выстреле равна 0,8.

Решение. $npq = 400 \cdot 0{,}8 \cdot 0{,}2 = 64 > 10$, следовательно, по формулам (2.7)–(2.9)

$$P_{400}(325) \approx \frac{1}{\sqrt{400 \cdot 0{,}8 \cdot 0{,}2}}\varphi(0{,}63) \approx \frac{1}{8} \cdot 0{,}3271 \approx 0{,}041,$$

где $x = \frac{k-np}{\sqrt{npq}} = \frac{325-400 \cdot 0{,}8}{\sqrt{400 \cdot 0{,}8 \cdot 0{,}2}} = 0{,}63.$

Пример 2.6. Завод отправил 5000 доброкачественных изделий. Вероятность того, что в пути разбили одно изделие, 0,0002. Найти вероятность того, что в пути будет повреждено:

а) три изделия; б) одно изделие; в) не более трех изделий.

Решение. $npq = 5000 \cdot 0{,}0002 \cdot 0{,}9998 = 0{,}9998 < 10$ и $p < 0{,}1$, поэтому применяем формулу Пуассона (2.20), где $\lambda = np = 5000 \cdot 0{,}0002 = 1$:

а) при $k = 3$: $P_{5000}(3) \approx \frac{1^3}{3!}e^{-1} = \frac{1}{6e} = 0{,}061,$

б) при $k = 1$: $P_{5000}(1) \approx \frac{1^1}{1!}e^{-1} = \frac{1}{e} \approx 0{,}368,$

в) $\mathrm{P}_{5000}\,(0 \le k \le 3) = \mathrm{P}_{5000}\,(0) + \mathrm{P}_{5000}\,(1) + \mathrm{P}_{5000}\,(2) + \mathrm{P}_{5000}\,(3) =$

$= \frac{1}{e} + \frac{1}{e} + \frac{1}{2e} + \frac{1}{6e} = \frac{16}{6e} \approx 0{,}981.$

3. При больших значениях n, для вычисления вероятности того, что произойдет от k_1 до k_2 событий по схеме Бернулли, используется *интегральная формула Муавра — Лапласа.*

Если вероятность p наступления события A в n независимых испытаниях постоянна, отлична от нуля и единицы, а число испытаний достаточно велико ($npq \ge 10$), то вероятность того, что событие A появится от k_1 до k_2 включительно раз приближенно равна

$$P_n(k_1 \le k \le k_2) \approx \Phi(x_2) - \Phi(x_1), \qquad (2.23)$$

где $x_1 = \frac{k_1-np}{\sqrt{npq}}$, $x_2 = \frac{k_2-np}{\sqrt{npq}}$,

Ф (x) — функция Лапласа (рис. 2.4).

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt. \qquad (2.24)$$

Доказательство. По условию схемы Бернулли число успехов k — переменная, принимающая целочисленные значения от k_1 до k_2. Введем новую переменную x, принимающую значения на непрерывном промежутке от x_1 до x_2, связанную с k по формуле

$$x = \frac{k - np}{\sqrt{npq}}. \qquad (2.25)$$

Выясним, какое приращение (Δx) получит переменная x, когда переменная k получит приращение 1 ($\Delta k = 1$).

$$k + 1 = np + (x + \Delta x)\sqrt{npq}, \qquad (2.26)$$

$$k = np + x\sqrt{npq}. \qquad (2.27)$$

Вычитая из формулы (2.26) формулу (2.27), получим, что

$$\Delta x = \frac{1}{\sqrt{npq}}. \qquad (2.28)$$

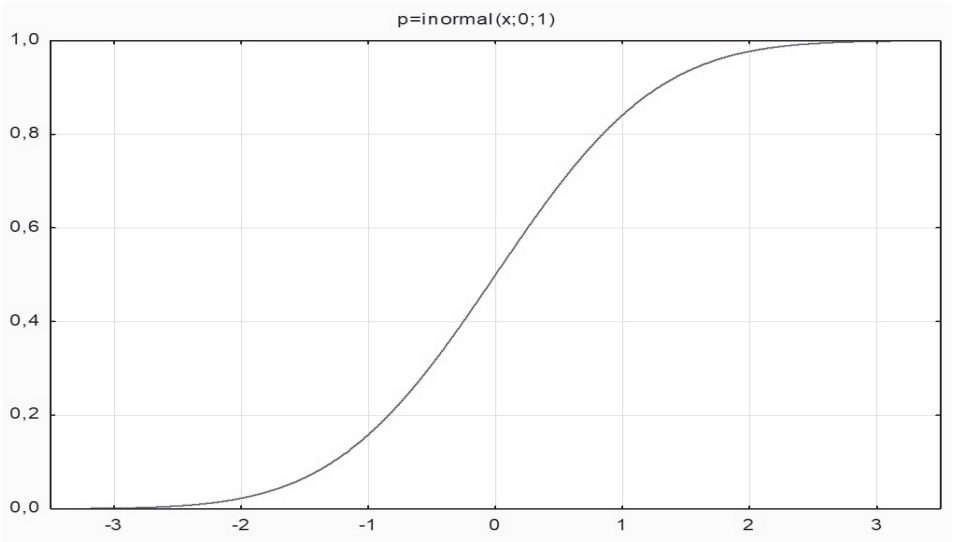


Рис. 2.4 — Функция Лапласа, сдвинутая вверх на 0,5 $(\Phi(x)+0{,}5)$

Используя формулу Бернулли и ее приближение при большом числе опытов, запишем вероятность получения в схеме Бернулли от k_1 до k_2 в виде

$$P_n(k_1 \le k \le k_2) = \sum_{k_1 \le k \le k_2} P_n(k) \approx \sum_{x_1 \le x \le x_2} \frac{1}{\sqrt{npq}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \quad (2.29)$$

В равенство (2.29) вместо $\frac{1}{\sqrt{npq}}$ подставим Δx:

$$P_n(k_1 \le k \le k_2) = \frac{1}{\sqrt{2\pi}} \sum_{x_1 \le x \le x_2} e^{-\frac{x^2}{2}} \Delta x. \quad (2.30)$$

По условию $n \to \infty$, следовательно, $\Delta x = \frac{1}{\sqrt{npq}} \to 0$. Таким образом, переходя к пределу при $\Delta x \to 0$, получим

$$P_n(k_1 \le k \le k_2) \approx \frac{1}{\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{x^2}{2}} dx = \Phi(x_2) - \Phi(x_1). \quad (2.31)$$

Функция Лапласа $\Phi(x)$ имеет следующие свойства:

1) $\Phi(-x) = -\Phi(x)$ — функция нечетная, поэтому достаточно применять ее для неотрицательных значений *x*;

2) функция $\Phi(x)$ возрастает на всей числовой оси;

3) при $x \ge 5$, $\Phi(x) \to \frac{1}{2}$ ($y = 0{,}5$ — горизонтальная асимптота при $x>0$), поэтому функция представлена для $0 \le x \le 5$ (приложение 1).

4. Если вероятность наступления события A в каждом из n независимых испытаний постоянна, отлична от нуля и единицы, а число испытаний достаточно велико, то *вероятность абсолютного отклонения относительной частоты от постоянной вероятности* в независимых испытаниях не более чем на некоторое число $\varepsilon > 0$ приближенно равна

$$P_n\left(\left|\frac{k}{n} - p\right| \le \varepsilon\right) = 2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right). \quad (2.32)$$

Доказательство. Из неравенства $\left|\frac{k}{n}-p\right| \leq \varepsilon$ получим равносильное неравенство $\left|\frac{k-np}{n}\right| \leq \varepsilon$. По свойству модуля $-\varepsilon \leq \frac{k-np}{n} \leq \varepsilon$. Умножая обе части равенства на $\sqrt{\frac{n}{pq}}$, получим

$$-\varepsilon\sqrt{\frac{n}{pq}} \leq \frac{k-np}{\sqrt{npq}} \leq \varepsilon\sqrt{\frac{n}{pq}}.$$

Обозначив $x_1 = -\varepsilon\sqrt{\frac{n}{pq}}, x_2 = \varepsilon\sqrt{\frac{n}{pq}}$ и применяя интегральную формулу Лапласа, имеем

$$P_n\left(\left|\frac{k}{n}-p\right| \leq \varepsilon\right) = \Phi(x_2) - \Phi(x_1) = \Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right) - \Phi\left(-\varepsilon\sqrt{\frac{n}{pq}}\right),$$

$$P_n\left(\left|\frac{k}{n}-p\right| \leq \varepsilon\right) = 2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right).$$

Замечание. 1. Формулы Муавра — Лапласа (2.7, 2.23) были доказаны П. Лапласом (1783 г.), который обобщил результаты Муавра (1730 г.), рассмотревшего опыт, аналогичный бросанию симметричной монеты, при $n = 2k, p = q = 0{,}5$, успех — появление герба:

а) вероятность появления k успехов,

б) вероятность $(k + \Delta)$ успехов.

2. Формула (2.20) была доказана С. Пуассоном (около 1837 г.) и часто называется «законом редких событий», так как приближенно описывает биномиальное распределение при $p \to 0$ $(p < 0{,}1)$. ■

Пример 2.7. Стрелок произвел 400 выстрелов, вероятность одного попадания 0,8. Найти вероятность того, что он попадет от 310 до 325 раз.

Решение. $\mathrm{P}_{400}(310 \leq k \leq 325) = \Phi(x_2) - \Phi(x_1)$, где

$$x_2 = \frac{k_2 - np}{\sqrt{npq}} = \frac{325-320}{\sqrt{400 \cdot 0{,}8 \cdot 0{,}2}} \approx 0{,}63;\ x_1 = \frac{k_1 - np}{\sqrt{npq}} = \frac{310-320}{\sqrt{64}} = -1{,}25.$$

$$\mathrm{P}_{400}(310 \leq k \leq 325) \approx \Phi(0{,}63) - \Phi(-1{,}25) = \Phi(0{,}63) + \Phi(1{,}25) =$$
$$= 0{,}2357 + 0{,}3944 = 0{,}6301.$$

Пример 2.8. В каждом из 10 000 независимых испытаний вероятность успеха $p = 0{,}75$. Найти вероятность того, что относительная частота появления события отклонится от постоянной вероятности по абсолютной величине не более чем на 0,01.

Решение. $n = 10000,\ p = 0{,}75, q = 1 - p = 1 - 0{,}75 = 0{,}25,\ \varepsilon = 0{,}01$, следовательно,

$$P_{10000}\left(\left|\frac{k}{n} - 0{,}75\right| \leq 0{,}01\right) = 2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right) = 2\Phi\left(0{,}01\sqrt{\frac{10000}{0{,}75 \cdot 0{,}25}}\right) =$$
$$= 2\Phi(2{,}31) = 2 \cdot 0{,}48955 = 0{,}9791.$$

Вероятность того, что абсолютное отклонение относительной частоты от постоянной вероятности в независимых испытаниях не превысит 0,01, равна 0,9791.

Пример 2.9. Вероятность появления события в каждом независимом испытании $p = 0{,}2$. Какое отклонение относительной частоты появления события от его вероятности может произойти с вероятностью 0,9128 при 5000 независимых испытаний по схеме Бернулли?

Решение. $P_{5000}\left(\left|\frac{k}{n} - p\right| \le \varepsilon\right) = 0{,}91282$ или $2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right) = 0{,}9128,$

отсюда $\Phi\left(\varepsilon\sqrt{\frac{5000}{0{,}2\cdot 0{,}8}}\right) = 0{,}4564,\ \varepsilon\sqrt{\frac{5000}{0{,}2\cdot 0{,}8}} = 1{,}71,\ \varepsilon \approx 0{,}00967.$

Пример 2.10. Сколько раз нужно бросить монету, чтобы с вероятностью 0,6 можно было ожидать, что отклонение относительной частоты появления герба от вероятности $p = 0{,}5$ окажется по абсолютной величине не больше 0,01?

Решение. По условию $2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right) = 0{,}6$, где $\varepsilon = 0{,}01; p = 0{,}5; q = 0{,}5.$

Отсюда $\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right) = 0{,}3$. Из приложения 1 имеем, что $\varepsilon\sqrt{\frac{n}{pq}} = 0{,}84$

или $0{,}01\sqrt{\frac{n}{0{,}5\cdot 0{,}5}} = 0{,}84.$ Следовательно, $n = \left(\frac{0{,}84}{0{,}01}\right)^2 \cdot 0{,}25 = 1764.$

Темы (вопросы) для самоконтроля

1. Повторные независимые испытания (схема Бернулли).
2. Схема Бернулли — постоянные условия опыта (формула Бернулли — биномиальное распределение вероятностей).
3. Схема Бернулли — постоянные условия опыта (вероятность «неудач», до появления нескольких «успехов» — отрицательное биномиальное распределение вероятностей).
4. Наивероятнейшее число успехов в схеме Бернулли.
5. Повторные независимые испытания — переменные условия опыта (производящая функция).
6. Формула Стирлинга.
7. Функция плотности распределения вероятностей стандартного нормального распределения.
8. Локальная формула Муавра — Лапласа.
9. Функция Лапласа.
10. Интегральная формула Муавра — Лапласа.
11. Приближение Пуассона.

Глава 3
Дискретные случайные величины

3.1. Закон распределения дискретной случайной величины

Случайной величиной называют такую величину, которая в результате опыта может принимать те или иные значения, причем до опыта мы не можем сказать какое именно значение она примет. Случайные величины обозначаются последними буквами латинского алфавита — *X, Y, Z*, а их возможные значения — *x, y, z*. (Более точно, случайная величина — это действительная функция, определенная на пространстве элементарных событий Ω: $X = X(\omega),\ Y = Y(\omega),\ Z = Z(\omega).$)

Случайные величины могут быть трех типов:

– дискретные,

– непрерывные,

– смешанные (дискретно-непрерывные).

Дискретная случайная величина может принимать конечное или бесконечное счетное число значений. Например, подбрасываем монету 4 раза. Случайная величина *X* — число появлений герба: 0, 1, 2, 3, 4. Примерами дискретной случайной величины может служить: число детей в семье, оценка на экзамене, число поломок автомобиля в течение года, количество вкладов в банке и т. п.

Непрерывная случайная величина, в отличие от дискретной случайной величины, принимает бесконечное несчетное число значений. Например, мишень имеет форму круга радиуса *R*. По этой мишени произвели выстрел с обязательным попаданием. Обозначим через *Y* расстояние от центра до точки попадания в мишень, $Y \in [0; R]$. *Y* — непрерывная случайная величина, так как она принимает бесконечное несчетное число значений.

Пусть *X* — дискретная случайная величина, которая принимает значения $x_1, x_2, \dots, x_n$ с вероятностями этих значений p_i, где $i = 1, 2, \dots, n$. Тогда можно говорить о вероятности того, что случайная величина *X* приняла значение x_i: $p_i = P(X = x_i)$. Значения x_i и соответствующие вероятности p_i представляют в виде таблицы.

x_i	x_1	x_2	x_3	...	x_n
p_i	p_1	p_2	p_3	...	p_n

Эта таблица является одной из форм задания дискретной случайной величины (*дискретного распределения*). Обычно значения случайной величины располагаются в возрастающем порядке. Так как события $X = x_1, X = x_2, \dots, X = x_n$ образуют полную группу, то сумма вероятностей значений равна 1:

$$\sum_{i=1}^{n} p_i = p_1 + p_2 + \cdots + p_n = 1. \qquad (3.1)$$

Пример 3.1. Монета бросается 5 раз. Составить закон распределения дискретной случайной величины *X* — числа появлений герба и представить его в виде таблицы.

Решение. Дискретная случайная величина X может принимать значения: 0, 1, 2, 3, 4, 5. Вероятность появления герба в одном испытании $p = 0{,}5$, непоявления герба $q = 0{,}5$, $n = 5$. Таким образом, выполняются условия применения формулы Бернулли.

Имеем:

$$P_5(X=0) = C_5^0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{5-0} = \frac{1}{32};$$

$$P_5(X=1) = C_5^1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^{5-1} = 5\left(\frac{1}{2}\right)^5 = \frac{5}{32};$$

$$P_5(X=2) = C_5^2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{5-3} = 10\left(\frac{1}{2}\right)^5 = \frac{10}{32};$$

$$P_5(X=3) = C_5^3\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^{5-3} = 10\left(\frac{1}{2}\right)^5 = \frac{10}{32};$$

$$P_5(X=4) = C_5^4\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^{5-4} = 5\left(\frac{1}{2}\right)^5 = \frac{5}{32};$$

$$P_5(X=5) = C_5^5\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^{5-5} = 1\left(\frac{1}{2}\right)^5 = \frac{1}{32}.$$

Полученные результаты представим в виде таблицы распределения.

x_i	0	1	2	3	4	5
p_i	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$

Графически дискретная случайная величина может быть представлена в виде многоугольника распределения — фигуры, состоящей из точек (x_i, p_i), соединенных отрезками (рис. 3.1).

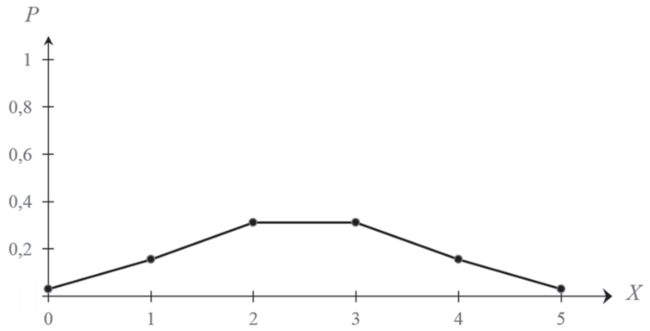


Рис. 3.1 — Многоугольник распределения

Операции сложения и умножения.

1. Суммой двух случайных величин X и Y называется случайная величина Z, которая получается в результате сложения каждого значения случайной величины X с каждым значением случайной величины Y, соответствующие вероятности перемножаются.

2. Произведением двух случайных величин X и Y называется случайная величина U, которая получается в результате перемножения каждого значения случайной величины X с каждым значением случайной величины Y, соответствующие вероятности перемножаются.

Пример 3.2. Случайные величины X и Y заданы таблицами.

x_i	0	1	2	3
p_i	0,1	0,4	0,3	0,2

y_i	–1	0	1
p_i	0,2	0,3	0,5

Найти: 1) $X + C$, где $C = 2$; 2) $X + Y$; 3) XY.

Решение. 1) $Z = X + C$, $C = 2$.

z_i	0+2	1+2	2+2	3+2
p_i	0,1	0,4	0,3	0,2

или

z_i	2	3	4	5
p_i	0,1	0,4	0,3	0,2

2) $U = X + Y$.

u_i	0–1	0+0	0+1	1–1	1+0	1+1	2–1	2+0	2+1	3–1	3+0	3+1
p_i	0,02	0,03	0,05	0,08	0,12	0,2	0,06	0,09	0,15	0,04	0,06	0,1

или

u_i	– 1	0	1	2	3	4
p_i	0,02	0,11	0,23	0,33	0,21	0,1

Одинаковые значения случайной величины Z необходимо записать один раз, предварительно сложив вероятности одинаковых значений.

3) $V = XY$.

v_i	0	–1	–2	–3	0	0	0	0	0	1	2	3
p_i	0,02	0,08	0,06	0,04	0,03	0,12	0,09	0,06	0,05	0,2	0,15	0,1

или

v_i	– 3	–2	–1	0	1	2	3
p_i	0,04	0,06	0,08	0,37	0,2	0,15	0,1

3.2. Числовые характеристики дискретных случайных величин

На практике нет необходимости характеризовать случайную величину полностью. Обычно достаточно указать только отдельные числовые параметры распределения. Такие числовые параметры принято называть числовыми характеристиками распределения. Прежде всего, это характеристики положения: математическое ожидание, медиана, мода; характеристики рассеяния: дисперсия, среднее квадратическое отклонение.

Математическим ожиданием $M(X) = m_x$ дискретной случайной величины *X* называется среднее значение случайной величины, представляет сумму произведений значений дискретной случайной величины на соответствующие им вероятности:

$$M(X) = \frac{x_1p_1+x_2p_2+ \dots +x_np_n}{p_1+p_2+ \dots +p_n} = \frac{\sum_{i=1}^{n} x_ip_i}{\sum_{i=1}^{n} p_i} = \frac{\sum_{i=1}^{n} x_ip_i}{1} = \sum_{i=1}^{n} x_ip_i. \qquad (3.2)$$

Мода $M_o(X)$ распределения — это значение случайной величины, имеющее наибольшую вероятность по сравнению с вероятностями других значений.

Медиана $M_e(X)$ — это значение случайной величины, которое делит значения случайной величины на две части таким образом, что вероятность попадания в одну из них равна 0,5.

Пусть случайная величина задана таблицей ($\sum p_i = 1$).

x_i	x_1	x_2	x_3	...	x_n
p_i	p_1	p_2	p_3	...	p_n

Рассмотрим свойства математического ожидания.

1) Математическое ожидание постоянной величины C равно самой постоянной

$$M(C) = C, \text{где } C = const.$$

Доказательство. $M(X) = Cp_1 + Cp_2 + \cdots + Cp_n = C(\sum_i p_i) = C.$

2) Математическое ожидание произведения постоянной *C* и случайной величины *X* равно произведению постоянной на математическое ожидание случайной величины

$$M(CX) = CM(X).$$

Доказательство. $M(CX) = Cx_1p_1 + Cx_2p_2 + \cdots + Cx_np_n =$
$= C(x_1p_1 + x_2p_2 + \cdots + x_np_n) = CM(X).$

Рассмотрим две случайные величины X и Y.

x_i	x_1	x_2	x_3	...	x_i	...	x_m
p_i	p_1	p_2	p_3	...	p_i	...	p_m

y_j	y_1	y_2	y_3	...	y_j	...	y_n
p_j	p_1	p_2	p_3	...	p_j	...	p_n

3) Математическое ожидание суммы (или разности) случайных величин *X* и *Y* равно сумме (разности) их математических ожиданий:

$$M(X \pm Y) = M(X) \pm M(Y).$$

Доказательство. 1. Рассмотрим произведение двух сумм. Пусть

$$(a_1 + a_2 + \cdots + a_i + \cdots)(b_1 + b_2 + \cdots + b_j + \cdots) = (\sum_i a_i)(\sum_j b_j),$$

тогда

$$(\sum_i a_i)(\sum_j b_j) = \sum_{i,j}(a_ib_j) = \sum_i \sum_j (a_ib_j)$$

— распределительный закон.

По определению суммы случайных величин:

$P(x_i + y_j) = p_{ij} = p_i p_j$. Используя распределительный закон, найдем математическое ожидание $\mathrm{M}(X \pm Y)$.

$$M(X \pm Y) = \sum_{i,j}\big(x_i \pm y_j\big)p_{ij} = \sum_i \sum_j \big(x_i \pm y_j\big)p_{ij} =$$
$$= \sum_i \sum_j (x_i \pm y_j)p_i\, p_j = \sum_j (\sum_i x_i\, p_i)p_j \pm \sum_i (\sum_j y_j\, p_j)p_i =$$
$$= \sum_j p_j \sum_i x_i p_i \pm \sum_i p_i \sum_j y_j p_j = \sum_i x_i p_i \pm \sum_j y_j p_j = M(X) \pm M(Y).$$

4) Если случайные величины X и Y, независимы, то

$$M(XY) = M(X)M(Y).$$

Доказательство. По условию случайные величины X и Y независимы, следовательно, $p_{ij} = P(x_i y_j) = P(x_i)P(y_j) = p_i p_j$.

$$M(XY) = \sum_i \sum_j (x_i\, y_j p_{ij}) = \sum_i \sum_j (x_i\, y_j p_i p_j) =$$
$$= (\sum_i x_i p_i)(\sum_j y_j p_j) = M(X)M(Y).$$

Пример 3.3. Случайные величины X и Y заданы законами распределения.

x_i	– 1	1
p_i	0,5	0,5

y_j	– 100	100
p_j	0,5	0,5

Найти математические ожидания случайных величин X и Y.

Решение.

$M(X) = -1 \cdot 0{,}5 + 1 \cdot 0{,}5 = 0,\ M(Y) = -100 \cdot 0{,}5 + 100 \cdot 0{,}5 = 0.$

Хотя математические ожидания случайных величин равны, однако случайные величины X и Y явно различны, поэтому для характеристики случайной величины одного математического ожидания недостаточно и необходимо ввести другие характеристики, одна из них дисперсия.

Дисперсия дискретной случайной величины служит для характеристики рассеяния значений случайной величины относительно ее математического ожидания. Она является более полной оценкой дискретной случайной величины X.

Дисперсией дискретной случайной величины X называется математическое ожидание квадрата отклонения случайной величины X от ее математического ожидания:

$$D(X) = M\big(X - M(X)\big)^2 = \sum_{i=1}^{n}\big(x_i - M(X)\big)^2 p_i. \qquad (3.3)$$

Свойства дисперсии.

1) Дисперсия постоянной величины равна нулю:

$$D(\mathrm{C}) = 0. \qquad (3.4)$$

Доказательство. $D(C) = \mathrm{M}(C - M(C))^2 = M(C - C)^2 = 0$, где $C = const$.

2) Константа выносится за знак дисперсии в квадрате:

$$D(CX) = C^2 D(X). \qquad (3.5)$$

Доказательство. $D(CX) = M\big(CX - M(CX)\big)^2 =$

$$= M(C(X - M(X))^2 = C^2 M\big(X - M(X)\big)^2 = C^2 D(X).$$

3) Дисперсия дискретной случайной величины X равна разности между математическим ожиданием квадрата случайной величины X и квадратом ее математического ожидания:

$$D(X) = M(X^2) - \left(M(X)\right)^2, \tag{3.6}$$

где $M(X^2) = x_1^2 p_1 + x_2^2 p_2 + \cdots + x_n^2 p_n$.

Доказательство. $D(X) = M(X - M(X))^2 =$

$$= M\left(X^2 - 2(X)M(X) + \left(M(X)\right)^2\right) =$$

$$= M(X^2) - 2M(X)M(X) + \left(M(X)\right)^2 = M(X^2) - \left(M(X)\right)^2.$$

4) Если случайные величины X и Y независимы, то дисперсия суммы или разности случайных величин равна сумме дисперсий этих величин:

$$D\left(X \pm Y\right) = D(X) + D(Y). \tag{3.7}$$

Доказательство. $D\left(X \pm Y\right) = M((X \pm Y)^2) - \left(M(X \pm Y)\right)^2 =$

$$= M(X^2) \pm 2M(X)M(Y) + M(Y^2) - (M^2(X) \pm 2M(X)M(Y) + M^2(Y)) =$$

$$= (M(X^2) - M^2(X)) + (M(Y^2) - (M^2(Y)) = D(X) + D(Y).$$

5) Дисперсия суммы постоянной величины и случайной равна дисперсии случайной величины:

$$D(C + X) = D(X). \tag{3.8}$$

Доказательство. $D(C + X) = D(C) + D(X) = 0 + D(X) = D(X)$.

6) Для любых случайных величин X и Y,

$$D(X \pm Y) = D(X) + D(Y) \pm 2cov(X, Y), \tag{3.9}$$

где $cov(X, Y) = M((X - \mathrm{M(X)})(Y - \mathrm{M}(Y)))$ — ковариация случайных величин X и Y.

7) Для независимых случайных величин X и Y

$$D(X \ Y) = D(X)D(Y) + m_x^2 D(Y) + m_y^2 D(X).$$

Дисперсия характеризует средний квадрат отклонения дискретной случайной величины X от математического ожидания, поэтому на практике часто используют в качестве характеристики разброса *среднее квадратическое отклонение,* которое имеет ту же размерность, что и случайная величина X.

$$\sigma(X) = \sqrt{D(X)}. \tag{3.10}$$

Пример 3.4. Найти дисперсии рассмотренных в примере 3.3 случайных величин X и Y.

Решение. $D(X) = M(X^2) - (M(X))^2 = 1 - 0^2 = 1$, $\sigma(X) = \sqrt{D(X)} = 1$, так как $M(X^2) = (-1)^2 \cdot 0{,}5 + (1)^2 \cdot 0{,}5 = 1$;

$$D(Y) = M(Y^2) - \left(M(Y)\right)^2 = 10\,000 - 0^2 = 10\,000,$$

$$\sigma(Y) = \sqrt{D(\mathrm{Y})} = 100,$$

так как $M(Y^2) = (-100)^2 \cdot 0{,}5 + (100)^2 \cdot 0{,}5 = 10\,000$.

Таким образом, вычисленные значения дисперсий и средних квадратических отклонений указывают на то, что несмотря на равенство математических ожиданий случайных величин X и Y — они различны.

Пример 3.5. Дискретная случайная величина задана таблицей распределения.

x_i	–1	0	1	2
p_i	0,1	0,2	0,1	0,6

Найти математическое ожидание, дисперсию и среднее квадратическое отклонение случайной величины *X*.

Решение. $\mathrm{M(X)} = (-1)\cdot 0{,}1 + 0\cdot 0{,}2 + 1\cdot 0{,}1 + 2\cdot 0{,}6 = 1{,}2;$

$$\mathrm{D(X)} = \mathrm{M}\big(x - M(X)\big)^2 = (-1-1{,}2)^2 0{,}1 + (0-1{,}2)^2 0{,}2 + (1-1{,}2)^2 0{,}1 + \\ +(2-1{,}2)^2 0{,}6 = 1{,}16,\ \sigma(X) = 1{,}077.$$

3.3. Законы распределения дискретных случайных величин

1. *Закон распределения Бернулли*. Пусть выполняются условия схемы Бернулли: некоторый опыт повторяется в неизменных условиях, причем событие A в каждом испытании может либо наступить (успех), либо не наступить, $P(A) = p$ — вероятность успеха, $P(\bar{A}) = 1 - p = q$ — вероятность неудачи. Рассмотрим *индикаторную случайную величину*

$$X = I_A(\omega) = \begin{cases} 1, \text{при } \omega \in A, \\ 0, \text{при } \omega \notin A. \end{cases} \qquad (3.11)$$

Случайная величина *X* (индикаторная случайная величина), распределенная по закону Бернулли, принимает значения 1 — успех или 0 — неудача с вероятностями p и q соответственно ($p + q = 1$).

x_i	0	1
p_i	q	p

$$\mathrm{M(X)} = p; \quad D(\mathrm{X}) = pq. \qquad (3.12)$$

Доказательство. $M(x) = 0\cdot q + 1\cdot p = p,$

$$M(X^2) = 0^2\cdot q + 1^2\cdot p = p,$$

$$D(X) = M(X^2) - \big(M(X)\big)^2 = p - p^2 = p(1-p) = pq.$$

2. *Биномиальный закон распределения*. Случайная величина *X* принимает значения: 0, 1, 2, 3, 4, 5, …, n с вероятностями, определяемыми по формуле Бернулли (2.1), где p — вероятность появления события A в одном испытании.

Для биномиального закона вводят обозначение

$$Bin(n; p).$$

x_i	0	1	…	k	…	n
p_i	$C_n^0 p^0 q^n$	$C_n^1 p^1 q^{n-1}$	…	$C_n^k p^k q^{n-k}$	…	$C_n^n p^n q^0$

$$M(X) = np;\ D(X) = npq, \sigma(X) = \sqrt{npq}\,. \qquad (3.13)$$

Доказательство. Рассмотрим случайную величину *X* как сумму *n* случайных величин, распределенных по закону Бернулли: $X = \sum_{i=1}^{n} X_i$. Тогда по свойству математического ожидания

$$M(X) = M(\textstyle\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} M(X_i) = p + p + \cdots + p = np.$$

По свойству дисперсии

$$D(X) = D\,(\textstyle\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} D(X_i) = pq + pq + \cdots + pq = npq.$$

Следовательно, $\sigma(X) = \sqrt{npq}$.

При $p = 0{,}5$ биномиальный закон $Bin(n; p)$ приближенно описывается нормальным законом[6] уже для $n = 10$ (рис. 3.2), а также непосредственно связан с треугольником Паскаля (п. 1.3, рис. 1.6) и доской Гальтона (рис. 5.5).

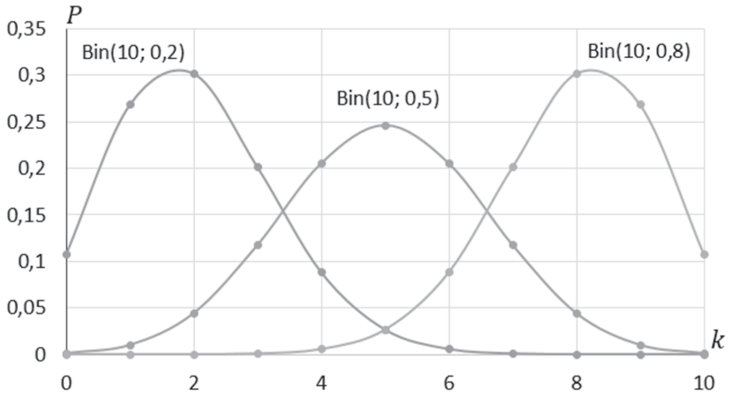


Рис. 3.2 — Биномиальный закон распределения

3. *Закон распределения Пуассона*. Случайная величина *X* принимает бесконечное счетное число значений: $0, 1, 2, 3, 4, 5, \ldots, k, \ldots$ с вероятностью, определяющейся по формуле Пуассона:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \tag{3.14}$$

где $\lambda > 0$ — параметр распределения Пуассона.

x_i	0	1	2	…	k	…
p_i	$e^{-\lambda}$	$\lambda e^{-\lambda}$	$\frac{\lambda^2}{2!} e^{-\lambda}$	…	$\frac{\lambda^k}{k!} e^{-\lambda}$	…

$$\lambda = np, M(X) = \lambda;\ D(X) = \lambda; \sigma(X) = \sqrt{\lambda}. \tag{3.15}$$

Доказательство. Воспользуемся разложением в ряд экспоненциальной функции: $e^t = \sum_{k=0}^{\infty} \frac{t^k}{k!}$.

$$M(X) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda}\lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = e^{-\lambda}\lambda\, e^{\lambda} = \lambda.$$

Для нахождения дисперсии найдем $M(X^2 - X) = M\big(X(X-1)\big)$.

$$M\big(X(X-1)\big) = \sum_{k=2}^{\infty} k(k-1) \frac{\lambda^k}{k!} e^{-\lambda} = \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{(k-2)}}{(k-2)!} e^{-\lambda} = \lambda^2.$$

Следовательно, $M(X^2) - M(X) = \lambda^2$, отсюда $M(X^2) = \lambda^2 + \lambda$.

Имеем $D(X) = M(X^2) - M^2(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$, $\sigma(X) = \sqrt{\lambda}$.

Только для распределения Пуассона всегда $M(X) = D(X) = \lambda$.

Покажем, что сумма вероятностей распределения Пуассона равна единице:

$$\sum P(X = k) = e^{-\lambda} \sum_0^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \left(1 + \frac{\lambda}{1} + \frac{\lambda^2}{2!} + \cdots + \frac{\lambda^k}{k!} + \cdots\right) = e^{-\lambda} e^{\lambda} = 1.$$

[6] см. разделы 2.2, 5.3, 8.3.

4. *Геометрический закон распределения*. Пусть $P(A)=p$ — вероятность наступления события A в каждом опыте, соответственно, $q = 1-p$ — вероятность ненаступления события A (схема Бернулли). Вероятность появления m — неудач до первого наступления события A определяется по формуле

$$P(X = m) = q^m p = pq^m. \quad (3.16)$$

Случайная величина X — распределенная по геометрическому закону, принимает значения: 0, 1, 2, …, m, … с вероятностью, определяемой по формуле (3.16).

x_i	0	1	2	...	m	…
p_i	p	pq	pq^2	...	pq^m	…

$$M(X) = \frac{q}{p};\ D(X) = \frac{q}{p^2}, \sigma(X) = \sqrt{\frac{q}{p^2}} = \frac{\sqrt{q}}{p}. \quad (3.17)$$

Доказательство. Воспользуемся формулой суммы бесконечной убывающей геометрической прогрессии с первым членом b_1 при $|q| < 1, m \to \infty$:

$$S = b_1 + b_1 q + b_1 q^2 + \cdots + b_1 q^m + \cdots = \frac{b_1}{1-q}. \quad (3.18)$$

Если считать b_1 постоянной величиной, а q — непрерывной переменной, то имеет смысл говорить о функции $S = S(q)$ и ее производных:

$$S'_q = \frac{b_1}{(1-q)^2};\ S''_{q^2} = \frac{2b_1}{(1-q)^3}. \quad (3.19)$$

Заметим, что $(q^m)'_q = mq^{m-1}$, $(q^m)''_{q^2} = m(m-1)q^{m-2}$. Тогда, так как производная суммы функций равна сумме их производных и наоборот, сумма производных функций равна производной их суммы

$$\sum_i {u_i}' = (\sum_i u_i)',$$

то

$$\sum_{m=0}^{\infty} mq^{m-1} = (\sum_{m=0}^{\infty} q^m)'_q = (\frac{1}{1-q})'_q = \frac{1}{(1-q)^2} = \frac{1}{p^2}, \quad (3.20)$$

$$\sum_{m=0}^{\infty} m(m-1)q^{m-2} = (\sum_{m=0}^{\infty} q^m)''_{q^2} = \left(\frac{1}{1-q}\right)''_{q^2} = \frac{2}{(1-q)^3} = \frac{2}{p^3}. \quad (3.21)$$

С учетом формул (3.18)–(3.21) найдем математическое ожидание случайной величины X и ее квадрата.

$$M(X) = \sum_{m=0}^{\infty} mpq^m = pq \sum_{m=0}^{\infty} mq^{m-1} = \frac{pq}{p^2} = \frac{q}{p}$$

$$M(X^2) = \sum_{m=0}^{\infty} m^2 pq^m = \sum_{m=0}^{\infty} m(m-1)\, pq^m + \sum_{m=0}^{\infty} m\, pq^m =$$

$$= pq^2 \sum_{m=0}^{\infty} m(m-1)\, pq^{m-2} + pq \sum_{m=0}^{\infty} m\, pq^{m-1} =$$

$$= pq^2 \frac{2}{p^3} + pq \frac{1}{p^2} = 2\left(\frac{q}{p}\right)^2 + \frac{q}{p}.$$

По свойству дисперсии

$$D(X) = 2\left(\frac{q}{p}\right)^2 + \frac{q}{p} - \left(\frac{q}{p}\right)^2 = \frac{q^2+pq}{p^2} = \frac{(p+q)q}{p^2} = \frac{q}{p^2}.$$

5. *Геометрический закон распределения, сдвинутый на единицу*.

Вероятность наступления события в m-ом опыте определяется по формуле

$$P(X = m) = pq^{m-1}. \quad (3.22)$$

Случайная величина X — распределенная по геометрическому закону, сдвинутому на 1 (*геометрический закон* +1), означает число опытов до первого появления события A и принимает значения: $1, 2, \ldots, m, \ldots$ с вероятностью, определяемой по формуле (3.22).

x_i	1	2	3	…	m	…
p_i	p	pq	pq^2	…	pq^{m-1}	…

$$M(X) = \frac{1}{p}; D(X) = \frac{q}{p^2}; \sigma(X) = \sqrt{\frac{q}{p^2}} = \frac{\sqrt{q}}{p}. \quad (3.23)$$

Доказательство. По определению $X = X_1 + 1$, где X_1 — случайная величина, подчиняющаяся геометрическому закону. Согласно свойствам математического ожидания и дисперсии:

$$M(X) = M(X_1 + 1) = M(X_1) + 1 = \frac{q}{p} + 1 = \frac{q+p}{p} = \frac{1}{p}.$$

$$D(X) = D(X_1 + 1) = D(X_1) = \frac{q}{p^2}.$$

Пример 3.6. Из орудия производили выстрелы по цели до первого попадания. Вероятность попадания в цель 0,6. Найти вероятность того, что попадание произойдет при первом, втором, третьем, …, k-ом выстреле.

Решение.

$\mathrm{P}(X = 1) = p = 0{,}6;$

$\mathrm{P}(X = 2) = qp = 0{,}6\ 0{,}4 = 0{,}24;$

$\mathrm{P}(X = 3) = q^2p = pq^2 = 0{,}6 \cdot 0{,}4 \cdot 0{,}4 = 0{,}096;$

……

$\mathrm{P}(X = k) = q^{k-1}p = pq^{k-1} = 0{,}6 \cdot 0{,}4^{k-1}.$

x_i	1	2	3	…	k	…
p_i	0,6	0,24	0,096	…	$0{,}6 \cdot 0{,}4^{k-1}$	…

Сумма вероятностей, как и для других законов, равна единице.

$S = \frac{p}{1-q} = \frac{0{,}6}{1-0{,}4} = 1$ — согласно формуле, суммы членов бесконечной геометрической прогрессии со знаменателем q меньше единицы.

6. *Отрицательное биномиальное распределение*. Если производится ряд независимых опытов, в каждом из которых событие A появляется с вероятностью p до получения k успехов (k = 1, 2, 3, …), то при этом вероятность X=m неудач можно определить по формуле

$$P_{m+k}(k, m) = C_{m+k-1}^{k-1} p^k q^m = C_{m+k-1}^{m} p^k q^m \ \ (m = 0, 1, 2, \ldots). \quad (3.24)$$

Формула для биномиальных коэффициентов для всех действительных x и целых r представляется в виде

$$C_x^r := \binom{x}{r} = \frac{x(x-1)\ldots(x-(r-1))}{r!} = \binom{x}{x-r}, \text{ отсюда}$$

$$\binom{-x}{r} = \frac{(-x)(-x-1)\ldots\left(-x-(r-1)\right)}{r!} = (-1)^r \frac{(x)(x+1)\ldots\left(x+(r-1)\right)}{r!} =$$

$$= (-1)^r \frac{(x+(r-1))(x+(r-2))\ldots(x-1)x}{r!} = (-1)^r \binom{x+r-1}{r}.$$

Вероятность появления m — неудач, до получения k — успехов совпадает с m −ым членом разложения выражения $p^k(1-q)^{-k}$ по степеням p, т. е. отрицательного бинома (отсюда и название):

$$1 = p^k(1-q)^{-k} = p^k \sum_{m=0}^{\infty}\binom{-k}{m}(-q)^m = \sum_{m=0}^{\infty}(-1)^m\binom{m+k-1}{m}(-q)^m = \\ = p^k \sum_{m=0}^{\infty}\binom{m+k-1}{m}q^k = p^k \sum_{m=0}^{\infty} C_{m+k-1}^{m}\, q^m. \qquad (3.25)$$

Коэффициенты C_{m+k-1}^{m} при разных значениях k появляются в виде рядов, параллельных боковой стороне треугольника Паскаля (рис. 1.6).

k=1: 1, 1, 1 (геометрическое распределение),

k=2: 1, 2, 3, 4, 5, 6…,

k=3: 1, 3, 6, 10, 15, 21…,

k=4: 1, 4, 10, 20, 35, 56….

Данное распределение определяется двумя параметрами «k» и «p».

$$M(X) = \frac{kq}{p};\ D(X) = \frac{kq}{p^2};\ \sigma(X) = \sqrt{\frac{kq}{p^2}} = \frac{\sqrt{kq}}{p}. \qquad (3.26)$$

Доказательство. Случайную величину *X*, подчиняющуюся отрицательному биномиальному распределению, можно представить как сумму *k* случайных величин, подчиняющихся геометрическому закону:

$$X = X_1 + X_2 + \cdots + X_i + \cdots + X_k.$$

Следовательно,

$$M(X) = M(X_1 + X_2 + \cdots + X_k) = M(X_1) + M(X_2) + \cdots + M(X_k) = k\frac{q}{p}.$$

$$D(X) = D(X_1 + X_2 + \cdots + X_k) = D(X_1) + D(X_2) + \cdots + D(X_k) = k\frac{q}{p^2}.$$

7. *Гипергеометрический закон распределения*. Пусть в урне *N*-шаров, из них *M* белых, а остальные (*N* – *M*) черные. Найдем вероятность того, что из извлеченных *n* шаров *m* белых и (*n* – *m*) черных.

$$N_{\text{шаров}} = M_{\text{белых шаров}} + (N–M)_{\text{ черных шаров}},$$
$$n_{\text{шаров}} = m_{\text{белых шаров}} + (n–m)_{\text{ черных шаров}};$$

C_M^m — число способов выбора *m* белых шаров из *M*;

C_{N-M}^{n-m} — число способов выбора $(n-m)$ черных шаров из (*N* – *M*);

C_N^n — общее число способов выбора n шаров из N.

Всего возможных наборов из *m* белых и (*n* – *m*) черных шаров, по правилу произведения, равно $C_M^m C_{N-M}^{n-m}$. Отсюда, по формуле классического определения вероятности,

$$P(A) = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n}. \qquad (3.27)$$

Ограничения на параметры:

$M \le N, m \le n;\ m = m_0, m_0+1, m_0+2, \ldots, \min(M,n),$

где $m_0 = max\{0, n-(N-M)\}$. Случайная величина $X = m$, распределенная по гипергеометрическому закону распределения, имеет следующий вид (при $m = 0,\ 1,\ 2,\ 3, \ldots,\ M$).

x_i	0	1	…	m	…	M
p_i	$\frac{C_M^0 C_{N-M}^n}{C_N^n}$	$\frac{C_M^1 C_{N-M}^{n-1}}{C_N^n}$	…	$\frac{C_M^m C_{N-M}^{n-m}}{C_N^n}$	…	$\frac{C_M^M C_{N-M}^{n-M}}{C_N^n}$

Гипергеометрический закон определяется тремя параметрами N, M, n. При $n < 0{,}1N$ этот закон стремится к биномиальному. Действительно,

$$P(X=m)=\frac{C_M^m C_{N-M}^{n-m}}{C_N^n}=\frac{\frac{M!}{m!(M-m)!}\frac{(N-M)!}{(n-m)!(N-M-(n-m))!}}{\frac{N!}{n!(N-n)!}}=\frac{n!}{m!\,(n-m)!}\times$$

$$\times\frac{(M-m+1)(M-m+2)\dots M(N-M-(n-m)+1)\dots(N-M)}{(N-n+1)(N-n+2)\dots N}.$$

При $N\to+\infty, M\to+\infty$ и конечных значениях чисел a и b

$$\frac{M+a}{N+b}\to\frac{M}{N}=p,\frac{(N-M)-a}{N-b}\to\left(1-\frac{M}{N}\right)=1-p=q.$$

Следовательно,

$P(X=m)\to C_n^m\left(\frac{M}{N}\right)^m\left(1-\frac{M}{N}\right)^{n-m}=C_n^m p^m q^{n-m}$, что и требовалось доказать.

Рассмотрим индикаторную случайную величину X_i.

x_i	0	1
p_i	$1-\frac{M}{N}$	$\frac{M}{N}$

Следовательно,

$$M(X_i)=\frac{M}{N}\,, M\left(X_i^2\right)=\frac{M}{N}\,, D(X_i)=\frac{M}{N}-\left(\frac{M}{N}\right)^2=\frac{M}{N}\left(1-\frac{M}{N}\right),$$

$$cov\left(X_i,X_j\right)=M\left(X_iX_j\right)-M(X_i)\mathrm{M}\left(X_j\right)=\frac{M}{N}\frac{M-1}{N-1}-\frac{M}{N}\frac{M}{N}=-\frac{M(N-M)}{N^2(N-1)}.$$

Математическое ожидание гипергеометрического закона распределения получается как математическое ожидание суммы n индикаторных случайных величин:

$$M(X)=M(X_1+X_2+\dots+X_n)=M(X_1)+M(X_2)+\dots+M(X_n),$$

$$M(X)=n\frac{M}{N}. \qquad (3.28)$$

Дисперсия гипергеометрического закона распределения получается как дисперсия суммы n индикаторных случайных величин:

$$D(X)=D(X_1+X_2+\dots+X_n)=\sum_i D(X_i)+2\sum_{i\neq j} cov(X_iX_j).$$

Имеем

$$D(X)=n\frac{M}{N}\left(1-\frac{M}{N}\right)-\frac{n(n-1)}{2}2\frac{M(N-M)}{N^2(N-1)}=$$

$$=n\frac{M}{N}\left(1-\frac{M}{N}\right)-n\frac{M}{N}\left(1-\frac{M}{N}\right)\frac{n-1}{N-1}.$$

Следовательно,

$$D(X)=n\frac{M}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}. \qquad (3.29)$$

С учетом ранее введенных обозначений

$$\left(\frac{M}{N}=p,\left(1-\frac{M}{N}\right)=1-p=q\right),$$

перепишем формулу (3.29) в виде

$$D(X)=npq\frac{N-n}{N-1}.$$

То есть дисперсия гипергеометрического распределения меньше дисперсии соответствующего биномиального, при $N \to +\infty$ дисперсии совпадают, так как $\frac{N-n}{N-1} \to 1$.

Для того чтобы показать, что сумма вероятностей, как и для других законов, равна единице, рассмотрим равенство

$$(1+Z)^N = (1+Z)^M (1+Z)^{N-M}.$$

Приравняем коэффициенты, полученные по формуле бинома Ньютона, в левой и правой части при Z^n:

$$C_N^n = \sum_{k=0}^{n} C_M^m C_{N-M}^{n-m}.$$

Получим известную формулу *свертки Вандермонда*, поэтому

$$\sum_{m=0}^{M} \frac{C_M^m C_{N-M}^{n-m}}{C_N^n} = 1,$$

что и требовалось доказать.

Замечание. 1. В теории вероятностей различают две основные схемы: выбора элементов с возвращением каждый раз обратно и выбора без возвращения, которые описываются соответственно биномиальным и гипергеометрическим законами. Если элементы возвращаются обратно, то математическое ожидание для биномиального и гипергеометрического закона совпадают. Дисперсия, за счет отрицательной корреляции, при выборе без возвращения меньше, чем при выборе с возвращением.

2. Геометрический закон описывает схему повторения опытов (в каждом из которых может наступить или не наступить событие A: $P(A) = p, q = 1-p$), до первого появления события A, то есть фактически это отрицательное биномиальное распределение при $m = 1$.

3. Закон распределения Пуассона обычно используют при изучении событий, вероятность которых близка к нулю (маловероятных событий), его иногда называют законом редких событий.

4. Известно также распределение Маркова — Пойа, включающее в себя биномиальное и гипергеометрическое распределения как частные случаи. После извлечения шара из урны (бесконечной вместимости), содержащей черные и белые шары, шар возвращается обратно и добавляется шар такого же цвета. Даже если первоначально было одинаковое число шаров разного цвета, то незначительное преимущество на начальном этапе приведет к доминированию в будущем (в экономике есть примеры, демонстрирующие «зависимости от пути»).

5. В литературе встречаются следующие обозначения для рассмотренных выше законов распределения: $Bern(p)$ — Бернулли; $Bin\,(n,p)$ — биномиального; $Mult(n;p)$ — полиномиального (мультиномиального), обобщающего биномиальное на n-мерный случай (формулы (1.12)–(2.5)), где $n = k_1 + \cdots + k_r$, $p = (p_1, \ldots, p_r)$, $\sum p_i = 1$, $i = \overline{1,r}$; $Poisson(\lambda)$ — Пуассона; $Geo(p)$ — геометрического; $NBin\,(k,p)$ — отрицательного биномиального; $Hyp(N,m,n)$ — гипергеометрического. ■

3.4. Независимые одинаково распределенные случайные величины

Случайные величины $X_1, X_2, \dots, X_n$ называются *одинаково распределенными*, если они имеют один и тот же закон распределения. Поэтому у них совпадают числовые характеристики: математические ожидания, дисперсии, средние квадратические отклонения.

Пусть $X_1, X_2, \dots, X_n$ одинаково распределенные, независимые дискретные случайные величины, тогда:

$$M(X_1) = M(X_2) = \dots = MX_n) = M(X),$$
$$D(X_1) = D(X_2) = \dots = D(X_n) = D(X).$$

Рассмотрим характеристики средней арифметической величины:

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \dots + X_n}{n}.$$

1) $M(\overline{X}) = M\left(\frac{X_1+X_2+X_3+\dots+X_n}{n}\right) = \frac{1}{n}\left(nM(X)\right) = M(X).$ (3.30)

2) $D(\overline{X}) = D(\frac{X_1+X_2+X_3+\dots+X_n}{n}) = \frac{1}{n^2}\left(D(X_1) + D(X_2) + \dots + D(X_n)\right) = \frac{D(X)}{n}.$ (3.31)

3) $\sigma(\bar{X}) = \sqrt{\frac{D(X)}{n}} = \frac{\sigma(X)}{\sqrt{n}}.$ (3.32)

$\sigma(\bar{X})$ — среднее квадратическое отклонение случайной величины $\overline{X}$ (*стандартное отклонение*).

Таким образом, математическое ожидание среднего арифметического n одинаково распределенных независимых случайных величин $X_1, X_2, \dots, X_n$ равно математическому ожиданию $M(X)$ каждой из этих величин, а дисперсия среднего арифметического в *n* раз меньше дисперсии $D(X)$ каждой из величин.

Дисперсия относительной частоты $\left(\frac{k}{n}\right)$ появления события A в n независимых испытаниях (в каждом из которых событие A появляется с вероятностью p и не появляется с вероятностью $q = 1 - p$) равна

$$D\left(\frac{k}{n}\right) = \frac{pq}{n}, \quad (3.33)$$

где k — число появлений события A в серии из *n* испытаний.

Следовательно,

$$\sigma\left(\frac{k}{n}\right) = \sqrt{\frac{pq}{n}}. \quad (3.34)$$

3.5. Производящие функции

Производящая функция является устройством, отчасти напоминающим мешок. Вместо того чтобы нести много предметов, что могло бы оказаться затруднительным, мы собираем их вместе, и тогда нам нужно нести лишь один предмет — мешок.

Д. Пойа

1) Пусть имеется некоторая (в общем случае бесконечная) последовательность $\{a_0, a_1, a_2, \dots, a_n, \dots\} = \{a_n\}$. Эту последовательность часто удобно выразить в виде степенного ряда по степеням вспомогательной переменной Z:

$$A(Z) = a_0 + a_1 Z + a_2 Z^2 + a_3 Z^3 + \dots = \sum_{k\geq 0} a_k Z^k. \qquad (3.35)$$

Теперь можно заняться изучением свойств функции $A(Z)$, которая представляет всю последовательность. Это особенно важно, если последовательность получена по индукции с использованием рекуррентных выражений. $A(Z)$ называется производящей функцией для последовательности $\{a_n\}$.

Замечание. В настоящее время самый мощный метод работы с последовательностями чисел — это преобразование бесконечных рядов (то есть производящих функций), которые «порождают» эти последовательности. Впервые метод производящих функций для решения линейных рекуррентных соотношений ввел в начале XVIII в. Де Муавр, затем в 1730 г. Дж. Стирлинг применил производящие функции для решения более сложных задач и показал, как применять при этом дифференцирование и интегрирование. Дальнейшее развитие метод производящих функций нашел в работах Л. Эйлера (1741–1750) и Пьера Лапласа (1812). Самое интересное заключается в том, что большинство операций, выполняемых над производящими функциями, можно обосновать, не затрагивая вопроса о сходимости соответствующего бесконечного ряда (в свое время это был один из опорных пунктов в критике Л. Эйлера). ■

Пусть для последовательности $\{b_0, b_1, b_2, \dots, b_m, \dots\} = \{b_m\}$ производящая функция

$$B(Z) = b_0 + b_1 Z + b_2 Z^2 + b_3 Z^3 + \dots = \sum_{m\geq 0} b_m Z^m. \qquad (3.36)$$

Рассмотрим произведение производящих функций *A(Z)B(Z) = C(Z).*

$$A(Z)B(Z) = a_0 b_0 + (a_0 b_1 + a_1 b_0)Z + (a_0 b_2 + a_1 b_1 + a_2 b_0)Z^2 + \cdots, \qquad (3.37)$$

следовательно,

$$A(Z)B(Z) = \sum_{k\geq 0} \sum_{m\geq 0} a_k Z^k b_m Z^m = C(Z) = c_0 + c_1 Z + c_2 Z^2 + \cdots \qquad (3.38)$$

Приравняем коэффициенты при Z^n и получим формулу

$$c_n = \sum_{k=0}^{n} a_k b_{n-k}. \qquad (3.39)$$

Последовательность, полученная по правилу (3.39), называется сверткой последовательностей $\{a_n\}$ и $\{b_m\}$.

Пример 3.7. Полученную в конце 3.3 формулу свертки Вандермонда теперь можно интерпретировать с точки зрения производящих функций. Пусть φ_1 и φ_2 — две производящие функции:

$$\varphi_1(Z) = (1+\mathrm{Z})^{\mathrm{m}} = \sum_{\mathrm{k}\geq 0} \mathrm{C}_{\mathrm{m}}^{\mathrm{k}} \mathrm{Z}^{\mathrm{k}},\ \varphi_2(Z) = (1+Z)^l = \sum_{\mathrm{k}\geq 0} \mathrm{C}_l^{\mathrm{k}} \mathrm{Z}^{\mathrm{k}}.$$

Тогда рассмотрим их произведение

$$\varphi_1(Z)\varphi_2(Z) = (1+Z)^m(1+Z)^l = (1+Z)^{m+l}$$

и приравняем коэффициенты при Z^n:

$$\sum_{k=0}^{n} C_m^k C_l^{n-k} = C_{m+l}^n. \qquad (3.40)$$

Как упоминалось выше, полученная формула носит название свертки Вандермонда (Александр Вандермонд написал по этому поводу в 1772 г. статью, хотя формула была известна еще в 1303 г. Чжу Ши-Цзе из Китая). Из свертки Вандермонда легко получить целый ряд частных случаев, например, пусть *m=l=n*, тогда

$$\sum_{k=0}^{n} C_n^k C_n^{n-k} = C_{2n}^n,$$

но по свойству биноминальных коэффициентов $C_n^k = C_n^{n-k}$, следовательно, имеем

$$\sum_{k=0}^{n} (C_n^k)^2 = C_{2n}^n. \qquad (3.41)$$

Пример 3.8. Два стрелка сделали по *n* выстрелов, по разным мишеням. Какова вероятность одинакового числа попаданий, если вероятность попадания каждого стрелка равна 0,5.

Решение. Пусть ДСВ X_1 — число попаданий первым стрелком в цель, ДСВ X_2 — число попаданий вторым стрелком в цель, P_k — вероятность того, что оба стрелка попали *k* раз. Тогда, учитывая, что вероятность *k* попаданий из n выстрелов находится по формуле Бернулли, получим

$$P_k = P(X_1 = k)P(X_2 = k) = C_n^k p^k q^{n-k} \cdot C_n^k p^k q^{n-k}.$$

По условию $p = q = \frac{1}{2}$, следовательно,

$$P_k = P(X_1 = k) \cdot P(X_2 = k) = C_n^k \left(\tfrac{1}{2}\right)^n \cdot C_n^k \left(\tfrac{1}{2}\right)^n = \left(\tfrac{1}{4}\right)^n (C_n^k)^2.$$

Суммируя P_k от нуля до *n* получим вероятность одинакового числа попаданий:

$$P = \sum_{k=0}^{n} P_k = \sum_{k=0}^{n} \left(\tfrac{1}{4}\right)^n (C_n^k)^2 = \left(\tfrac{1}{4}\right)^n \sum_{k=0}^{n} (C_n^k)^2.$$

Учитывая формулу (3.41), получим выражение для вероятности одинакового числа попаданий в замкнутом виде (т. е. не в виде суммы, а в виде конечной формулы):

$$P = \left(\tfrac{1}{4}\right)^n \sum_{k=0}^{n} (C_n^k)^2 = \left(\tfrac{1}{4}\right)^n C_{2n}^n.$$

2) Одним из важнейших в дискретной математике является понятие *производящей функции* (вероятностей) неотрицательной целочисленной *случайной величины X* (ПФСВ) с вероятностями $P_0, P_1, P_2, \ldots, P_k, \ldots$

Если случайная величина X принимает только целые неотрицательные значения, то распределение ее вероятностей можно представить как степенной ряд (в частном случае многочлен) по степеням Z:

$$\varphi(Z) = \sum_{k\geq 0} P(X = k)Z^k = M(Z^x),$$

где $P(X = k) = P_k,\ |Z| \leq 1$.

$\varphi(Z)$ — производящая функция неотрицательной целочисленной случайной величины X.

Коэффициенты φ неотрицательны и их сумма равна 1:

$$(\sum_{k\geq 0} P(X = k) = 1), \text{ следовательно, } \varphi(1) = 1.$$

И обратно, любой степенной ряд $\varphi(Z)$ с неотрицательными коэффициентами и свойством $\varphi(1) = 1$ является производящей функцией случайной величины. Он содержит всю информацию о случайной величине X.

Свойства производящей функции неотрицательной целочисленной случайной величины, $\varphi_x(Z)$.

1) $M(x) = \sum_{k\geq 0} P(X = k)kZ^{k-1} = \varphi'_x;$ (3.42)

2) $M(x^2) = \sum_{k\geq 0} P(X = k)k(k-1)Z^{k-2} + \sum_{k\geq 0} P(X = k)kZ^{k-1} =$
$= \varphi''_x(1) + \varphi'_x(1), \text{следовательно,}$

3) $D(x) = \varphi''_x(1) + \varphi'_x(1) - [\varphi'_x(1)]^2.$ (3.43)

4) Если случайные величины X и Y независимы и принимают только целые неотрицательные значения, то распределение их суммы будет определяться формулой *свертки вероятностей* дискретных распределений:

$$P(X + Y = r) = \sum_k P(X = k, Y = r - k) = \sum_k P(X = r - k)P(Y = k), \quad (3.44)$$

которая называется формулой полной вероятности.

Свертке этих последовательностей отвечает произведение производящих функций

$$\varphi_{X+Y}(Z) = \varphi_x(Z)\varphi_Y(Z). \quad (3.45)$$

Замечание. 1. Из свойств 1, 4 следует, что для поиска числовых характеристик результата суммы независимых случайных величин $Y = X_1 + X_2 + X_3 + \cdots X_n$ — необходимо представить производящую функцию вероятностей СВ Z в виде произведения производящих функций вероятностей слагаемых $X_1, X_2, X_3, \ldots X_n$.

$$\varphi_y = \prod_{i=1}^{n} \varphi_{x_i}(Z) \text{ и найти } \varphi'_y(1),$$

или, найдя с помощью производящих функций числовые характеристики $X_1, X_2, X_3, \ldots X_n$ (свойства 1, 2), легко перейти к числовым характеристикам их суммы по свойствам математического ожидания и дисперсии, то есть просто сложить числовые характеристики СВ $X_1, X_2, X_3, \ldots, X_n$.

2. Математическое ожидание и дисперсия — это лишь первые члены из бесконечного ряда так называемых кумулянтов (или семиинвариантов, связанных с начальными и центральными моментами (4.15)–(4.22)), введенных в 1903 г. датским астрономом Т. Н. Тиеле:

$$k_1, k_2, \dots;\ k_1 = M(X), k_2 = D(X).$$

Кумулянты более высоких порядков выражают более тонкие свойства распределения. Если $\varphi(Z)$ — производящая функция вероятностей случайной величины, то производящая функция моментов

$$ln\,\varphi(e^t) = \frac{k_1}{1!}t + \frac{k_2}{2!}t^2 + \frac{k_3}{3!}t^3 + \frac{k_4}{4!}t^4 + \cdots$$

3. Производящая функция случайной величины для часто встречающихся законов распределения дискретной случайной величины.

1) Закон распределения Бернулли: $\varphi(Z) = q + pZ$.

2) Биномиальный закон распределения: $\varphi(Z) = (q + pZ)^n$.

3) Геометрический закон распределения: $\varphi(Z) = \sum_{m\geq 0} pq^m\, Z^m = \frac{P}{1-qZ}$.

4) Геометрический закон распределения, сдвинутый на единицу:

$$\varphi(Z) = pZ + pZ^2 + \cdots = \frac{pZ}{1-qZ}.$$

5) Закон отрицательного биномиального распределения:

$$\varphi(Z) = (p/(1-qZ))^k = \sum_m C^m{}_{m+k-1}\, p^k q^m Z^m.$$

6) Закон распределения Пуассона: $\varphi(Z) = \sum_{k\geq 0} \frac{\lambda^k}{k!} e^{-\lambda} Z^k = e^{\lambda(Z-1)}$.

7) Гипергеометрический закон распределения: $\varphi(Z) = \sum_k \frac{C_M^k C_{N-M}^{n-k}}{C_N^n} Z^k$. ■

Пример 3.9. Случайные величины *X* и *Y* независимы и имеют биноминальное распределение с параметрами (*m, p*), (*n, p*) соответственно. Найдите распределение случайной величины $(X + Y)$.

Решение. Согласно формуле свертки вероятностей для суммы независимых случайных величин, имеем

$$P(X + Y = r) = \sum_{k=0}^{n+m} C_n^k p^k (1-p)^{n-k}\, C_m^{r-k} p^{r-k} (1-p)^{m-(r-k)} =$$
$$= p^r (1-p)^{n+m-r} \sum_{k=0}^{n+m} C_n^k C_m^{r-k} = C_{n+m}^r\, p^r (1-p)^{n+m-r},$$

что соответствует произведению производящих функций φ_X и φ_Y:

$$\varphi_{X+Y}(Z) = ((1-p) + pZ)^{m+n}.$$

Пример 3.10. Игрок поочередно покупает билеты двух разных лотерей до первого выигрыша. Вероятность выигрыша по одному билету первой лотереи составляет 0,2, а второй — 0,3. Игрок вначале покупает билет первой лотереи. Составить закон распределения и найти математическое ожидание случайной величины *X* — числа купленных билетов, если он имеет возможность купить: а) только 5 билетов; б) неограниченное число билетов.

Решение. Пусть событие A_i — выигрыш *i*-го билета первой из лотерей, B_i — выигрыш *i*-го билета второй из лотерей. По условию известно, что $P(A_i) = 0{,}2$ и $P(B_i) = 0{,}3$.

Рассмотрим дискретную случайную величину *X* — число билетов, купленных до выигрыша: $X = \{1, 2, 3, 4, 5\}$.

а) Представим события $X=k$ (k=1, 2, 3, 4, 5) и вероятности в виде таблицы.

X	1	2	3	4	5
Событие $X=k$	A_1	$\bar{A}_1B_1$	$\bar{A}_1\bar{B}_1A_2$	$\bar{A}_1\bar{B}_1\bar{A}_2B_2$	$\bar{A}_1\bar{B}_1\bar{A}_2\bar{B}_2A_3 + +\bar{A}_1\bar{B}_1\bar{A}_2\bar{B}_2\bar{A}_3$
Формула вероятности: $P(X=k)$	0,2	0,8· 0,3	0,8·0,7·0,2	$0{,}8^2 \cdot 0{,}7 \cdot 0{,}3$	$0{,}8^2 \cdot 0{,}7^2 \cdot 0{,}2 + +0{,}8^3 \cdot 0{,}7^2$
Значение вероятности: $P(X=k)$	0,2	0,24	0,112	0,1344	0,3136

$M(X) = 1 \cdot 0{,}2 + 2 \cdot 0{,}24 + 3 \cdot 0{,}112 + 4 \cdot 0{,}1344 + 5 \cdot 0{,}3136 = 3{,}1216$ — математическое ожидание числа купленных билетов до первого выигрыша.

б) Рассмотрим производящую функцию случайной величины X — числа купленных билетов до первого выигрыша при возможности покупки неограниченного числа билетов. Учитывая результаты подсчета вероятностей в пункте *а*), легко получить производящую функцию $\varphi_x(Z)$ случайной величины X.

Она будет состоять из двух сумм:

$$\sum_{k\geq 0} P(x = 2k+1)Z^{2k+1} \text{ и } \sum_{k\geq 1}(x = 2k)2^{2k}$$

или

$$P(x=2k+1) = 0{,}8^k 0{,}7^k 0{,}2 \text{ и } P(x=2k) = 0{,}8^k 0{,}7^{k-1} 0{,}3,$$

соответствующих выигрышу на нечетном ($X=2k+1$) или четном ($X=2k$) билете.

$$\varphi_x(Z) = \sum_{k\geq 0} P(X = 2k+1)Z^{2k+1} + \sum_{k\geq 1} P(X = 2k)Z^{2k} =$$

$$= \sum_{k\geq 0} 0{,}8^k 0{,}7^k 0{,}2Z^{2k+1} + \sum_{k\geq 1} 0{,}8^k 0{,}7^{k-1} 0{,}3Z^{2k} =$$

$$= 0{,}2Z \sum_{k\geq 0}(0{,}56Z^2)^k + \frac{3}{7}\sum_{k\geq 1}(0{,}56Z^2)^k =$$

$$= 0{,}2Z \frac{1}{1-0{,}56Z^2} + \frac{3}{7}\frac{0{,}56Z^2}{1-0{,}56Z^2} = \frac{0{,}24Z^2+0{,}2Z}{1-0{,}56Z^2}.$$

Таким образом, производящая функция случайной величины X — числа купленных билетов до первого выигрыша имеет вид

$$\varphi_x(Z) = \frac{0{,}24Z^2+0{,}2Z}{1-0{,}56Z^2}.$$

Найдем производную $\varphi'_x(Z) = \frac{(0{,}48Z+0{,}2)(1-0{,}56Z^2)+2\cdot 0{,}56Z(0{,}24Z^2+0{,}2Z)}{(1-0{,}56Z^2)^2}$.

Математическое ожидание по первому свойству производящей функции (10.8) равно $\varphi'(1)$, следовательно,

$$M(X) = \frac{0{,}68\cdot 0{,}44+1{,}12\cdot 0{,}44}{0{,}44^2} = \frac{0{,}68+1{,}12}{0{,}44} = \frac{1{,}8}{0{,}44} = \frac{180}{44} \approx 4{,}091.$$

Темы (вопросы) для самоконтроля

1. Закон распределения дискретной случайной величины.
2. Математическое ожидание и его свойства.
3. Дисперсия и ее свойства.
4. Распределение Бернулли.

5. Биномиальное распределение.
6. Распределение Пуассона.
7. Геометрическое распределение.
8. Геометрическое распределение, сдвинутое на единицу.
9. Отрицательное биномиальное распределение.
10. Гипергеометрическое распределение.
11. Полиномиальное распределение.
12. Числовые характеристики независимых одинаково распределенных случайных величин.
13. Производящая функция последовательности, свертка последовательностей.
14. Производящая функция неотрицательной целочисленной случайной величины и ее свойства.
15. Производящие функции основных законов распределения дискретных случайных величин.

Глава 4
Непрерывные случайные величины

4.1. Функция распределения и ее свойства

Если дискретная случайная величина принимает отдельные, изолированные значения, то значения непрерывной случайной величины полностью заполняют какой-то промежуток. Для непрерывной случайной величины X вероятность $P(X = x_i) \to 0$, поэтому для нее используется вероятность того, что случайная величина $X < x$, где $x = x_i$ текущее значение переменной. Вероятность того, что случайная величина X примет значение, меньшее значения x, называется *функцией* распределения случайной величины (интегральной функцией):

$$P(X < x) = F(x). \tag{4.1}$$

Функция распределения является универсальным способом задания случайной величины (как дискретной, так и для непрерывной).

Свойства функции распределения:

1) значения функции распределения принадлежат отрезку [0,1]:

$$0 \le F(x) \le 1;$$

2) функция распределения является неубывающей функцией, т. е. если $x_2 > x_1$, то $F(x_2) \ge F(x_1)$;

3) $F(-\infty) = 0$ как вероятность невозможного события;

4) $F(+\infty) = 1$ как вероятность достоверного события;

5) вероятность попадания случайной величины X в интервал от a до b равна приращению функции распределения на этом интервале:

$$P(a \le X < b) = F(b) - F(a). \tag{4.2}$$

Для непрерывной случайной величины верны неравенства

$$\mathrm{P}(a < \mathrm{X} < b) = \mathrm{P}(a \le \mathrm{X} < b) = \mathrm{P}(a < \mathrm{X} \le b) = \mathrm{P}(a \le \mathrm{X} \le b),$$

так как $P(X = x) \to 0$.

Функция распределения дискретной случайной величины X, принимающей значения $x_1, x_2, \ldots, x_n$, имеет вид

$$F(x) = \sum_{X<x_i} P(X = x_i). \tag{4.3}$$

Пример 4.1 Дискретная случайная величина X задана таблицей:

x_i	1	3	5	8
p_i	0,1	0,3	0,4	0,2

Составить функцию распределения случайной величины X и начертить ее график.

Решение. Пусть $x \le 1$, тогда $F(x) = 0$, так как событие $\mathrm{X} < x$ будет невозможным. На основании равенства (4.3) имеем:

если $1 < x \le 3$, то $F(x) = p_1 = 0{,}1$;

если $3 < x \le 5$, то $F(x) = p_1 + p_2 = 0{,}1 + 0{,}3 = 0{,}4$;

если $5 < x \le 8$, то $F(x) = p_1 + p_2 + p_3 = 0{,}1 + 0{,}3 + 0{,}4 = 0{,}8$;

если $x > 8$, то $F(x) = p_1 + p_2 + p_3 + p_4 = 0{,}1 + 0{,}3 + 0{,}4 + 0{,}2 = 1$.

Окончательно получаем

$$F(x) = \begin{cases} 0, & \text{при } x \le 1, \\ 0{,}1, & \text{при } 1 < x \le 3, \\ 0{,}4, & \text{при } 3 < x \le 5, \\ 0{,}8, & \text{при } 5 < x \le 8, \\ 1, & \text{при } x > 8. \end{cases}$$

График функции $F(x)$изображен на рисунке 4.1 и представляет собой разрывную скачкообразную функцию.

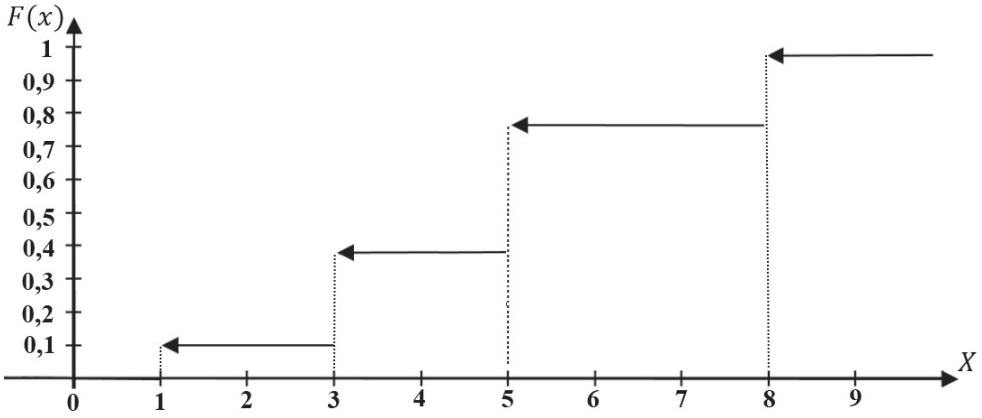


Рис. 4.1 — График функции распределения дискретной случайной величины X

4.2. Плотность распределения вероятностей непрерывной случайной величины

Рассмотрим непрерывную случайную величину X. Если ее функция распределения $F(x)$ непрерывна и дифференцируема всюду, за исключением конечного числа точек на любом конечном промежутке, то с физической точки зрения вероятность можно интерпретировать как массу, равную 1, распределенную в области определения функции $F(x)$ на оси X. На непрерывном промежутке Δx масса вероятности равна

$$F(x + \Delta x) - F(x).$$

Тогда плотность вероятности на промежутке Δx:

$$\frac{F(x + \Delta x) - F(x)}{\Delta x}.$$

Переходя к пределу при $\Delta x \to 0$, получим, что плотность вероятности в точке x равна значению производной:

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

Функцией плотности распределения вероятностей (дифференциальной функцией) непрерывной случайной величины X называется производная ее функции распределения:

$$f(x) = F'(x). \qquad (4.4)$$

Свойства функции плотности распределения вероятностей:

1) плотность распределения вероятностей является неотрицательной функцией, так как функция распределения $F(x)$ является неубывающей функцией, т. е.

$$f(x) \geq 0;$$

2) несобственный интеграл от плотности вероятности в бесконечных пределах равен единице:

$$\int_{-\infty}^{+\infty} f(x)dx = 1. \quad (4.5)$$

Геометрически свойства означают, что график плотности распределения вероятностей, называемый кривой распределения, лежит над осью абсцисс или на оси абсцисс. Площадь, ограниченная кривой распределения и осью абсцисс, равна единице.

Зная плотность распределения вероятностей, учитывая формулу (4.2), можно получить формулу вероятности попадания случайной величины X в заданный интервал (a, b):

$$P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a). \quad (4.6)$$

Если известна плотность распределения $f(x)$, то функция распределения находится по формуле

$$F(x) = P(X < x) = P(-\infty < X < x) = \int_{-\infty}^{x} f(x)dx,$$

то есть

$$F(x) = \int_{-\infty}^{x} f(x)dx. \quad (4.7)$$

4.3. Числовые характеристики непрерывных случайных величин

1) *Математическое ожидание* непрерывной случайной величины X определяется по формуле

$$M(X) = \int_{-\infty}^{+\infty} xf(x)dx, \quad (4.8)$$

если интеграл абсолютно сходится (является конечным числом)

$$\int_{-\infty}^{+\infty} |x| f(x)dx < \infty.$$

Если непрерывная случайная величина X определена на интервале (a; b), то

$$M(X) = \int_a^b xf(x)dx. \quad (4.9)$$

2) *Мода* непрерывной случайной величины X будет определяться как значение, доставляющее максимум ее функции плотности распределения вероятностей:

$$M_o(X) = \underset{(-\infty;+\infty)}{argmax}\, f(x). \quad (4.10)$$

Если непрерывная случайная величина X определена на интервале (a; b), то

$$M_o(X) = \underset{(a;b)}{argmax}\, f(x). \quad (4.11)$$

3) *Медиана* — это значение случайной величины, которое делит площадь под функцией плотности вероятности на две равные части.

$$P(X < M_e(X)) = P\big(X > M_e(X)\big) = \frac{1}{2}. \quad (4.12)$$

4) *Дисперсия* непрерывной случайной величины:

$$D(X) = \int_{-\infty}^{+\infty} \big(x - M(X)\big)^2 f(x)dx = \int_{-\infty}^{+\infty} x^2 f(x)dx \; - \; M^2(X). \quad (4.13)$$

Если непрерывная случайная величина *X* определена на интервале (*a; b*), то:

$$D(X) = \int_{a}^{b} \big(x - M(X)\big)^2 f(x)dx = \int_{a}^{b} x^2 f(x)dx \; - \; M^2(X). \quad (4.14)$$

Все свойства дисперсии и математического ожидания, установленные для дискретных случайных величин, сохраняются для непрерывных случайных величин. Если распределение случайной величины X симметрично, то $M(X) = M_o(X) = M_e(X)$.

5) *Моменты случайных величин*.

Кроме характеристик положения и рассеяния применяется ряд других числовых характеристик распределения, например моменты.

Начальным моментом порядка *s* называется математическое ожидание степени *s* случайной величины *X*:

$$\alpha_s = M(X^s). \quad (4.15)$$

Для дискретной случайной величины:

$$\alpha_s = \sum_{i=1}^{n} x_i^s \, p_i = x_1^s p_1 + x_2^s p_2 + \cdots + x_n^s p_n. \quad (4.16)$$

Для непрерывной случайной величины:

$$\alpha_s = \int_{-\infty}^{+\infty} x^s \, f(x)dx. \quad (4.17)$$

При $s = 1$: $\alpha_1 = M(\mathrm{X}) = m_x$, то есть первый начальный момент — это математическое ожидание случайной величины.

Отклонение случайной величины от ее математического ожидания называется *центрированной* случайной величиной $\dot{X} = X - m_x$, что соответствует переносу начала координат в математическое ожидание системы точек (или центр тяжести).

Центральным моментом порядка *s* случайной величины *X* называется математическое ожидание степени *s*, соответствующей центрированной случайной величины:

$$\mu_s = M\big(\dot{X}^s\big) = M\left((x - m_x)^s\right); \quad (4.18)$$

для дискретной случайной величины:

$$\mu_s = \sum_{i=1}^{n} (x_i - m_x)^s p_i = (x_1 - m_x)^s p_1 + \cdots + (x_n - m_x)^s p_n; \quad (4.19)$$

для непрерывной случайной величины:

$$\mu_s = \int_{-\infty}^{+\infty} (x - m_x)^s \, f(x)dx. \quad (4.20)$$

При вычислении центральных моментов пользуются формулами связи между центральными и начальными моментами:

$\mu_1 = 0,$

$\mu_2 = \alpha_2 - m_x^2,$

$\mu_3 = \alpha_3 - 3m_x\alpha_2 + 2m_x^3,$ (4.21)

$\mu_4 = \alpha_4 - 4m_x\alpha_3 + 6m_x^2\alpha_2 - 3m_x^4.$ (4.22)

Обычно рассматривают первые четыре центральных момента:

1) $\mu_1 = M(x - m_x) = 0$ — математическое ожидание центрированной случайной величины равно нулю;

2) $\mu_2 = M(x - m_x)^2 = D(X) = \sigma_x^2$ — второй центральный момент — это дисперсия случайной величины;

3) $\mu_3 = M(x - m_x)^3$ — третий центральный момент может служить для характеристики асимметрии (скошенности распределения), обычно рассматривают безразмерный коэффициент асимметрии:

$$Ka = \frac{\mu_3}{\sigma^3}; \quad (4.23)$$

4) $\mu_4 = M(x - m_x)^4$ — четвертый центральный момент, может служить для характеристики «крутости» или островершинности распределения, описывающейся с помощью эксцесса:

$$Ex = \frac{\mu_4}{\sigma^4} - 3. \quad (4.24)$$

Основным моментом порядка *s* называется нормированный центральный момент порядка *S*:

$$r_s = \frac{\mu_S}{\sigma^S}, \quad (4.25)$$

то есть $Ka = r_3$, $Ex = r_4 - 3$.

r_1 — нормированная (стандартизированная) случайная величина *X*.

Замечание. 1) $Ka = 0$ — распределение симметрично,

$$M_o(X) = M_e(X) = M(X),$$

$Ka > 0$ — распределение имеет правостороннюю асимметрию,

$$M_o(X) < M(X),$$

$Ka < 0$ — распределение имеет левостороннюю асимметрию,

$$M_o(X) > M(X).$$

Асимметрия увеличивает влияние хвостов распределения. Отрицательная асимметрия показывает смещение моды вправо и доминирование левого хвоста распределения, положительная асимметрия показывает смещение моды влево и доминирование правого хвоста распределения.

2) Распределение имеет вершину:

при $Ex = 0$ — типа $\varphi(x)$ (рис. 2.1),

при $Ex > 0$ — более заостренную, чем $\varphi(x)$,

при $Ex < 0$ — более плоскую, чем $\varphi(x)$.

Введем понятие *«хвостов распределения» при* $x \to \infty$: $F(-x)$ *и* $1 - F(x)$.

Эксцесс описывает, насколько тяжелые хвосты распределения: положительный эксцесс — длинные, тяжелые хвосты; отрицательный эксцесс — короткие, легкие хвосты.

3) Фактически начальные и центральные моменты служат для вычисления основных моментов, представляющих вполне определенные численные характеристики отражения различных свойств случайных величин.

4) С точки зрения механики, если рассматривать дискретные случайные величины как точки на оси OX с координатами x_i, в которых размещены массы p_i, то первый начальный момент $m_x = a$ — координата (абсцисса) центра тяжести системы точек (x_i, p_i). Как известно, масса характеризует инертность системы при прямолинейном движении. Второй центральный момент $\mu_2 = D(X) = \sigma_x^2 = I_{m_x}$ — момент инерции системы точек (x_i, p_i) при вращении относительно оси, проходящей через центр тяжести m_x, характеризует инертность системы при вращательном движении (рис. 4.2).

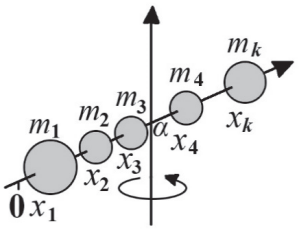


Рис. 4.2 — Иллюстрация момента инерции I_{m_x}

5) В случае непрерывной случайной величины механистические рассуждения приводят к пониманию математического ожидания и дисперсии как центра тяжести и соответствующего момента инерции геометрической фигуры, ограниченной функцией плотности вероятности $f(x)$ и осью OX.

6) В литературе встречается формализованная запись, позволяющая объединить изучение непрерывных и дискретных величин в виде интеграла Стилтьеса:

$$P(a \le X < b) = \int_a^b dF(x), \tag{4.26}$$

где $F(x) = P(X < x)$.

В случае непрерывного распределения $dF(x) = f(x)dx$, интеграл (4.26) сводится к стандартному случаю — интегралу Римана. Для дискретного распределения на интервале $a = x_0 < x_1 < x_2 < \cdots < x_n = b$

$$\int_a^b dF(x) = \sum_{i=1}^n [F(x_i) - F(x_{i-1})] = \sum_i P(x_i). \tag{4.27}$$

Таким образом, можно записать общие формулы для дискретных и непрерывных распределений случайных величин:

$$M(X) = \int_{-\infty}^{+\infty} x dF(x),$$
$$M\big(\varphi(x)\big) = \int_{-\infty}^{+\infty} \varphi(x) dF(x),$$
$$D(X) = \int_{-\infty}^{+\infty} (x - m_x)^2 dF(x). \quad \blacksquare$$

Пример 4.2. Непрерывная случайная величина *X* задана функцией распределения:

$$F(x) = \begin{cases} 0, \text{при } x < 0, \\ \frac{x^3}{125}, \text{при } 0 \le x < 5, \\ 1, \text{при } x \ge 5. \end{cases}$$

Определить: а) вероятность попадания случайной величины в интервал (2;3); б) математическое ожидание, дисперсию и среднее квадратическое отклонение случайной величины *X*; в) функции распределения изобразить графически.

Решение. По четвертому свойству функции распределения:

$$P(2 \le X < 3) = F(3) - F(2) = \frac{x^3}{125}\Big|_{x=3} - \frac{x^3}{125}\Big|_{x=2} = \frac{27}{125} - \frac{8}{125} = \\ = \frac{19}{125} = 0{,}152.$$

Найдем функцию плотности вероятности (дифференциальную функцию):

$$f(x) = F'(x) = \begin{cases} 0, \text{при } x < 0, \\ \frac{3x^2}{125}, \text{при } 0 \le x < 5, \\ 0, \text{при } x \ge 5. \end{cases}$$

Вероятность попадания случайной величины в интервал (2,3) также можно найти, зная функцию плотности вероятности (4.6):

$$P(2 < X < 3) = \int_2^3 \frac{3x^2}{125} dx = \frac{3}{125}\frac{x^3}{3}\Big|_2^3 = \frac{27}{125} - \frac{8}{125} = \frac{19}{125} = 0{,}152.$$

Найдем числовые характеристики непрерывной случайной величины X. Следует обратить внимание, что случайная величина задана на интервале (0;5).

$$M(X) = \int_0^5 x \frac{3}{125} x^2 dx = \frac{3}{125}\int_0^5 x^3 dx = \frac{3}{125}\frac{x^4}{4}\Big|_0^5 = \\ = \frac{3}{5^3} \cdot \frac{5^4}{4} = \frac{15}{4} = 3{,}75.$$

$$D(X) = \int_0^5 x^2 \frac{3x^2}{125} dx - 3{,}75^2 = \frac{3}{125}\int_0^5 x^4 dx - 14{,}0625 = \\ = \frac{3}{125}\frac{x^5}{5}\Big|_0^5 - 14{,}0625 = 15 - 14{,}0625 = 0{,}9375.$$

$$\sigma(X) = \sqrt{D(x)} = \sqrt{0{,}9375} = 0{,}9682.$$

Построим графики функций $F(x)$ и $f(x)$.

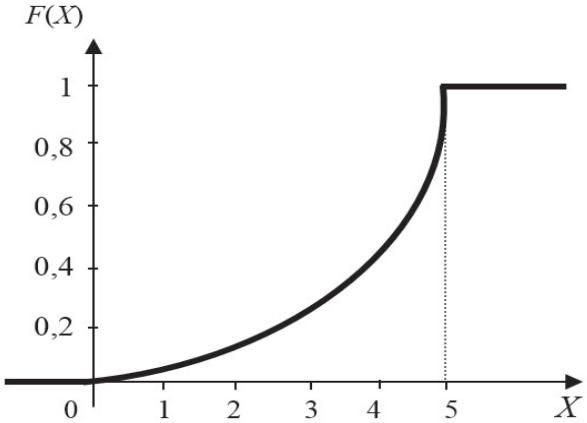


Рис. 4.3 — Функция распределения вероятностей случайной величины *X*

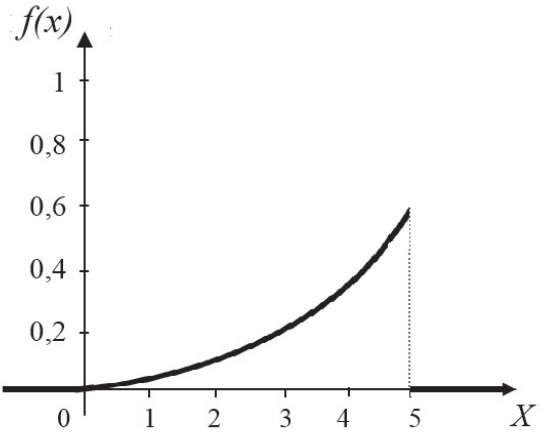


Рис. 4.4 — Функция плотности распределения вероятностей случайной величины X

Пример 4.3. Непрерывная случайная величина X задана плотностью распределения вероятностей:

$$f(x) = \begin{cases} 0, & \text{при } x < 3, \\ \frac{x-3}{32}, & \text{при } 3 \le x < 11, \\ 0, & \text{при } x \ge 11. \end{cases}$$

Определить: а) функцию распределения; б) математическое ожидание, дисперсию и среднее квадратическое отклонение.

Решение. Зная плотность распределения, функция распределения случайной величины находится по формуле

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

При $x < 3, F(x) = \int_{-\infty}^{x} 0\,dx = 0;$

при $3 \le x < 11, F(x) = \int_{-\infty}^{3} 0\,dx + \int_{3}^{x} \frac{x-3}{32}dx = \frac{x^2}{64}\Big|_3^x - \frac{3x}{32}\Big|_3^x = \frac{x^2-6x+9}{64};$

при $x \ge 11, F(x) = \int_{-\infty}^{3} 0\,dx + \int_{3}^{11} \frac{x-3}{32}dx + \int_{11}^{x} 0dx = 1.$

Функция распределения примет вид

$$F(x) = \begin{cases} 0, & \text{при } x < 3, \\ \frac{x^2 - 6x + 9}{64}, & \text{при } 3 \le x < 11, \\ 1, & \text{при } x \ge 11. \end{cases}$$

$$M(X) = \int_3^{11} x\frac{x-3}{32}dx = 8\frac{1}{3},$$

$$D(X) = \int_3^{11} x^2\frac{x-3}{32}dx - \left(8\frac{1}{3}\right)^2 = 3{,}399;\ \sigma(X) = 1{,}844.$$

Темы (вопросы) для самоконтроля

1. Функция распределения непрерывной случайной величины.
2. Функция плотности вероятностей.
3. Числовые характеристики положения центра распределения.
4. Числовые характеристики рассеяния относительно центра распределения.
5. Начальные, центральные и основные моменты.
6. Формулы связи начальных и центральных моментов.
7. Асимметрия распределения.
8. Эксцесс распределения.
9. Физический смысл первых моментов.

Глава 5
Основные законы распределения непрерывных случайных величин

5.1. Равномерное распределение

Случайная величина X распределена по *равномерному (прямоугольному) закону*, если на заданном интервале плотность распределения вероятностей принимает постоянное значение. Например, если весы имеют точность 1 г и полученное значение округляется до ближайшего целого числа k, то точный вес можно считать равномерно распределенной случайной величиной на интервале $(k-0{,}5; k+0{,}5)$.

Плотность распределения равномерного закона на интервале (α, β) имеет вид

$$f(x)=\begin{cases}0, \text{при } x<a,\\ \frac{1}{\beta-\alpha}, \text{при } \alpha\le x<\beta,\\ 0, \text{при } x\ge\beta.\end{cases} \tag{5.1}$$

Найдем функцию распределения равномерного закона $F(x)$ на интервале (α, β):

а) при $x<\alpha, F(x)=\int_{-\infty}^{x} f(x)dx=\int_{-\infty}^{x} 0dx=0,$

б) при $\alpha\le x<\beta, F(x)=\int_{-\infty}^{\alpha} f(x)dx+\int_{\alpha}^{x} f(x)\,dx=$

$$=\int_{-\infty}^{\alpha} 0dx+\int_{\alpha}^{x}\frac{1}{\beta-\alpha}dx=\frac{x-\alpha}{\beta-\alpha},$$

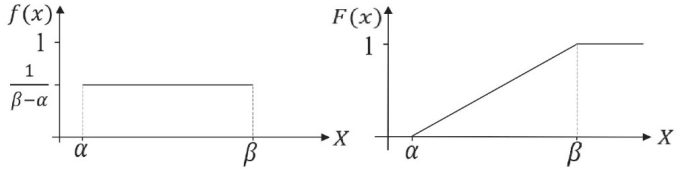
в) при при $x\ge\beta, F(x)=\int_{-\infty}^{\alpha} f(x)dx+\int_{\alpha}^{\beta} f(x)\,dx+\int_{\beta}^{+\infty} f(x)\,dx=$

$$=\int_{-\infty}^{\alpha} 0dx+\int_{\alpha}^{\beta}\frac{1}{\beta-\alpha}dx+\int_{\beta}^{+\infty} 0\,dx=\left.\frac{x-\alpha}{\beta-\alpha}\right|_{\alpha}^{\beta}=1.$$

Имеем

$$F(x)=\begin{cases}0, \text{при } x<\alpha,\\ \frac{x-\alpha}{\beta-\alpha}, \text{при } \alpha\le x<\beta,\\ 1, \text{при } x\ge\beta.\end{cases} \tag{5.2}$$

Графики функций приведены на рисунке 5.1



Функция плотности распределения Функция распределения вероятностей

Рис. 5.1 — Равномерный закон распределения

Основные числовые характеристики равномерного закона.

1. Математическое ожидание:

$$M(X) = \int_{\alpha}^{\beta} x \frac{1}{\beta-\alpha} dx = \frac{1}{\beta-\alpha} \frac{x^2}{2}\Big|_{\alpha}^{\beta} = \frac{\beta+\alpha}{2}. \quad (5.3)$$

Математическое ожидание равномерного распределения, в силу симметрии распределения, совпадает с медианой.

2. Моды равномерное распределение не имеет.

3. Дисперсия:

$$D(X) = \int_{\alpha}^{\beta} x^2 \frac{1}{\beta-\alpha} dx - \left(\frac{\beta+\alpha}{2}\right)^2 = \frac{1}{\beta-\alpha} \frac{x^3}{3}\Big|_{\alpha}^{\beta} - \left(\frac{\beta+\alpha}{2}\right)^2 = \frac{(\beta-\alpha)^2}{12}. \quad (5.4)$$

Отсюда среднее квадратическое отклонение:

$$\sigma(X) = \sqrt{D(X)} = \frac{|\beta-\alpha|}{2\sqrt{3}}. \quad (5.5)$$

4. Третий центральный момент:

$$\mu_3 = M\left(\left(X - \frac{\beta+\alpha}{2}\right)^3\right) = 0 \Rightarrow Ka = 0, \quad (5.6)$$

поэтому распределение симметрично относительно $M(X)$.

5. Четвертый центральный момент:

$$\mu_4 = M\left(\left(X - \frac{\beta+\alpha}{2}\right)^4\right) = \frac{(\beta+\alpha)^4}{80} \Rightarrow \quad (5.7)$$

$$Ex = \frac{\mu_4}{\sigma^4} - 3 = -1{,}2.$$

6. Вероятность попадания случайной величины в заданный интервал (a; b). Пусть случайная величина X распределена по равномерному закону, тогда (рис. 5.2):

$$P(a \le X \le b) = \int_{a}^{b} \frac{1}{\beta-\alpha} dx = \frac{b-a}{\beta-\alpha}. \quad (5.8)$$

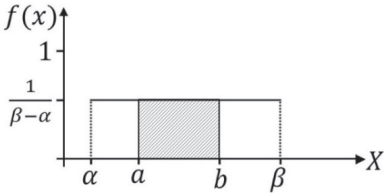


Рис. 5.2 — Вероятность попадания равномерно распределенной случайной величины X в интервал (a; b)

Пример 5.1. Случайная величина X распределена по равномерному закону, у которой $M(X) = 6$, $D(X) = 3$. Составить функции распределения случайной величины. Найти вероятность того, что случайная величина примет значение на интервале (5;8).

Решение. Воспользуемся формулами (5.3) и (5.4):

$$\begin{cases} \frac{\alpha+\beta}{2} = 6; \\ \frac{(\beta-\alpha)^2}{12} = 3; \end{cases} \Rightarrow \alpha = 3;\ \beta = 9.$$

По формулам (5.1) и (5.2) функции распределения будут иметь следующий вид:

$$f(x) = \begin{cases} 0, & \text{при } x < 3, \\ \frac{1}{6}, & \text{при } 3 \le x < 9, \\ 0, & \text{при } x \ge 9, \end{cases} \quad F(x) = \begin{cases} 0, & \text{при } x < 3, \\ \frac{x-3}{6}, & \text{при } 3 \le x < 9, \\ 1, & \text{при } x \ge 9. \end{cases}$$

$$P(5 < x < 8) = \int_5^8 \frac{1}{6}\,dx = \frac{1}{6}x\Big|_5^8 = \frac{8}{6} - \frac{5}{6} = \frac{1}{2} \text{ или}$$

$$P(5 < x < 8) = F(8) - F(5) = \left(\frac{x-3}{6}\right)_{x=8} - \left(\frac{x-3}{6}\right)_{x=5} = \frac{8-3}{6} - \frac{5-3}{6} = \frac{1}{2}.$$

5.2. Показательное распределение

Непрерывная случайная величина X, принимающая неотрицательные значения, имеет показательное (экспоненциальное) распределение, если ее плотность распределения имеет вид (рис. 5.3):

$$f(x) = \begin{cases} 0, & \text{при } x < 0, \\ \lambda e^{-\lambda x}, & \text{при } x \ge 0, \end{cases} \tag{5.9}$$

где $\lambda = const, \lambda > 0$.

Функция распределения показательного закона (рис. 5.3):

$$F(X) = \begin{cases} 0, & \text{при } x < 0, \\ 1 - e^{-\lambda x}, & \text{при } x \ge 0. \end{cases} \tag{5.10}$$

Числовые характеристики показательного закона

1. Математическое ожидание:

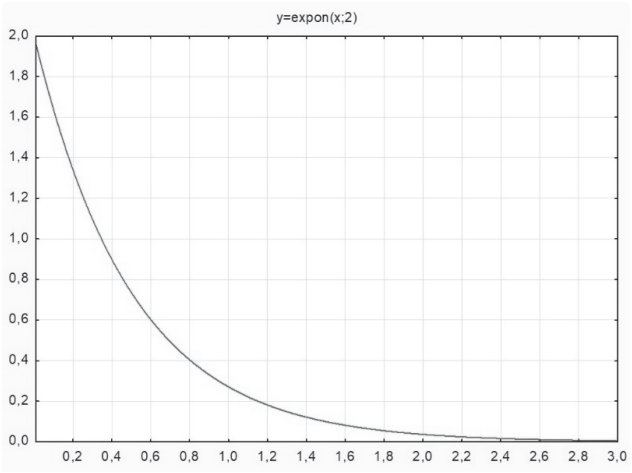
$$M(X) = \frac{1}{\lambda}. \tag{5.11}$$

2. Дисперсия:

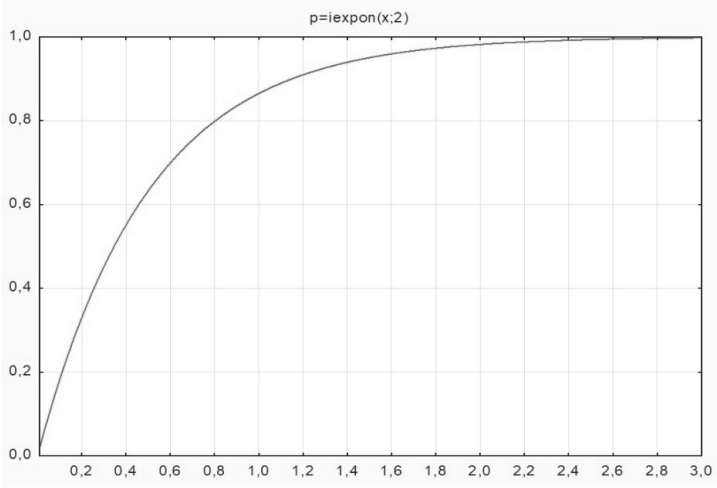
$$D(X) = \frac{1}{\lambda^2}, \tag{5.12}$$

среднее квадратическое отклонение:

$$\sigma(x) = \sqrt{D} = \frac{1}{\lambda}. \tag{5.13}$$



Функция плотности вероятностей



Функция распределения

Рис. 5.3 — Показательный закон распределения ($\lambda = 2$)

Доказательство. Воспользовавшись формулой интегрирования по частям

$$\int_a^b udv = uv|_a^b - \int_a^b vdu,$$

считая, что $u = x, v = e^{-\lambda x}$ и, следовательно, $du = dx, dv = -\lambda e^{-\lambda x}dx$, получим

$$M(X) = \int_0^{+\infty} x\lambda e^{-\lambda x}dx =$$
$$= -\int_0^{+\infty} xde^{-\lambda x} = -xe^{-\lambda x}\Big|_0^{+\infty} + \int_0^{+\infty} e^{-\lambda x}dx =$$
$$= -xe^{-\lambda x}\Big|_0^{+\infty} - \frac{1}{\lambda}\int_0^{+\infty} de^{-\lambda x} = -xe^{-\lambda x}\Big|_0^{+\infty} - \frac{1}{\lambda}e^{-\lambda x}\Big|_0^{+\infty} =$$
$$= -\Big(\lim_{b\to+\infty} b\, e^{-\lambda b} - 0\Big) - \frac{1}{\lambda}\Big(\lim_{b\to+\infty} e^{-\lambda b} - 1\Big) =$$
$$= -(0-0) - \frac{1}{\lambda}(0-1) = \frac{1}{\lambda}.$$

Аналогично, дважды интегрируя по частям, найдем, что

$$M(X^2) = \int_0^{+\infty} x^2 \lambda e^{-\lambda x} dx = \frac{2}{\lambda^2}.$$

Следовательно,

$$D(X) = M(X^2) - M^2(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}, \sigma(x) = \sqrt{D} = \frac{1}{\lambda}.$$

Вероятность попадания непрерывной случайной величины X в заданный интервал:

$$P(a \le x < b) = \int_a^b \lambda e^{-\lambda x} dx = e^{-\lambda a} - e^{-\lambda b}. \qquad (5.14)$$

Пример 5.2. Непрерывная случайная величина распределяется по показательному закону с параметром $\lambda = 3$. Составить функции распределения случайной величины. Найти вероятность того, что случайная величина примет значение на интервале (1;2). Определить $\mathrm{M(X)}, D(X), \sigma(X)$.

Решение. Функции распределения показательно распределенной случайной величины определяются формулами (5.9) и (5.10).

$$f(x) = \begin{cases} 0, & \text{при } x < 0, \\ 3e^{-3x}, & \text{при } x \ge 0, \end{cases} \qquad F(x) = \begin{cases} 0, & \text{при } x < 0, \\ 1 - e^{-3x}, & \text{при } x \ge 0. \end{cases}$$

$$\mathrm{M(X)} = \frac{1}{\lambda} = \frac{1}{3}; \quad D(X) = \frac{1}{\lambda^2} = \frac{1}{9}; \quad \sigma(X) = \frac{1}{\lambda} = \frac{1}{3}.$$

$$P(1 < x < 2) = e^{-3\cdot 1} - e^{-3\cdot 2} = 0{,}04954.$$

Замечание. Показательное распределение играет большую роль в теории массового обслуживания (ТМО), теории надежности. В ТМО λ — среднее число событий, приходящихся на единицу времени. При определенных условиях число событий, произошедших за промежуток времени τ, распределено по закону Пуассона с математическим ожиданием $a = \lambda\tau$. Длина промежутка t между произвольными двумя соседними событиями подчиняется показательному закону:

$$P(T < t) = F(t) = 1 - e^{-\lambda t}.$$

Функцию $R(t) = P(T \ge t) = 1 - F(t)$, называют законом надежности. Если $F(t) = 1 - e^{-\lambda t}$, где λ — интенсивность отказов, то $R(t)$ — показательный закон надежности. ■

Пример 5.3. Рассмотрим функцию распределения неотрицательной случайной величины $T(T \ge 0)$, характеризующей, например, время ожидания или жизни: $F(t) = P(T < t)$.

Пусть $G(t) = 1 - F(t) = P(T \ge t)$ — функция, характеризующая «хвост распределения» вероятностей. Найти закон распределения, который удовлетворяет условию

$$G(t + s) = G(t)G(s)$$

или, иначе,

$$P(T \ge t + s) = P(T \ge s)P(T \ge t + s \,/\, T \ge s). \qquad (5.15)$$

Решение. 1) Пусть рассматриваются дискретные моменты времени T, тогда условие задачи можно переписать в терминах схемы Бернулли, как вероятность того, что наступает некоторое событие A с постоянной вероятностью $P(A) = p$,

либо не наступает с вероятностью $q = 1 - p,$ причем наступление события прекращает опыт. Вероятность наступления события через время $T = s + t$ будет удовлетворять условию

$$P((T = s + t)/\,(T \geq s)) = P(T = t). \tag{5.16}$$

Действительно,

$$P\big((T = s + t)\,/(T \geq s)\big) = \frac{P\big((T=s+t),\ T\geq s\big)}{P(T\geq s)} =$$
$$= \frac{P(T=s+t)}{P(T\geq s)} = \frac{pq^{s+t}}{\sum_{m=s}^{\infty} pq^m} = \frac{q^{s+t}}{\frac{q^s}{1-q}} = pq^t. \tag{5.17}$$

Таким образом, если время (m дискретных единиц) ожидания наступления «успеха» описывается геометрическим распределением вида

$$P(T = m) = pq^m \quad (m = 0, 1, 2, \dots),$$

то условная вероятность события $P((T = s + t)/\,(T \geq s))$ равна безусловной вероятности $P(T = t)$, что означает отсутствие последействия.

Например, если длительность телефонного разговора оценивается целым числом минут со временем ожидания, подчиняющимся геометрическому закону, то вероятность окончания разговора не зависит от предыстории.

2) В случае непрерывной случайной величины, считая, что $G(t) \neq 0$ при всех t, положим, что $s = t = \frac{x}{2}$, имеем

$$G(x) = \left(G\left(\frac{x}{2}\right)\right)^2.$$

Следовательно, $G(x) > 0$. Поэтому, логарифмируя (5.15), получим функциональное уравнение Коши:

$$ln\,G(t + s) = ln\,G(t) + ln\,G(s). \tag{5.18}$$

Из курса математического анализа известно, что решение (5.18) имеет вид:

$$G(t) = e^{-\lambda t} = P(T \geq t), \quad t \geq 0.$$

Действительно, так как $[t + s; b + s] \cap [s; +\infty] = [t + s; b + s]$, где $0 \leq t \leq b$ некоторый промежуток времени, то

$$P(t + s \leq T \leq b + s\ /\ T \geq s) = \frac{P((t+s\leq T\leq b+s)\cap(T\geq s))}{P(T\geq s)} =$$
$$= \frac{P(t+s\leq T\leq b+s)}{P(T\geq s)} = \frac{e^{-\lambda(t+s)} - e^{-\lambda(b+s)}}{e^{-\lambda s}} = e^{-\lambda t} - e^{-\lambda b}, \tag{5.19}$$

но $P(t \leq T \leq b) = e^{-\lambda t} - e^{-\lambda b}$, что и отражает свойство «отсутствия последействия». При $b \to +\infty$ из полученного выше равенства (5.19) следует формула (5.15).

Если $T = mh,$ то время ожидания первого успеха в схеме Бернулли подчиняется геометрическому закону распределения:

$$P(T = m) = pq^{mh}, \tag{5.20}$$

где $h > 0$ — время одного испытания, $m = 0, 1, 2, \dots$

Пусть $t = nh,$ найдем вероятность

$$P(T \geq t) = \sum_{m=n}^{\infty} pq^m = pq^n \frac{1}{1-q} = q^n = e^{-\lambda t},$$

где $e^{-\lambda h} = q$.

Поэтому геометрическое распределение рассматривают как дискретное показательное распределение.

Таким образом, если некоторое явление характеризуется отсутствием памяти (последействия), то распределение вероятностей его длительности описывается геометрическим или показательным законом распределения (что, например, характерно для радиоактивных атомов).

Указанные свойства *отсутствия последействия* (памяти или старения) характерны для цепей Маркова (*свойство марковости*, см. гл. 9). Если рассматривать распределение времени ожидания T, то это может быть время ожидания «поклевки» при ловле на удочку в не зарыбленном пруду, либо окончания телефонного разговора «болтливой дамы» (наступление действия — поклевка или окончание разговора соответственно, никак не связано с предшествующим событием).

5.3. Нормальное распределение

Нормальное распределение вероятностей играет исключительно важную роль в теории вероятностей и математической статистике. Это наиболее часто встречающийся закон распределения различных явлений и процессов, главная особенность которого заключается в том, что он является предельным законом, к которому, при определенных условиях, приближаются другие законы распределения.

Функция плотности распределения вероятностей нормального закона (*функция Гаусса, гауссиан*) имеет вид (рис. 5.4; ось OX — горизонтальная асимптота для $f(x)$ и $F(x)$, прямая $y = 1$ горизонтальная асимптота для $F(x)$):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-a)^2}{2\sigma^2}}. \qquad (5.21)$$

Покажем, что функция $f(x)$ удовлетворяет условию (4.5). Сделаем замену переменной $(x-a)/(\sqrt{2}\sigma) = t$. Тогда

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right) dx = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} exp(-t^2)dt.$$

Интеграл Эйлера — Пуассона:

$$\int_{-\infty}^{+\infty} exp(-t^2)dt = \sqrt{\pi},$$

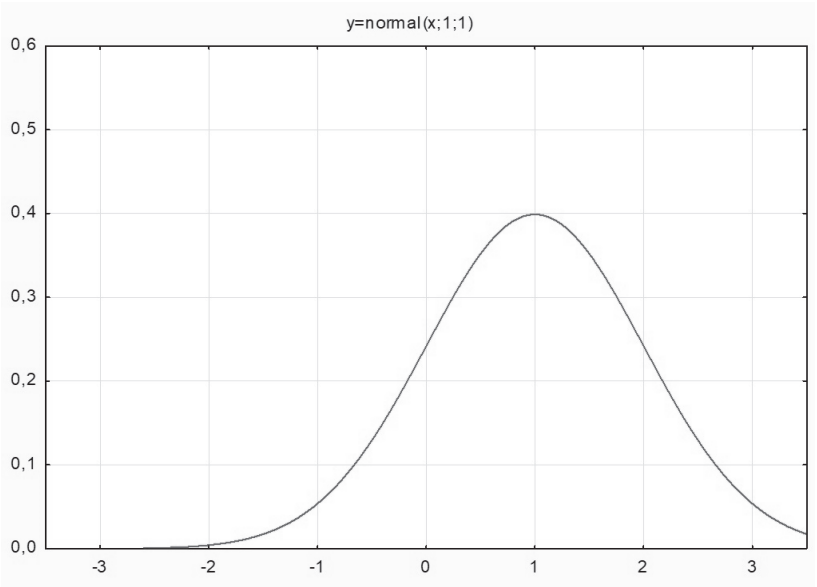
следовательно,

$$\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-a)^2}{2\sigma^2}} dx = 1.$$

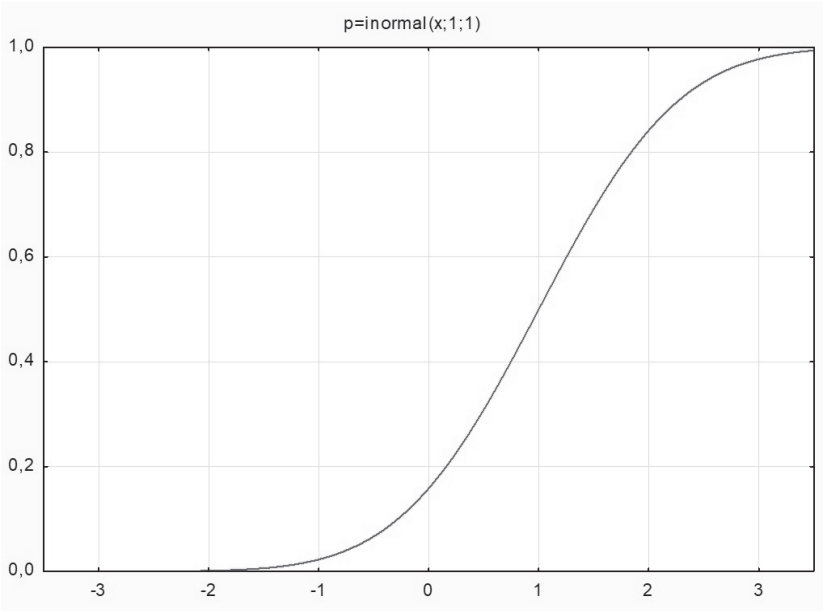
Если a и σ принимают произвольные значения, то распределение (5.21) называется общим и обозначается как $N(a, \sigma^2)$.

Если $a = 0, \sigma = 1$, то распределение называется нормированным ($N(0,1)$), значения которого представлены в приложении 1:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \qquad (5.22)$$



Плотность распределения



Функция распределения

Рис. 5.4 — Нормальный закон распределения $(a = 1, \sigma = 1)$

Замечание. Интеграл Эйлера — Пуассона $I = \int_{-\infty}^{+\infty} \exp(-x^2)\, dx = \sqrt{\pi}$. В силу четности функции под знаком интеграла перейдем к удвоенному интегралу от 0 до $+\infty$:

$$I = 2 \int_0^{+\infty} \exp(-x^2)\, dx.$$

Рассмотрим несобственный двойной интеграл

$$K = \iint_{\sigma} \exp(-(x^2 + y^2))\, d\sigma,$$

где область σ — первая четверть координатной плоскости. Выражая двойной интеграл через повторный, получим

$$K = \iint_{\sigma} \exp(-(x^2 + y^2))\, d\sigma = \int_0^{+\infty} dx \int_0^{+\infty} \exp(-(x^2 + y^2))dy =$$
$$= \int_0^{\infty} \exp(-x^2)\, dx \int_0^{+\infty} \exp(-y^2))dy = \frac{I^2}{4},$$

так как интеграл не зависит от обозначения переменной интегрирования. Для вычисления интеграла K перейдем к полярным координатам.

Пусть $x = \rho sin\varphi, y = \rho cos\varphi$, $\rho = \sqrt{x^2 + y^2}$, учитывая, что элемент площади $d\sigma = dxdy = \rho d\rho d\varphi$ (ρ — якобиан[7] преобразования, см. пример 7.6) и переходя от двойного интеграла к повторному, получим

$$K = \iint_{\sigma} \exp(-(x^2 + y^2))\, d\sigma = \int_0^{\pi/2} d\varphi \int_0^{+\infty} exp\,(-\rho^2)\rho d\rho.$$

Вычислим внутренний интеграл

$$\int_0^{+\infty} \rho \exp(-\rho^2)d\rho = \lim_{b\to+\infty} \int_0^b \rho \exp(-\rho^2)d\rho = -\frac{1}{2}\lim_{b\to+\infty} \int_0^b d(\exp(-\rho^2)) =$$
$$= \lim_{b\to\infty} (\frac{1}{2} - \exp(-b^2)) = \frac{1}{2}.$$

Следовательно,

$$K = \frac{I^2}{4} = \frac{1}{2}\int_0^{\frac{\pi}{2}} d\varphi = \frac{\pi}{4}.$$

Значит,

$$\int_{-\infty}^{+\infty} exp(-x^2)\, dx = \sqrt{\pi}.$$

Легко заметить, что

$$\int_{-\infty}^{+\infty} exp\left(\frac{-x^2}{2}\right) dx = \sqrt{2\pi}. \quad \blacksquare$$

Числовые характеристики нормального закона.

1. Математическое ожидание характеризует центр распределения и равно параметру a:

$$M(X) = \int_{-\infty}^{+\infty} x \frac{1}{\sqrt{2\pi}} exp\left(-\frac{(x-a)^2}{2\sigma^2}\right) dx = a, \qquad (5.23)$$

где $e^x = exp(x)$.

Введем новую переменную $z = \frac{x-a}{\sigma\sqrt{2}}$, тогда $x = z\sigma\sqrt{2} + a, dx = \sigma\sqrt{2}dz$ (по свойству дифференциала — если $y = f(x)$, то $dy = f'(x)dx$).

Следовательно,

$$M(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \left(z\sigma\sqrt{2} + a\right) exp(-z^2)\, \sigma\sqrt{2}dz =$$
$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \left(z\sigma\sqrt{2}\right) exp(-z^2)\, dz + \frac{a}{\sqrt{\pi}} \int_{-\infty}^{+\infty} exp(-z^2)\, dz = 0 + a = a.$$

[7] Коэффициент искажения или коэффициент растяжения плоскости (в данной точке), задаваемой вектором (x,y), при преобразовании его в плоскость, задаваемую полярными координатами (ρ, φ).

Первый интеграл равен нулю как интеграл нечетной функции с симметричными пределами, а второй — интеграл Эйлера — Пуассона равен $\sqrt{\pi}$.

2. Дисперсия характеризует форму распределения:

$$D(X) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}(x - M(x))^2 e^{-\frac{(x-a)^2}{2\sigma^2}}dx = \sigma^2, \qquad (5.24)$$

полагая $x = \sigma y + a$, получим

$$D(X) = \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} y^2 e^{-\frac{y^2}{2}}dy =$$

$$= -\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} y de^{-\frac{y^2}{2}} = -\frac{\sigma^2}{\sqrt{2\pi}}\left(ye^{-\frac{y^2}{2}}\Big|_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}}dy\right) =$$

$$= -\frac{\sigma^2}{\sqrt{2\pi}}\left(\lim_{b\to+\infty} be^{-\frac{b^2}{2}} - \lim_{b\to-\infty} be^{-\frac{b^2}{2}} - \int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}}dy\right) =$$

$$= -\frac{\sigma^2}{\sqrt{2\pi}}\left(0 - 0 - \sqrt{2\pi}\right) = \sigma^2.$$

Среднее квадратическое отклонение равно параметру σ.

Свойства функции плотности распределения нормального закона:

1) область определения: $D_f = R$;

2) ось абсцисс является горизонтальной асимптотой функции;

3) функция имеет максимум в точке с координатами: $(a; \frac{1}{\sigma\sqrt{2\pi}})$;

4) функция имеет две точки перегиба $(x = a \pm \sigma, y = \frac{1}{\sigma\sqrt{2\pi e}})$;

5) график функции симметричен относительно прямой $x = a$;

6) моменты:

$$\mu_1 = \cdots = \mu_{2k+1} = \cdots = 0 —$$

все нечетные центральные моменты равны нулю, $\mu_2 = \sigma^2, \mu_4 = 3\sigma^4$;

$$Ka = \frac{\mu^3}{\sigma^3} = 0; Ex = \frac{\mu_4}{\sigma_4} - 3 = 0 —$$

коэффициент асимметрии и эксцесс равны нулю.

Функция распределения имеет вид

$$F(x) = P(X < x) = \int_{-\infty}^{x}\frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-a)^2}{2\sigma^2}}\, dx. \qquad (5.25)$$

Введя переменную $z = \frac{x-a}{\sigma}$, $x = \sigma z + a, dx = \sigma dz$, получим

$$F(x) = \int_{-\infty}^{\frac{x-a}{\sigma}}\frac{1}{\sqrt{2\pi}}\, e^{-\frac{z^2}{2}}\, dz = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} e^{-\frac{z^2}{2}}\, dz + \frac{1}{\sqrt{2\pi}}\int_{0}^{\frac{x-a}{\sigma}} e^{-\frac{z^2}{2}}\, dz,$$

$$F(x) = \frac{1}{2} + \Phi\left(\frac{x-a}{\sigma}\right), \qquad (5.26)$$

где

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{\frac{z^2}{2}}\, dz — \qquad (5.27)$$

функция распределения стандартного $(a = 0, \sigma = 1)$ *нормального закона* или, иначе, функция Лапласа (*интеграл вероятностей* или интеграл ошибок).

Свойства функции распределения нормального закона:

1) $F(-\infty) = 0$;

2) $F(+\infty) = 1$;

3) $F(-x) = \frac{1}{2} - \Phi(\frac{x-a}{\sigma})$.

С изменением математического ожидания «a» и неизменностью дисперсии нормальная кривая сдвигается вправо с ростом a и влево с уменьшением a. На форму нормальной кривой влияет среднее квадратическое отклонение. При одинаковом математическом ожидании, с увеличением среднего квадратического отклонения нормальная кривая растягивается по оси абсцисс и распределение становится плосковершинным, с его уменьшением распределение сжимается и становится островершинным.

Вероятность попадания нормально распределенной случайной величины в заданный интервал определяется по свойству функции распределения непрерывной случайной величины:

$$P(\alpha < X < \beta) = \Phi\left(\frac{\beta-a}{\sigma}\right) - \Phi\left(\frac{\alpha-a}{\sigma}\right). \quad (5.28)$$

Вероятность заданного отклонения. Правило трех сигм.

Найдем вероятность того, что случайная величина *X*, распределенная по нормальному закону, отклонится от математического ожидания $M(x) = a$ не более чем на величину $\delta > 0$.

$$P(|X - a| < \delta) = P(-\delta < X - a < +\delta) = P(a - \delta < X < a + \delta) =$$

$$= \Phi\left(\frac{a+\delta-a}{\sigma}\right) - \Phi\left(\frac{a-\delta-a}{\sigma}\right) = \Phi\left(\frac{\delta}{\sigma}\right) - \Phi\left(\frac{-\delta}{\sigma}\right) = \Phi\left(\frac{\delta}{\sigma}\right) + \Phi\left(\frac{\delta}{\sigma}\right) = 2\,\Phi\left(\frac{\delta}{\sigma}\right),$$

таким образом,

$$P(|X - a| < \delta) = 2\,\Phi\left(\frac{\delta}{\sigma}\right). \quad (5.29)$$

Найдем вероятность того, что нормально распределенная случайная величина *X* отклонится от *M*(*X*) = *a* на σ, 2σ, 3σ:

$$\mathrm{P}(|\,\mathrm{X} - a|\, < \sigma) = 2\Phi\left(\frac{\sigma}{\sigma}\right) = 2\Phi(1) = 2{\cdot}0{,}3413 = 0{,}6826,$$

$$\mathrm{P}(|\mathrm{X} - a| < 2\sigma) = 2\Phi\left(\frac{2\sigma}{\sigma}\right) = 2\Phi(2) = 2{\cdot}0{,}4772 = 0{,}9544,$$

$$\mathrm{P}(|\mathrm{X} - a|\, < 3\sigma) = 2\Phi\left(\frac{3\sigma}{\sigma}\right) = 2\Phi(3) = 2{\cdot}0{,}49865 = 0{,}9973.$$

Отсюда следует правило 3σ: если случайная величина *X* имеет нормальное распределение, то абсолютное отклонение этой случайной величины от ее математического ожидания не превышает утроенное среднее квадратическое отклонение (3σ) и является событием почти достоверным.

Замечание. Нормальный закон часто иллюстрируют опытом из молекулярной физики с использованием доски Гальтона, представляющей собой ящик с передней прозрачной стенкой. В заднюю стенку в шахматном порядке вбиты параллельные ряды штырей (гвоздей, иголок). Внизу расположены ячейки, вмещающие шарики (зерна пшеницы, песок), вверху воронка (рис. 5.5).

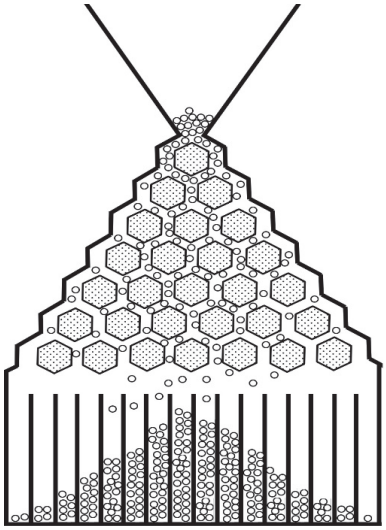


Рис. 5.5 — Доска Гальтона

Шарики (зерна, песок), подаваемые через воронку, испытывая случайные столкновения с препятствиями будут падать вниз, описывая кривую, которая при большом числе шариков будет асимптотически приближаться к функции плотности вероятностей нормального закона распределения (кривой ошибок Гаусса). Теоретическая интерпретация опыта опирается на идею из комбинаторики о числе траекторий (путей) на бесконечном ориентированном графе-решетке (рис. 1.6) и биномиальный закон распределения $Bin(n; 0{,}5)$ (рис. 3.2) — чем ближе к центру, тем шариков больше, а чем ближе к краям — меньше. Наклон доски даст представление о биномиальном законе распределения при $p < 0{,}5$ или $p > 0{,}5$ (рис. 3.2), а увеличение наклона — об экстремальных распределениях (см. 7, раздел 7.3). ■

Пример 5.4. Цена 1 кг конфет в магазинах торговой сети распределяется по нормальному закону. Средняя цена реализации составила 400 руб. за 1 кг, при среднем квадратическом отклонении 150 руб. за 1 кг. Какой процент конфет реализуется с ценой за 1 кг: а) от 250 до 600 руб.; б) свыше 400 руб.; в) менее 560 руб.; г) абсолютное отклонение цены реализации от средней цены не превысит 250 руб.?

Решение. По условию задачи: $a = 400$, $\sigma = 150$.

а) $\alpha = 200, \beta = 600$. Воспользуемся формулой (5.28):

$$P(250 < x < 600) = \Phi\left(\frac{600 - 400}{150}\right) - \Phi\left(\frac{250 - 400}{150}\right) = \Phi(1{,}33) - \Phi(-1) = \\ = \Phi(1{,}33) + \Phi(1).$$

Значения функции Ф(*x*)представлены в приложении 1.

При $x = 1, \Phi(1) = 0{,}3413$, при $x = 1{,}33$, $\Phi(1{,}33) = 0{,}4082$.

$P(250 < x < 600) = 0{,}4082 + 0{,}3413 = 0{,}7495$. Значит, 74,95% конфет в торговой сети реализуется по цене от 250 до 600 руб. за 1 кг.

б) $P(X > 400) = P(400 < x < +\infty) = \Phi\left(\frac{\infty - 400}{150}\right) - \Phi\left(\frac{400 - 400}{150}\right) = \\ = \Phi(+\infty) - \Phi(0) = 0{,}5 + 0 = 0{,}5.$

Значит, 50% конфет реализуется по цене свыше 400 руб./кг.

в) $P(X < 560) = P(0 < x < 560) = \Phi\left(\frac{560-400}{150}\right) - \Phi\left(\frac{0-400}{150}\right) = \Phi(1{,}07) - -\Phi(-2{,}67) = \Phi(1{,}07) + \Phi(2{,}67) = 0{,}3577 + 0{,}4962 = 0{,}8539.$

Следовательно, 85,4% конфет реализуется с ценой до 560 руб./кг.

г) $\delta = 250, P(|X - 400| \leq 250) = 2\,\Phi\left(\frac{250}{150}\right) = 2\,\Phi(1{,}67) = 2 \cdot 0{,}4525 = = 0{,}905.$

Таким образом, 90,5% конфет в торговой сети реализуется с ценой 400 ± 250, т. е. от 150 до 650 руб./кг.

5.4. Логарифмически нормальное распределение

Непрерывная случайная величина X имеет логарифмически нормальное распределение с параметрами μ и σ, если $X = e^Y$, причем $Y = \ln X$ — случайная величина, распределенная по нормальному закону с параметрами μ и σ. Логарифмически нормальная случайная величина принимает только положительные значения.

Если $X < x$, то $\ln X < lnx$, используя формулу (5.26), получим

$$F(x) = P(X < x) = P(lnX < lnx) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{lnx} e^{-\frac{(z-lna)^2}{2\sigma^2}}\, dz,$$

$$F(x) = \frac{1}{2} + \Phi\left(\frac{lnx-lna}{\sigma}\right). \qquad (5.30)$$

Плотность распределения вероятностей логнормального распределения имеет вид

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{(lnx-lna)^2}{2\sigma^2}}. \qquad (5.31)$$

График плотности вероятности и функции распределения непрерывной случайной величины X, распределенной по логарифмически нормальному закону при $\mu = 1, \sigma = 0{,}4$, представлен на рисунке 5.6.

Чем больше σ, тем значительнее различаются математическое ожидание, мода и медиана распределения. Каждая кривая логнормального распределения имеет один максимум, соответствующий модальному значению. Асимметрия распределения возрастает с увеличением значений *a* и σ.

Числовые характеристики непрерывной случайной величины *X,* распределенной по логарифмически нормальному закону, определяются по следующим формулам:

математическое ожидание $M(X) = ae^{\frac{\sigma^2}{2}}$;

дисперсия $D(X) = a^2 e^{\sigma^2}(e^{\sigma^2} - 1)$;

мода $Mo(X) = ae^{-\sigma^2}$;

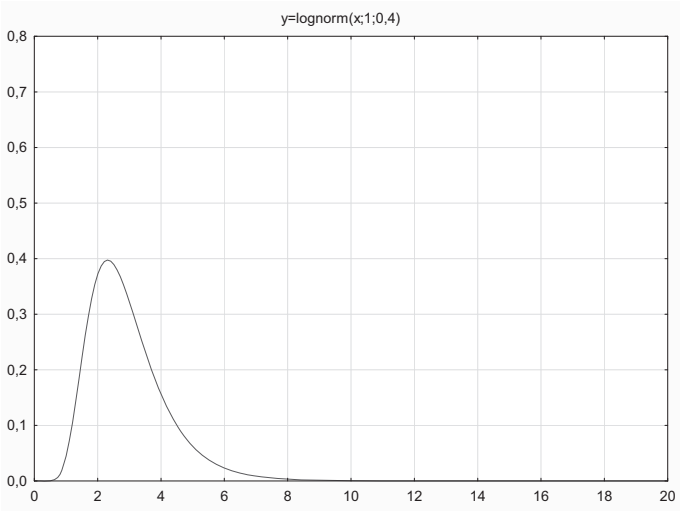
медиана $Me(X) = a$.

В качестве среднего значения в нормальном распределении служит параметр *a*, в логарифмически нормальном распределении этот параметр является медианой.

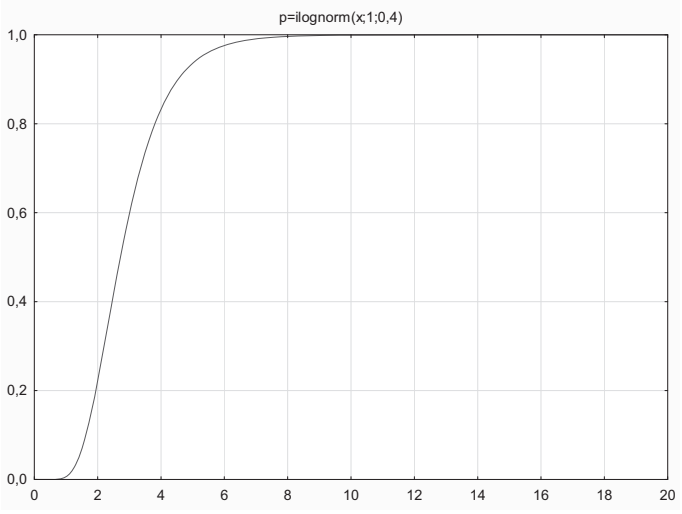
На практике логарифмически нормальное распределение встречается довольно часто. По данному закону распределяются: доходы населения; предприятия сельского хозяйства по земельной площади, выручке от реализации продукции, стоимости основных фондов, численности работников; размеры вкладов населения в банках; цены на однородные товары и т. п.

Вероятность попадания непрерывной случайной величины в заданный интервал определяется по формуле

$$P(b < x < c) = P(lnb < lnx < lnc) = \Phi\left(\frac{lnc - \overline{lnx}}{\sigma_{lnx}}\right) - \Phi\left(\frac{lnb - \overline{lnx}}{\sigma_{lnx}}\right). \quad (5.32)$$



Плотность распределения



Функция распределения

Рис. 5.6 — Логарифмически-нормальное распределение ($\mu = 1, \sigma = 0{,}4$)

Пример 5.5. Личные подсобные хозяйства населения региона по размеру площади земельных участков распределяются по логарифмически нормальному закону с параметрами $a = 12{,}62$ сотки земли и $\sigma^2 = 0{,}478$; $\overline{lnx} = 2{,}654$.

Определить: а) средний размер приусадебного участка; б) дисперсию и среднее квадратическое отклонение; в) медианное и модальное значения площади участков; г) какой процент личных подсобных хозяйств населения имеет площадь от 6 до 15 соток.

Решение. а) $M(X) = ae^{\frac{\sigma^2}{2}} = 12{,}62e^{0{,}478/2} = 16{,}03$.

б) $D(X) = a^2e^{\sigma^2}\left(e^{\sigma^2} - 1\right) = 12{,}62^2e^{0{,}478}(e^{0{,}478} - 1) = 157{,}476$;

$\sigma = \sqrt{157{,}476} = 12{,}549$;

в) $Me(X) = a = 12{,}62$; $Mo(X) = ae^{-\sigma^2} = 12{,}62e^{-0{,}478} = 7{,}81$;

г) $P(6 < x < 15) = P(ln6 < lnx < ln15) = P(1{,}792 < lnx < 2{,}708) =$

$= \Phi\left(\frac{2{,}708-2{,}654}{0{,}694}\right) - \Phi\left(\frac{1{,}792-2{,}654}{0{,}694}\right) = \Phi(0{,}0778) - \Phi(-1{,}242) = \Phi(0{,}08) +$

$+\Phi(1{,}24) = 0{,}0359 + 0{,}3925 = 0{,}42845$.

Значит, средний размер одного приусадебного участка составил 16,03 соток земли. Размер участков в среднем колебался в границах $16{,}03 \pm 12{,}55$, т. е. от 3,48 до 28,58 соток. Половина приусадебных участков населения имела площадь до 12,62 соток, а другая половина свыше 12,62 соток. Наиболее часто встречаются участки с площадью 7,8 соток. Вероятность того, что случайно взятый участок имеет площадь от 6 до 15 соток земли составляет 0,428, т. е. 42,8% участков имеет площадь землепользования от 6 до 15 соток.

Темы (вопросы) для самоконтроля

1. Равномерный (прямоугольный) закон распределения и его свойства.
2. Показательное распределение и его свойства.
3. Законы распределения с отсутствием последействия.
4. Нормальный закон распределения и его свойства.
5. Доска Гальтона и другие иллюстрации нормального распределения.
6. Правило трех сигм.
7. Логарифмически-нормальное распределение и его свойства.

Глава 6
Система двух случайных величин

6.1. Понятие и закон распределения двумерной случайной величины

В практических задачах приходится сталкиваться со случаями, когда результат описывается двумя и более случайными величинами ($X_1,\ X_2, \dots,\ X_n$). Если они рассматриваются вместе, то образуют систему случайных величин (случайный вектор). Например: вес и рост человека; цена и себестоимость единицы продукции; точка попадания снаряда имеет две координаты: *x* и *y*, которые можно принять за систему случайных величин (x, y), определенных на одном и том же пространстве элементарных событий Ω. Многомерная случайная величина есть функция элементарных событий. Примером многомерной случайной величины может быть стаж работы, возраст и месячная заработная плата работника и др.

Случайные величины, входящие в систему, могут быть как дискретными, так и непрерывными. Если одномерная случайная величина задается одним числом, то двумерная — парами чисел. Двумерная случайная величина определяется на одном пространстве элементарных событий и обозначается (X, Y), а ее возможные значения (x, y).

Закон распределения *дискретной двумерной случайной* величины можно представить в виде таблицы, характеризующей совокупность всех пар значений случайных величин $\left(X = x_i, Y = y_j\right)$ и соответствующих вероятностей

$$p\left(x_i, y_j\right), i = 1, 2, \dots, n; j = 1, 2, \dots, m.$$

Случайные величины X и Y называются составляющими системы случайных величин (X, Y). Так как события $(X = x_i; Y = y_j)$ образуют полную группу, то

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) = 1.$$

Если число возможных значений, составляющих X и Y, конечно, то двумерная случайная величина (X, Y) представляется в виде таблицы 6.1.

Таблица 6.1

Закон распределения двумерной дискретной случайной величины

X	Y						$\sum p(x_i)$
	y_1	y_2	…	y_j	…	y_m	
x_1	$p(x_1, y_1)$	$p(x_1, y_2)$	…	$p(x_1, y_j)$	…	$p(x_1, y_m)$	$p(x_1)$
x_2	$p(x_2, y_1)$	$p(x_2, y_2)$	…	$p(x_2, y_j)$	…	$p(x_2, y_m)$	$p(x_2)$
…	…	…	…	…	…	…	…
x_i	$p(x_i, y_1)$	$p(x_i, y_2)$		$p(x_i, y_j)$	…	$p(x_i, y_m)$	$p\ (x_i)$
…	…	…	…	…	…	…	…
x_n	$p(x_n, y_1)$	$p(x_n, y_2)$	…	$p(x_n, y_j)$	…	$p(x_n, y_m)$	$p(x_n)$
$\sum p(y_j)$	$p\ (y_1)$	$p\ (y_2)$	…	$p\ (y_j)$	…	$p\ (y_m)$	1

Итоговые столбец и строка представляют вероятности составляющих X и Y (*маргинальные распределения*):

$$p(x_i) = \sum_{j=1}^{m} p(x_i, y_j);\ \ p(y_j) = \sum_{i=1}^{n} p(x_i, y_j); \tag{6.1}$$

$$\sum p(y_j) = 1;\ \sum p(x_i) = 1.$$

По таблице распределения можно найти условные законы распределения составляющих, характеризующих законы распределения одной составляющей при фиксированном значении другой составляющей.

Условным распределением составляющей X называется совокупность значений случайной величины X и условных вероятностей

$$p(x_i/y_j), i = 1, 2, \dots, n;\ j = 1, 2, .., m,$$

где $p(x_i/y_j)$ — это условная вероятность того, что случайная величина X примет i −е значение при условии, что случайная величина Y приняла j −ое значение.

Аналогично, $p(y_j/x_i)$ — условная вероятность того, что случайная величина Y примет j −е значение при условии, что случайная величина X приняла i −е значение. Условные вероятности находятся по формулам:

$$p(x_i/y_j) = \frac{p(x_i, y_j)}{p(y_j)},\ p(y_j/x_i) = \frac{p(x_i, y_j)}{p(x_i)}, \tag{6.2}$$

$$\sum_{i=1}^{n} p(x_i/y_j) = 1, \sum_{j=1}^{m} p(y_j/x_i) = 1.$$

Условное математическое ожидание случайной величины X при условии, что $Y = y_j$ определяется как

$$M(X/\ Y = y_j) = \sum_{i=1}^{n} x_i \frac{p(x_i, y_j)}{p(y_j)} = \sum_{i=1}^{n} x_i p(x_i/y_j). \tag{6.3}$$

Условное математическое ожидание имеет следующие свойства.

1. $M(\varphi(Y)/Y) = \varphi(Y)$.
2. $M(\varphi(Y)X/Y) = \varphi(Y)M(X/Y)$.
3. $M(aX_1 + bX_2/Y) = aM(X_1/Y) + bM(X_2/Y)$, a и b константы.
4. Правило повторного математического ожидания.

$$M(M(Y/X)/g(X)) = M(Y/g(X)),$$

$$M(M(Y/X)) = M(Y).$$

5. Если случайные величины X и Y независимы, то

$$M(X/Y) = M(X).$$

Можно записать формулу полного математического ожидания:

$$M(X) = \sum_{j=1}^{m} p(Y = y_j)\, M(X/\ Y = y_j).$$

Если случайные величины X, Y независимы, то для любой их комбинации выполняется равенство $p(x_i y_j) = p(x_i)p(y_j)$.

Если на одном пространстве элементарных событий рассматривается *система случайных величин* $(X, Y, \dots, W)$, то эти величины *взаимно независимы, если для любой комбинации их значений верно равенство*

$$P(X = x, Y = y, \dots, W = w) = P(X = x)P(Y = y) \dots P(W = w).$$

6.2. Функции распределения и плотности вероятности двумерной случайной величины

В общем случае двумерная случайная величина задается в виде функции распределения: $F(x, y) = P(X < x, Y < y)$, которая означает вероятность попадания двумерной случайной величины в квадрант левее и ниже точки с координатами $(x; y)$.

Свойства функции распределения:

1) значения функции распределения заключены между нулем и единицей:

$$0 \le F(x, y) \le 1;$$

2) функция распределения не убывает и непрерывна слева по каждому аргументу, т. е. если $x_2 > x_1$, то $F(x_2, y) \ge F(x_1, y)$, и если $y_2 > y_1$, то

$$F(x, y_2) \ge F(x, y_1);$$

3) $F(x, +\infty) = F_1(x)$ — функция распределения случайной величины X; $F(+\infty, y) = F_2(y)$ — функция распределения случайной величины Y;

4) $F(-\infty, y) = F(x, -\infty) = F(-\infty, -\infty) = 0$; $F(+\infty, +\infty) = 1$.

Вероятность попадания двумерной случайной величины в прямоугольник находится исходя из определения функции распределения двумерной случайной величины (рис. 6.1):

$$P(x_1 \le X \le x_2, y_1 \le Y \le y_2) = \\ = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1). \qquad (6.4)$$

Случайные величины X, Y *независимы*, если

$$F(x, y) = F_1(x)F_2(y), \qquad (6.5)$$

где $F_1(x)$ и $F_2(y)$ — функции распределения составляющих.

Плотность вероятности системы двух непрерывных случайных величин определяется как вторая смешанная частная производная ее функции распределения:

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} = F''(x, y). \qquad (6.6)$$

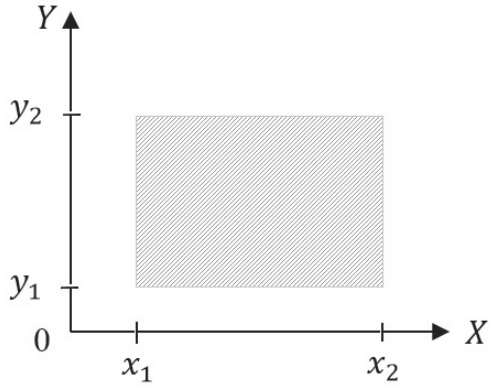


Рис. 6.1 — Вероятность попадания точки (x, y) в прямоугольник

Свойства функции плотности вероятности:

1) плотность вероятности двумерной случайной величины неотрицательна:

$$f\ (x,y) \geq 0; \tag{6.7}$$

2) двойной несобственный интеграл двумерной случайной величины с бесконечными пределами равен единице:

$$\iint_{-\infty-\infty}^{+\infty+\infty} f(x,y)dxdy = 1; \tag{6.8}$$

3) функция распределения двумерной случайной величины через плотность распределения выражается формулой

$$F(x,y) = \iint_{-\infty-\infty}^{x\ y} f(x,y)dxdy. \tag{6.9}$$

Геометрически свойство 2 означает, что объем тела, ограниченного поверхностью $f(x,y)$ и плоскостью XOY, равен 1.

Если случайные величины X и Y независимы, то

$$f(x,y) = f_1(x)f_2(y), \tag{6.10}$$

где $f_1(x) = F_1'(x), f_2(y) = F_2'(y)$, — плотности распределения составляющих X и Y.

В противном случае

$$f(x,y) = f_1(x)f(y/x) \quad \text{или} \quad f(x,y) = f_2(y)f(x/y), \tag{6.11}$$

где $f(x/y) = \frac{f(x,y)}{f_1(x)}$ — условная плотность распределения случайной величины Y при заданном значении случайной величины $X = x$; $f(x/y) = \frac{f(x,y)}{f_2(y)}$ – условная плотность распределения случайной величины X при заданном значении $Y = y$.

Маргинальные распределения случайных величин X и Y, согласно свойствам функции двумерного распределения, соответственно равны

$$F_1(x) = \iint_{-\infty-\infty}^{x\ +\infty} f(x,y)dydx, \quad F_2(y) = F(+\infty,y) = \\ = \iint_{-\infty-\infty}^{y\ +\infty} f(x,y)dxdy. \tag{6.12}$$

Учитывая, что производная функции с переменным верхним пределом равна подынтегральному выражению, получим

$$\frac{dF_1(x)}{dx} = f_1(x) = \int_{-\infty}^{+\infty} f(x,y)dy, \quad \frac{dF_2(y)}{dx} = f_2(y) = \int_{-\infty}^{+\infty} f(x,y)dx \text{ —} \tag{6.13}$$

плотности вероятности одномерных непрерывных случайных величин X и Y, входящих в систему, характеризующих предельную (*маргинальную*) вероятность.

Условное математическое ожидание непрерывной случайной величины X при условии, что $Y = y$ определяется как

$$M(X/\ Y = y) = \int_{-\infty}^{+\infty} xf(x/Y = y)dx \tag{6.14}$$

и имеет те же свойства, что и в дискретном случае (6.3).

Если известна плотность вероятности двумерной случайной величины, то вероятность ее попадания в область D определяется по формуле

$$P\big((X,Y) \subset D\big) = \iint_D f(x,y)dxdy. \tag{6.15}$$

6.3. Числовые характеристики системы двух случайных величин. Коэффициент корреляции

Начальным моментом порядка s, h системы двух случайных величин X, Y называется математическое ожидание произведения степени s случайной величины X и степени h случайной величины Y:

$$a_{s,h} = M(X^s Y^h). \tag{6.16}$$

Центральным моментом порядка s, h системы случайных величин (X, Y) называется математическое ожидание произведения степеней s, h соответствующих центрированных случайных величин:

$$\mu_{s,h} = M\left(\dot{X}^s \dot{Y}^h\right), \tag{6.17}$$

где $\dot{X} = X - M(X), \dot{Y} = Y - M(Y)$ — центрированные случайные величины X и Y.

Основным моментом порядка s, h системы случайных величин (X, Y) называется нормированный центральный момент порядка s, h:

$$\rho_{s,h} = \frac{\mu_{s,h}}{\sigma_x^s \sigma_y^h}. \tag{6.18}$$

Начальные моменты $a_{1,0}, a_{0,1}$:

$$\alpha_{1,0} = M(X^1 Y^0) = M(X) = m_x, \alpha_{0,1} = M(X^0 Y^1) = M(Y) = m_y.$$

Вторые центральные моменты:

$$\mu_{2,0} = M\left(\dot{X}^2 \dot{Y}^0\right) = M\left(\dot{X}^2\right) = M\left(X - M(X)\right)^2 = D(X) = \sigma_x^2 —$$

характеризует рассеяние случайных величин в направлении оси *0X*.

$$\mu_{0,2} = M\left(\dot{X}^0 \dot{Y}^2\right) = M\left(\dot{Y}^2\right) = M(Y - M(Y))^2 = D(Y) = \sigma_y^2 —$$

характеризует рассеяние случайных величин в направлении оси *0Y*.

Особую роль в качестве характеристики совместной вариации случайных величин *X* и *Y* играет второй смешанный центральный момент, который называется корреляционным моментом (ковариацией):

$$\mu_{1,1} = M\left(\dot{X}^1 \dot{Y}^1\right) = cov(X, Y) = M(XY) - M(X)M(Y) = m_{xy}. \tag{6.19}$$

Корреляционный момент является мерой связи случайных величин. *Если случайные величины* X и Y *независимы, то математическое ожидание произведения случайных величин равно произведению их математических ожиданий:*

$$M(XY) = M(X)M(Y), \tag{6.20}$$

отсюда $cov\,(X, Y) = 0$.

Случайные величины

$$\rho_{1,0} = \frac{X - m_x}{\sigma_x}, \qquad \rho_{0,1} = \frac{Y - m_y}{\sigma_y} \tag{6.21}$$

называют *стандартизированными* (*нормированными*), так как их математические ожидания равны нулю, а дисперсии — единице.

Если ковариация случайных величин не равна нулю, то случайные величины коррелированы. Ковариация может принимать значения на всей числовой

оси, поэтому в качестве меры связи используют основной момент порядка $s = 1, h = 1$, который называют *коэффициентом корреляции*:

$$\rho_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y}. \tag{6.22}$$

Пример 6.1. Докажем, что если случайные величины X и Y линейно зависимы, то коэффициент корреляции равен ± 1.

Решение. Пусть между случайными величинами X и Y имеет место связь, выражаемая линейным уравнением $Y = AX + B$,
где $M(X) = a,\ D(X) = \sigma^2,\ A$ и B — константы. Тогда имеем:

$$M(Y) = M(AX + B) = A\,M(X) + B = Aa + B,$$

$$D(Y) = D(AX + B) = D(AX) + D(B) = A^2\,D(X) = A^2\sigma^2,$$

следовательно,

$$\sigma(Y) = \sqrt{A^2\sigma^2} = |A|\sigma,$$

$$cov(X,Y) = \mathrm{M}\big((X - a)(Y - A\,a - B)\big) =$$

$$= M\left((X - a)(AX + B - A\,a - B)\right) = AM\left((X - a)^2\right) = A\,D(X) = A\sigma^2.$$

Отсюда $\rho_{xy} = \frac{A\sigma^2}{\sigma(|A|\sigma)} = \frac{A}{|A|} = \pm 1$, что и требовалось доказать.

Покажем, что коэффициент корреляции не превышает по модулю единицы. Для этого, предполагая, что зависимость между случайными величинами X и Y не является в точности линейной, рассмотрим случайную величину $\varepsilon = Y - AX - B$ и найдем ее дисперсию, учитывая свойство дисперсии (3.9):

$$D(\varepsilon) = D(Y - AX - B) = D(Y) - 2cov(X,Y)A + D(X)A^2. \tag{6.23}$$

Дисперсия неотрицательна, следовательно, выражение (6.23) больше либо равно нулю. Поэтому, если рассмотреть (6.23) как квадратное уравнение относительно параметра A, то его дискриминант будет меньше либо равен нулю:

$$[cov(X,Y)]^2 - D(X)D(Y) \le 0, \tag{6.24}$$

отсюда

$$\rho_{xy}^2 = \left[\frac{cov(x,y)}{\sigma_x \sigma_y}\right]^2 \le 1.$$

Свойства коэффициента корреляции можно получить, рассмотрев математическое ожидание квадрата разности (суммы) стандартизированных случайных величин X и Y (6.18), учитывая его неотрицательность:

$$M\left(\frac{X - m_x}{\sigma_x} \pm \frac{Y - m_y}{\sigma_y}\right)^2 \ge 0, \tag{6.25}$$

отсюда

$$M\left(\frac{X - m_x}{\sigma_x}\right)^2 \pm 2M\left(\left(\frac{X - m_x}{\sigma_x}\right)\left(\frac{Y - m_y}{\sigma_y}\right)\right) + M\left(\frac{Y - m_y}{\sigma_y}\right)^2 \ge 0$$

или

$$2 \pm 2\rho_{xy} \ge 0.$$

Откуда следует, что

$$-1 \le \rho_{xy} \le 1. \tag{6.26}$$

Заметим, что $\rho_{xy} = \pm 1$ только в случае, если в (6.23) выполняется равенство, поэтому

$$\frac{Y-m_y}{\sigma_y} = \pm \frac{X-m_x}{\sigma_x}$$

или

$$Y = m_y \mp \frac{\sigma_y}{\sigma_x} m_x \pm \frac{\sigma_y}{\sigma_x} X. \qquad (6.27)$$

То есть если $\rho_{xy} = +1$, то система случайных величин (X, Y) лежит на возрастающей прямой, если $\rho_{xy} = -1$, то на убывающей прямой.

Равенство коэффициента корреляции единице, как было показано ранее, возможно только при линейной зависимости. Значит, если между случайными величинами X и Y существует линейная функциональная связь, то коэффициент корреляции равен $\pm$ 1, в других случаях $\rho_{xy}^2 < 1$.

Таким образом, коэффициент корреляции служит мерой *линейной зависимости* между случайными величинами.

Свойства коэффициента корреляции:

1) значения коэффициента корреляции принадлежат отрезку [–1; 1], т. е.

$$-1 \le \rho_{xy} \le 1;$$

2) если $\rho_{xy} = \pm 1$, то случайные величины линейно зависимы;

3) если $\rho_{xy} = 0$, то случайные величины не коррелированы, что не означает их независимости вообще.

Первые моменты случайных величин:

а) дискретных:	б) непрерывных:
$M(X) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_i p_{ij}$,	$M(X) = \iint_{-\infty-\infty}^{+\infty+\infty} x f(x,y) dxdy,$
$M(Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} y_j p_{ij}$,	$M(Y) = \iint_{-\infty-\infty}^{+\infty+\infty} y f(x,y) dxdy,$
$D(X) = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_i - M(X))^2 p_{ij},$	$D(X) = \iint_{-\infty-\infty}^{+\infty+\infty} (x - M(X))^2 f(x,y) dxdy,$
$D(Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} (y_j - M(Y))^2 p_{ij},$	$D(Y) = \iint_{-\infty-\infty}^{+\infty+\infty} (y - M(Y))^2 f(x,y) dxdy,$
$cov(X,Y) =$ $= \sum_{i=1}^{n} \sum_{j=1}^{m} (x_i - M(X))(y_j - M(Y)) p_{ij}$	$cov(X,Y) =$ $= \iint_{-\infty-\infty}^{+\infty+\infty} \big(x - M(X)\big)\big(y - M(Y)\big) f(x,y) dxdy$

Замечание. 1) Если случайные величины X и Y подчиняются нормальному закону распределения, то некоррелированность случайных величин X и Y означает их независимость.

2) Обобщая механистические рассуждения 4.3 (см. замечание) для дискретных случайных величин в двумерном случае, рассмотрим систему случайных величин (*X, Y*) как множество точек плоскости (x_i, y_j) с массами p_{ij}. Тогда (m_x, m_y) — координаты центра тяжести системы точек (x_i, y_j).

3) Вторые центральные моменты

$\mu_{2,0} = D(X) = \sigma_{11} = \sigma_x^2 = I_{m_x}$ и $\mu_{0,2} = D(Y) = \sigma_{22} = \sigma_y^2 = I_{m_y}$ —

моменты инерции системы точек (x_i, y_j) при вращении относительно осей, проходящих через центр тяжести перпендикулярно осям *OX* и *OY* соответственно.

$$\mu_{1,1} = cov(X,Y) = \sigma_{12} = \sigma_{21} = \sigma_{xy} = I_{m_x m_y} = K_{xy} —$$

центробежный (корреляционный) момент инерции относительно оси, проходящей через центр тяжести системы точек (x_i, y_j). В механике из моментов формируют и изучают матрицу (тензор инерции системы точек).

$$J = \begin{pmatrix} I_{m_x} & -I_{m_x m_y} \\ -I_{m_x m_y} & I_{m_y} \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & -\sigma_{xy} \\ -\sigma_{xy} & \sigma_y^2 \end{pmatrix}.$$

В теории вероятностей и математической статистике рассмотрение моментов сводится к формированию ковариационной матрицы

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \sigma_x\sigma_y\rho \\ \sigma_x\sigma_y\rho & \sigma_y^2 \end{pmatrix}, \quad (6.28)$$

для которой обратная имеет вид

$$\Sigma^{-1} = \frac{1}{\sigma_x^2\sigma_y^2 - \sigma_{xy}\sigma_{xy}} \begin{pmatrix} \sigma_y^2 & -\sigma_{xy} \\ -\sigma_{xy} & \sigma_x^2 \end{pmatrix} = \frac{1}{1-\rho^2} \begin{pmatrix} \frac{1}{\sigma_x^2} & -\frac{\rho}{\sigma_x\sigma_y} \\ -\frac{\rho}{\sigma_x\sigma_y} & \frac{1}{\sigma_y^2} \end{pmatrix},$$

где величина $\rho = \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$ — коэффициент корреляции между X и Y.

Инвариантами ковариационной матрицы Σ являются ее определитель ($|\Sigma|$) и след ($Sp = \sigma_x^2 + \sigma_y^2$).

Матрица

$$R = \begin{pmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{pmatrix} \quad (6.29)$$

называется корреляционной.

Двумерный случай (моментов и ковариационных матриц) естественным образом обобщается на многомерный.

4) Для k-мерной нормально распределенной случайной величины $X(x_1, x_2, \dots, x_k), X \in R^k$ с вектором математических ожиданий $M(\mu_1, \mu_2, \dots, \mu_k)$, ковариационной матрицей Σ

$$\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kk} \end{bmatrix}, \quad (6.30)$$

обратной ковариационной матрицей Σ^{-1} и корреляционной матрицей

$$R = \begin{pmatrix} 1 & \cdots & \rho_{1k} \\ \vdots & \ddots & \vdots \\ \rho_{k1} & \cdots & 1 \end{pmatrix}, \quad (6.31)$$

плотность распределения имеет вид

$$f(X) = [(2\pi)^k|\Sigma|]^{-1/2} \exp\left[-\frac{1}{2}(X-M)^T\Sigma^{-1}(X-M)\right]. \quad (6.32)$$

5) Рассмотрев случай *k=2,* получим плотность двумерного нормального закона распределения $f(x,y) =$

$$= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{(1-\rho^2)}} exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-m_x)^2}{\sigma_x^2} - 2\rho\frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \frac{(y-m_y)^2}{\sigma_y^2}\right)\right]. \quad (6.33)$$

6) В терминах многомерного (в том числе и бесконечномерного) евклидова пространства L, элементами которого являются случайные вектора, для любых двух ненулевых векторов $X, Y \in L$ вводится скалярное произведение

$$(X,Y) = |X||Y|cos\varphi,$$

где

$$(X,X) = |X|^2 = \sigma_x^2,\ (Y,Y) = |Y|^2 = \sigma_y^2.$$

Поэтому длины векторов $|X|, |Y|$ совпадают с соответствующими средними квадратическими отклонениями σ_x и σ_y.

Таким образом,

$$\rho_{xy} = \frac{(X,Y)}{|X||Y|} = \cos\varphi. \qquad (6.34)$$

То есть коэффициент корреляции $\rho = \rho_{xy}$ равен косинусу между многомерными векторами. Если $\rho = 0$, то вектора перпендикулярны, если $\rho = +1$, то вектора сонаправлены, если $\rho = -1$, то вектора противоположно направлены.

7) Формула (6.24) представляет собой известное в математическом анализе неравенство Коши — Шварца — Буняковского для двух случайных величин X и Y, которое для двух векторов $X(x_1, x_1, \dots, x_n), Y(y_1, y_1, \dots, y_n) \in L$ произвольного евклидова пространства имеет вид

$$|(X,Y)| \le |X||Y| \qquad (6.35)$$

или

$$\sqrt{\sum_i x_i y_i} \le \sqrt{\sum_i x_i{}^2}\sqrt{\sum_i y_i{}^2}. \qquad (6.36)$$

В непрерывном случае (квадрируемых в евклидовом пространстве функций), неравенство Коши — Шварца — Буняковского имеет вид

$$(\int f(x)g(x)dx)^2 \le \int f^2(x)dx \int g^2(x). \qquad (6.37)$$

Неравенство Коши — Шварца — Буняковского является обобщением известного в геометрии неравенства треугольника, равенство достигается в случае коллинеарности векторов X и Y. ■

Пример 6.2. Двумерная дискретная случайная величина задана таблицей совместного распределения. Составить условные законы распределения составляющих: случайной величины X, если случайная величина Y примет значение *15;* случайной величины Y, если случайная величина X приняла значение *6.* Определить основные числовые характеристики двумерной случайной величины (X, Y).

Решение. Случайная величина X при условии, что случайная величина y=15 $(X/y = 15)$ принимает значения 4; 6; 8 с вероятностями $\frac{0,05}{0,45}; \frac{0,3}{0,45}; \frac{0,1}{0,45}$.

X	Y			$p(x)$
	10	15	20	
4	0,2	0,05		0,25
6	0,1	0,3	0,1	0,5
8		0,1	0,15	0,25
$p(y)$	0,3	0,45	0,25	1,0

Тогда условный закон распределения случайной величины $X/Y = 15$ будет иметь следующий вид.

$X/y = 15$	4	6	8
$p(x/y = 15)$	$\frac{1}{9}$	$\frac{2}{3}$	$\frac{2}{9}$

Случайная величина $Y/x = 6$ принимает значения $10; 15; 20$ с вероятностями $\frac{0,1}{0,5}; \frac{0,3}{0,5}; \frac{0,1}{0,5}$. Тогда условный закон распределения случайной величины $Y/x = 6$ будет иметь следующий вид.

$Y/x = 6$	10	15	20
$p(y/x = 6)$	0,2	0,6	0,2

Найдем математические ожидания, дисперсии и средние квадратические отклонения случайных величин X и Y.

$$M(Y) = 10 \cdot 0,3 + 15 \cdot 0,45 + 20 \cdot 0,25 = 14,75;$$

$$D(\mathrm{Y}) = 10^2 \cdot 0,3 + 15^2 \cdot 0,45 + 20^2 \cdot 0.25 - 14,75^2 = 13,6875;$$

$$\sigma_y = \sqrt{13,6875} = 3,7.$$

$$M(X) = 4 \cdot 0,25 + 6 \cdot 0,5 + 8 \cdot 0,25 = 6;$$

$$D(X) = 4^2 \cdot 0,25 + 6^2 \cdot 0,5 + 8^2 \cdot 0,2 - 36 = 2;\ \sigma_x = \sqrt{2} = 1,414;$$

$$M(XY) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j p_{x_i y_j} = 4 \cdot 10 \cdot 0,2 + 4 \cdot 15 \cdot 0,05 + 6 \cdot 10 \cdot 0,1 +$$

$$+ 6 \cdot 15 \cdot 0,3 + 6 \cdot 20 \cdot 0.1 + 8 \cdot 15 \cdot 0,1 + 8 \cdot 20 \cdot 0,15 = 92;$$

$$r_{yx} = \frac{M(XY) - M(X)M(Y)}{\sigma_x \sigma_y} = \frac{92 - 6 \cdot 14,75}{1,414 \cdot 3,7} = 0,669.$$

Значение коэффициента корреляции показывает, что связь между переменными довольно сильная.

Темы (вопросы) для самоконтроля

1. Система двух случайных величин (случайный вектор).
2. Закон распределения и числовые характеристики двумерных дискретных случайных величин.
3. Маргинальное распределение и его свойства.
4. Условное распределение вероятностей.
5. Условные функции (плотности) распределения вероятностей.
6. Условное математическое ожидание и его свойства.
7. Вероятность попадания в область.
8. Моменты двумерных случайных величин.
9. Ковариация.
10. Коэффициент корреляции и его свойства.
11. Двумерный нормальный закон распределения.
12. Геометрический смысл коэффициента корреляции.

Глава 7
Функции случайных величин

7.1. Закон распределения функции случайных величин и генерация случайных чисел (сэмплирование)

Часто возникают задачи, когда необходимо найти закон распределения одной случайной величины, зная закон распределения другой, предполагая, что они связаны функционально. Если каждому значению одной случайной величины X соответствует одно определенное значение другой случайной величины Y, то случайная величина Y называется функцией одного случайного аргумента X, т. е. $Y = \varphi(X)$.

Если случайная величина X дискретная, то и случайная величина Y также дискретная. Пусть дискретная случайная величина X принимает значения $x_i,\ \ i = 1,2, \dots, n$ с вероятностями p_i. Тогда случайная величина $Y = \varphi(X)$ будет принимать значения $y_i = \varphi(x_i)$ с вероятностями $P(Y = y_i) = P(X = x_i)$.

Числовые характеристики случайной величины Y находятся по следующим формулам:

$$M(Y) = \sum_{i=1}^{n} y_i p_i = \sum_{i=1}^{n} \varphi(x_i) p_i; \tag{7.1}$$

$$D(Y) = M(y - M(Y))^2 = M(Y^2) - M^2(Y). \tag{7.2}$$

Пример 7.1. Дискретная случайная величина X задана таблицей.

X	–1	1	2	3
p	0,1	0,3	0,5	0,1

Найти $M(Y)$, $D(Y)$ и $\sigma(Y)$, если $Y = 2X^2 + 1$.

Решение. Случайная величина Y примет значения:
$y_1 = 2(-1)^2 + 1 = 3; y_2 = 2(1)^2 + 1 = 3;\ y_3 = 2(2)^2 + 1 = 9;\quad y_4 = 2(3)^2 + +1 = 19$, с теми же вероятностями, что и случайная величина X. Одинаковые значения случайной величины Y, объединяются, а вероятности этих значений складываются $(0,1 + 0,3 = 0,4)$. Тогда закон распределения случайной величины Y имеет следующий вид.

Y	3	9	19
p	0,4	0,5	0,1

$$M(Y) = 3 \cdot 0,4 + 9 \cdot 0,5 + 19 \cdot 0,1 = 7,6;$$

$$D(Y) = 3^2 \cdot 0,4 + 9^2 \cdot 0,5 + 19^2 \cdot 0,1 - 7,6^2 = 22,5;\ \sigma(Y) = 4,74.$$

Пусть имеется непрерывная случайная величина X с функцией плотности вероятности $f(x)$. Другая случайная величина Y связана со случайной величиной X функциональной зависимостью: $Y = \varphi(X)$.

Случайная точка (X, Y) может находиться только на кривой $y = \varphi(X)$. Плотность вероятности случайной величины Y определяется при условии, что $\varphi(x)$ является монотонной функцией на интервале (a, b), тогда для функции $\varphi(x)$ существует обратная функция:

$$\varphi^{-1} = \psi, \qquad x = \psi(y).$$

Рассмотрим два случая — монотонного возрастания и монотонного убывания функции.

1. Если функция $\varphi(x)$ монотонно возрастает, то функцию распределения случайной величины Y (подкрепляя визуальным представлением) можно представить в виде

$$G(y) = P(Y < y) = P(a < X < x) = \int_a^x f(x)dx = \int_a^{\psi(y)} f(x)dx.$$

Найдем производную $G(y)$ как производную интеграла с переменным верхним пределом, получим

$$g(y) = G'(y) = f(\psi(y))\psi'(y).$$

2. Если функция $\varphi(x)$ монотонно убывает, то функцию распределения случайной величины Y (подкрепляя визуальным представлением) можно представить в виде

$$G(y) = P(Y < y) = P(x < X < b) = \int_x^b f(x)dx = \int_{\psi(y)}^b f(x)dx.$$

Найдем производную $G(y)$ как производную интеграла с переменным нижним пределом, имеем

$$g(y) = G'(y) = -f(\psi(y))\psi'(y).$$

Объединяя полученные формулы в одну, получим, что плотность распределения случайной величины Y находится по формуле

$$g(y) = f\big(\psi(y)\big)|\psi'(y)|. \tag{7.3}$$

Если случайная величина $Y = \varphi(X)$ имеет несколько промежутков монотонности, то числовая прямая разбивается на n промежутков монотонности и обратная функция находится на каждом из них, поэтому

$$g(y) = \sum_{i=1}^n f(\psi_i(y))\,|\psi_i{}'(y)|. \tag{7.4}$$

Математическое ожидание и дисперсию непрерывной случайной величины Y — функции случайной величины X, можно определить по формулам:

$$M(Y) = M\big(\varphi(x)\big) = \int_{-\infty}^{+\infty} \varphi(x)f(x)dx, \tag{7.5}$$

$$D(Y) = D\big(\varphi(x)\big) = \int_{-\infty}^{+\infty} \varphi^2(x)f(x)dx - M^2(Y). \tag{7.6}$$

Пример 7.2**.** Найти функцию плотности вероятностей случайной величины $Y = bX + a(b > 0,\ a \in R)$, если случайная величина X имеет непрерывную функцию распределения $F(x) = P(X \le x)$.

Решение. Плотность вероятностей непрерывной случайной величины X имеет вид

$$f(x) = F'(x),$$

$\psi = x = \frac{y-a}{b}$ — обратная функция на $x \in (-\infty; \infty)$, следовательно, по формуле (7.3) получим

$$g(y) = f(\psi(y)|\psi'(y)| = \frac{1}{b} f\left(\frac{y-a}{b}\right). \tag{7.7}$$

Модель (7.7) называют сдвигмасштабной, a — параметр положения (сдвига), b — параметр масштаба.

Полагается, что сдвигмасштабная модель получается преобразованием стандартной. Например, стандартное нормальное распределение $N(0,1)$ преобразуясь по формуле $Y = \sigma X + a$, превращается в сдвигмасштабную нормальную модель $N(a, \sigma^2)$.

Пример 7.3. Найти функцию плотности вероятностей случайной величины $Y = X^2$. Случайная величина X подчиняется нормальному закону распределения с математическим ожиданием a и дисперсией σ^2, то есть плотность распределения имеет вид

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

Определить $M(Y), D(Y), \sigma(Y)$, если $a = 0$ и $\sigma = 1$.

Решение. На промежутке $(0; +\infty)$, для $y = x^2$,

обратная функция $x = \sqrt{y} = \psi_1$, на промежутке $(-\infty; 0)$ — обратная функция $x = -\sqrt{y} = \psi_2$. $|\psi_1'(y)| = |\psi_2'(y)| = \frac{1}{2\sqrt{y}}$. По формуле (7.4):

$$g(y) = f(\psi_1(y)|\psi_1'(y)| + f(\psi_2(y)|\psi_2'(y)| =$$
$$= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\sqrt{y}-a)^2}{2\sigma^2}} \frac{1}{2\sqrt{y}} + \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(-\sqrt{y}-a)^2}{2\sigma^2}} \frac{1}{2\sqrt{y}}.$$

При $a = 0$ и $\sigma = 1$: $g(y) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}}$.

Найдем математическое ожидание и дисперсию.

$$M(Y) = \int_0^{+\infty} y \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}} dy = \int_0^{+\infty} \sqrt{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2}} dy.$$

Пусть $\sqrt{y} = t$, тогда $y = t^2, dy = 2tdt$. Имеем:

$$I_1 := \int_0^{+\infty} t \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} 2tdt = -\frac{2}{\sqrt{2\pi}} \int_0^{+\infty} t\, de^{-\frac{t^2}{2}} =$$
$$= -\frac{2}{\sqrt{2\pi}} \left(te^{-\frac{t^2}{2}}\Big|_0^{+\infty} - \int_0^{+\infty} e^{-\frac{t^2}{2}} dt \right) =$$
$$= -\frac{2}{\sqrt{2\pi}} \left(\lim_{b\to+\infty} be^{-\frac{b^2}{2}} - 0 - \frac{\sqrt{2\pi}}{2} \right) = -\frac{2}{\sqrt{2\pi}} \left(0 - 0 - \frac{\sqrt{2\pi}}{2} \right) = 1.$$

Пусть

$$I_2 := \int_0^{+\infty} y^2 \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}} dy = \int_0^{+\infty} y^{\frac{3}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2}} dy,$$

сделаем замену переменной $\sqrt{y} = t$, тогда $y = t^2, dy = 2tdt$. Имеем:

$$I_2 = \frac{1}{\sqrt{2\pi}} \int_0^{+\infty} t^3 e^{-\frac{t^2}{2}} 2tdt = -\frac{2}{\sqrt{2\pi}} \int_0^{+\infty} t^3\, de^{-\frac{t^2}{2}} =$$
$$= -\frac{2}{\sqrt{2\pi}} \left(t^3 e^{-\frac{t^2}{2}}\Big|_0^{+\infty} - 3 \int_0^{+\infty} t^2 e^{-\frac{t^2}{2}} dt \right) =$$
$$= -\frac{2}{\sqrt{2\pi}} \left(\lim_{b\to+\infty} b^3 e^{-\frac{b^2}{2}} - 0 + 3 \int_0^{+\infty} t de^{-\frac{t^2}{2}} \right) =$$
$$= -\frac{2}{\sqrt{2\pi}} \left(0 - 0 + 3 \left(-\frac{\sqrt{2\pi}}{2} \right) I_1 \right) = 3.$$

Следовательно,

$$D(Y) = \int_0^{+\infty} y^2 \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}} dy - 1^2 = 3 - 1 = 2.$$

Пример 7.4. Случайная величина X распределяется по равномерному закону на интервале $\left(-\frac{\pi}{2}; \frac{\pi}{2}\right)$. Найти функцию плотности вероятности g(y), где $Y = \sin X$.

Решение. Плотность вероятности равномерного закона на интервале $(-\frac{\pi}{2}; \frac{\pi}{2})$ имеет вид

$$f(x) = \begin{cases} 0, \text{при } x < -\frac{\pi}{2}, \\ \frac{1}{\pi}, \text{при } -\frac{\pi}{2} < x < \frac{\pi}{2}, \\ 0, \text{при } x > \frac{\pi}{2}. \end{cases}$$

На интервале $(-\frac{\pi}{2}; \frac{\pi}{2})$ функция $\sin x$ монотонна, поэтому имеет обратную функцию:

$$x = arcsin\, y, \quad x' = \frac{1}{\sqrt{1-y^2}}.$$

По формуле (7.3):

$$g(y) = f(\arcsin y) \frac{1}{\sqrt{1-y^2}} = \frac{1}{\pi} \frac{1}{\sqrt{1-y^2}},$$

где $-1 < y < 1, (-\frac{\pi}{2} < x < \frac{\pi}{2})$.

Пример 7.5. Два человека договорились встретиться в определенном месте в промежутке времени от 9^{00} до 10^{00}. Каждый из них приходит на место встречи независимо от другого, с постоянной функцией плотности вероятности в любой момент назначенного промежутка времени. Пришедший раньше ожидает другого. Найти распределение вероятностей времени ожидания и вероятность того, что ожидать придется менее десяти минут.

Решение. Пусть моменты прихода двух лиц T_1 и T_2 соответственно. За начало отсчета времени выберем девять часов. Тогда каждая из независимых случайных величин равномерно распределена в промежутке (0;1). Случайная величина T — время ожидания: $T = |T_1 - T_2|$. Найдем функцию распределения этой величины $F(t) = P(T < t)$. Рассмотрим на плоскости $t_1 O\, t_2$ область $D(t)$, в которой $|t_1 - t_2| < t$ (рис. 7.1).

Функция распределения $F(t)$ равна площади области $D(t)$ (рис. 7.1).

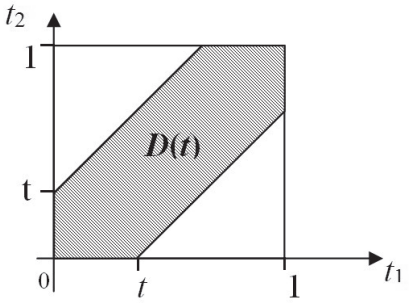


Рис. 7.1 — Область $D(t)$

$F(t) = 1-(1-t)^2 = (2-t)t$. Вероятность того, что ожидать придется менее десяти минут, определим как

$$F\left(\frac{1}{6}\right) = P(T < \frac{10}{60}) = \mathrm{P}(\mathrm{T} < \frac{1}{6}) = (2 - \frac{1}{6}) \cdot \frac{1}{6} = \frac{11}{36}.$$

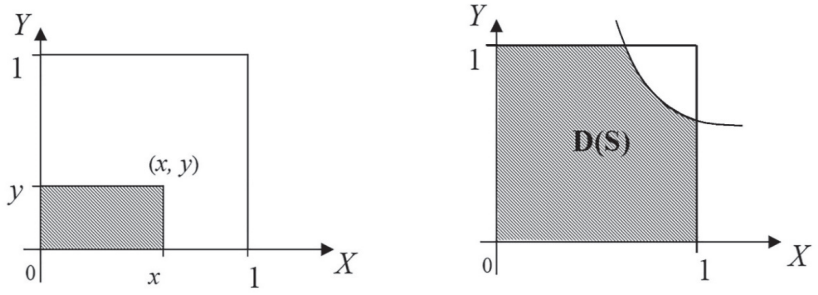
(Сравните с примером 1.3.)

Пример 7.6. Сторона квадрата равномерно распределена на отрезке [0; 1] оси OY. Найти закон распределения площади S прямоугольника со сторонами (x, y), расположенного внутри квадрата, причем две стороны прямоугольника лежат на сторонах квадрата.

Решение. Выберем в качестве точки отсчета точку с координатами (0; 0). Тогда площадь прямоугольника со сторонами x, y: $S = xy$ (рис. 7.2*а*). Выделим на плоскости XOY область $D(s)$, в которой $xy < s$ (заштрихованная область на рисунке 7.2*б*). Функция распределения $H(s)$ равна площади этой области:

$$H(s) = \iint_{D(s)} dxdy = 1 - \int_s^1 (1 - \frac{s}{x})dx = 1 - (x - slnx)\,|_s^1 = s(1 - lns).$$

Отсюда, $h(s) = H'(s) = -lns$, где $0 < s < 1$.



а *б*

Рис. 7.2 — Закон распределения площади прямоугольника, определенного системой двух случайных величин (X, Y), равномерно распределенных в единичном квадрате

Рассмотрим *функции случайных величин в пространстве*. При выполнении исходных условий данной главы случайная величина

X — случайный вектор размерности n $(X = (X_1, X_2, \ldots, X_n))$,

Y — случайный вектор размерности m $(Y = (Y_1, Y_2, \ldots, Y_m))$,

$$Y = \varphi(\mathrm{X}).$$

Пусть $\varphi(\mathrm{X})$ имеет кусочно-непрерывные первые производные по всем координатам вектора X и непостоянна ни на каком множестве значений аргумента X, имеющем отличную от нуля вероятность. Тогда плотность вероятности случайной величины $Y = \varphi(X)$ при $n = m$:

$$g(y) = f(\varphi^{-1}(y)|J(y)|, \tag{7.8}$$

где f — функция плотности вероятности вектора X, $|J(y)|$ — абсолютная величина якобиана координат[8] вектора $\mathrm{X} = \varphi^{-1}(Y)$ по координатам *вектора Y*:

[8] $|J(y)|$ — коэффициент искажения или коэффициент растяжения пространства (в данной точке), задаваемого вектором X, при преобразовании его в пространство, задаваемое вектором Y (в двумерном случае плоскости x_1x_2 в плоскость y_1y_2).

$$J(y) = \frac{\partial(\varphi_1^{-1},\dots,\varphi_n^{-1})}{\partial(y_1,\dots,y_n)} = \begin{vmatrix} \frac{\partial\varphi_1^{-1}}{\partial y_1} \dots \frac{\partial\varphi_1^{-1}}{\partial y_n} \\ \dots\dots\dots\dots\dots \\ \frac{\partial\varphi_n^{-1}}{\partial y_1} \dots \frac{\partial\varphi_n^{-1}}{\partial y_n} \end{vmatrix} \tag{7.9}$$

или

$$J(y) = \frac{\partial(x_1,x_2,\dots,x_n)}{\partial(y_1,y_2,\dots,y_n)} = \begin{vmatrix} \frac{\partial x_1}{\partial y_1}\frac{\partial x_1}{\partial y_2} \dots \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1}\frac{\partial x_2}{\partial y_2} \dots \frac{\partial x_2}{\partial y_n} \\ \dots\dots\dots\dots\dots\dots \\ \frac{\partial x_n}{\partial y_1}\frac{\partial x_n}{\partial y_2} \dots \frac{\partial x_n}{\partial y_n} \end{vmatrix}$$

(или $J(y)$ — это производная функции $\varphi^{-1}(Y)$, в случае скалярных X и Y).

Отсюда (7.8) можно записать в виде функции распределения Y:

$$G(y) = \int_D f\left(\varphi^{-1}(y)\right)|J(y)|dy, \tag{7.10}$$

где D — область задания случайной величины Y.

Замечание [113]. Якобианы имеют свойства, аналогичные свойствам обычных производных (что особенно проявляется в теории неявных функций и при замене переменных в кратных интегралах):

$$\frac{\partial(x_1,x_2,\dots,x_n)}{\partial(y_1,y_2,\dots,y_n)}\frac{\partial(y_1,y_2,\dots,y_n)}{\partial(t_1,t_2,\dots,t_n)} = \frac{\partial(x_1,x_2,\dots,x_n)}{\partial(t_1,t_2,\dots,t_n)}, \tag{7.11}$$

$$\frac{\partial(x_1,x_2,\dots,x_n)}{\partial(y_1,y_2,\dots,y_n)}\frac{\partial(y_1,y_2,\dots,y_n)}{\partial(x_1,x_2,\dots,x_n)} = 1.$$

Пусть имеется система n уравнений с $2n$ переменными

$$F_i(y_1,y_2,\dots,y_n;\ x_1,x_2,\dots,x_n) = 0\ (i=\overline{1,n}),$$

якобиан $\frac{\partial(F_1,F_2,\dots,F_n)}{\partial(y_1,y_2,\dots,y_n)}$ отличен от нуля; $y_1,y_2,\dots,y_n$ — функции от $x_1,x_2,\dots,x_n$, определяемые системой уравнений и обращающие их в тождества. Тогда

$$\frac{\partial(y_1,y_2,\dots,y_n)}{\partial(x_1,x_2,\dots,x_n)} = (-1)^n \frac{\frac{\partial(F_1,F_2,\dots,F_n)}{\partial(x_1,x_2,\dots,x_n)}}{\frac{\partial(F_1,F_2,\dots,F_n)}{\partial(y_1,y_2,\dots,y_n)}}. \ \blacksquare \tag{7.12}$$

Пример 7.7. По заданной функции плотности распределения $f(x_1,x_2)$ двумерной случайной величины $(X_1,\ X_2)$ найти плотность распределения $g(Y_1,Y_2)$ двумерной случайной величины (Y_1,Y_2), связанной взаимно однозначно с $(X_1,\ X_2)$ указанными ниже соотношениями:

$$f(x_1,x_2) = \frac{1}{2\pi\cdot3\cdot4}\,e^{-\frac{1}{2}\left(\frac{X_1^2}{3^2}+\frac{X_2^2}{4^2}\right)},$$

$X_1 = 3Y_1 cos2Y_2\ , X_2 = 4Y_1 sin2Y_2\ , 0 \le Y_1 < \infty\ , 0 \le\ Y_2 < \pi.$

Решение. Из (7.7) следует, что плотность вероятности $g(y) = f(x_1,\ x_2)\left|J(y)\right|$, определена в области D, заданной неравенствами: $0 \le Y_1 < \infty, 0 \ \le\ Y_2 < \pi.$

Имеем

$$J(y) = \begin{vmatrix} 3\cos 2y_2 & -6y_1 \sin 2y_2 \\ 4\sin 2y_2 & 8y_1 \cos 2y_2 \end{vmatrix} = 24y_1 \cos^2 2y_2 + 24y_1 \sin^2 2y_2 = 24y_1.$$

Отсюда

$$g(y_1, y_2) = \frac{1}{2\pi \cdot 3 \cdot 4} e^{-\frac{1}{2}\left(\frac{3^2 y_1^2 \cos^2 2y_2}{3^2} + \frac{4^2 y_1^2 \sin^2 2y_2}{4^2}\right)} \cdot 24y_1,$$

$$g(y_1, y_2) = y_1 e^{-\frac{y_1^2}{2}} \frac{1}{\pi},$$

где $g_1(y_1) = y_1 e^{-\frac{y_1^2}{2}}$ — функция плотности вероятности закона Рэлея $0 \le Y_1 < \infty$, $g_2(y_2) = \frac{1}{\pi}$ — функция плотности вероятности равномерного закона вероятности $0 \le Y_2 < \pi$.

Генерация законов распределения (*сэмплирование*) *известного вида из равномерного* осуществляется методом обратного преобразования. Для требуемого распределения находится обратная функция, в качестве аргумента которой рассматривается случайная величина с равномерным законом. В результате получается величина с требуемым распределением.

Как известно, функция распределения случайной величины X задается как $P(X < x) = F(x)$, где $F(x) \in [0, 1]$.

Если $F(x)$ непрерывная, возрастающая функция, то существует обратная F^{-1} (то есть если $y = F(x)$, то $x = F^{-1}(y)$). Значение x можно выбрать из распределения $F(x)$ с помощью обратной функции F^{-1} и случайного числа r, из равномерно распределенной совокупности:

$$x = F^{-1}(r), \tag{7.13}$$

$$P(X < x) = P(F^{-1}(r) < x) = P\big(r < F(x)\big) = F(x). \tag{7.14}$$

Иногда обратную функцию F^{-1} легко вычислить (примеры 7.7, 7.8), либо приходится использовать специальные приемы моделирования распределения, так, например, в случае стандартного нормального распределения может использоваться метод полярных координат (Дж. Бокс, М. Мюллер, Дж. Марсалья; пример 7.9) [50, 59].

Пример 7.8. Случайная величина R равномерно распределена на (0;1] и связана со случайной величиной X функциональной зависимостью

$$x = -\frac{1}{\lambda} ln(1 - r), (\lambda = const > 0, x > 0).$$

Найти функцию плотности вероятности случайной величины X.

Решение. Из (7.3) следует, что искомая функция плотности вероятности

$$g(x) = f\big(\psi(x)\big)|\psi'(x)|.$$

Дифференциальная функция равномерно распределенной случайной величины R на (0;1] равна 1. В силу монотонности логарифмической функции обратная функция функции $Y = \varphi(X)$ будет иметь вид

$$\psi(x) = 1 - e^{-\lambda x};$$

$$\psi'(x) = \lambda e^{-\lambda x}.$$

Следовательно, согласно формуле (7.3) искомая функция плотности вероятности имеет вид

$$g(x) = \lambda e^{-\lambda x},$$

то есть случайная величина X подчиняется показательному распределению.

Таким образом, можно задать алгоритм получения (моделирования) показательного закона распределения из чисел, подчиняющихся равномерному закону, подавая на вход функции $x = -\frac{1}{\lambda} ln(1-r)$ значения случайной величины r, подчиняющейся равномерному закону на (0;1].

Действительно, рассмотрим экспоненциальное распределение

$$g(x) = \lambda e^{-\lambda x}, (\lambda = const > 0, y > 0),$$
$$F(x) = \int_0^x \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda x}.$$

Применим (7.13)–(7.14), то есть

$$F(x) = r,$$

где r — равномерно распределена на (0;1].

Следовательно,

$$1 - e^{-\lambda x} = r,$$
$$x = -\frac{1}{\lambda} ln(1-r),$$

где случайная величина X будет подчиняться экспоненциальному закону распределения.

Пример 7.9. Случайные величины Y_1, Y_2 независимы и равномерно распределены на [–1;1].

Пусть $S \coloneqq Y_1^2 + Y_2^2 < 1$, тогда

$$X_1 \coloneqq Y_1\sqrt{\frac{-2\,ln(S)}{S}}, \qquad X_2 \coloneqq Y_2\sqrt{\frac{-2\,ln(S)}{S}}.$$

Найти функцию плотности вероятности системы случайных величин (X_1, X_1).

Решение. 1. Точка плоскости с декартовыми координатами (Y_1, Y_2) — случайная точка, равномерно распределенная внутри единичного круга радиуса $r = \sqrt{S} = \sqrt{Y_1^2 + Y_2^2}$. Поэтому можно принять, что

$$cos\,\theta = \frac{Y_1}{\sqrt{S}},\ \sin\theta = \frac{Y_2}{\sqrt{S}}, \text{где } \theta \in [0;\ 2\pi].$$

Имеем

$$X_1 \coloneqq \cos\theta\sqrt{-2\,ln(S)}, \qquad X_2 \coloneqq \sin\theta\sqrt{-2\,ln(S)}.$$

Перейдем к полярным координатам: $X_1 = Rcos\theta, X_2 = Rsin\theta$. Случайная величина θ независима от R и равномерно распределена на [0; 2π] ($q(\theta) = 1/2\pi$).

То есть

$$r = \sqrt{-2\,ln\,s},$$

отсюда

$$s = \psi(r) = e^{-\frac{r^2}{2}},$$

где случайная величина S равномерно распределена на (0; 1) и ее функция плотности вероятности $f(s) = 1$. Следовательно, по формуле (7.3) функция плотности вероятности случайной величины R имеет вид

$$h(r) = f\big(\psi(r)\big)|\psi'(r)| = re^{-\frac{r^2}{2}}.$$

Функцию распределения случайной величины R можно оценить как

$$H(r) = P(R \le r) = P\big(\sqrt{-2\,ln\,S} \le r\big) = P(-2\,ln\,S \le r^2) = P(S \ge e^{-\frac{r^2}{2}}),$$

или

$$H(r) = \int_o^r te^{-\frac{t^2}{2}}dt = -\int_o^r de^{-\frac{t^2}{2}} = 1 - e^{-\frac{r^2}{2}}.$$

Итак, закон распределения системы независимых случайных величин (R, θ) в виде функции плотности вероятности можно представить как

$$g(r, \theta) = q(\,\theta)h(r) = \frac{1}{2\pi}re^{-\frac{r^2}{2}}.$$

По формуле (7.8) переход к системе координат (X_1, X_2) осуществляется по формуле

$$g(x_1, x_2) = g(r, \theta)|J(x_1, x_2)|.$$

Имеем

$$J(r, \theta\,) = \begin{vmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \theta} \\ \frac{\partial x_2}{\partial r} & \frac{\partial x_2}{\partial \theta} \end{vmatrix} = \begin{vmatrix} cos\theta & rsin\theta \\ sin\theta & rcos\theta \end{vmatrix} = r[(cos\theta)^2 + (sin\theta)^2] = r.$$

По свойству якобиана (7.11)

$$J(x_1, x_2\,) = \frac{1}{J(r,\theta\,)} = \frac{1}{r}.$$

Имеем

$$g(x_1, x_2) = \frac{1}{2\pi}re^{-\frac{r^2}{2}}\left|\frac{1}{r}\right| = \frac{1}{2\pi}e^{-\frac{r^2}{2}} = \frac{1}{2\pi}e^{-\frac{x_1^2+x_2^2}{2}}.$$

Таким образом, совместная плотность пары (X_1, X_2) представляет собой систему независимых стандартных нормальных случайных величин X_1, X_2:

$$g(x_1, x_2) = g_1(x_1)g_2(x_2) = \left(\frac{1}{\sqrt{2\pi}}e^{-\frac{x_1^2}{2}}\right)\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{x_2^2}{2}}\right).$$

2. Итак, если $(X_1, X_2) \to N(0, 1)$, то переходя к полярным координатам $X_1 = Rcos\theta$, $X_2 = Rsin\theta$ получим систему независимых случайных величин (R, θ), причем радиус R распределен с функцией распределения

$$H(r) = 1 - e^{-\frac{r^2}{2}},$$

а угол θ равномерно распределен на $[0, 2\pi]$. Действительно,

$$H(r) = P(R < r) = \iint_{x^2+y^2\le r^2}\frac{1}{2\pi}e^{-\frac{x_1^2+x_2^2}{2}} = \frac{1}{2\pi}\int_0^r\int_0^{2\pi}e^{-\frac{r^2}{2}}|J(r, \theta\,)|drd\theta,$$

$$H(r) = 1 - e^{-\frac{r^2}{2}}.$$

И наоборот, если случайные величины (R, θ) распределены как указано выше, то случайные величины X_1 и X_2 независимы и распределены по стандартному нормальному закону:

$$g(r, \theta) = \frac{1}{2\pi}re^{-\frac{r^2}{2}} = \frac{1}{2\pi}re^{-\frac{x_1^2+x_2^2}{2}}|\,J(X_1, X_2)|,$$

$$g(X_1, X_2) = \left(\frac{1}{\sqrt{2\pi}}e^{-\frac{x_1^2}{2}}\right)\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{x_2^2}{2}}\right).$$

Замечание. 1) Опираясь на метод обратной функции, можно *генерировать*, используя числа, распределенные по равномерному закону, пары чисел, *распределенные по стандартному нормальному закону* (*сэмплировать из гауссиана*) (примеры 7.9, 10.5) и другим законам с известной функцией плотности вероятности. Если закон распределения не известен, но есть наблюдения, то генерация (сэмплирование) в машинном обучении и байесовской статистике осуществляется, например, с использованием динамических методов Монте-Карло или иначе *MCMC — Marcov Chain Monte Carlo* (генерации Марковских цепей по методу Монте-Карло), позволяющих «размножать» выборку в соответствии с распределением имеющихся данных (см. гл. 18 в [53]). ■

7.2. Композиция законов распределения

В приложениях часто рассматривается вопрос о распределении суммы нескольких случайных величин. Например, пусть $Z=X+Y$, тогда $G(z)$ — функцию распределения случайной величины Z можно определить по формуле

$$G(z) = \int_{(D)} f(x,y)dxdy = \int_{-\infty}^{+\infty} dx = \int_{-\infty}^{+\infty} dy \int_{-\infty}^{z-x} f(x, z-x)dx, \quad (7.15)$$

где $f(x,y)$ — функция плотности распределения системы случайных величин (X,Y); область D — полуплоскость, ограниченная сверху прямой $y = z - x$.

Отсюда

$$g(z) = G'(z) = \int_{-\infty}^{+\infty} f(x, z-x)dx.$$

Если X и Y независимы, то говорят о композиции законов распределения случайных величин, и плотность вероятности случайной величины Z определяется как

$$g(z) = \int_{-\infty}^{+\infty} f_1(x) f_2(z-x)dx = \int_{-\infty}^{+\infty} f_1(z-y) f_2(y)dy, \quad (7.16)$$

где $f_1(x)$ и $f_2(y)$ функции плотности вероятности случайных величин X и Y соответственно.

Если возможные значения аргументов неотрицательны, то функция плотности распределения вероятностей случайной величины Z определяется по формуле

$$g(z) = \int_0^z f_1(x) f_2(z-x)dx \quad (7.17)$$

или

$$g(z) = \int_0^z f_1(z-y) f_2(y)dy. \quad (7.18)$$

Пример 7.10. Случайные величины X и Y распределены равномерно: случайная величина X на интервале (0;5), случайная величина Y на интервале (0;2). Найти функцию распределения и плотность вероятности случайной величины $Z = X + Y$.

Решение. Найдем функцию распределения $G(Z)$ величины $Z = X + Y$. Система двух случайных величин (X, Y) равномерно распределена в прямоугольнике со сторонами 2 и 5 (рис. 7.3).

$$G(\mathrm{z}) = \iint_D f(x,y)dxdy = \iint_D f_1(x)f_2(y)dxdy,$$

где D — часть прямоугольника, лежащая левее и ниже прямой $y = z - x$, $f(x,y)$ — плотность вероятности системы двух случайных величин (X,Y), f_1 — плотность вероятности случайной величины X, f_2 — плотность вероятности случайной величины Y.

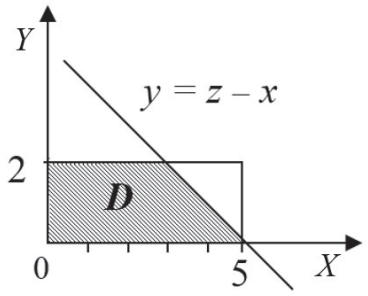


Рис. 7.3 — Область D

В силу независимости X и Y: $f(x,y) = f_1(x)f_2(y) = \frac{1}{5}\cdot\frac{1}{2} = \frac{1}{10}$. То есть $G(z) = S(D)\cdot\frac{1}{10}$, где $S(D)$ — площадь области D, $\frac{1}{10}$ — плотность вероятности на единицу площади прямоугольника.

В зависимости от значения z в уравнении $y = z - x$ область D будет принимать различный вид, отсюда имеем:

1) $z \le 0, G(z) = 0$;

2) $0 < z \le 2, G(z) = \frac{z^2}{2}\frac{1}{10} = \frac{z^2}{20}$;

3) $2 < z \le 5, G(z) = \frac{1}{10}\big(2(z-1)\big) = \frac{z-1}{5}$;

4) $5 < z \le 7, G(z) = \frac{1}{10}\left(10 - \frac{(7-z)^2}{2}\right) = 1 - \frac{(7-z)^2}{20}$;

5) $z > 7, G(z) = 1$.

Отсюда $g(z) = G'(z) = \begin{cases} 0, \text{при } z \le 0, \\ \frac{z}{10}, \quad \text{при } 0 < z \le 2 \\ \frac{1}{5}, \quad \text{при } 2 < z \le 5, \\ \frac{7-z}{10}, \text{при } 5 < z \le 7, \\ 0, \quad \text{при } z > 7. \end{cases}$

Замечание. Если рассматривается композиция нормальных законов:

$$X = \sum_{i=1}^{n} X_i,$$

где случайные величины X_i независимы $(i = 1,2,\dots,n)$ $X_i \in N\big(a_i, \sigma_i^2\big)$, то $X \in N(a,\sigma^2)$, где $M(X) = a = \sum_i a_i$, $D(X) = \sigma^2 = \sum_i \sigma_i^2$. ■

7.3. Специальные законы распределения

В математической статистике и различных приложениях используют специальные функции — Эйлеровы интегралы первого и второго рода (по определению Лежандра), зависящие от параметра — бета- и гамма-функции, соответственно:

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx,\ \ (\alpha>0,\beta>0), \tag{7.19}$$

$$\Gamma(\alpha)=\int_0^{+\infty} x^{\alpha-1}e^{-x}dx,\ \ (\text{при } R_+,\alpha>0). \tag{7.20}$$

Отметим некоторые свойства бета-функции.

1) $B(\alpha,\beta)=B(\beta,\alpha)$ — симметричность.

2) $B(\alpha,\beta)=\frac{\alpha-1}{\alpha+\beta-1}\,B(\alpha-1,\beta)$ — формула понижения, в частности

$$B(\alpha,1)=\frac{1}{\alpha};\ B(\alpha,n)=\frac{(n-1)!}{\alpha(\alpha+1)\cdot\ldots\cdot(\alpha+n-1)},$$

где $n\in N$;

$$B(m,n)=\frac{(m-1)!(n-1)!}{(m+n-1)!},\qquad (m,n\in N). \tag{7.21}$$

3) Между функциями B и Γ имеет место взаимосвязь

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{7.22}$$

Бета-функция выражается через гамма-функцию (7.22), кроме того, ее используют законы распределения, рассматриваемые ниже, поэтому рассмотрим ее подробнее.

Гамма-функция удовлетворяет следующим свойствам.

1) $\Gamma(1)=1$.

Действительно, интегрируя, получим

$$\Gamma(1)=\int_0^{+\infty} e^{-x}\,dx=-\int_0^{+\infty} de^{-x}=-(\lim_{b\to+\infty} e^{-b}-1)=1.$$

2) $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$ при $\alpha>0$.

Интегрируя по частям, имеем

$$\Gamma(\alpha+1)=\int_0^{+\infty} x^{\alpha}e^{-x}dx=-\int_0^{+\infty} x^{\alpha}de^{-x}=$$

$$=-(\lim_{b\to+\infty} b^{\alpha}e^{-b}-0)+\alpha\int_0^{+\infty} x^{\alpha-1}e^{-x}dx=\alpha\Gamma(\alpha).$$

Значит, при больших значениях α гамма-функция может вычисляться по формуле

$$\Gamma(\alpha)=(\alpha-1)\Gamma(\alpha-1)=(\alpha-1)(\alpha-2)(\Gamma(\alpha-2)=\ldots$$

Следовательно, если α — целое положительное число, то

$$\Gamma(\alpha+1)=\alpha!$$

Отсюда, при $\alpha=0$, имеем

$$\Gamma(1)=0!=1.$$

3) $\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$.

Действительно,

$$\Gamma\left(\frac{1}{2}\right)=\int_0^{+\infty} x^{-\frac{1}{2}}e^{-x}dx=\left|\begin{matrix}x=t^2\\ dx=2tdx\end{matrix}\right|=$$

$$=\int_0^{+\infty} t^{-1}e^{-t^2}2tdt=2\int_0^{+\infty} e^{-t^2}dt=\sqrt{\pi}\,.$$

То есть $\Gamma\left(\frac{1}{2}\right)$ — равно значению интеграла Эйлера-Пуассона (см. раздел 5.3). Если α кратно $\frac{1}{2}$, то значение гамма-функции легко вычисляется, например:

$$\Gamma\left(\frac{3}{2}\right) = \Gamma\left(\frac{1}{2}+1\right) = \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{1}{2}\sqrt{\pi},\ \Gamma\left(\frac{5}{2}\right) = \frac{3}{2}\Gamma\left(\frac{3}{2}\right) = \frac{3}{4}\sqrt{\pi}.$$

1. χ^2 *–распределение Пирсона*. Пусть $X_1, X_2, \ldots, X_n$ одинаково распределенные по нормальному закону случайные величины, являющиеся взаимно независимыми, для которых математические ожидания равны нулю, а средние квадратические отклонения единице, тогда сумма квадратов этих случайных величин носит название случайной величины χ_n^2 (хи-квадрат) с $k = n$ степенями свободы (число степеней свободы обозначают буквами: k, ν или df — *degrees of freedom*):

$$\chi_n^2 = \sum_{i=1}^{n} X_i^2. \tag{7.23}$$

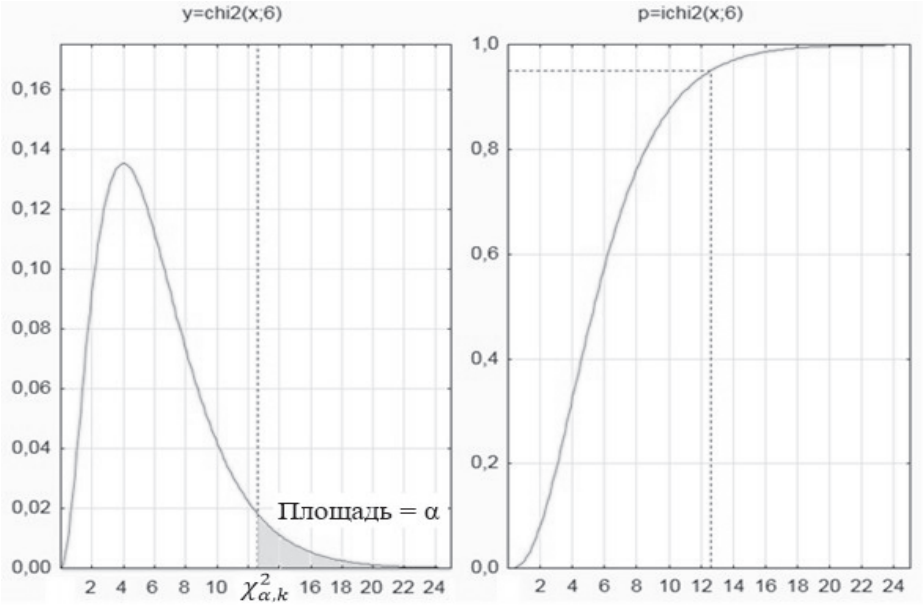


Рис. 7.4 — Распределение χ^2 с k степенями свободы
(плотность и функция распределения)

Плотность распределения χ_n^2 (рис. 7.4) задается формулой

$$f(\chi^2) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} (\chi^2)^{\frac{n}{2}-1} e^{-\frac{\chi^2}{2}}, & \text{при } \chi^2 \geq 0, \\ 0, & \text{при } \chi^2 < 0. \end{cases} \tag{7.24}$$

Распределение χ^2, как и нормальный закон распределения, обладает свойством *воспроизводимости* (результат суммы случайных величин, подчиняющихся распределению χ^2, также подчиняется распределению χ^2). В частности, можно показать, что $\chi_n^2 + \chi_m^2 = \chi_{n+m}^2$. Как следует из примера 7.3:

$$M(\chi_n^2) = \sum_{i=1}^{n} X_i^2 = 1 \cdot n = n, D(\chi_n^2) = \sum_{i=1}^{n} X_i^2 = 2 \cdot n = 2n.$$

С возрастанием числа степеней свободы $k = n$ распределение χ_n^2 медленно приближается к нормальному закону распределения. На практике используют обычно не плотность вероятности, а квантили распределения.

Квантилью χ_n^2-распределения, отвечающей заданному уровню значимости α (альфа), называется такое значение $\chi^2 = \chi_{\alpha,k}^2$, при котором вероятность того, что χ^2 превысит значение $\chi_{\alpha,n}^2$, равна α (рис. 7.4):

$$P\left(\chi^2 > \chi_{\alpha,n}^2\right) = \int_{\chi_{\alpha,n}^2}^{+\infty} f(\chi^2) d\chi^2 = \alpha. \qquad (7.25)$$

С геометрической точки зрения нахождение квантили χ_n^2 заключается в выборе такого значения $\chi^2 = \chi_{\alpha,n}^2$, при котором площадь криволинейной трапеции, ограниченной функцией плотности распределения, была бы равна α. Значения квантилей затабулированы (приложение 2). При $n > 30$ распределение практически не отличается от нормального.

Замечание. Квантиль случайной величины X порядка α — это такое значение случайной величины X, что $F(x_\alpha) = \alpha$ (или $F(x_{1-\alpha}) = 1 - \alpha$), где $F(x) = P(X < x)$. Например, медиана — это квантиль $x_{0,5}$. ■

2. *t-распределение Стьюдента*. Это распределение имеет большое значение при статистических вычислениях, связанных с нормальным законом распределения, где σ — неизвестный параметр распределения и подлежит определению из опытных данных, например, при статистической обработке наблюдений с неизвестной точностью.

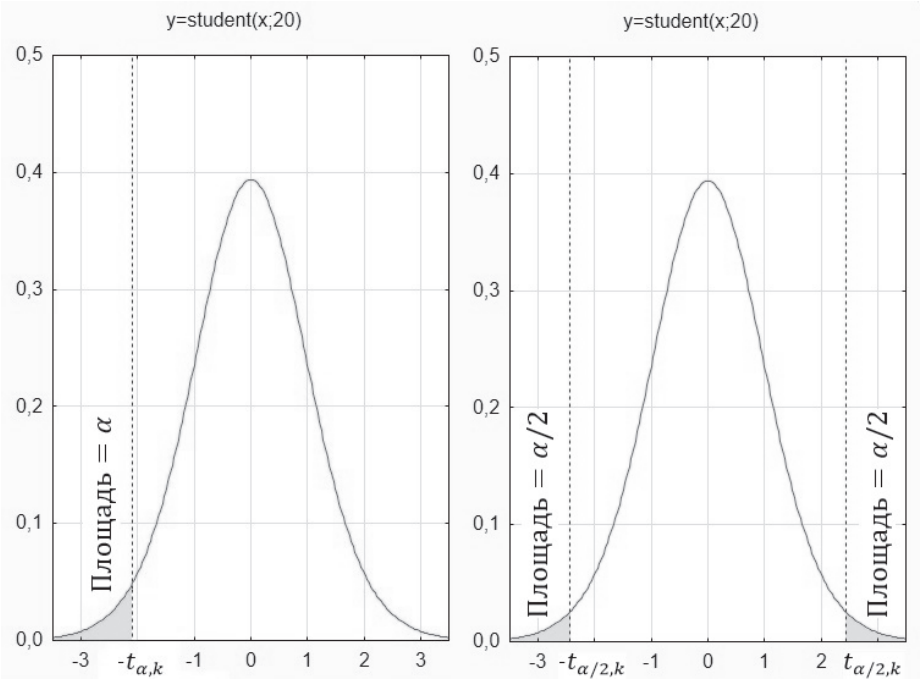


Рис. 7.5 — t −распределение Стьюдента с k степенями свободы (плотность распределения с левосторонней и двусторонней критической областью)

Пусть $Z, X_1, X_2, \ldots, X_k$ — независимые нормально распределенные случайные величины с нулевыми математическими ожиданиями и одинаковыми дисперсиями. Безразмерная случайная величина t:

$$t = \frac{Z}{\sqrt{\frac{1}{k}\sum_{i=1}^{k} X_i^2}} = \frac{Z}{\sqrt{\frac{\chi_k^2}{k}}}, \tag{7.26}$$

называется дробью Стьюдента.

Распределение t не зависит от σ в силу его безразмерности. Функция плотности вероятности t-распределения с $\nu = k$ степенями свободы имеет вид

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)}\left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}. \tag{7.27}$$

$$M(t) = 0, (k > 1); \quad D(t) = \frac{k}{k-2}, (k > 2).$$

t-распределение Стьюдента с увеличением числа степеней свободы быстрее, чем χ^2 стремится к нормальному закону распределения.

t-распределение Стьюдента при $k = 1$ называется распределением Коши — это отношение двух независимых случайных величин, распределенных по стандартному нормальному закону (см. ниже).

На практике используют квантили распределения в зависимости от числа степеней свободы и уровня значимости α. С геометрической точки зрения нахождение квантилей, например, для левосторонней (симметричной ей правосторонней) или двусторонней области, заключается в выборе такого значения t, при котором площадь криволинейной трапеции была бы равна α (рис. 7.5), соответственно:

$$P\left(t < -t_{\alpha,k}\right) = \int_{-\infty}^{-t_{\alpha,k}} f(t)dt = \alpha, P\left(|t| < t_{\frac{\alpha}{2},k}\right) = 2\int_{-\infty}^{-t_{\alpha/2,k}} f(t)dt = \alpha. \tag{7.28}$$

3. *F-распределение Фишера — Снедекора.*

Пусть $X_1, X_2, \ldots, X_m$ и $Y_1, Y_2, \ldots, Y_n$ одинаково распределенные по нормальному закону случайные величины, являющиеся взаимно независимыми, для которых математическое ожидание равно нулю, а среднее квадратическое отклонение равно единице.

Рассмотрим безразмерную случайную величину:

$$F(m, n) = \frac{\chi_m^2/m}{\chi_n^2/n}, \tag{7.29}$$

она распределена по закону Фишера — Снедекора с $k_1 = m$ — числом степеней свободы числителя, и $k_2 = n$ — числом степеней свободы знаменателя ((m, n) степенями свободы).

$$f(F) = \begin{cases} \frac{\Gamma\left(\frac{k_1+k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)}\left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}} - \frac{F^{\frac{k_1}{2}-1}}{\left(1+\frac{k_1}{k_2}F\right)^{\frac{k_1+k_2}{2}}}, \text{при } F > 0, \\ 0, \qquad\qquad \text{при } F \le 0, \end{cases} \tag{7.30}$$

$$M(F) = \frac{k_2}{k_2-2}, (k_2 > 2); D(F) = \frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)}, (k_2 > 4).$$

Обычно используют квантили распределения в зависимости от числа степеней свободы (m, n) и уровня значимости α (рис. 7.6):

$$P\big(F > F_\alpha(k_1, k_2)\big) = \int_{F_\alpha(k_1,k_2)}^{+\infty} f(F)dF. \qquad (7.31)$$

Для квантилей распределения Фишера — Снедекора геометрический смысл аналогичен другим распределениям. Имеет место равенство:

$$F_{1-\alpha}(m, n) = \frac{1}{F_\alpha(n,m)}. \qquad (7.32)$$

Распределения χ^2 Пирсона, t Стьюдента, F Фишера — Снедекора нашли широкое применение в математической статистике, в частности, при проверке статистических гипотез, в дисперсионном и корреляционно-регрессионном анализе.

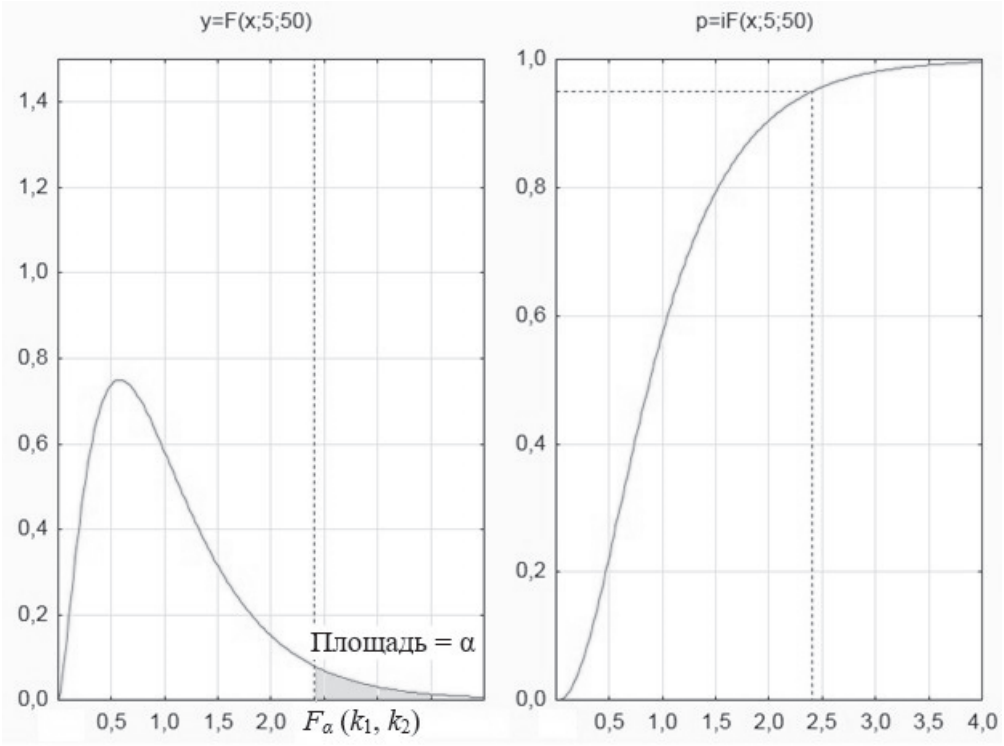


Рис. 7.6 — F-распределение Фишера — Снедекора
с k_1=5, k_2=50 степенями свободы (плотность и функция распределения)

Замечание. 1) В 1908 г. Стьюдент открыл новое направление в анализе данных, введя распределение, основанное на теоретическом нормальном и зависящее только от n. До этого, как пишут Мостеллер и Тьюки [82], «...анализ данных … напоминал длинную лестницу от очевидного первого шага до туманных высот», требующих рассмотрения последовательных статистик. После введения распределения Стьюдента достаточно стало рассмотрения оценок математического ожидания, дисперсии и объема выборки. Причем отличие в разы соответствующих

статистик (квантилей) для нормального закона и распределения неявно рекомендует исследователям работать с большими совокупностями ($n \geq 30$).

2) Запись $r^2 = \sum_{i=1}^{n} X_i^2$ позволяет заключить, что r — радиус сферы в n-мерном пространстве. Так как $X_i \in N(0, \sigma^2)$, то обычно полагают, что $r^2 = \chi_n^2$.

3) В математической статистике имеют дело с выборочными характеристиками параметров генеральной совокупности. Пусть, например, $\bar{x} = m$ и s^2 — выборочное среднее и выборочная дисперсия. Они *независимы, если исходные величины распределены нормально, обратное также верно — если* m *и* s^2 *независимы, то образующие их случайные величины подчиняются нормальному закону*.

Для упрощения полагают, что математическое ожидание совокупности совпадает с началом отсчета, поэтому выборочная средняя подчиняется нормальному закону: $m \to N\left(0; \frac{\sigma^2}{n}\right)$; выборочная дисперсия подчиняется распределению хи-квадрат: $s^2 \to \chi_{n-1}^2$.

4) Отношение выборочной средней к ее среднему квадратическому отклонению стремится к случайной величине, подчиняющейся распределению Стьюдента с $k = n - 1$ степенями свободы:

$$\frac{m}{s(m)} = \frac{\sqrt{n}m}{s} \to t(k = n - 1).$$

Таким образом, можно рассмотреть тест для проверки гипотезы о равенстве нулю средних n случайных величин с равными, но неизвестными дисперсиями. Если $P\left(\left|\frac{\sqrt{n}m}{s}\right| > t(k = n - 1)\right) < \alpha$, то гипотеза о равенстве нулю средних отвергается (обычно $\alpha = 0{,}01; 0{,}05$). Аналогично рассматривается статистика $\frac{\sqrt{n}(m-a)}{s}$, если среднее равно a.

5) Рассмотрим геометрический смысл отношения m/s (а заодно и распределения Стьюдента) в трехмерном евклидовом пространстве. Пусть в прямоугольной декартовой системе координат $OX_1X_2X_3$ имеем три наблюдения x_1, x_2, x_3, которые можно представить как координаты вектора $\overline{\mathrm{OM}}(x_1, x_2, x_3)$. Рассмотрим биссектрису первого координатного угла (первого октанта, характеризующегося положительными координатами) l с направляющим вектором $\bar{e}(1, 1, 1)$ (рис. 7.7).

Найдем вектор $\overline{ON}$ — проекцию вектора $\overline{\mathrm{OM}}$ на прямую l по известной формуле аналитической геометрии:

$$\overline{ON} = \mathrm{Пр}_l \overline{OM} = \frac{(\overline{OM}, \bar{e})}{e^2}\bar{e} = \frac{x_1 \cdot 1 + x_2 \cdot 1 + x_3 \cdot 1}{\left(\sqrt{3}\right)^2}\bar{e} = \frac{x_1 + x_2 + x_3}{3}\bar{e} = m\bar{e},$$

$$|\overline{ON}| = |\mathrm{Пр}_l \overline{OM}| = m|\bar{e}| = \sqrt{3}m.$$

Можно проверить, рассмотрев скалярное произведение, что вектор $\overline{MN}$ ортогонален вектору $\overline{ON}$, длина вектора $\overline{MN}$ равна

$$|\overline{MN}| = \sqrt{(x_1 - m)^2 + (x_2 - m)^2 + (x_3 - m)^2} = \sqrt{3}s.$$

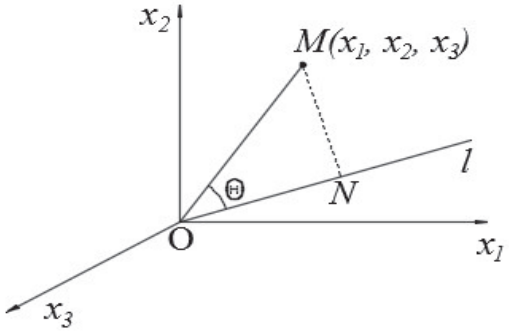


Рис. 7.7 — Геометрический смысл отношения m/s

Следовательно, котангенс угла θ между прямой l и $\overline{OM}$ равен m/s:

$$\frac{m}{s} = t = ctg\theta.$$

Аналогично в *n*-мерном евклидовом пространстве можно получить

$$|\overline{ON}| = |\text{Пр}_l\overline{OM}| = m|\bar{e}| = \sqrt{n}m; \overline{MN} = \sqrt{n}s.$$

То есть *t*-распределение принимает значение $t = m/s$, причем $t = ctg\theta$.

6) Геометрический смысл распределения Фишера — Снедекора рассмотрен в 14.1.

7) *Нецентральные выборочные распределения* используются в приложениях при рассмотрении функций мощности (чувствительности). Хи-квадрат Пирсона. Пусть $x_i \longrightarrow N(a_i, 1)$, тогда рассматривается сумма квадратов с числом степеней свободы $k = n$ и параметром нецентральности $\delta = \sum_{i=1}^{n} a_i^2$. Если $a_i = 0$, то распределение сводится к обычному.

F-распределение Фишера — Снедекора представляет собой отношение нецентрального распределения хи-квадрат с параметром δ и m степенями свободы и обычного распределения χ^2 с n степенями свободы.

t −распределение Стьюдента. Величина $\frac{U+\delta}{\sqrt{V/m}}$ с m степенями свободы и параметром δ имеет нецентральное распределение Стьюдента, где $U \to N(0,1)$, V — центральная хи-квадрат переменная с m степенями свободы, *U* и *V* взаимно независимы.

Таблицы квантилей для нецентральных распределений рассмотрены в [7]. ■

4. *Гамма-распределение* с параметрами $\alpha > 0, \lambda > 0$ имеет функцию плотности вероятности вида

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, x > 0. \qquad (7.33)$$

$$M(x) = \frac{\alpha}{\lambda}, \qquad D(x) = \frac{\alpha}{\lambda^2}.$$

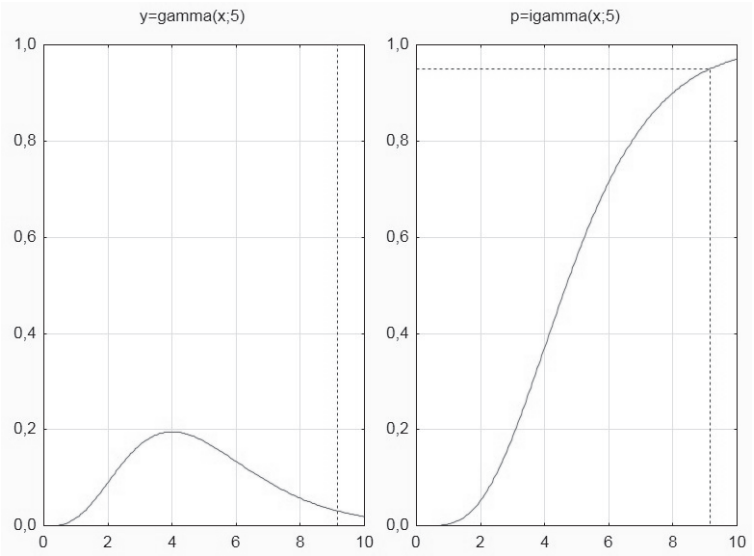


Рис. 7.8 — *Гамма*-распределение (плотность и функция распределения)

При $\alpha = 1$ гамма-распределение превращается в показательное.

5. *Бета-рапределение* с параметрами $\alpha > 0, \beta > 0$ имеет функцию плотности вероятности вида

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)},\ 0 < x < 1. \qquad (7.34)$$

$$M(x) = \frac{\alpha}{\alpha+\beta},\ D(x) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

При $\alpha = \beta = 1$ бета-распределение превращается в равномерное на $0 < < x < 1$.

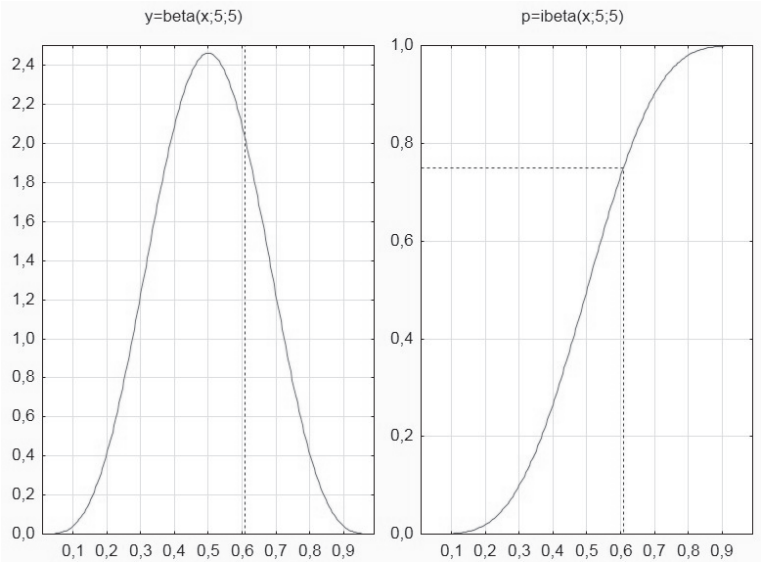


Рис. 7.9 — *Бета*-распределение (плотность и функция распределения)

Замечание. Если рассматривать текстовый документ *t* как вектор вероятностей из r слов, то сумма координат этого вектора равна 1:

$$\sum_{i=1}^{r} p_{ti} = 1. \qquad (7.35)$$

Уравнение (7.35) задает в *r*-мерном пространстве (*r-1*)-мерный симплекс (0-мерный симплекс — точка, одномерный симплекс — отрезок, двумерный симплекс — треугольник, трехмерный симплекс — тетраэдр и т. д.). Распределение вероятностей по симплексу — это и есть распределение Дирихле, которое обобщает бета-распределение (распределение Дирихле на отрезке) на многомерный случай, $Dir(\alpha)$ [36, 111]:

$$f(p,\alpha) = \frac{\prod p_i^{\alpha_i - 1}}{B(\alpha)}, \qquad (7.36)$$

где $p_i \in [0;1]$, $B(\alpha) = \frac{\prod \Gamma(\alpha_i)}{\Gamma(\sum \alpha_i)}$ — многомерная бета-функция, $\alpha = (\alpha_1, \ldots, \alpha_i, \ldots, \alpha_r)$, $\alpha_i > 0$. ■

6. *Распределение Коши* с параметрами α, β имеет функцию плотности вероятности вида

$$f(x) = \frac{1}{\pi\alpha} \frac{1}{1+\left(\frac{x-\beta}{\alpha}\right)^2}, \qquad (7.37)$$

где α — параметр масштаба, β — параметр сдвига.

Пусть случайная величина y равномерно распределена на $\left(-\frac{\pi}{2}; \frac{\pi}{2}\right)$, тогда случайная величина

$$x = \beta + \alpha tg(y),\ y \in \left(-\frac{\pi}{2}; \frac{\pi}{2}\right)$$

имеет распределение Коши.

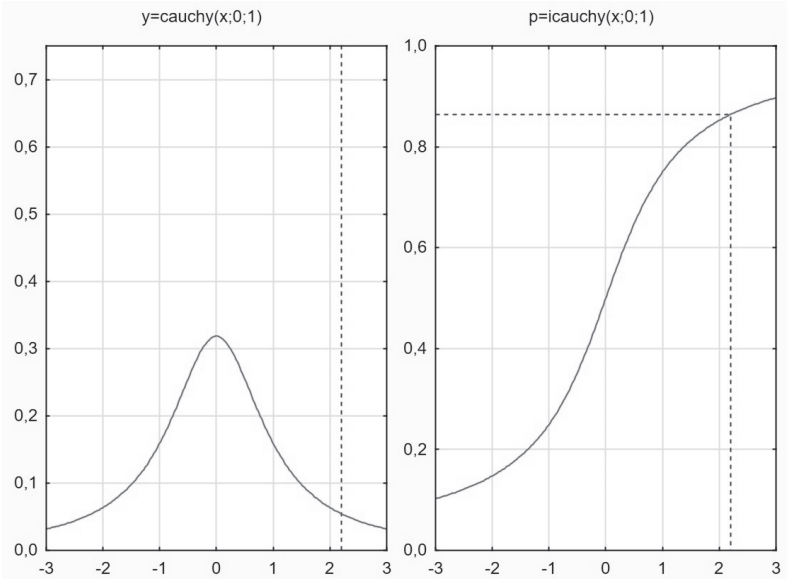


Рис. 7.10 — Распределение Коши (плотность и функция распределения)

Напомним понятие *«хвостов распределения» при* $x \to \infty$:

$$F(-x) \text{ и } 1 - F(x).$$

Закон больших чисел и центральные предельные теоремы (см. главу 8) остаются верными, если хвосты распределения быстро убывают.

Распределение Коши часто приводится в качестве примера *распределения с* «*тяжелыми хвостами*», что объясняет, несмотря на симметрию относительно $x = \beta$, отсутствие у распределения Коши математического ожидания и дисперсии, а также обязательное появление (с существенной вероятностью) «*выбросов*» — сильно отличающихся наблюдений.

7. *Законы распределения статистик экстремальных значений*. Если возможно подобрать такие сдвигмасштабные преобразования (см. пример 7.2), что, например, распределение $X_{max} = X_{(n)}$ с «тяжелым правым хвостом» можно «вернуть» в конечную область, то такой закон распределения максимума (минимума) выборки принадлежит одному из трех типов распределений экстремальных значений [33]:

1) тип *I* — *распределение Гумбеля*:

$$F(x) = e^{-e^{-(x-\mu)/\beta}}, -\infty < x < \infty, \tag{7.38}$$

где μ — параметр центра, β — параметр масштаба.

При $\mu = 0, \beta = 1$ — стандартное распределение Гумбеля (примеры использования распределения: длительность жизни долголетних жителей, максимальные паводки, испытания материалов);

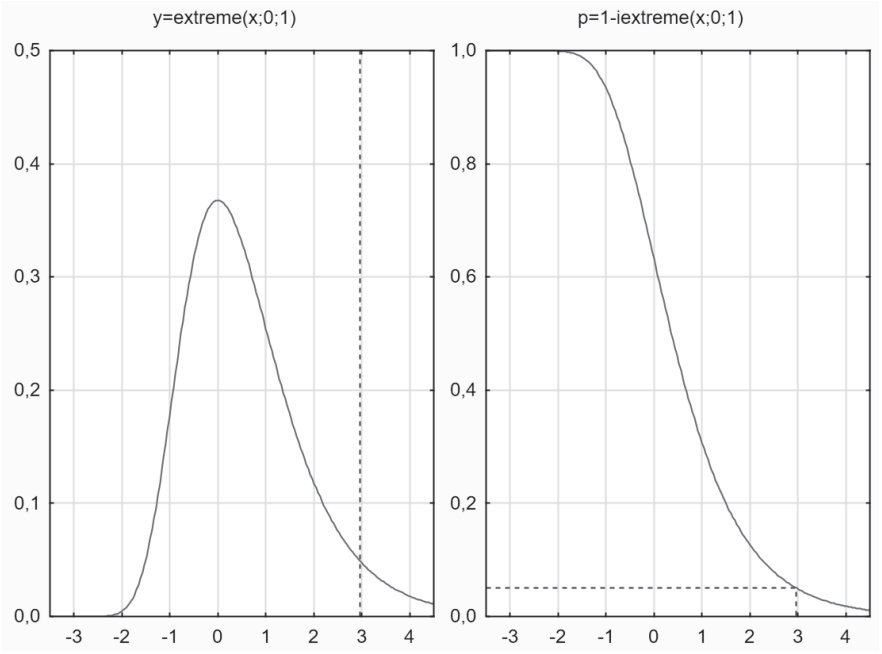


Рис. 7.11 — Распределение экстремальных значений (тип *I*) Гумбеля (плотность и функция надежности)

2) тип *II — распределение Фреше*:

$$F(x) = \begin{cases} 0, & x \le 0, \\ e^{-x^{-\alpha}}, & x > 0, \end{cases} \tag{7.39}$$

где $\alpha > 0$ (примеры использования распределения в гидрологии: максимальные осадки, речные сбросы);

3) тип *III — распределение Вейбулла*:

$$F(x) = \begin{cases} 0, & x \le 0, \\ 1 - e^{-(x)^{\alpha}}, & x > 0, \end{cases} \tag{7.40}$$

где $\alpha > 0$ (примеры использования распределения: прочность стали на разрыв; отказ шарикоподшипников, вакуумных приборов, элементов электроники; надежность машин; распределение скоростей ветра).

Интерес к экстремальным распределениям возрос в связи с проблемами техносферной безопасности и стихийных бедствий. В исследованиях часто рассматривается функция надежности $R(x) = 1 - F(x)$.

Условия сходимости экстремальных статистик наиболее полно в 1943 г. исследовал Б. В. Гнеденко. Распределения *II* и *III* типа можно путем преобразований свести к *I*, поэтому иногда его считают основным.

Темы (вопросы) для самоконтроля

1. Закон распределения функции одного случайного аргумента.
2. Сдвигмасштабная модель.
3. Закон распределения функции случайных величин в пространстве.
4. Генерация законов распределения (сэмплирование) известного вида из равномерного закона методом обратного преобразования.
5. Генерация нормально распределенных случайных величин методом полярных координат.
6. Композиция законов распределения независимых случайных величин.
7. Квантиль распределения.
8. Гамма- и бета-функции.
9. Распределение хи-квадрат Пирсона.
10. t-распределение Стьюдента.
11. F-распределение Фишера — Снедекора.
12. Гамма-распределение.
13. Бета-распределение.
14. Распределение Дирихле.
15. Распределение Коши.
16. Законы распределения статистик экстремальных значений (Гумбеля, Фреше, Вейбулла).

Глава 8
Закон больших чисел

— Законы статистики везде одинаковы, — продолжал Николай Петрович солидно. — Утром, например, гостей бывает меньше, потому что публика еще исправна; но чем больше солнце поднимается к зениту, тем наплыв делается сильнее. И, наконец, ночью, по выходе из театров — это почти целая оргия!

— И заметьте, — пояснил Семен Иванович, — каждый день, в одни и те же промежутки времени, цифры всегда одинаковые. Колебаний — никаких! Такова незыблемость законов статистики!

М. Е. Салтыков-Щедрин
«За рубежом»

8.1. Сущность закона больших чисел

Под законом больших чисел в теории вероятностей понимается совокупность теорем, в которых устанавливается связь между средним арифметическим достаточно большого числа случайных величин и средним арифметическим их математических ожиданий.

На протяжении веков математическая наука, основываясь на логике Аристотеля, развивалась под влиянием основной потребности — исключить неточность и неопределенность в человеческих суждениях о природе наблюдаемых социально-экономических, технических и других процессов и явлений.

В повседневной жизни, бизнесе, научных исследованиях мы постоянно сталкиваемся с процессами и явлениями с неопределенным исходом. Например: торговец не знает, сколько посетителей придет к нему в магазин; бизнесмен не знает курс доллара через 1 день или год; банкир — вернут ли ему заем в срок; страховые компании — когда, кому и в каком размере придется выплачивать страховое вознаграждение.

Развитие любой науки предполагает установление основных закономерностей и причинно-следственных связей в виде определений, правил, аксиом, теорем. Однако более углубленное рассмотрение указанных закономерностей показывает неточность первоначальных утверждений. Последнее связано с открытием в 1927 г. Гейзенбергом «принципа неопределенности», заключающегося в ограниченности «измерительного познания» окружающего нас мира.

Таким образом, неопределенность может служить философским и методологическим объяснением ограниченности традиционных подходов к моделированию экономических, биологических и других процессов. Математическая статистика, бурно развивавшаяся последние 200 лет, позволяет учитывать случайность и неопределенность при анализе статистических данных для принятия

решений. Это послужило толчком к развитию теории ошибок наблюдений, теории связи, радиотехники, теории автоматического управления, социологии, экономики, психологии и других наук.

Основой для математической статистики служит математический аппарат и выводы теории вероятностей, изучающей закономерности, происходящие в массовых, однородных случайных явлениях и процессах.

Связующим звеном между теорией вероятностей и математической статистикой являются так называемые предельные теоремы, к которым относится закон больших чисел. Закон больших чисел — это общий принцип, согласно которому совместное действие большого числа случайных факторов приводит, при некоторых весьма общих условиях, к результату, практически не зависящему от случая. В самом общем виде закон больших чисел сформулировал П. Чебышёв. Большой вклад в изучение закона больших чисел внесли А. Колмогоров, А. Хинчин, Б. Гнеденко, В. Гливенко.

К предельным теоремам относится также так называемая Центральная предельная теорема А. Ляпунова, устанавливающая условия, при которых сумма случайных величин будет стремиться к случайной величине с нормальным законом распределения. Эта теорема позволяет обосновать методы проверки статистических гипотез, корреляционно-регрессионный анализ и др. методы классической статистики.

Дальнейшее развитие центральной предельной теоремы связано с именами Я. Линденберга, С. Бернштейна, А. Хинчина, П. Леви.

Замечание. 1) Теория вероятностей и классическая математическая статистика, получившая развитие в XIX и первой половине XX веков, трактуют понятие неопределенности только с точки зрения вероятности (вероятностная неопределенность). Между тем *вероятность имеет место* на практике (равно как и законы распределения вероятностей) только *при наличии устойчивой частоты* появления события, стремящейся к некоторому числу (об этом писали С. Н. Бернштейн, В. Н.Тутубалин, С. А. Айвазян, А. И. Орлов и др.). В других случаях говорить о вероятностной неопределенности нельзя (говорят о неопределенности II рода, которая является предметом теории игр, теории возможностей и теории нечетких множеств). А. Эйнштейн вообще не верил в существование вероятности, считая, что все процессы детерминированы. С другой стороны, многие ученые (физики-теоретики) считают вероятность неотъемлемой частью нашей жизни. С 1990-х годов возрос интерес к *субъективной вероятности*, которая является основой байесовского подхода к статистическому выводу, опирающемуся на формулу Байеса (хотя сам подход рассматривался еще Лапласом). Байесовская статистика сегодня является основой машинного обучения и рассматривает частотную интерпретацию вероятности в качестве частного случая (см. часть III в [53]).

2) Практическое применение методов теории вероятностей и математической статистики основано на двух принципах, фактически основывающихся на предельных теоремах:

- ✓ *принцип невозможности наступления маловероятного события*;
- ✓ *принцип достаточной уверенности в наступлении события, вероятность которого близка к единице*.

3) Следует отметить, что в XX веке была известна ограниченность предельных теорем в силу того, что выборки, имеющие место на практике, — конечны, а зашумленность реальных данных ведет к существенным отклонениям от предполагаемых законов распределения и искажению числовых оценок их параметров. Однако в связи с ростом объема данных и развитием информационных технологий их анализа, в частности успехом машинного обучения (глубокого обучения), объясняющимся эффектом концентрации меры функций большого числа переменных, интерес к законам больших чисел и предельным теоремам снова возрос (см. часть III в [53]).

4) В последние несколько десятков лет осознана роль степенных (экстремальных, см. 7.3) законов распределения в различных социально-экономических, биологических и других системах, которые имеют выражение в ряде известных эмпирических законов, например Ципфа, Парето. Так последний закон утверждает, что в экономике на 80% результатов влияет 20% затрат. ■

8.2. Неравенство и теорема Чебышёва

Рассмотрим закон больших чисел в форме Чебышёва.

Лемма Чебышёва (*Маркова*). Если случайная величина X принимает только неотрицательные значения и имеет математическое ожидание $M(X)$, то для любого $t > 0$ имеет место неравенство

$$P(X \geq t) \leq \frac{M(X)}{t}. \tag{8.1}$$

Доказательство. Пусть случайная величина X принимает значения $x_1, x_2, \dots, x_n$ с вероятностями, соответственно $p_1, p_2, \dots, p_n$, причем $x_i \geq 0$, которые расположены в порядке возрастания их значений, т. е. $x_1 < x_2 < \dots < x_n$.

Возьмем некоторое число $t > 0$. Тогда часть значений случайной величины X будет больше или равна t, а часть — меньше t. Допустим, что первые k значений меньше t. Найдем математическое ожидание случайной величины X:

$$M(X) = \sum_{i=1}^{n} x_i p_i = \sum_{i=1}^{k} x_i p_i + \sum_{i=k+1}^{n} x_i p_i. \tag{8.2}$$

Так как $x_i \geq 0$, $p_i > 0$, то $x_i p_i \geq 0$. В (8.2) отбросим те слагаемые, при которых $x_i < t$, то есть первые k слагаемых. От этого сумма (8.2) уменьшится.

$$M(X) \geq \sum_{i=k+1}^{n} x_i p_i. \tag{8.3}$$

Если в (8.3) $x_{k+1}, x_{k+2}, \dots, x_n$ заменить на число t, то неравенство только усилится, т. е. имеем:

$$M(X) \geq \sum_{i=k+1}^{n} t p_i. \quad M(X) \geq t \sum_{i=k+1}^{n} p_i, \quad \sum_{i=k+1}^{n} p_i \leq \frac{M(X)}{t}. \tag{8.4}$$

Так как $x_{k+1}, x_{k+2}, \dots, x_n$ представляют собой значения случайной величины $X \geq t$, то $\sum_{i=k+1}^{n} p_i$ есть вероятность того, что случайная величина X примет значения большее или равное t. Следовательно,

$$P(X \geq t) \leq \frac{M(X)}{t}.$$

Неравенство доказывают аналогично и для непрерывных случайных величин, взяв вместо сумм соответствующие интегралы.

Неравенство Чебышёва. Если случайная величина X имеет математическое ожидание $M(X)$ и дисперсию $D(X)$, то для любого $\varepsilon > 0$ имеет место неравенство:

$$P(|x - M(x)| < \varepsilon) \geq 1 - \frac{D(X)}{\varepsilon^2}. \quad (8.5)$$

Доказательство. Пусть X — произвольная случайная величина, которая имеет конечное математическое ожидание и дисперсию. Заметим, что неравенство $(X - M(X))^2 < \varepsilon^2$ и $|X - M(X)| < \varepsilon$ равносильны. Воспользуемся неравенством Маркова (8.1), в котором вместо случайной величины X возьмем $(X - M(X))^2$, а t заменим на ε^2.

$$P\{(X - M(X))^2 \geq \varepsilon^2\} = P(|X - M(X)| \geq \varepsilon) \leq \frac{M(X - M(X))^2}{\varepsilon^2}$$

или

$$P(|X - M(X)| \geq \varepsilon) \leq \frac{D(X)}{\varepsilon^2}. \quad (8.6)$$

Так как события $|X - M(X)| \geq \varepsilon$ и $|X - M(X)| < \varepsilon$ несовместные и образуют полную группу, то

$$P(|X - M(X)| \geq \varepsilon) + P(|X - M(X)| < \varepsilon) = 1.$$

Отсюда получаем

$$P(|X - M(X)| < \varepsilon) = 1 - P(|X - M(X)| \geq \varepsilon). \quad (8.7)$$

Применив неравенство (8.6) к уравнению (8.7), получаем неравенство:

$$P(|X - M(X)| < \varepsilon) \geq 1 - \frac{D(X)}{\varepsilon^2}.$$

Неравенство Чебышёва является в теории вероятностей общим фактом и позволяет оценить нижнюю границу вероятности.

Следствие. Если произведено n независимых испытаний по схеме Бернулли, где p — вероятность успеха, q — вероятность неудачи, n — число опытов, k — число успехов, то вероятность того, что абсолютное отклонение числа появления события A в независимых испытаниях от $M(X) = np$ будет меньше некоторого положительного числа ε, не меньше, чем $1 - \frac{npq}{\varepsilon^2}$:

$$P(|k - np| < \varepsilon) \geq 1 - \frac{npq}{\varepsilon^2}. \quad (8.8)$$

Для относительной частоты появления события в независимых испытаниях $\left(\frac{k}{n}\right)$ аналогичное неравенство имеет вид

$$P\left(\left|\frac{k}{n} - p\right| < \varepsilon\right) \geq 1 - \frac{pq}{n\varepsilon^2}. \quad (8.9)$$

Теорема Чебышёва. Если $X_1, X_2, \ldots, X_n$ — последовательность попарно независимых случайных величин, дисперсии каждой из которых ограничены сверху числом C, то каково бы ни было постоянное число $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} M(X_i)\right| < \varepsilon\right\} = 1. \quad (8.10)$$

Доказательство. Пусть $X_1, X_2, \ldots, X_n$ — попарно независимые случайные величины. Найдем среднее арифметическое этих величин:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}, M(\bar{X}) = M\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right).$$

Воспользуемся свойствами математического ожидания: математическое ожидание суммы независимых случайных величин равно сумме их математических ожиданий; постоянный множитель можно выносить за знак математического ожидания.

$$M(\bar{X}) = \frac{M(X_1) + M(X_2) + \ \dots \ + M(X_n)}{n}. \qquad (8.11)$$

Используем свойства дисперсии случайной величины: дисперсия суммы независимых случайных величин равна сумме дисперсий этих величин; постоянный множитель можно выносить за знак дисперсии, возведя его при этом в квадрат. Тогда дисперсия среднего арифметического *n* попарно независимых случайных величин будет равна

$$D(\bar{X}) = D\left(\frac{X_1+X_2+ \dots +X_n}{n}\right) = \frac{D(X_1)+D(X_2)+ \dots +D(X_n)}{n^2}. \qquad (8.12)$$

По условию $D(X_i) \le C$, поэтому в (8.12), заменив $D(X_i)$ на число *C*, получим

$$D(\bar{X}) \le \frac{nC}{n^2}, D(\bar{X}) \le \frac{C}{n}. \qquad (8.13)$$

К величине $\bar{X}$ применим неравенство Чебышёва:

$$P(|\bar{X} - M(\bar{X})| < \varepsilon) \ge 1 - \frac{D(\bar{X})}{\varepsilon^2}. \qquad (8.14)$$

Учитывая (8.11)–(8.13), получим

$$P\left(\left|\frac{X_1 + X_2 + \ \dots \ + X_n}{n} - \frac{M(X_1) + M(X_2) + \ \dots \ + M(X_n)}{n}\right| < \varepsilon\right) \ge 1 - \frac{D(\bar{X})}{\varepsilon^2}$$

или

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} M(X_i)\right| < \varepsilon\right) \ge 1 - \frac{C}{n\varepsilon^2}. \qquad (8.15)$$

Перейдем к пределу при $n \to \infty$. Тогда получим

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} M(X_i)\right| < \varepsilon\right) \ge 1.$$

Известно, что вероятность любого события не может быть больше единицы, поэтому

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} M(X_i)\right| < \varepsilon\right) = 1.$$

Таким образом, среднее арифметическое достаточно большого числа независимых случайных величин с вероятностью, сколь угодно близкой к единице, будет сколь угодно мало отличаться от среднего арифметического математических ожиданий этих величин. Значит, если число случайных величин велико, а случайные величины имеют ограниченные сверху дисперсии, то среднее арифметическое этих величин является устойчивым и отражает общую закономерность данного явления. Имеющиеся отклонения от среднего взаимно погашаются.

Следствие 1. Если $X_1, X_2, \dots, X_n$ — последовательность попарно независимых случайных величин, имеющих одинаковые математические ожидания ($M(X_i)=a$) и равномерно ограниченные дисперсии, то, как бы ни было мало постоянное число $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - a\right| < \varepsilon\right) = 1. \qquad (8.16)$$

Следствие 1 является наиболее простой формой закона больших чисел, отражает устойчивость средних арифметических большого числа одинаково распределенных взаимно независимых случайных величин, сходимость средних арифметических случайных величин к общему математическому ожиданию. Например, если производится многократное измерение прибором некоторой величины, то результаты отдельных измерений будут случайными. Если отсутствуют систематические ошибки, то математические ожидания результатов измерений будут постоянны и равны определенному числу.

Исходя из закона больших чисел, среднее арифметическое большого числа измерений будет сколь угодно мало отличаться от «истинного» значения, точнее при $n \to \infty$ среднее арифметическое результатов измерений «сходится по вероятности» к постоянной величине, выражаемой математическим ожиданием. *Этот частный случай закона больших чисел позволяет обосновать правило средней арифметической.*

Следствие 2 (**теорема Бернулли**). Если вероятность наступления события *A* в каждом из *n* независимых испытаний постоянна и равна *p*, а число испытаний достаточно велико, то сколь угодно близка к единице вероятность того, что относительная частота события будет сколь угодно мало отличаться от постоянной вероятности:

$$\lim_{n\to\infty} P\left(\left|\frac{k}{n} - p\right| < \varepsilon\right) = 1. \qquad (8.17)$$

Доказательство. Пусть производится *n* независимых испытаний. Вероятность появления события *A* в каждом испытании постоянна и равна *p*, тогда вероятность непоявления события *A* также постоянна и равна *q*. Обозначим X_i — случайную величину — число появления события *A* в *i*-ом испытании, $i = 1, 2, \ldots, n$. В одном испытании каждая из величин принимает только два значения 0 или 1. Как было показано, $M(X_i) = p,\ D(X_i) = pq$.

Если случайные величины независимы и имеют ограниченные сверху дисперсии, то к ним можно применить теорему Чебышёва. Независимость величин X_i вытекает из независимости событий. А так как $p + q = 1$, то можно показать, что при любых значениях *p* и *q* произведение $pq \leq 0{,}25$. Поэтому дисперсии случайных величин ограничены числом 0,25.

Применим следствие 1 теоремы Чебышёва. Учитывая, что математическое ожидание каждой из величин равно *p*, получим

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - p\right| < \varepsilon\right) = 1. \qquad (8.18)$$

В формуле (8.18) $\sum_{i=1}^{n} X_i$ есть сумма числа появления события *A* в каждом из *n* независимых испытаний и равна *k* — общему числу появления события *A* в *n* независимых испытаниях, т. е.

$$X_1 + X_2 + \ \ldots\ + X_n = k.$$

Тогда

$$\frac{X_1 + X_2 + \ \ldots\ + X_n}{n} = \frac{k}{n}.$$

Подставив это значение в (8.18), получим

$$\lim_{n\to\infty} P\left(\left|\frac{k}{n} - p\right| < \varepsilon\right) = 1.$$

Теорема Бернулли устанавливает связь между относительной частотой появления события A и постоянной вероятностью. Если число испытаний достаточно велико, то относительная частота обладает свойством устойчивости и становится практически неслучайной. Так, если многократно подбрасывать монету в одних и тех же условиях, то доля выпадений «герба» будет стремиться к 0,5. Например, английский статистик К. Пирсон подбрасывал монету 24 000 раз, при этом герб выпал 12 012 раз, т. е. относительная частота выпадения герба составила 0,5005.

На практике часто встречаются события с разной вероятностью появления в каждом испытании.

Следствие 3 (**теорема Пуассона**). Если в последовательности независимых испытаний вероятность появления события A в i-м испытании равна p_i, $i = 1, 2, 3, \dots, n$, а число испытаний достаточно велико, то сколь угодно близка к единице вероятность того, что относительная частота события A будет сколь угодно мало отличаться от среднего арифметического вероятностей p_i:

$$\lim_{n\to\infty} P\left(\left|\frac{k}{n} - \frac{p_1+p_2+\dots+p_n}{n}\right| < \varepsilon\right) = 1. \tag{8.19}$$

Доказательство. Пусть X_i — случайная величина, число появления события A в i-ом испытании, $i = 1, 2, \dots, n$. $M(X_i) = p_i$, $D(X_i) = p_i q_i$, так как $p_i + q_i = 1$, то $D(X_i) \le 0{,}25$. $k = X_1 + X_2 + \dots + X_n$ есть число появления события A в n независимых испытаниях, и $\frac{k}{n}$ — относительная частота появления события A. Отсюда вытекает, что теорема Пуассона является частным случаем теоремы Чебышёва.

Замечание. 1. *Законы больших чисел не позволяют уменьшить неопределенность в каждом конкретном случае*, они утверждают лишь о существовании закономерности при достаточно большом числе опытов. Например, если при подбрасывании монеты 100 раз появился герб, то это не означает, что в следующий — 101-й раз появится цифра.

2. Сходимость случайной последовательности $\{X_n\}$ к пределу X при $n \to \infty$ называется *сходимостью по вероятности* и обозначается как

$$X_n \xrightarrow{P} X, \tag{8.20}$$

если для любого $\varepsilon > 0$:

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1 \tag{8.21}$$

или

$$\lim_{n\to\infty} P(|X_n - X| \ge \varepsilon) = 0. \tag{8.22}$$

3. Случайная последовательность $\{X_n\}$ *почти наверное* сходится к пределу X при $n \to \infty$ и обозначается так

$$X_n \xrightarrow{\text{п.н.}} X,$$

если

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1. \tag{8.23}$$

4. В случае непрерывной случайной величины X говорят о *сходимости последовательности* $\{X_n\}$ *по распределению* (либо в точках непрерывности дискретной случайной величины):

$$X_n \xrightarrow{d} X, \tag{8.24}$$

если

$$P\left(\lim_{n\to\infty} X_n < x\right) = P(X < x). \tag{8.25}$$

Понятия сходимости по вероятности, сходимости почти наверное, сходимости по распределению используются в математической статистике, причем из (8.24) следует (8.20)–(8.23), обратное верно не всегда. Сходимость по вероятности (и сходимость почти наверное) влечет сходимость по распределению. ■

Пример 8.1. С помощью неравенства Чебышёва оценить вероятность того, что при подбрасывании 12 игральных костей сумма очков на верхних гранях отклонится от математического ожидания меньше, чем на 15.

Решение. X_i — случайная величина — число очков на i-й кости ($i = 1, 2, \dots, 12$). Тогда $X = X_1 + X_2 + \cdots + X_{12}$, где X — сумма числа очков при подбрасывании 12 игральных костей.

Случайная величина X_i имеет закон распределения.

X_i	1	2	3	4	5	6
p_i	1/6	1/6	1/6	1/6	1/6	1/6

Так как кости одинаковы, то

$$\mathrm{M}(X_i) = \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = \frac{7}{2},$$
$$M\left(X_i^2\right) = \frac{1}{6}(1^2+2^2+3^2+4^2+5^2+6^2) = \frac{91}{6},$$
$$\mathrm{M}(X) = 3{,}5 \cdot 12 = 42, D(X_i) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12},$$
$$D(X) = (35/12)12 = 35.$$

Согласно неравенству Чебышёва, имеем

$$P(|X-42| < 15) \geq 1 - \frac{35}{225}, P(|X-42| < 15) \geq 0{,}844.$$

8.3. Понятие о центральной предельной теореме

В теории вероятностей и математической статистике большое значение имеют предельные теоремы, из которых наиболее общей является центральная предельная теорема Ляпунова, в которой утверждается, что если сложить большое число случайных величин, имеющих один и тот же или различные законы распределения, то случайная величина, являющаяся результатом суммы, при некоторых условиях, будет иметь нормальный закон распределения.

При доказательстве предельных теорем используется *характеристическая функция* $g_X(t)$:

$$g_X(t) = M\left(e^{itx}\right) = \int_{-\infty}^{+\infty} e^{itx} f(x)dx =$$
$$= \int_{-\infty}^{+\infty} \cos(tx)\, f(x)dx + i\int_{-\infty}^{+\infty} \sin(tx)\, f(x)dx. \tag{8.26}$$

Формула (8.26) представляет собой известное в математическом анализе преобразование Фурье функции плотности вероятности $f(x)$ случайной величины X.

Свойства $g(t)$:

1) $g_x(t)$ — непрерывная функция;

2) если $Y = a + bX$, то $g_y(t) = g_x(bt)e^{iat}$;

3) если $X_1,\ X_2$ — независимые случайные величины и $Y = X_1 + X_2$, то

$$g_y(t) = g_{X_1}(t)g_{X_2}(t);$$

4) если непрерывная случайная величина X имеет начальный момент порядка n (α_n), то она дифференцируема k раз $(k \le n)$:

$$f^{(k)}(0) = i^k\alpha_k,$$

следовательно, допускается разложение по формуле Маклорена:

$$g_x(t) = 1 + \alpha_1 it + \alpha_1\frac{(it)^2}{2!} + \cdots + \alpha_n\frac{(it)^n}{n!} + O(t^{n+1}).$$

Свойство 3 объясняет широкое применение характеристических функций — при суммировании независимых случайных величин, перемножив их характеристические функции, мы получим характеристическую функцию суммы.

Пусть случайная величина X подчиняется стандартному нормальному закону: $X \to N(0,1)$. Тогда

$$g_X(t) = M\left(e^{itx}\right) = \int_{-\infty}^{+\infty} e^{itx}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx = |y = x - it| =$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty-it}^{+\infty-it} e^{-\frac{y^2+t^2}{2}}dy = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\int_{-\infty-it}^{+\infty-it} e^{-\frac{y^2}{2}}dy = e^{-\frac{t^2}{2}}. \qquad (8.27)$$

$\int_{-\infty-it}^{+\infty-it} e^{-\frac{y^2}{2}}dy = \int_{-\infty}^{+\infty} e^{-\frac{u^2}{2}}du = \sqrt{2\pi}$ — как следует из теории функций комплексного переменного.

Итак, для стандартного нормального закона распределения, характеристическая функция имеет вид

$$g_X(t) = e^{-\frac{t^2}{2}}. \qquad (8.28)$$

Замечание. 1) Комплексная случайная величина, определяется так же, как и действительная, но каждому событию ставится в соответствие комплексное число $z = x + iy$, где i — комплексная единица $(i^2 = -1)$.

По формуле Эйлера $e^{ix} = \cos(x) + isin(x)$. При $x = \pi$ получим замечательную формулу $e^{i\pi} + 1 = 0$, в которой объединяются логика, геометрия, арифметика, алгебра и анализ.

2) Доказательство формулы (8.27) следует из теоремы Коши для аналитической функции в теории функций комплексного переменного: интеграл по замкнутому контуру равен нулю. Можно показать, рассмотрев контур $(-R - it;\ -R), (-R; R), (R; R - it), (R - it;\ -R - it)$ при $R \to +\infty$, что

$$\int_{-\infty-it}^{+\infty-it} e^{-\frac{y^2}{2}}dy = \lim_{R\to+\infty}\int_{-R}^{R} e^{-\frac{u^2}{2}}du = \sqrt{2\pi},$$

где второй интеграл — вариант формулы Эйлера — Пуассона (см. 5.3). ■

Примерами центральных предельных теорем являются теорема Линденберга — Леви (для суммы независимых одинаково распределенных случайных величин) и ее частный случай (для последовательности независимых случайных величин) — интегральная теорема Муавра — Лапласа.

Теорема 1 (**Муавра — Лапласа**). Пусть производится n независимых опытов в каждом из которых вероятность наступления события A равна p (не наступления $q = 1 - p, p \neq 0, p \neq 1$). Если K — число появлений события A в серии из n испытаний, то при достаточно больших n случайную величину K можно считать нормально распределенной:

$$M(\text{К}) = np, \sigma(K) = \sqrt{D(K)} = \sqrt{npq};$$
$$P(K < k) \to P(X < x_0) = \int_{-\infty}^{x_0} \varphi(x)dx = \frac{1}{2} +, \Phi(x_0), \qquad (8.29)$$

где $x_0 = \frac{k-np}{\sqrt{npq}}, \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \Phi(x_0)$ — функция Лапласа.

Теорема 2 (**Линденберга — Леви**). Если случайные величины X_1, $\text{X}_2, \ldots, \text{X}_n$ независимы, одинаково распределены и имеют конечную дисперсию, то при $n \to \infty$

$$P\left(\frac{X_1 + X_2 + \cdots + X_n - na}{\sigma\sqrt{n}} < t\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{u^2}{2}} du, \qquad (8.30)$$

где $M(\text{X}_i) = a$, $\sigma^2 = D(\text{X}_i)$; $\frac{\sum X_i - na}{\sigma\sqrt{n}} = U$ — нормально распределенная случайная величина, $M(U) = 0, D(U) = 1$.

Доказательство. Пусть

$$U = \frac{\sum(X_i - a)}{\sigma\sqrt{n}}.$$

Рассмотрим величину $Y_i = (X_i - a)$ и ее характеристическую функцию

$$g_{Y_i}(t) = g_{X_i - a}(t).$$

Согласно свойству 4 характеристической функции, имеем

$$g_{X_i - a}(t) = 1 + M(X_i - a)\frac{(it)^1}{1!} + M((X_i - a)^2)\frac{(it)^2}{2!} + O(t^3). \qquad (8.31)$$

Учитывая, что $M(X_i - a) = 0$, $M((X_i - a)^2) = \sigma^2$ перепишем (8.25) в виде

$$g_{X_i - a}(t) = 1 + \sigma^2 \frac{(it)^2}{2!} + O(t^3). \qquad (8.32)$$

Согласно свойству 3 характеристической функции, для суммы независимых слагаемых ($\sum(X_i - a)$), получим

$$g_{\sum(X_i - a)}(t) = \left[g_{X_i - a}(t)\right]^n.$$

Для величины $U = \frac{\sum(X_i - a)}{\sigma\sqrt{n}}$, согласно свойству 2 характеристической функции и формуле (8.32), имеем

$$g_U(t) = g_{\sum(X_i - a)}\left(\frac{t}{\sigma\sqrt{n}}\right) = \left[g_{X_i - a}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n =$$
$$= \left[1 + \sigma^2 \frac{(i)^2}{2!}\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + O\left(\left(\frac{t}{\sigma\sqrt{n}}\right)^3\right)\right]^n =$$
$$= \left[1 - \frac{\sigma^2}{2!}\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + O\left(\left(\frac{t}{\sigma\sqrt{n}}\right)^3\right)\right]^n = \left[1 - \frac{t^2}{2n} + O\left(\left(\frac{t}{\sigma\sqrt{n}}\right)^3\right)\right]^n =$$
$$= e^{\lim\limits_{n\to\infty} ln\left[1 - \frac{t^2}{2n} + O\left(\left(\frac{t}{\sigma\sqrt{n}}\right)^3\right)\right]^n}.$$

Следовательно,

$$\lim_{n\to\infty} g_U(t) = e^{\lim\limits_{n\to\infty} ln\left[1 - \frac{t^2}{2n} + O\left(\left(\frac{t}{\sigma\sqrt{n}}\right)^3\right)\right]^n} = e^{\lim\limits_{n\to\infty} n\left[-\frac{t^2}{2n} + O\left(\left(\frac{t}{\sigma\sqrt{n}}\right)^3\right)\right]} = e^{-\frac{t^2}{2}},$$

то есть характеристическая функция нормированной (стандартизированной) случайной величины U стремится при $n \to \infty$ к характеристической функции стандартного нормального закона распределения (8.28). Значит, для функции распределения нормированной суммы независимых одинаково распределенных случайных величин при достаточно больших значениях n (числа переменных) справедлива приближенная формула (8.30), что и требовалось доказать.

Для схемы Бернулли из теоремы 2 (Линденберга — Леви) следует теорема 1 (Муавра — Лапласа).

Центральная предельная теорема показывает, что явления и процессы, подверженные большому числу случайных воздействий, распределены по нормальному закону.

Замечание. Используя теорему Линденберга — Леви, можно показать, что сумма 12 равномерно распределенных на [0, 1] случайных величин стремится к стандартному нормальному закону (см. пример 8.1). Это можно сделать как формально, так и опираясь на численный эксперимент, например с использованием генератора случайных чисел в *MS Excel*. ■

Пример 8.2. На отрезке [0; 1] случайным образом выбрано 100 чисел, точнее рассматриваются 100 независимых средних $X_1, \ldots, X_n$, равномерно распределенных на отрезке [0; 1]. Найти вероятность того, что их сумма заключена между 51 и 60, т. е. $P(51 \le \sum X_i \le 60)$.

Решение. В силу теоремы 2

$$\sum X_i = \sigma\sqrt{n}U + na.$$

Кроме того, линейная комбинация случайных величин, подчиняющихся нормальному закону, будет подчиняться нормальному закону. У нас полагается, что $U \to N(0,1)$, следовательно,

$$\sum X_i \to N(na, n\sigma^2).$$

Из условия, в силу равномерности случайной величины X_i,

$$M(X_i) = \frac{1+0}{2} = \frac{1}{2}, \quad \sigma^2 = \frac{(1-0)^2}{12} = \frac{1}{12}.$$

Следовательно,

$$M(\sum X_i) = 50, D(\sum X_i) = \frac{100}{12}.$$

Итак, $\sum X_i \to N(50, \frac{100}{12})$ — сумма, нормально распределенная случайная величина с математическим ожиданием $M(\sum X_i) = na = 50$ и дисперсией $D(\sum X_i) = n\sigma^2 = 100/12$.

Отсюда

$$P(51 \le \sum X_i \le 60) = \Phi\left(\frac{60-na}{\sqrt{n}\sigma}\right) - \Phi\left(\frac{51-na}{\sqrt{n}\sigma}\right) =$$

$$= \Phi\left(\frac{60-50}{\frac{10}{\sqrt{12}}}\right) - \Phi\left(\frac{51-50}{\frac{10}{\sqrt{12}}}\right) = \Phi(\sqrt{12}) - \Phi\left(\frac{\sqrt{12}}{10}\right) =$$

$$= \Phi(3{,}464) - \Phi(0{,}3464) \approx 0{,}49971 - 0{,}1353 = 0{,}3644.$$

То есть вероятность того, что сумма 100 независимых средних $X_1, X_2, \ldots, X_n$, равномерно распределенных на отрезке [0; 1], заключена между 51 и 60, равна 0,3644.

Темы (вопросы) для самоконтроля

1. Сущность закона больших чисел и его значение в науке и технике.
2. Неравенство Маркова.
3. Неравенство Чебышёва.
4. Теорема Чебышёва и следствия из нее (теорема об устойчивости средних одинаково распределенных случайных величин, теорема Бернулли, теорема Пуассона).
5. Характеристическая функция.
6. Теорема Муавра — Лапласа.
7. Теорема Линденберга — Леви.

Глава 9
Цепи Маркова

Протекающий в системе S случайный процесс называется марковским, если для каждого момента времени вероятность любого состояния системы в будущем зависит только от ее состояния в настоящем и не зависит от того, когда и каким образом система пришла в это состояние.

Случайный процесс называется процессом с дискретными состояниями, если возможные состояния системы $S_1, S_2, S_3, \ldots$ можно перечислить (занумеровать) одно за другим, а сам процесс состоит в том, что время от времени система S скачком (мгновенно) перескакивает из одного состояния в другое.

При анализе случайный процесс с дискретными состояниями удобно представить в виде графа состояний. Каждое состояние можно изображать прямоугольником, а возможные переходы из состояния в состояние — стрелками, соединяющими эти переходы.

Случайный процесс называется процессом с дискретным временем, если переходы из одного состояния в другое возможны только в строго определенные, заранее фиксированные моменты времени: $t_1, t_2, t_3, \ldots$ В промежутки времени между этими моментами система S сохраняет свое состояние.

Пусть имеется физическая система S, которая может находиться в состояниях: $S_1, S_2, \ldots, S_n$, причем переходы системы из состояния в состояние возможны только в моменты времени: $t_1, t_2, \ldots, t_k, \ldots$ Будем называть эти моменты «шагами» или «этапами» процесса. В общем случае система может не только менять свое состояние, но и оставаться в прежнем.

Будем обозначать через $S_i^{(k)}$ событие, заключающееся в том, что после k шагов система находится в состоянии S_i.

Последовательность состояний $S_1^{(k)}, S_2^{(k)}, \ldots, S_i^{(k)}, \ldots, S_n^{(k)}$ образует полную группу событий при любом k (s_i — состояние системы или, иначе, значение случайной переменной $S_i^{(k)}$ в момент времени k).

Происходящий в системе процесс можно представить как последовательность (цепочку) событий, например: $S_2^{(0)}, S_1^{(1)}, S_4^{(2)}, S_2^{(3)}, \ldots$ Такая случайная последовательность событий называется марковской цепью (*Marcov chain*) с дискретным временем, если для каждого шага вероятность перехода из любого состояния S_i в состояние S_j не зависит от того, когда и как система пришла в состояние S_i, а зависит только от предыдущего значения:

$$P(S_j^{(k+1)}) = s_j / S_i^{(k)} = s_i, \ldots, S_i^{(0)} = s_i) = P(S_j^{(k+1)} = s_j / S_i^{(k)} = s_i). \quad (9.1)$$

Наиболее важной ее характеристикой являются безусловные вероятности нахождения системы S на любом −ом шаге в состоянии S_j. Обозначим эту вероятность $p_j(k)$:

$$p_j(k) = P\left\{S_j^{(k)}\right\}, j = 1, 2, \ldots, n;\ k = 0, 1, 2, \ldots, \quad (9.2)$$

при этом для каждого номера шага k выполняется

$$\sum_{j=1}^{n} p_j(k) = 1. \quad (9.3)$$

Вероятности $p_1(k), p_2(k), \dots, p_n(k)$ будем называть вероятностями состояний. Для нахождения этих вероятностей необходимо знать условные вероятности перехода системы S на $-$ом шаге в состояние S_j, если известно, что на предыдущем $(k-1)$-ом шаге она была в состоянии S_i. Обозначим эту вероятность:

$$P_{ij}^{(k)} = P\{S_j^{(k)}/S_i^{(k-1)}\},\ i,j = 1,2,\dots,n. \tag{9.4}$$

Вероятности $P_{ij}^{(k)}$ называются переходными вероятностями марковской цепи на k-ом шаге. Вероятность $P_{ii}^{(k)}$ есть вероятность того, что на k-ом шаге система задержится (останется) в состоянии S_i.

Цепь называется *однородной*, если переходные вероятности $P_{ij}^{(k)}$ не зависят от номера шага k: $P_{ij}^{(k)} = P_{ij}$. В противном случае цепь называется неоднородной.

Рассмотрим однородную марковскую цепь. Переходные вероятности P_{ij} можно записать в виде матрицы размерности $n \times n$:

$$P = \begin{pmatrix} P_{11} & P_{12} & \dots & P_{1j} & \dots & P_{1n} \\ P_{21} & P_{22} & \dots & P_{2j} & \dots & P_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ P_{i1} & P_{i2} & \dots & P_{ij} & \dots & P_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ P_{n1} & P_{n2} & \dots & P_{nj} & \dots & P_{nn} \end{pmatrix}. \tag{9.5}$$

Так как на каждом шаге система S может находиться только в одном из взаимно исключающих состояний, то для любой i-ой строки матрицы (9.5) сумма всех стоящих в ней переходных вероятностей P_{ij} равна единице:

$$\sum_{j=1}^{n} P_{ij} = 1, i = 1, 2, \dots, n, \tag{9.6}$$

такая строка P_i назывется вероятностной.

Матрица, удовлетворяющая условию (9.6), называется стохастической. Задание матрицы переходов P_{ij} и начальных вероятностей $P_{ij}^{(0)}$ задает цепь Маркова.

Вероятностный процесс, в котором вероятность перехода в любое данное состояние одна и та же на всех этапах процесса, называется *стационарным марковским процессом*.

Основной задачей исследования марковской цепи является нахождение безусловных вероятностей нахождения системы S на любом $k-$ом шаге в состоянии S_i $(i = 1, 2, \dots, n;\ k = 0, 1, 2, \dots)$, если задана матрица переходных вероятностей (9.5) и $p_i(0)$ $(i = 1, 2, \dots, n)$ — начальное распределение вероятностей.

Пусть система S в начальный момент $k = 0$ находилась в каком-то состоянии S_m. Тогда для начального момента (0) будем иметь

$$p_1(0) = 0, p_2(0) = 0, \dots, p_m(0) = 1, \dots, p_n(0) = 0. \tag{9.7}$$

За первый шаг система перейдет в состояния $S_1, S_2, \dots, S_m, \dots, S_n$ с вероятностями $P_{m1}, P_{m2}, \dots, P_{mm}, \dots, P_{mn}$ соответственно, составляющими $m-$ую строку матрицы переходных вероятностей (9.5). Таким образом, после первого шага имеем

$$p_1(1) = P_{m1},\ p_2(1) = P_{m2}, \dots,\ p_m(1) = P_{mm}, \dots,\ p_n(1) = P_{mn}. \tag{9.8}$$

Найдем вероятности состояний после второго шага по формуле полной вероятности с коэффициентами:

– после первого шага система была в состоянии S_1;

– после первого шага система была в состоянии S_2;

.....

– после первого шага система была в состоянии S_i;

.....

– после первого шага система была в состоянии S_n,

вероятности которых известны (9.8), соответствующие условные вероятности перехода в состояние S_i записаны в матрице переходных вероятностей (9.5).

По формуле полной вероятности имеем

$$p_i(2) = \sum_{j=1}^{n} p_j(1)\, P_{ji},\ \ i = 1,2, \dots, n. \qquad (9.9)$$

Переходя таким же способом от $k = 2$ к $k = 3$ и т. д., получим рекуррентные формулы:

$$p_i(k) = \sum_{j=1}^{n} p_j(k-1)\, P_{ji}, i = 1,2, \dots, n. \qquad (9.10)$$

Фактически в формуле (9.10) учитываются те слагаемые, для которых переходные вероятности отличны от нуля.

Замечание. В общем случае для неоднородной марковской цепи, рассматриваются на каждом k −ом шаге матрицы вероятностей перехода, состоящие из условных вероятностей перехода $P_{ij}^{(k)}$ системы S на k -ом шаге в состояние S_j, если известно, что на предыдущем $(k-1)$-ом шаге она была в состоянии S_i (9.4). Тогда безусловная вероятность $p_i(k)$ нахождения системы S на любом k -ом шаге в состоянии S_i выразится формулой

$$p_i(k) = \sum_{j=1}^{n} p_j(k-1)\, P_{ji}^{(k)}, i = 1,2, \dots, n. \qquad (9.11)$$

Итак, в общем виде

$$p_i(k) = \sum_{j=1}^{n} p_j(k-1)\, P_{ji}^{(k)}, i = 1,2, \dots, n. \ \blacksquare \qquad (9.12)$$

Пример 9.1. По некоторой цели производятся выстрелы в моменты времени $t_1, t_2, t_3, \dots$ Возможные состояния системы (цели):

S_1 — цель невредима,

S_2 — цель повреждена (но может функционировать),

S_3 — цель полностью поражена (не может функционировать).

Граф состояний представлен на рисунке 9.1.

Матрица переходных вероятностей:

$$(P) = \begin{pmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{pmatrix} = \begin{pmatrix} 0{,}1 & 0{,}7 & 0{,}2 \\ 0 & 0{,}6 & 0{,}4 \\ 0 & 0 & 1 \end{pmatrix}. \qquad (9.13)$$

Сколько надо сделать выстрелов, чтобы вероятность поражения цели оказалась не менее 0,6?

Решение. До того, как начались выстрелы, система (цель) находилась в состоянии s_1. После первого выстрела (после первого шага) вероятности состояний системы определяются из первой строки матрицы (9.13).

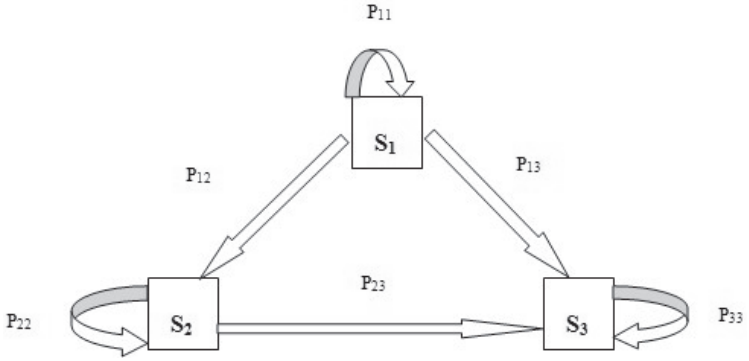


Рис. 9.1 — Граф состояний системы (цели)

$$p_1(1) = 0{,}1,\ \ p_2(1) = 0{,}7,\ \ p_3(1) = 0{,}2. \qquad (9.14)$$

Таким образом, после первого выстрела вероятность полного поражения цели составляет 0,2.

Делаем второй шаг (второй выстрел). Используя формулы (9.9) и (9.13), (9.14), находим:

$$p_1(2) = p_1(1)P_{11} + p_2(1)P_{21} + p_3(1)P_{31} = 0{,}1 \cdot 0{,}1 + 0{,}7 \cdot 0 + \\ + 0{,}2 \cdot 0 = 0{,}01;$$

$$p_2(2) = p_1(1)P_{12} + p_2(1)P_{22} + p_3(1)P_{32} = 0{,}1 \cdot 0{,}7 + 0{,}7 \cdot 0{,}6 + \\ + 0{,}2 \cdot 0 = 0{,}49;$$

$$p_3(2) = p_1(1)P_{13} + p_2(1)P_{23} + p_3(1)P_{33} = 0{,}1 \cdot 0{,}2 + 0{,}7 \cdot 0{,}4 + \\ + 0{,}2 \cdot 1 = 0{,}5.$$

Видим, что вероятность поражения цели равна теперь 0,5.

Делаем третий шаг (третий выстрел). Используя (9.9), (9.12) и полученные безусловные вероятности, находим:

$$p_1(3) = p_1(2)P_{11} + p_2(2)P_{21} + p_3(2)P_{31} = \\ = 0{,}01 \cdot 0{,}1 + 0{,}49 \cdot 0 + 0{,}5 \cdot 0 = 0{,}001;$$

$$p_2(3) = p_1(2)P_{12} + p_2(2)P_{22} + p_3(2)P_{32} = \\ = 0{,}01 \cdot 0{,}7 + 0{,}49 \cdot 0{,}6 + 0{,}5 \cdot 0 = 0{,}301;$$

$$p_3(3) = p_1(2)P_{13} + p_2(2)P_{23} + p_3(2)P_{33} = \\ = 0{,}01 \cdot 0{,}2 + 0{,}49 \cdot 0{,}4 + 0{,}5 \cdot 1 = 0{,}698.$$

Мы видим, что вероятность поражения цели равна теперь 0,698, т. е. больше 0,6. Значит, требуется сделать не меньше трех выстрелов, чтобы вероятность поражения цели превысила 0,6.

Рассмотрим основные свойства дискретных однородных цепей Маркова.

1) Матрица вероятностных переходов однородной дискретной цепи Маркова за n шагов получается возведением исходной матрицы в степень n:

$$P(n) = P^n.$$

Обозначим через $p_{ik}(n)$ вероятность перехода из состояния S_i в S_k за n шагов, которая равна вероятности суммы всех путей начинающихся в S_i и заканчивающихся в S_k. Например, $p_{ik}(1) = p_{ik}$,

$$p_{ij}(2) = \sum_{k=0}^{n} p_{ik}\, p_{kj}.$$

Рекуррентная формула, полученная по индукции, будет иметь вид

$$p_{ij}(n+1) = \sum_{k=0}^{n} p_{ik}\, p_{kj}(n).$$

Дальнейшее обобщение по индукции приводит к (дискретным) *уравнениям Колмогорова — Чепмена:*

$$p_{ij}(m+n) = \sum_{k=0}^{n} p_{ik}(m)\, p_{kj}(n), \qquad (9.15)$$

или в матричной форме

$$P(m+n) = P(m)P(n). \qquad (9.16)$$

Уравнение Колмогорова — Чепмена показывает, что первые m шагов приводят систему из состояния S_i в промежуточное состояние S_k, вероятность перехода из которого в S_n не зависит от того, как система попала в S_k.

2) Если система была в состоянии $p_i(0)$, заданном стохастическим вектором, i-ая координата которого равна 1, а остальные 0, то стохастический вектор состояния системы через *n* шагов можно получить как

$$p(n) = p_i(0)P^n. \qquad (9.17)$$

3) Длительность пребывания системы в любом состоянии подчиняется геометрическому закону распределения.

Пусть система перешла в состояние S_i, тогда с вероятностью p_{ii} она останется в этом состоянии и с вероятностью $q_{ii} = (1 - p_{ii})$ перейдет в другое состояние. Отсюда вероятность остаться в состоянии S_i еще m шагов равна

$$P(X = m) = q_{ii}p_{ii}^{m}, \qquad (9.18)$$

что соответствует геометрическому закону распределения.

Замечание. 1) В теории вероятностей утверждается, что в природе известно два закона распределения вероятностей, удовлетворяющих условию марковости, то есть отсутствия последействия (памяти). Для дискретных случайных величин — это геометрический закон распределения, для непрерывных случайных величин — экспоненциальный (показательный) закон распределения (см. раздел 5.2). ■

2) Так как вероятности переходов за n шагов в дискретном случае представляют собой геометрические прогрессии, то имеется возможность при анализе цепей Маркова использовать аппарат производящих функций (см. раздел 10.1).

Пример 9.2. Всем хорош альплагерь «Таймази» (ЮФУ), но только не погодой и не сложным характером начальника лагеря «Бизона». В начале августа здесь не бывает двух дней без дождя подряд. Если с утра пасмурно: может пойти дождь с вероятностью 0,5; если туристы до 9 утра не ушли в поход, а дождь не пошел — они по указанию начальника лагеря работают на кухне (или на субботнике) с вероятностью 0,25. Если сегодня утром был дождь, то с вероятностью 0,7 погода не изменится и с вероятностью 0,1 будет просто пасмурно. Если сегодня туристы работали на кухне, то с равной вероятностью завтра будет пасмурно или пойдет дождь. Матрица переходов P представлена далее в виде таблицы.

Решение. Сегодня туристы работали на кухне. Представим марковскую цепь в виде дерева, задав состояния в лагере для туристов буквами П, К, Д. Построим дерево логических возможностей состояния лагеря на три дня и соответствующее вероятностное дерево, задав на ребрах графа вероятностную меру перехода в новое состояние (рис. 9.2).

	Пасмурно	Кухня	Дождь
Пасмурно	0,5	0,25	0,25
Кухня	0,5	0	0,5
Дождь	0,1	0,2	0,7

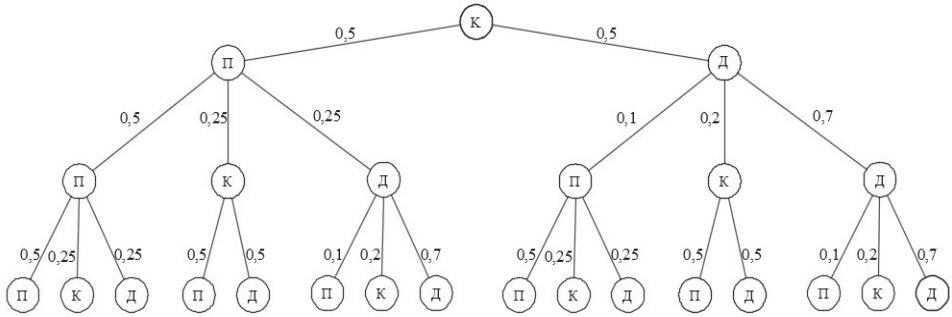


Рис. 9.2

Начальное состояние в лагере задается вектором $p(0) = [0 \quad 1 \quad 0]$, а матрица переходов по условию имеет вид

$$P = \begin{pmatrix} 0{,}5 & 0{,}25 & 0{,}25 \\ 0{,}5 & 0 & 0{,}5 \\ 0{,}1 & 0{,}2 & 0{,}7 \end{pmatrix}.$$

По формулам (9.9)–(9.10) получим вероятности состояния через 1, 2, 3 дня соответственно в виде стохастических векторов:

$$p(1) = p(0)P = [0 \quad 1 \quad 0]\begin{pmatrix} 0{,}5 & 0{,}25 & 0{,}25 \\ 0{,}5 & 0 & 0{,}5 \\ 0{,}1 & 0{,}2 & 0{,}7 \end{pmatrix} = [0{,}5 \quad 0 \quad 0{,}5],$$

$$p(2) = p(1)P = [0{,}5 \quad 0 \quad 0{,}5]\begin{pmatrix} 0{,}5 & 0{,}25 & 0{,}25 \\ 0{,}5 & 0 & 0{,}5 \\ 0{,}1 & 0{,}2 & 0{,}7 \end{pmatrix} =$$

$$= [0 \quad 1 \quad 0]\begin{pmatrix} 0{,}5 & 0{,}25 & 0{,}25 \\ 0{,}5 & 0 & 0{,}5 \\ 0{,}1 & 0{,}2 & 0{,}7 \end{pmatrix}^2 = [0{,}3 \quad 0{,}225 \quad 0{,}475],$$

$$p(3) = p(2)P = [0{,}3 \quad 0{,}225 \quad 0{,}475]\begin{pmatrix} 0{,}5 & 0{,}25 & 0{,}25 \\ 0{,}5 & 0 & 0{,}5 \\ 0{,}1 & 0{,}2 & 0{,}7 \end{pmatrix} =$$

$$= [0 \quad 1 \quad 0]\begin{pmatrix} 0{,}5 & 0{,}25 & 0{,}25 \\ 0{,}5 & 0 & 0{,}5 \\ 0{,}1 & 0{,}2 & 0{,}7 \end{pmatrix}^3 = [0{,}31 \quad 0{,}17 \quad 0{,}52].$$

Таким образом, если сегодня туристы трудились на кухне, то через три дня с вероятностью 0,31 их ожидает пасмурный день, с вероятностью 0,17 они попадут опять на кухню или на субботник и с вероятностью 0,52 туристов ожидает дождь.

Пример 9.3. Маркетолог формирует расположение товаров на витрине магазина. В простейшем случае существует два состояния в работе магазина: S_1 — товары пользуются спросом, S_2 — товары не пользуются спросом. Если магазин находится в состоянии S_1, то с вероятностью 0,5 он в нем останется в течение недели и с вероятностью 0,5 он перейдет в состояние S_2. В результате работы

маркетолога из состояния S_2 магазин может перейти в состояние S_1 с вероятностью 0,75 и остаться в состоянии S_2 с вероятностью 0,25. Таким образом, условие задачи приводит нас к последовательности состояний, которая является цепью Маркова. Графическое изображение этой цепи в виде графа представлено на рисунке 9.3.

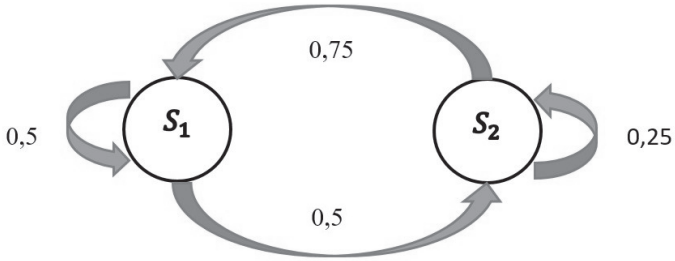


Рис. 9.3 — Граф состояний системы

Матрица вероятностей перехода дискретной однородной цепи Маркова из i-го состояния в j-ое за один шаг:

$$P = \begin{pmatrix} 0{,}5 & 0{,}5 \\ 0{,}75 & 0{,}25 \end{pmatrix}.$$

Если первоначально магазин находился в состоянии S_1, то вектор начального состояния $p(0) = (1 \quad 0)$, следовательно,

$$p(1) = p(0)P = (1 \quad 0)\begin{pmatrix} 0{,}5 & 0{,}5 \\ 0{,}75 & 0{,}25 \end{pmatrix} = (0{,}5 \quad 0{,}5),$$

$$p(2) = p(1)P = (0{,}5 \quad 0{,}5)\begin{pmatrix} 0{,}5 & 0{,}5 \\ 0{,}75 & 0{,}25 \end{pmatrix} = (0{,}625 \quad 0{,}375),$$

$$p(3) = p(2)P = (0{,}625 \quad 0{,}375)\begin{pmatrix} 0{,}5 & 0{,}5 \\ 0{,}75 & 0{,}25 \end{pmatrix} =$$
$$= (0{,}59375 \quad 0{,}40625).$$

Через k шагов вектор состояния магазина будет иметь вид

$$p(k) = p(k-1)P = p(0)P^k = (1 \quad 0)\begin{pmatrix} 0{,}5 & 0{,}5 \\ 0{,}75 & 0{,}25 \end{pmatrix}^k.$$

Соответственно, если вектор начального состояния $p(0) = (0 \quad 1)$, то формула состояния магазина через k шагов:

$$p(k) = p(k-1)P = p(0)P^k = (0 \quad 1)\begin{pmatrix} 0{,}5 & 0{,}5 \\ 0{,}75 & 0{,}25 \end{pmatrix}^k.$$

Объединив результаты расчетов в таблице, получим следующее.

k / $p(k)$	0	1	2	3	4	5	…	10
$p_1(k)$	1	0,5	0,625	0,59375	0,6015625	0,599609375	…	0,6000003815
$p_2(k)$	0	0,5	0,375	0,40625	0,3984375	0,400390625	…	0,3999996185
$p_1(k)$	0	0,75	0,5625	0,609375	0,5976563	0,600585938	…	0,5999994278
$p_2(k)$	1	0,25	0,4375	0,390625	0,4023437	0,399414062	…	0,4000005722

Из таблицы видно, что с увеличением k — числа шагов, $p_1 \to \frac{3}{5};\ p_2 \to \frac{2}{5}$. То есть при увеличении числа шагов, вероятности состояний системы становятся (*стационарными*) независимыми от первоначального состояния.

Иногда возможно найти вероятностный вектор $z(z_1, z_2, \dots, z_n)$ такой, что $z = zP$ (вектор z называется вероятностным, если $\sum z_i = 1$). Такой вектор z называется *неподвижным вектором матрицы преобразования* P (собственным вектором матрицы преобразования P), соответствующим значению 1.

Если $z := \pi$, то

$$p^{(n)} = \pi P^n.$$

Определение. Состояние марковской цепи называется *возвратным* (*рекуррентным*), если оно посещается ею бесконечное число раз с вероятностью, равной 1.

Если вероятность меньше 1, то состояние называется *транзитивным* и оно повторяется конечное число раз. В случае нескольких возвратных и транзитивных состояний процесс перемещается только между возвратными состояниями.

Определение. Марковская цепь называется *регулярной*, если какая-либо степень матрицы перехода P не содержит нулевых элементов.

Теорема 1. Если P — матрица перехода регулярной цепи, то

$$P^n \to Z,$$

где Z — матрица, строки которой образуют вероятностный вектор π, неподвижный вектор преобразования P, все компоненты которого положительны.

Теорема 2. Если P — матрица перехода регулярной цепи, то для любого вероятностного вектора p

$$pP^n \to \pi,$$

где π — единственный вероятностный вектор, что $\pi P = \pi$.

Регулярные цепи являются эргодическими, обратное верно не всегда.

Для нахождения вероятностного вектора примера 9.3 составим систему:

$$\begin{cases} 0{,}5\pi_1 + 0{,}5\pi_2 = \pi_1, \\ 0{,}75\pi_1 + 0{,}25\pi_2 = \pi_2, \\ p_1 + p_2 = 1. \end{cases}$$

Решением системы будут значения, которые ранее уже были найдены в таблице как предельные:

$$\pi_1 = \frac{3}{5};\ \pi_2 = \frac{2}{5}.$$

Во многих приложениях именно эти значения интересуют исследователей. То есть с вероятностью $\frac{3}{5}$ работа маркетолога приводит к увеличению спроса в магазине и с вероятностью $\frac{2}{5}$ — не приводит.

В случае большого числа состояний трудности решения систем уравнений можно обойти возведением в степень матрицы переходов в *Mathcad*, до установления стационарного состояния, когда все строки сравняются.

Определение. Марковская цепь называется: 1) *несократимой*, если все ее состояния связаны, то есть если из каждого ее состояния можно попасть в любое другое, 2) *эргодической,* если время, проведенное в состоянии i, соответствует p_i.

Теоремы 1, 2 составляют содержание следующей теоремы А. А. Маркова.

Теорема 3 (*Маркова о предельных вероятностях*). Пусть существует t шагов эргодической и несократимой цепи Маркова, таких, что все элементы матрицы P положительны ($p_{ij}(t) > 0$). Тогда для каждого состояния S_j существует предельная вероятность его наступления ($\pi_j = const$), которая не зависит от числа шагов t, имеет место предел

$$\lim_{n\to\infty} p_{ij}(n) = \pi_j. \qquad (9.19)$$

$\pi = \{\pi_j\}$ — единственное *стационарное* (*инвариантное*) *распределение* для положительно возвратной цепи Маркова с дискретным временем.

Среднее время возврата в состояние j равно $1/\pi_j$.

Для отыскания стационарного распределения используют *уравнения детального баланса*:

$$\pi_i p_{ij} = \pi_j p_{ji}, \qquad (9.20)$$

означающие, что в состоянии равновесия (симметрии) частота прямого и обратного перехода совпадают.

Суть уравнений (9.20) поясним в примере 9.4, а пока отметим, что они играют основополагающую роль в *алгоритме Метрополиса — Гастингса*, позволяющего осуществлять *сэмплирование* (выборку) (см. гл. 18) по реальным многомерным данным (закон распределения которых, конечно, не известен), что отражено в следующем утверждении.

Теорема 4. Цепь Маркова обратима, если она находится в состоянии равновесия и имеет место уравнения детального баланса.

Это позволяет утверждать, что за достаточно большое время цепь Маркова может адекватно «*просэмплировать*» неизвестное распределение данных, то есть сделать выборку, соответствующую неизвестному распределению, имеющихся данных. Причем *распределение полученной стационарной цепи Маркова будет соответствовать неизвестному распределению*.

Пример 9.4 (условия из [136]). Путь марковская цепь (рис. 9.4) имеет три состояния, причем, например, состояние S_1 получается из S_1 или S_2, или S_3:

S_1: | 1 | 3 | |
|---|---|---|
| 4 | 2 | 6 |
| 7 | 5 | 8 |

S_2: | 1 | | 3 |
|---|---|---|
| 4 | 2 | 6 |
| 7 | 5 | 8 |

S_3: | 1 | 3 | 6 |
|---|---|---|
| 4 | 2 | |
| 7 | 5 | 8 |

Рис. 9.4 — Головоломка с тремя связанными состояниями $S_1,\ S_2,\ S_3$

Головоломка (игра в 8) показывает, что пустой квадрат нельзя случайно переместить в произвольном направлении — это противоречит условию баланса. Действительно, если пустой квадрат находится в середине, то он свободно может перемещаться в любом из четырех направлений (рис. 9.4). Однако если пустой квадрат находится на границе, то направления перемещения не равновероятны. Рассмотрим уравнение связи между вероятностями состояний и вероятностями переходов изучаемой Марковской цепи, так как состояние S_1 — пустая клетка в левом верхнем углу может получиться только из состояний S_1, S_2, S_3, то

$$P(S_1) = P(S_2)P(S_2 \to S_1) + P(S_3)P(S_3 \to S_1) + P(S_1)P(S_1 \to S_1)^9$$

или

$$P(S_1)\big(1 - P(S_1 \to S_1)\big) = P(S_2)P(S_2 \to S_1) + P(S_3)P(S_3 \to S_1). \quad (9.21)$$

События, составляющие возможные переходы из состояния в состояние, образуют полную группу событий, следовательно,

$$P(S_1 \to S_1) + P(S_1 \to S_2) + P(S_1 \to S_3) = 1. \quad (9.22)$$

Отсюда

$$P(S_1 \to S_2) + P(S_1 \to S_3) = 1 - P(S_1 \to S_1).$$

Имеем

$$P(S_1)\big(P(S_1 \to S_2) + P(S_1 \to S_3)\big) = P(S_2)P(S_2 \to S_1) + P(S_3)P(S_3 \to S_1).$$

Раскрывая скобки в последнем равенстве и группируя связанные формулы, получим так называемые *условия детального баланса:*

$$P(S_1)P(S_1 \to S_2) = P(S_2)P(S_2 \to S_1), \quad (9.23)$$

$$P(S_1)P(S_1 \to S_3) = P(S_3)P(S_3 \to S_1). \quad (9.24)$$

Согласно условиям детального баланса для замкнутых систем, вероятность прямого перехода равна вероятности обратного перехода. Пусть все конфигурации имеют равные вероятности, то есть вероятность перехода пустой клетки внутри головоломки равна $1/4$, тогда из рисунка 9.4 следует, что

$$P(S_1 \to S_2) = P(S_1 \to S_3) = P(S_3 \to S_1) = \ P(S_1 \to S_3) = \frac{1}{4},$$

поэтому вероятность

$$P(S_1 \to S_1) = \frac{1}{2},$$

так как актуальны всего два перемещения (S_2, S_3) из потенциальных четырех, то в остальных случаях мы полагаем, что пустая клетка останется на месте. Аналогично, в одном из четырех случаев пустая клетка останется на месте для S_2 и S_3:

$$P(S_2 \to S_2) = P(S_3 \to S_3) = \frac{1}{4}.$$

Рассмотрим следующий алгоритм действий с головоломкой, на каждом шаге $t = 0, 1, 2, \dots, n$.

Шаг 1. Выбрать одно из четырех возможных направлений с равной вероятностью.

Шаг 2. Если конфигурация головоломки допускает, то переместить пустой квадратик в выбранном направлении, иначе — оставаться на месте.

Шаг 3. Пока $t < n, t =: t + 1$, шаг 1. Иначе — *end.*

Выполнение условия баланса (9.20) эквивалентно следующему равенству:

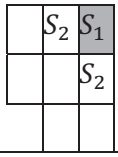
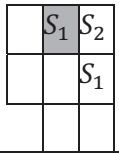
$$P(S_1 \to S_2) = min\left(1, \frac{P(S_2)}{P(S_1)}\right). \quad (9.25)$$

Таким образом, алгоритм показывает, что если мы находимся в состоянии S_i, то либо мы переходим в состояние S_j с вероятностью 1, либо остаемся на месте с вероятностью, зависящей от положения S_i: $1/2, 1/4$. Передвигаясь достаточно

[9] $S_2 \to S_1$ означает переход из состояния S_2 в состояние S_1.

большое количество раз и выбирая, для обеспечения независимости наблюдений, например, каждый 50-й или 100-й шаг мы получим частоты попадания пустого квадратика в ту или иную ячейку, что позволит оценить двумерное распределение пустой ячейки в рассматриваемой головоломке и в силу выполнения условий стационарности (уравнений баланса) будет соответствовать единственному стационарному распределению цепи Маркова. В этом состоит суть *алгоритма Метрополиса — базового алгоритма сэмплирования* (осуществления выборки по имеющимся данным с неизвестным законом распределения).

Если сторона квадрата $n \geq 3$ и пустая клетка, находясь внутри, случайным образом передвигается, то условия баланса, очевидно, выполняются. Если пустая клетка находится на границе, проверим условия (9.20, 9.25), заполнив таблицу (а) пустая клетка двигается из угла, б) от границы в угол).

№ п/п	Конфигурация	а) 	б) 
		$P(S_1) = 0{,}5 > > P(S_2) = 0{,}25$	$P(S_2) = 0{,}5 > > P(S_1) = 0{,}25$
1	$P(S_1 \to S_2)$	$0{,}5$	1
2	$P(S_1)P(S_1 \to S_2)$	$0{,}5 \cdot 0{,}5$	$0{,}25 \cdot 1$
3	$P(S_2 \to S_1)$	1	$0{,}5$
4	$P(S_2)P(S_2 \to S_1)$	$0{,}25 \cdot 1$	$0{,}5 \cdot 0{,}5$

Совпадение в таблице строк 2 и 4 подтверждает выполнение формул условия баланса (9.20), (9.25).

В общем случае роль границ головоломки играет неизвестное распределение данных. Благодаря выполнению описанного выше алгоритма и условий детального баланса — первоначальное случайное блуждание преобразуется в стационарное распределение цепи Маркова, соответствующее неизвестному распределению, что является основой алгоритма сэмплирования *MCMC* (*Markov Chain Monte Carlo* (метод Монте Карло по схеме марковских цепей)).

Задача принятия решений и дискретные цепи Маркова с доходами.

Изучение управляемых цепей Маркова в экономике началось в 1950–1960-е годы в работах Р. Ховарда, использовавшего (итеративные) идеи динамического программирования Р. Беллмана и предложившего осуществимую аналитическую модель процесса принятия решений для получения дохода в условиях описания работы систем с помощью математического аппарата цепей Маркова [101, 102, 121].

Пусть система описывается процессом Маркова и может иметь состояния

$$S_1, S_2, \dots, S_N.$$

Переходы из состояния S_i в состояние S_j описываются матрицей

$$P = \{p_{ij}\}, i, j = \overline{1, N},$$

получаемые доходы, соответственно, описываются матрицей

$$R = \{r_{ij}\}.$$

Таким образом Марковский генерирует последовательность доходов при каждом переходе от i к j. Оценим ожидаемый доход $f_i(n)$ через n шагов, если система находилась в состоянии *i*. При переходе на первом шаге из состояния *i* в состояние j система получит доход r_{ij}, далее из этого состояния доход системы за $(n-1)$ шаг составит $f_j(n-1)$. Значит, полный ожидаемый доход за n шагов:

$$f_j(n) = r_{ij} + f_j(n-1). \tag{9.26}$$

Математическое ожидание дохода:

$$M\big(f_i(n)\big) = \sum_{j=1}^{N} p_{ij}(r_{ij} + f_j(n-1)) \tag{9.27}$$

или

$$M\big(f_i(n)\big) = q_i + \sum_{j=1}^{N} p_{ij} f_j(n-1)), \tag{9.28}$$

где $i = \overline{1, N}$; $n = 1, 2, \dots, n_{max}$; $f_j(0)$ — цена при окончании функционирования в состоянии j; $q_i = \sum_{j=1}^{N} p_{ij} r_{ij}$ — средний одношаговый доход из S_i.

Задача оптимального управления. Цепь Маркова называется управляемой, если на каждом шаге n и в каждом состоянии S_i имеется возможность выбора строк матрицы P и R, $p_i^{k_i}$ и $r_i^{k_i}$ соответственно. Величина k_i называется стратегией управления в i −м состоянии (вектор стратегий $k = (k_1, k_2, \dots, k_N)$ называют политикой); последовательность политик на каждом шаге формируют управление $\overline{k} = (k_1, k_2, \dots, k_N)$, стратегия k_i или политика k на шаге *n,* обозначается добавлением соответствующего индекса (k_{in} и k_n). Если обозначить эффективность функционирования системы, описывающейся цепью Маркова с доходами, как функцию от реализации управления

$$\text{Э} = \text{Э}(\overline{k}), \tag{9.29}$$

то возникает задача поиска управления $\overline{k}^*$, доставляющего функции эффективности максимума:

$$\text{Э}\left(\overline{k}^*\right) = \max_{\overline{k}^*} \text{Э}\big(\overline{k}\big). \tag{9.30}$$

Иногда в расчетах учитывают коэффициент переоценки $\beta \in (0, 1]$ (приведения) будущих доходов:

$$\beta = \frac{1}{1+t},$$

где t — норма прибыли (%) или процентная ставка.

Различают два варианта поиска оптимального управления цепи Маркова с доходами без учета или с учетом β: 1) с конечным горизонтом управления, 2) с бесконечным горизонтом управления. Последний сводится к поиску решения в классе политик *стационарных управлений,* не зависящих от шага n, и рассматривается в работах Р. Ховарда, Г. Соколова и др. [101, 102, 121].

В настоящей книге рассмотрим простейший иллюстративный пример управления с конечным горизонтом без переоценки.

Пусть $f_n(i)$ — оптимальный ожидаемый доход, полученный на этапах от n до N включительно, если система находится в состоянии i в начале этапа n. Тогда рассматриваемая задача может рассматриваться как задача динамического программирования (ДП) с конечным числом этапов. Суть которой в декомпозиции N-мерной задачи в N — этапов относительно одной переменной, то есть решается N одномерных задач. Вычисления проводятся рекуррентно — предыдущее решение используется на следующем этапе. Пусть

$$q_i^k =: \sum_{j=1}^{N} p_{ij}^k r_{ij}^k — \tag{9.31}$$

средний ожидаемый доход на шаге i для стратегии k, тогда рекуррентные уравнения динамического программирования можно записать как

$$f_N(i) = \max_k \{q_i^k\}, \tag{9.32}$$

$$f_n(i) = \max_k \{q_i^k + \sum_{j=1}^{N-1} p_{ij}^k f_{n+1}(j)\}. \tag{9.33}$$

Пример 9.5. Пусть фирма регулярно оценивает положение сбыта продукции и дает ему удовлетворительную (состояние 1) или неудовлетворительную оценку (состояние 2). Требуется принять решение о необходимости рекламы для улучшения сбыта. Имеются матрицы переходных вероятностей с рекламой (P_1) и без (P_2) и соответствующие матрицы доходов (R_1) и (R_2) в течение месяца:

$$P_1 = \begin{pmatrix} 0{,}9 & 0{,}1 \\ 0{,}6 & 0{,}4 \end{pmatrix}, \quad R_1 = \begin{pmatrix} 2 & -1 \\ 1 & -3 \end{pmatrix},$$
$$P_2 = \begin{pmatrix} 0{,}7 & 0{,}3 \\ 0{,}2 & 0{,}8 \end{pmatrix}, \quad R_2 = \begin{pmatrix} 4 & 1 \\ 2 & -1 \end{pmatrix}.$$

Найти оптимальные решения на три месяца вперед.

Решение. По формуле (9.31):

$$q_1^1 = 0{,}9 \cdot 2 - 0{,}1 \cdot 1 = 1{,}7,$$
$$q_2^1 = 0{,}6 \cdot 1 - 0{,}4 \cdot 3 = -0{,}6,$$
$$q_1^2 = 0{,}7 \cdot 4 + 0{,}3 \cdot 1 = 3{,}1,$$
$$q_2^2 = 0{,}2 \cdot 2 - 0{,}8 \cdot 1 = -0{,}4.$$

Эти значения показывают, что если состояние фирмы удовлетворительное ($k = 1$), то при одном переходе применение рекламы дает доход 1,7, а в случае неудовлетворительного состояния приводит к убытку 0,6. При начальном удовлетворительном состоянии, в отсутствии рекламы, один переход дает доход 3,1, а в неудовлетворительном состоянии — убыток 0,4. Итак, имеем

i	q_i^1	q_i^2
1	1,7	3,1
2	$-0{,}6$	$-0{,}4$

Рассмотрим решение нашей задачи методом полного перебора. Для этого рассмотрим последовательно три этапа.

Этап 3.

	q_i^k		Оптимальное решение	
i	$k=1$	$k=2$	$f_3(i)$	k^*
1	1,7	3,1	3,1	2
2	$-0{,}6$	$-0{,}4$	$-0{,}4$	2

Этап 2.

	$q_1^k + p_{i1}^k f_3(1) + p_{i2}^k f_3(2) + p_{i3}^k f_3(3)$		Оптимальное решение	
i	$k=1$	$k=2$	$f_2(i)$	k^*
1	$1{,}7 + 0{,}9 \cdot 3{,}1 - 0{,}1 \cdot 0{,}4 = 4{,}45$	$3{,}1 + 0{,}7 \cdot 3{,}1 - 0{,}3 \cdot 0{,}4 = 5{,}15$	5,15	2
2	$-0{,}6 + 0{,}6 \cdot 3{,}1 - 0{,}4 \cdot 0{,}4 = 1{,}10$	$-0{,}4 + 0{,}2 \cdot 3{,}1 - 0{,}8 \cdot 0{,}4 = -0{,}1$	1,10	1

Этап 1.

	$q_1^k + p_{i1}^k f_2(1) + p_{i2}^k f_2(2) + p_{i3}^k f_2(3)$		Оптимальное решение	
i	$k=1$	$k=2$	$f_1(i)$	k^*
1	$1{,}7 + 0{,}9 \cdot 5{,}15 + +0{,}1 \cdot 1{,}10 = 6{,}445$	$3{,}1 + 0{,}7 \cdot 5{,}15 + +0{,}3 \cdot 1{,}10 = 7{,}035$	7,035	2
2	$-0{,}6 + 0{,}6 \cdot 5{,}15 + +0{,}4 \cdot 1{,}10 = 2{,}930$	$-0{,}4 + 0{,}2 \cdot 5{,}15 - -0{,}8 \cdot 1{,}10 = 1{,}510$	2,930	1

Суммарный ожидаемый доход за три месяца составит $f_1(1) = 7{,}035$ — при удовлетворительном состоянии фирмы в первый месяц, $f_1(2) = 2{,}93$ — при неудовлетворительном состоянии фирмы в первый месяц. Анализ этапов показывает, что при снижении доходов в первый и второй месяц рекламную кампанию следует проводить, а на третий месяц нет.

Скрытые марковские модели (СММ или *HMM — Hidden Markov Model*). СММ прежде всего используются при анализе нуклеотидных последовательностей в биологии (например, при изучении вопроса — родственны ли две последовательности?) и являются одним из продуктивных подходов к решению задач биоинформатики, а также распознавания образов (речи, лиц и т. д.). Теория скрытых марковских моделей выходит за рамки нашего изложения, но мы рассмотрим некоторые понятия СММ и дадим необходимые ссылки на первоисточники, так как сегодня это один из успешных трендов в машинном обучении, связанный непосредственно с теорией вероятностей.

Пример 9.6. Рассмотрим задачу № 21 из раздела 1.5 — о (эпизодически нечестном) казино, в котором иногда правильную игральную кость подменяют неправильной. Подмена игральных костей является марковским процессом. Игральная кость может находиться в одном из двух состояний: $S_0 =: S_{\text{прав}}$ — кость правильная, $S_1 =: S_{\text{неправ}}$ — кость неправильная. Тогда если события A_i обозначают выпадение цифры $i = \overline{1,6}$, то

$$P(A_i/S_0) = 1/6,\ P(A_6/S_1) = 0{,}5 \text{ и } P(A_1/S_1) = \cdots = P(A_5/S_1) = 0{,}1.$$

Положим, что в результате испытаний получены оценки вероятностей перехода из одного состояния в другое: $P(S_0/S_1) = 0{,}1$; $P(S_1/S_0) = 0{,}05$ (рис. 9.6).

Если рассмотреть последовательность результатов бросков игральной кости, то мы не можем утверждать, в каком испытании кость была правильной, а в каком — неправильной, так как последовательность состояний скрыта и, следовательно, мы имеем дело со скрытой марковской моделью (первого порядка) (HMM).

Вероятность состояния игральной кости в момент времени (t) зависит от того, в каком состоянии она была в предыдущий момент времени $(t-1)$:

$$a_{lj} = P\big(S(t) = s_j / S(t-1) = s_l\big) — \qquad (9.34)$$

путь цепи Маркова ($l, j = \{0, 1\}$).

В каждом состоянии $S(t) = s_j$ игральной кости в момент времени t при подбрасывании с определенной вероятностью генерируются символы $A(t) = A_i$ $(i = \overline{1,6})$ (рис. 9.5):

$$e_j(A_i) = P\big(A_i / s_j\big) — \qquad (9.35)$$

вероятность того, что символ A_i появится в состоянии s_j, или, иначе, *эмиссионная*[10] *вероятность*.

Эмиссионные вероятности для примера с казино представлены в рамках на рисунке (9.6).

Обычно наблюдаются последовательности символов (в нашем случае цифр от одного до шести), глядя на них, требуется восстановить последовательность скрытых состояний (дешифровать).

Вероятность увидеть определенную последовательность длины $N: A = A(1), A(2), \ldots, A(N)$ можно оценить по формуле полной вероятности как

$$P(A) = \sum_S P(A/S)P(S), \qquad (9.36)$$

сложив вероятности всех возможных путей.

При изучении СММ предполагается, что имеются условные законы распределения вероятностей

$$f(A(t)/S(t), \theta_1), f(S(t)/S(t-1), \theta_2)$$

и априорное распределение

$$f(S(1), \theta_3),$$

где $\theta = (\theta_1, \theta_2, \theta_3)$ — вектор параметров. В этом контексте можно поставить несколько задач.

1. Есть обучающая последовательность $(S(t), A(t))$, требуется оценить вектор параметров распределений θ (*обучение с учителем*).

2. Есть последовательность $A(t)$ и закон распределения $f(A,\ \theta_1)$, требуется найти наиболее вероятную последовательность $S(t)$. Экспоненциальный рост количества путей (9.36) приводит к изучению наиболее вероятного пути с использованием *алгоритма динамического программирования Витерби*, который генерирует вероятностный выбор правильной и неправильной игральной кости. Путь с наибольшей вероятностью может предсказать последовательность состояний кости (*сегментация*).

[10] Emit (*англ.*) — выпускать.

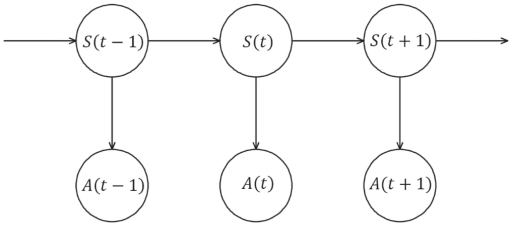


Рис. 9.5 — Структура скрытой марковской модели

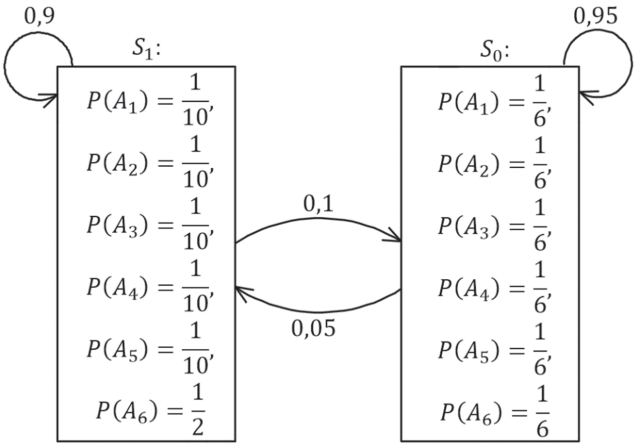


Рис. 9.6 — Граф состояний системы

3. Есть последовательность $A(t)$, но породившая их последовательность сотояний неизвестна, требуется найти наиболее вероятную последовательность $S(t)$. Наиболее общим подходом к решению этой задачи является *EM* (*Expectation Maximization*) алгоритм — метод получения оценок максимального правдоподобия в условиях неполных данных (частным случаем которого является *алгоритм обучения без учителя Баума-Уэлча*).

4. Есть последовательность $A = \{A(\tau)\}$ за t промежутков времени ($\tau = \overline{1,t}$). Требуется найти вероятность $P(A(t+1)/A(t))$ (*прогноз*).

Большие объемы данных и вычислений требуют использования методов машинного обучения (например, существует возможность реализации обучения и вывода скрытых марковских моделей в *Python*).

Скрытые марковские модели используются для анализа последовательностей (назначения правильной последовательности меток последовательностям данных, оценки вероятности последовательности или метки) при: распознавании речи, распознавании объектов и их поведения, анализе видео, анализе фондовых рынков, анализе текстов на плагиат, анализе последовательностей нуклеотидов в биоинформатике[11] (и, по мнению ученых-биологов, СММ могут приоткрыть для нас понимание феномена жизни[12]) и т. д. [150].

[11] Сайт учебно-научного центра «биоинформатика» института проблем передачи информации им. А. А. Харкевича РАН. URL: http://sector3.iitp.ru/teaching_r.html.

[12] «Но мы, те кто понимает, что такое жизнь, — мы конечно смеемся над номерами и цифрами!» (А. де Сент-Экзюпери).

Замечание. 1. Скрытая марковская модель — частный случай байесовской сети (см. раздел 10.5). В обычной марковской цепи можно наблюдать состояния системы и оценивать вероятности переходов. В скрытой марковской цепи мы наблюдаем лишь последовательность символов, на которые оказывают влияние скрытые состояния. Обобщение на случай нескольких измерений: цепь Маркова — Марковское случайное поле; скрытая марковская цепь — скрытое Марковское случайное поле.

2. Можно представить последовательность вложенных вероятностных графовых моделей: Марковские случайные поля (*Markov Random Fields — MRF, ненаправленные графические модели*) ⊃ условные случайные поля (*Conditional Random Fields — CRF, условные распределения по направленному графу*) ⊃ скрытые марковские модели (*HMM*, рис. 9.5).

3. Для *CRF* отсутствует свойство независимости от предыдущих результатов. *HMM — причинно-следственные*, или *порождающие* (генеративные), модели, *CRF — дискриминантные модели*, использующиеся для маркировки и сегментирования последовательностей данных. Модели *HMM, CRF* и *MRF* активно используются в машинном обучении для решения прикладных задач. ■

Темы (вопросы) для самоконтроля

1. Цепь Маркова.
2. Однородная цепь Маркова.
3. Стационарный Марковский процесс.
4. Уравнения Колмогорова — Чепмена.
5. Эргодическая цепь Маркова.
6. Уравнение детального баланса.
7. Скрытые марковские модели.

Глава 10
Приложения теории вероятностей в компьютерных науках (*computer science*)

... нам пришлось потратить годы на то, чтобы самым тщательным образом проверить и перепроверить бесконечное множество рецептов и отобрать для вас самые лучшие, самые интересные, самые совершенные.

Теперь без тени сомнения мы можем сказать, что если вы будете следовать инструкциям, то каждое блюдо получится таким же, как и у нас, даже если вы никогда не занимались приготовлением пищи (*поваренная книга Мак-Колла*).

…

Кулинария — это искусство, благородная наука; все кулинары — джентльмены (*Тит Ливий*).

(Дональд Кнут, Искусство программирования)

Говорят, что XXI век — век генетики, информатики и т. д. Не секрет, что наша страна утратила большинство из ведущих позиций в мире за исключением космических технологий ряда военных разработок. Профессиональные пользователи — математики, физики и астрофизики, генетики и метеорологи, военные и контрразведчики давно пришли к выводу, что вычислительные возможности современного компьютера ограничены, все чаще они сталкиваются с задачами, для решения которых требуется больше ресурсов, чем имеется в наличии на всем земном шаре. Как панацея от неотвратимо надвигающегося мрака вычислительной беспомощности, на пороге XXI века появились слова «параллельные и распределенные вычисления», «суперкомпьютер» и «*petaflops*» (петафлопс — миллион миллиардов операций с плавающей запятой в секунду). Задачи, эффективное решение которых под силу исключительно суперкомпьютеру с производительностью порядка одного петафлопса, распадаются на два класса: задачи с преобладанием целочисленных вычислений и задачи с преобладанием вычислений с плавающей запятой. В каждом классе, в свою очередь, легко выделить подклассы военно-прикладных и научно-практических приложений. К первому классу относятся криптография (например, взламывание кодов) и создание полноценного искусственного интеллекта, ко второму — моделирование ядерных взрывов, долгосрочный прогноз погоды и вычислительные задачи гидродинамики, оптимизация металлургических процессов (в частности производства алюминия) и т. д.

В связи с указанными выше проблемами возникает ряд задач, которые требуют оценки времени работы алгоритмов и поиска путей их ускорения на базе компьютеров со стандартной архитектурой, что дает неограниченное поле деятельности для приложения теории вероятностей.

10.1. Вероятностный анализ скорости выполнения алгоритмов

Лектор пользуется известными привилегиями, в пользе которых, надеюсь, нет никаких оснований сомневаться. Так, встретив у меня непонятное место, читатель должен предположить, что под ним кроется нечто весьма полезное и глубокомысленное.

(1704 г., Дж. Свифт
(в интерпретации *Lectora*)

Алгоритм — это последовательность инструкций для выполнения отдельного задания. При этом предполагается, что субъект, выполняющий алгоритм, следует инструкциям и умеет их выполнять. Компьютер имеет ограниченный словарь (язык программирования), поэтому для написания алгоритмов используют формализованный стиль. 300 лет до н. э. Евклид изложил алгоритм решения целого ряда геометрических задач. Сначала излагается «словарь» — неопределяемые понятия и аксиомы, а затем на их основе строятся алгоритмы решения сложных задач (алгоритмы деления угла (отрезка) в данном отношении, равенства (подобия) фигур и т. д.). Формализованные алгоритмы такого типа обычно используются для проверки истинности определенных положений или возможности выполнения каких-либо действий, скорость же работы алгоритма не важна.

При решении реальных задач на компьютере, например, сортировки записей о миллионе покупателей, эффективность становится основным критерием оценки алгоритма. Оценка сложности алгоритмов складывается из:

а) скорости алгоритма (например, 1 алгоритм 1000 записей сортирует за 1 с; 1 млн — 10 с, 2 алгоритм 1000 записей сортирует за 2 с; 1 млн — 5 с), т. е. 1 алгоритм лучше для малых списков, 2 алгоритм лучше для больших списков;

б) требований к размеру памяти (от быстрого алгоритма нет толка, если мало памяти);

в) свободного места на диске.

Производительность алгоритма можно оценивать по порядку величины N — размерности исходных данных. Алгоритм имеет сложность порядка $O(f(N))$, если с увеличением размерности исходных данных N время выполнения алгоритма растет пропорционально функции $f(N)$.

Например, для алгоритма, содержащего вложенный цикл:

$For\ I = 1$ to$N$
  $For\ J = 1$ to$N$
  ……..
  $Next\ J$
$Next\ I$

сложность алгоритма будет порядка $O(N^2)$.

Часто встречающиеся функции оценки порядка сложности алгоритма ($C = const, C > 1$):

а) работают с достаточной скоростью:

$$f(N) = C,$$
$$f(N) = log_2(log_2 N),$$
$$f(N) = (log_2 N),$$
$$f(N) = NC,$$
$$f(N) = N,$$
$$f(N) = N^C;$$

б) пригодны только для решения задач с небольшими N:

$$f(N) = C^N, f(N) = N!$$

Время выполнения сложных алгоритмов на компьютере со скоростью 1 млн операций в секунду (1*MFLOPS*).

	$N = 10$	$N = 20$	$N = 30$	$N = 40$	$N = 50$
N^3	0,001 с	0,008 с	0,027 с	0,064 с	0,125 с
2^N	0,001 с	1,05 с	17,9 мин	1,27 дня	35,7 лет
3^N	0,059 с	58,1 мин	6,53 года	$3{,}68 \cdot 10^5$ лет	$8{,}20 \cdot 10^{10}$ лет
$N!$	3,63 с	$7{,}71 \cdot 10^4$ лет	$8{,}41 \cdot 10^{18}$ лет	$2{,}59 \cdot 10^{34}$ лет	$9{,}64 \cdot 10^{50}$ лет

Сегодня можно предположить два пути решения сложных задач.

1) Совершенствование аппаратной базы — для решения задач с порядком сложности $O(N!)$ при $N = 24$ требуется больше времени, чем существует вселенная. Поэтому алгоритмы со сложностью порядка $O(C^N)$ и $O(N!)$ и пригодны только при малых значениях N. При больших значениях N возможны следующие направления реализации решения: совершенствование стандартных компьютеров с архитектурой фон Неймана, создание принципиально новых компьютеров с параллельной обработкой информации.

2) Разработка рандомизированных алгоритмов, содержащих элемент случайности (см. далее «хеширование»).

Анализ сложности алгоритмов — необходимая часть разработки приложения, часто это достигается тестированием (с использованием генератора случайных чисел). Важным фактором также является частота обращения к файлу подкачки, т. е. необходимо экономно расходовать оперативную память (например, задавая тип переменных).

При решении практических задач возникают следующие задачи.

1. *Задача построения больших массивов данных и их переупорядочивания* (базы данных, списки, бинарные и другие деревья, связи, структуры).

2. *Рекурсия* — вызов функции или подпрограммы самой себя.

3. *Моделирование реальных задач с помощью дерева решений*. Поиск наилучшего решения соответствует поиску наилучшего пути на дереве (метод полного перебора, ветвей и границ, эвристические методы, случайный поиск и последовательное приближение).

4. *Сортировка* (вставкой, выбором, пузырьковая, слиянием, пирамидальная, подсчетом, большая сортировка).

5. *Поиск* (полный перебор, двоичный поиск, интерполяционный поиск).

Сложность алгоритма оценивается по затратам времени и ресурсов, обычно находится компромисс между ними. Проведем анализ следующего алгоритма.

Пример 10.1. *Алгоритм M* (*нахождение максимума*). Для данных n элементов $x[1], x[2], \ldots, x[n]$ необходимо найти такие величины m и j, что $m = x(j) = \max\limits_{1 \le i \le n} x[i]$, где j — наибольший индекс, удовлетворяющий этому соотношению (символ ← означает операцию присвоения, $j \leftarrow n$ означает, что переменной j присвоено значение n).

M 1. (Инициализация.) Положим $j \leftarrow n$, $k \leftarrow n-1, m \leftarrow x[n]$

(во время выполнения алгоритма будем иметь

$m = [j] = \max\limits_{k \le i \le n} x[i]$).

M 2. (Все проверено?) Если $k = 0$, то работа алгоритма заканчивается.

M 3. (Сравнение.) Если $x[k] \le m$, перейти к шагу M 5.

M 4. (Замена m.) Положим $j \leftarrow k,\ m \leftarrow x[k]$. (Это значение m является новым текущим максимумом.)

M 5. (Уменьшение k.) Уменьшим k на единицу и вернемся к шагу M 2.

Зная, сколько раз выполняется каждый шаг, можно оценить время выполнения алгоритма на конкретном компьютере.

Для выполнения алгоритма M требуется фиксированный объем памяти, поэтому будем анализировать время, необходимое для его выполнения. Для этого подсчитаем, сколько раз выполняется каждый шаг.

№ шага	Количество выполнений
M 1	1
M 2	n
M 3	$n-1$
M 4	A
M 5	$n-1$

В приведенной таблице значение A неизвестно (A определяет, сколько раз необходимо изменить значение текущего max). Для проведения анализа требуется оценить значение A и его $\sigma(A)$.

$A \to min$ (для оптимистов) $= 0: x[n] = \max\limits_{1 \le k \le n} x[k]$,

$A \to max$ (для пессимистов) $= (n-1): x[1] > x[2] > \cdots > x[n]$,

$A \to M(x)$, *к среднему* — для специалистов по теории вероятностей и математической статистике.

Пусть все значения $x[k]$ — различны и все значения их $n!$ перестановок равновероятны.

Например, если $n = 3$, то следующие 6 возможностей равновероятны.

Ситуация	*Значение A*
$x[1] < x[2] < x[3]$	0
$x[1] < x[3] < x[2]$	1
$x[2] < x[1] < x[3]$	0
$x[2] < x[3] < x[1]$	1
$x[3] < x[1] < x[2]$	1
$x[3] < x[2] < x[1]$	2

$$\bar{A} = \frac{0+1+0+1+1+2}{6} = \frac{5}{6}.$$

Вероятность того, что A имеет значение k, равна

$$P_n(A = k) = \frac{\text{число перестановок, для которых } A = k}{n!}$$

при $n = 3$: $P_3(A = 0) = \frac{2}{6} = \frac{1}{3}, P_3(A = 1) = \frac{3}{6} = \frac{1}{2}, \ P_3(A = 2) = \frac{1}{6}$.

Чтобы определить поведение A найдем вероятности $P_n(A{=}k)$ по индукции. Обозначим число перестановок, для которых $A = k$ через P_{nk}. Оно равно $P_{nk} = n!\,P_n(A = k)$. Рассмотрим перестановки $x_1, x_2, \ldots, x_n$, элементов $\{1, 2, \ldots, n\}$. Если $x_1 = n$, то значение A на единицу больше, чем значение для перестановки $x_2, x_3, \ldots, x_n$. Если $x_1 \neq n$, то значение A точно такое, как для перестановки $x_2, x_3, \ldots, x_n$. Следовательно, по индукции получим, что

$$P_{nk} = P_{(n-1)(k-1)} + (n-1)P_{(n-1)k},$$

$$P_n(A = k) = \frac{P_{n-1}(A=k-1)}{n} + \frac{n-1}{n}P_{n-1}(A = k).$$

К аналогичному выводу можно прийти с использованием формулы полной вероятности. Пусть событие B означает, что число изменений текущего максимума равно k $(A = k)$.

Гипотеза H_1: $x_1 = n$, ее вероятность $P(H_1) = 1/n$, а условная вероятность $P(B/H_1) = P_{(n-1)}(A = k-1)$.

Гипотеза H_2: $x_1 \neq n$, ее вероятность $P(H_2) = (n-1)/n$, а условная вероятность $P(B/H_2) = P_{(n-1)}(A = k)$. Тогда по формуле полной вероятности

$$P_n(A = k) = \frac{P_{n-1}(A=k-1)}{n} + \frac{n-1}{n}P_{n-1}(A = k).$$

Пусть $G_n(Z) = \sum_k P_n(A = k)\,Z^k$ — производящая функция случайной величины, тогда

$$G_n(Z) = \frac{Z}{n}G_{n-1}(Z) + \frac{n-1}{n}G_{n-1}(Z) = \frac{Z+n-1}{n}G_{n-1}(Z),$$

$$G_n'(Z) = \frac{1}{n}G_{n-1}(Z) + \frac{Z+n-1}{n}G_{n-1}'(Z),$$

$$G_n'(1) = \frac{1}{n} + G_{n-1}'(1).$$

Отсюда найдем, что математическое ожидание A — это сумма ряда:

$$M(A) = G_n'(1) = \frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{3} + \frac{1}{2} + \frac{1}{1} - 1$$

(учитывая, что $G_1'(1) = 0$). Это и есть среднее число выполнений шага M 4. При $n \to \infty, M(A) \approx \ln n$.

Замечание. Гармонические числа $H_n = \sum_{k=1}^{n}\frac{1}{k} = \ln(n) + \gamma + O(\frac{1}{n})$, где $\gamma = 0{,}5772156649\ldots$ — постоянная Эйлера (1735). ■

6. *Хеширование* [32, 59, 60]. Ряд важных алгоритмов хранения и выборки информации на ЭВМ основаны на методе, который называется «хеширование». Анализ алгоритмов — раздел информатики, занимающийся выводом количественной информации об эффективности компьютерных методов, его основателем является известный американский математик Д. Кнут, который занимается им с 1962 г. Вероятностный анализ алгоритмов — это изучение времени работы алгоритмов, рассматриваемого как случайная величина, зависящая от предполагаемых

характеристик исходных данных. Хеширование особенно подходит для вероятностного анализа, так как метод хеширования исключительно эффективен в среднем, хотя наихудший вариант ужасен, когда все ключи имеют одинаковый хеш-код. Общая задача заключается в хранении некоторого множества записей, каждая из которых содержит значения «ключа» K и некоторые данные об этом ключе $D(K)$. Цель находить $D(K)$ по заданному K. Хеширование реализует подход аналогичный интерполяционному поиску, создающему функцию соответствия между искомым значением и индексом позиции, где он должен находиться.

Алгоритм хеширования использует *функцию, определяющую вероятное положение элемента в таблице* на основе значения искомого элемента. Хеш-таблицы обычно используют, если требуется часто вставлять или удалять элементы.

1. Пусть требуется заполнить несколько значений с ключом от 1 до 100, тогда формируют массив со 100 ячейками и устанавливают соответствие.

2. Реально диапазон ключа выше. Например, при использовании в качестве ключа идентификационные номера социального страхования, состоящего из 9 цифр (1 млрд). Если одна запись 1 Кб, то весь массив — 1 млн Мб (1 Гигабайт). При штате менее 10 млн, 99 % массива не востребовано (пустые значения).

3. Поэтому схемы хеширования, отображают большое количество возможных ключей, отображают на достаточно компактную хеш-таблицу. Однако если номер социального страхования 1 млрд записей, то для таблицы с 1000 позиций в среднем одной ячейке будет соответствовать млн записей. Поэтому для решения этой проблемы схема хеширования предполагает наличие алгоритма разрешения конфликтов.

Для реализации хеширования необходимы:

а) хеш-таблица для хранения данных;

б) функция хеширования (элемент $\leftrightarrow$ место);

в) алгоритм решения конфликтов, определяющих последовательность действий, если несколько ключей соответствуют одной ячейке таблицы.

Общая задача: хранение некоторого множества записей, каждая из которых содержит значение «ключа» K и некоторые данные $D(K)$ об этом ключе.

Одно из решений. Хранятся две таблицы, $KEY(j)$ — для ключей и $DATA\ (j)$ для данных $1 \leq j \leq\ N$, где N — общее число записей, которые могут быть размещены, n — обозначает фактическое число записей.

Поиск ключа K можно осуществить последовательно, просматривая таблицу.

$S1$. Установить $j = 1$ (уже просмотрены позиции $<\ j$).

$S2$. Если $j\ >\ n$, остановиться (безуспешный поиск).

$S3$. Если $KEY(j)\ =\ K$, остановиться (успешный поиск).

$S4$. Увеличить j на 1 и вернуться к шагу 2 (следующая попытка).

Метод работает правильно, но его работа может быть ужасающе медленной; при безуспешном поиске шаг $S2$ приходится повторять $(n + 1)$ раз, а n может быть большим.

Замечание. 1. Обычно заранее не известно, какие ключи будут в таблице, но часто существует возможность выбора хеш-функции h, чтобы значения $h(K)$ можно было считать случайной величиной, равномерно распределенной

в интервале от 1 до m и независимой от хеш-кодов других присутствующих ключей. В этом случае вычисления хеш-функции подобно бросанию кубика с m гранями.

2. Все записи могут попасть в один список, точно так же как на кубике может выпадать цифра 6. Однако закон больших чисел говорит о том, что списки почти всегда будут хорошо сбалансированы. ■

Анализ хеширования. Хеширование было придумано для ускорения поиска в базах и хранилищах данных. Например, используется m отдельных списков вместо одного огромного. «Хеш-функция» преобразует любой возможный ключ K в номер списка $h(K)$, лежащий в диапазоне от 1 до m. Используются вспомогательные таблицы.

1) $First\,[i]$ — содержит для каждого $i, 1 \le i \le m$ указатель на первый элемент в списке i.

2) $NEXT\,[j], 1 \le j \le N$, указывает на запись, следующую за записью j в списке, которому эта запись принадлежит; $First\,[i] = -1$, если список i пуст; $NEXT\,[j] = 0$, если j — последняя запись в своем списке; n — общее количество записей, n содержит информацию о числе записей.

Пример 10.2. Пусть ключи — имена и имеется $m = 4$ списка, разделяемых по первой букве имени:

$$h(\textit{имя}) = \begin{cases} 1 & \textit{при} \quad А - Ж, \\ 2 & \textit{при} \quad З - О, \\ 3 & \textit{при} \quad П - Ц, \\ 4 & \textit{при} \quad Ч - Я. \end{cases}$$

Вначале имеется 4 пустых списка и $n = 0$. Если ключ первой записи будет, скажем, имя Даниил, то $h(\text{Даниил}) = 1$, и поэтому «Даниил» станет ключом первого элемента в списке 1. Если следующими именами окажутся Инна и Оксана, то они попадут в список 2.

Таблица памяти выглядит так:

$First\,[1] = 1, First\,[2] = 2, First\,[3] = -1, First\,[4] = -1.$

Ключи: $KEY\,[1] =$ Даниил, $NEXT\,[1] = 0,$

$KEY\,[2] =$ Инна, $NEXT\,[2] = 2,$

$KEY\,[3] =$ Оксана, $NEXT\,[3] = 0, n = 3.$

(Значения $DATA$ (К) содержат секретную информацию и здесь не приводятся.) После вставки 16 имен списки могли бы содержать следующие записи.

Список 1	Список 2	Список 3	Список 4
Даниил	Инна	Петр	Эльвира
Елена	Оксана	Роман	Юлия
Василий	Ирина	Светлана	
Владимир	Ольга	Тамара	
	Кристина		
	Леонид		

В массиве KEY те же имена записаны вперемешку, но значения $NEXT$ позволяют разделить списки. Если нужно найти имя Кристина, то нужно просмотреть 6 имен списка 2, но это не идет в сравнение с просмотром 16 имен.

Алгоритм поиска ключа в соответствии с описанной схемой.

Н 1. Установить $i := h(K)$ и $j := FIRST[i]$.

Н 2. Если $j \leq 0$, остановиться (безуспешный поиск). (*)

Н 3. Если $KEY\ [j] = K$, остановиться (успешный поиск).

Н 4. Установить $i := j$, затем $j := NEXT[i]$ и вернуться к шагу Н 2.

Например, при поиске имени Кристина установим на шаге Н1 $i := 2$, $j := 2$; на шаге Н3 найдем, что Инна ≠ Кристина; на шаге Н4 установим $j := 3$ и на шаге Н3 найдем, что Оксана ≠ Кристина; выполнив шаги Н4 и Н3 еще 3 раза, мы найдем Кристину в таблице.

После успешного поиска данные $D(K)$ содержатся в $DATA\ [j]$.

Рассмотрим *вероятностный анализ хеширования*. Пусть r — число «проб» в таблице, или иначе — количество выполнений шага Н 3 при работе алгоритма (*).

Шаг	Безуспешный поиск	Успешный поиск
1	1	1
2	$r+1$	r
3	r	r
4	r	$r-1$

Таким образом, главная характеристика, определяющая время работы процедуры поиска, есть число проб r. Например, существует некоторая книга, в которой на каждой странице одна запись, на обложке — номер страницы первой записи в каждом из m списков; каждому ключу K соответствует список $h(K)$, которому тот принадлежит. На каждой странице книги содержится ссылка на следующую страницу в том списке, к которому эта страница относится. Число проб — это число страниц, которые придется просмотреть.

Случай 1. Ключ отсутствует.

Безуспешный поиск требует по одной пробе для каждого элемента списка h_{n+1}.

Вероятность того, что $h_j = h_{n+1}$ равна $1/m$ для $1 \leq j \leq N$:

$$P\left(h_j = h_{n+1}\right) = \frac{1}{m}.$$

Математическое ожидание числа «проб r»:

$$M(r) = \frac{n}{m},$$

т. е. среднее число проб в m раз меньше, чем без хеширования.

Процесс поиска можно описать биноминальным распределением:

$$p = \frac{1}{m},\ q = 1 - p = \frac{m-1}{m},$$

следовательно, $npq = \frac{n(m-1)}{m^2}$ — дисперсия. $G_r(Z) = \left(\frac{m-1+Z}{m}\right)^n$ — производящая функция общего числа проб в безуспешном поиске.

Итак, анализ алгоритма в случае безуспешного поиска дает следующие результаты:

$min\, r = 0$ (минимальное число проб $r = 0$);

$max\, r = n$ (максимальное число проб $r = n$ — числу записей в списке);

$M(r) = \frac{n}{m}$ — математическое ожидание числа проб;

$D(Z) = \frac{n(m-1)}{m^2}$ — дисперсия числа проб; при $m \to \infty$;

$D(r) \approx \frac{n}{m}, \sigma(r) = \sqrt{\frac{n}{m}}$.

Случай 2. Ключ присутствует.

Пусть s_j — вероятность того, что ищется j −ый ключ. Тогда можно показать [32], что производящая функция числа проб в случае успешного поиска

$$G_r(Z) = zS\left(\frac{m-1+Z}{m}\right),$$

где $S(Z) = s_1 + s_2 Z + s_3 Z^2 + \cdots + s_n Z^{n-1}$ — ПФСВ для поиска вероятностей s_j. Математическое ожидание числа проб $M(r) = \frac{n-1}{2m} + 1$. С ростом n, число проб стремится к $0{,}5\, ln\, n$, а среднее квадратическое отклонение к $(ln\, n)/\sqrt{12}$.

В обоих случаях приведенный анализ позволяет спать спокойно пессимистам, которые опасаются наихудшего случая. Из неравенства Чебышёва следует, что списки будут хорошими, за исключением крайне редких случаев, для любого $\varepsilon > 0$:

1) $P\left(\left|r - \frac{n}{m}\right| < \varepsilon\right) \geq 1 - \frac{n(m-1)}{m^2\varepsilon^2}$,

2) $P\left(\left|r - \frac{1}{2} ln\, n\right| < \varepsilon\right) \geq 1 - \frac{(ln\, n)^2}{12\varepsilon^2}$.

10.2. Случайные числа, генераторы случайных чисел

Defindit numerus (В числах ты найдешь покой) — это истина дураков;
Deperdit numerus (В числах ты найдешь погибель) — истина мудрых.
Ч. Колтон, 1820 г.

Числа, которые выбираются случайным образом, находят множество полезных применений. Общее название всех алгоритмов, использующих случайные числа, — метод Монте-Карло.

1. *Моделирование*. Кажущиеся случайности могут иметь закономерности при углубленном рассмотрении, например, рассмотрим задачу 12 из раздела 1.2. *Вероятность «черной пятницы».*

Пример 10.3. Доказать, что 13 число месяца с большей вероятностью приходится на пятницу, чем на другие дни недели.

Решение. Известно, что период обращения Земли вокруг Солнца и обращения вокруг собственной оси точно несоизмеримы, а имеет место приближенное соотношение, согласно которому год по григорианскому календарю приближенно равен

$$365{,}2425 = \frac{146097}{400} \text{ суток.}$$

Ошибка в сутки накапливается за 10 000 лет. Причем все года кратные 4, 400 — високосные, кроме кратных 100, 200, 300. Таким образом, календарь повторяется с периодичностью 400 лет. Учитывая это, студент Р. Пропищин предложил программу на языке *turbo pascal*:

```
procedure TForm1.PDJXPButton1Click(Sender: TObject);
//  В любых идущих подряд 400 годов
// имеется 400/4-4+1 високосных годов, т. е. 97
// значит, дней будет 365*400+97=146097
// достаточно вычислить вероятность попадания 13 числа
// на пятницу в любые 400 лет идущие подряд
var
// номер дня недели (от 1 до 7)
Den:byte;
// переменная в которую сохраняется 13 + месяц + год
DatePR:TDate;
// считаем в массиве выпадание 13 на день недели
mas:array[1..7] of integer;
// используем для цикла
i:integer;
// год, месяц
year,month:word;
begin
// обнуляем счетчики
for i:=1 to 7 do mas[i]:=0;
// цикл по году (400 лет)
for year:=3000 to 3399 do
// цикл по месяцу
for month:=1 to 12 do
begin
// заносим в переменную дату
DatePR:=EncodeDate(year, month, 13);
// функция DayOfWeek возвращает от 1 (воскресенье) до 7 (суббота)
Den:=DayOfWeek(DatePR);
// увеличиваем счетчик
inc(mas[Den]);
end;
// выводим данные
Memo1.Lines.Add('Воскресенье'+' - '+inttostr(mas[1]));
Memo1.Lines.Add('Понедельник'+' - '+inttostr(mas[2]));
Memo1.Lines.Add('Вторник'+' - '+inttostr(mas[3]));
Memo1.Lines.Add('Среда'+' - '+inttostr(mas[4]));
Memo1.Lines.Add('Четверг'+' - '+inttostr(mas[5]));
Memo1.Lines.Add('Пятница'+' - '+inttostr(mas[6]));
Memo1.Lines.Add('Суббота'+' - '+inttostr(mas[7]));
end;
```

В результате получим следующие результаты.

День недели	*Число попаданий на 13 число*
Понедельник	685
Вторник	685
Среда	687
Четверг	684
Пятница	688
Суббота	684
Воскресенье	687

Таким образом, частота попадания 13 числа на пятницу является наибольшей и равна 688, следовательно, 13 число месяца с большей вероятностью приходится на пятницу, чем на другие дни недели.

Имитационное моделирование. Компьютер используется для моделирования естественных явлений или состояний различных систем, когда на вход подаются случайные числа, подчиняющиеся тому или иному закону распределения, и изучаются параметры выходных переменных. Например, исследования в ядерной физике (первое применение случайных чисел на компьютере связано с именем Д. фон Неймана, который в годы Второй мировой войны предложил использовать их при исследовании проблемы создания атомного оружия); теории массового обслуживания и т. д. Классическим примером является модель «ядерной зимы», полученная в 1970-е годы на ВЦ АН СССР под руководством академика Н. Н. Моисеева. В современном пакете расширений *MatLab — Simulink* имеется прекрасная возможность построения имитационных моделей в режиме языка визуального программирования.

С начала 1990-х гг. группа ученых из Санкт-Петербургского Политехнического университета Петра Великого (СПбПУ) разрабатывает и поддерживает программное обеспечение для имитационного моделирования. В настоящее время продукт называется *AnyLogic*, обладает графическим интерфейсом и позволяет быстро, экономично и безопасно формировать прогнозы сценариев развития в социально-экономических, технических и других системах. Большим плюсом является наглядность и возможность интеграции с различными источниками данных, предлагается облачный онлайн-сервис. *AnyLogic* поддерживает три наиболее известных направления имитационного моделирования (рис. 10.1):

– *системную динамику Дж. Форрестера,* позволяющую учитывать нелинейность поведения систем и обратную связь в динамике;

– *дискретно-событийное (процессное) моделирование,* которое часто используется для моделирования последовательности событий некоторого непрерывного процесса в производстве, логистике и т. д.;

– *агентное моделирование* опирается на индивидуальное поведение агентов, которое выливается в глобальное поведение всей системы в целом.

Рис. 10.1 — Направления имитационного моделирования, реализованные в *AnyLogic*[13]

Стохастическое программирование. В общей постановке задача оптимизации включает в себя: целевую функцию, ограничения, граничные условия, соответственно:

$$\left.\begin{array}{c} F = f(x_j) \to \max(min, const), \\ g_i(x_j) \le b_i, \\ d_j \le x_j \le D_j, \end{array}\right\} \qquad (10.1)$$

где f — целевая функция от оптимизируемых переменных x_j; g_i — i-ая функция ограничений; $d_j \le x_j \le D_j$ — границы переменной x_j; $i = 1, \dots, m; j = 1, \dots, n$.

Для задачи линейного программирования условия (10.1) примут вид

$$\left.\begin{array}{c} F = \sum_{j=1}^{n} c_j\, x_j \to \max(min, const), \\ \sum_{j=1}^{n} a_{ij}\, x_j \le b_i, \\ d_j \le x_j \le D_j. \end{array}\right\} \qquad (10.2)$$

Если коэффициенты целевой функции (c_j) и параметры ограничений (a_{ij}, b_i) являются случайными величинами, а также задается α_i — вероятность выполнения i-го ограничения, то возникает задача стохастического программирования, которая может рассматриваться в двух постановках.

M — постановка (оптимизация математического ожидания или среднего значения целевой функции):

$$\left.\begin{array}{c} M[F] \to \max(min), \\ P\left[\sum_{j=1}^{n} a_{ij}\, x_j \le b_i\right] \ge \alpha_i, \\ d_j \le x_j \le D_j. \end{array}\right\} \qquad (10.3)$$

[13] URL: https://www.anylogic.ru.

P — постановка (максимизация вероятности получения максимального (минимального) значения целевой функции):

$$\left.\begin{array}{c} P[Fmax(min)] \to max, \\ P\left[\sum_{j=1}^{n} a_{ij}\, x_j \le b_i\right] \ge \alpha_i, \\ d_j \le x_j \le D_j. \end{array}\right\} \qquad (10.4)$$

2. *Выборка* (*sample*). Часто невозможно или нецелесообразно исследовать всю совокупность данных, например, базу данных сети предприятий или информацию в сети Интернет. Но случайный репрезентативный отбор элементов совокупности позволяет изучить ее «типичные свойства» (см. гл. 12).

3. *Сэмплинг* (*sampling*). Если имеются данные и требуется научиться генерировать сэмплы (выборки), соответствующие их распределению, оценивать ожидания функций, то используются Марковские методы алгоритма Монте-Карло (алгоритм Метрополиса — Гастингса и его упрощенная версия — алгоритм Гиббса, см. гл. 18) или вариационные методы, лучше работающие с большими данными (*big data*). Сэмплинг используется в байесовской статистике и машинном обучении.

4. *Численный анализ*. Для решения сложных задач не всегда подходят точные методы, которые часто и разработать невозможно. Поэтому используют различные приближенные методы. Причем наиболее эффективными из них являются методы, использующие случайные числа.

Например, в пакете анализа *Excel* опция *генерация случайных чисел* заполняет диапазон случайными числами, заданными по одному из законов: равномерному; нормальному; Бернулли; биномиальному; Пуассона; модельному (позволяющему генерировать последовательности случайных чисел от *a* до *b* с шагом *c* и возможностью повторения каждого числа и последовательности); дискретному (решающему задачу, получения по имеющемуся распределению, новых значений того же распределения).

Замечание. Инструмент генерации случайных чисел позволяет решать целый ряд задач: численных методов (например, приближенного вычисления определенных интегралов методом статистических испытаний — методом Монте-Карло, имитационного моделирования изучаемых процессов и т. д.). ■

Пример 10.4. Вычислить определенный интеграл $I = \int_0^1 x^2 dx$ методом Монте-Карло.

Решение. Вычисление определенного интеграла I равносильно нахождению площади D криволинейной трапеции функции

$$Y = f(x) = x^2 \text{ (рис. 10.2).}$$

Рассмотрим систему двумерных равномерно распределенных случайных величин *(X, Y)* на интервале от 0 до 1. При достаточно большом числе опытов *N* площадь *D* будет приближенно равна относительной частоте попадания точек $M_i(x_i, y_i)$ в область *D* (в силу закона больших чисел):

$$I \approx \frac{n}{N}\,.$$

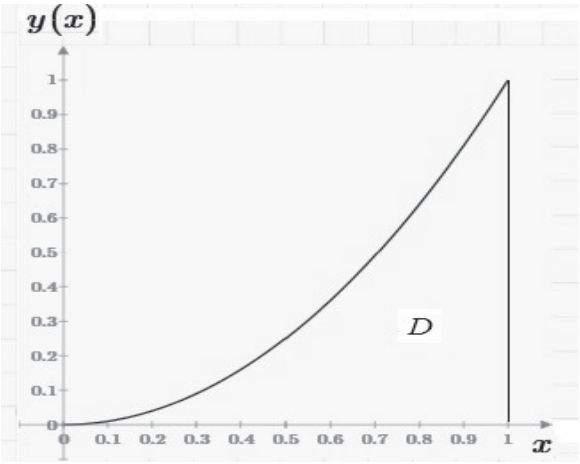


Рис. 10.2 — Область D

Для генерации системы двух равномерно распределенных на интервале от 0 до 1 случайных величин используем инструмент табличного процессора *MS Excel — Анализ данных — Генерация случайных чисел*. Заполним диалоговое окно для генерации 10 000 пар указанных случайных чисел (рис. 10.3).

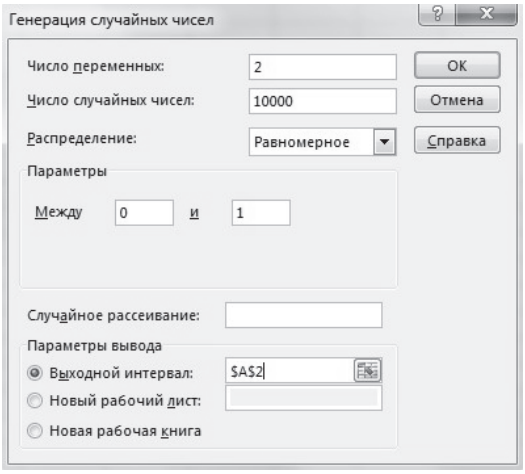


Рис. 10.3 — Диалоговое окно генерации двумерных, равномерно распределенных случайных величин

В результате в диапазоне $A1{:}B10001$ получим искомые пары случайных чисел.

В ячейке $C2$ введем формулу: $= A1\text{^}2$;

в ячейке $D2$: $= \text{Если}\,(B2 > C2;\ 0;\ 1)$.

Последняя формула присваивает ячейке значение 0 ($m_i = 0$), если точка M_i не попадает в область D и значение 1 ($m_i = 1$) в противном случае. Выделим диапазон $C2{:}D2$ и скопируем вниз до строки 10001. Найдем сумму значений в диапазоне $D2{:}D10001$, в результате получим, что $n = \sum m_i = 3334$ (рис. 10.4).

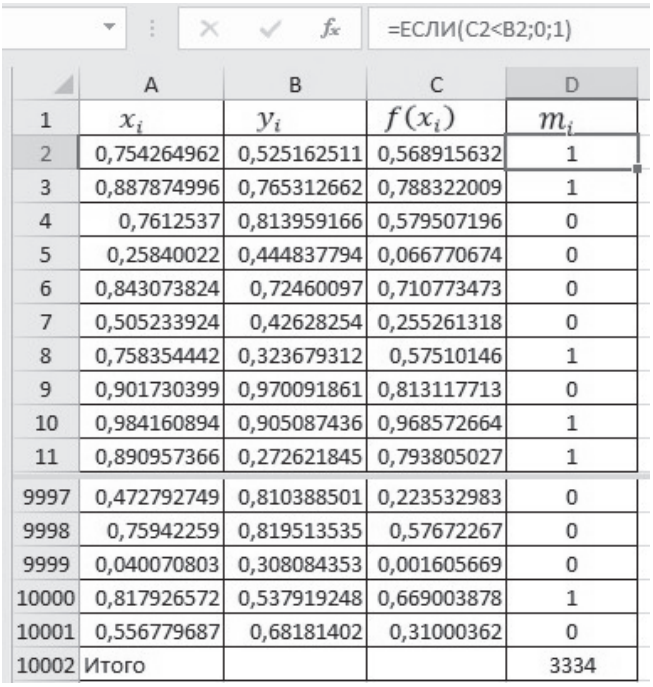
Отсюда, $I \approx \frac{n}{N} = \frac{3334}{10000} = 0{,}3334$.

Применяя неравенство Чебышёва, имеем

$$P\left(\left|\frac{n}{N} - I\right| < \varepsilon\right) \geq 1 - \frac{I(1-I)}{\varepsilon^2 N} \geq 1 - \frac{1}{4\varepsilon^2 N}.$$

Если мы зададим уровень значимости α, то неравенство, приведенное выше, будет всегда верно с гарантийной вероятностью

$$p = 1 - \alpha, \text{ при } \alpha = \frac{1}{4\varepsilon^2 N}.$$



=ЕСЛИ(C2<B2;0;1)

	A	B	C	D
1	x_i	y_i	$f(x_i)$	m_i
2	0,754264962	0,525162511	0,568915632	1
3	0,887874996	0,765312662	0,788322009	1
4	0,7612537	0,813959166	0,579507196	0
5	0,25840022	0,444837794	0,066770674	0
6	0,843073824	0,72460097	0,710773473	0
7	0,505233924	0,42628254	0,255261318	0
8	0,758354442	0,323679312	0,57510146	1
9	0,901730399	0,970091861	0,813117713	0
10	0,984160894	0,905087436	0,968572664	1
11	0,890957366	0,272621845	0,793805027	1
9997	0,472792749	0,810388501	0,223532983	0
9998	0,75942259	0,819513535	0,57672267	0
9999	0,040070803	0,308084353	0,001605669	0
10000	0,817926572	0,537919248	0,669003878	1
10001	0,556779687	0,68181402	0,31000362	0
10002	Итого			3334

Рис. 10.4 — Результат применения метода Монте-Карло

При заданных значениях ε и α можно определить необходимое число испытаний:

$$N = \frac{1}{4\varepsilon^2 \alpha}.$$

В силу того, что неравенство Чебышёва дает нижнюю оценку вероятности, значение *N* будет завышено, например, в нашем случае при $\varepsilon = 0{,}001$ и $\alpha = 0{,}01$: $N = 25\,000\,000$. Точнее значение $I = 0,(3)$. В рассматриваемом примере точность ε = 0,001 достигается уже при 10 000 испытаний. (Существуют более точные методы оценки *N*[14], основывающиеся на предельных теоремах теории вероятностей.)

5. *Компьютерное программирование*. С помощью случайных чисел часто тестируют эффективность компьютерных алгоритмов, а также создают рандомизированные (случайные) алгоритмы (например, хеширование), которые часто превосходят свои детерминированные аналоги.

[14] *Демидович Б. П.* Основы вычислительной математики / Б. П. Демидович, И. А. Марон. — М. : Физматгиз, 1963. — 660 с.

6. *Байесовская статистика* ([53], часть III).

7. *Статистическое (машинное) обучение* ([53], часть IV).

8. *Принятие решений.* Иногда важно принять полностью беспристрастное решение, тогда говорят, что некоторые преподаватели кафедры статистики бросают монеты или игральные кости… Случайность — важная часть оптимальных стратегий в теории матричных игр.

9. *Эстетика.* Небольшая добавка случайности в живопись, музыку часто их оживляет.

10. *Развлечения.* Люди любят проводить время, играя в карты, нарды, вращая рулетку и т. д. Такие традиционные способы использования случайных чисел получили название метод Монте-Карло.

Каковы источники случайных чисел?

1) Вытаскивание шаров из урны, бросание костей, вращение рулетки и т. д.

2) Существуют *таблицы случайных чисел*. Первая, содержащая 40 000 значений, взятых наудачу из отчетов о переписи, опубликована в 1927 г. Типпетом.

3) *Механические генераторы случайных чисел.* Первая такая машина была использована в 1939 г. Кендаллом и Бабингтоном-Смитом для построения таблицы, содержащей 100 000 случайных чисел.

4) *Встроенные программы*. Первая, запущенная в 1951 г., использовала резисторный генератор шума, поставляющий 20 случайных битов на сумматор (Тьюринг). В 1955 г. были опубликованы таблицы с 1 000 000 случайных чисел.

Использование таблиц было ограничено, но в 1990-е гг. интерес к ним вернулся. Д. Марсалья в 1995 г. подготовил демо-диск с 650 Мбайт случайных чисел, при генерировании которых запись шума диодной цепи сочеталась с музыкой в стиле «рэп» (он назвал это «черным и белым шумом»).

Несовершенство первых механических методов побудило интерес к получению случайных чисел с помощью арифметических операций, заложенных в компьютер.

Джон фон Нейман в 1946 г. первым предложил такой алгоритм. Идея — возводим в квадрат k значное число и берем среднее k цифр и т. д.

Например, для десятизначного числа 5772156649, возводим в квадрат, получим 33317792380594909201; значит, следующее число 7923805949. Однако! При небольших значениях k алгоритм части приводит к циклу (!), т. е. числа не являются случайными! «Каждый, кто использует арифметические методы генерации случайных чисел, безусловно, грешит» (Дж. фон Нейман). Действительно, если их можно рассчитать, то они не случайны. Поэтому генераторы случайных чисел часто называют генераторами «псевдослучайных» чисел. В языке программирования ПАСКАЛЬ, в основе генераторов случайных чисел лежат встроенные функции *RANDOM* и *RANDOMIZE*, генерирующие равномерно распределенные числа на [0;1]. На их основе строятся случайные числа, подчиняющиеся другим законам распределения.

Пример 10.5. Рассмотрим алгоритм вычисления двух независимых нормально распределенных случайных величин X_1, X_2.

1. Сгенерируем случайные величины Z_1, Z_2, подчиняющиеся равномерному закону распределения на $[0; 1]$. Тогда случайные величины

$$Y_1 = 2Z_1 - 1,\ Y_2 = 2Z_2 - 1$$

равномерно распределены на $[-1; 1]$.

2. Присвоим S значение суммы квадратов Y_1, Y_2.

$$S \coloneqq Y_1^2 + Y_2^2.$$

3. Если $S \geq 1$, то возврат на предыдущие этапы (1, 2).

4. Присвоим X_1 и X_2 значения:

$$X_1 \coloneqq Y_1\sqrt{\frac{-2\,ln(S)}{S}},\quad X_2 \coloneqq Y_2\sqrt{\frac{-2\,ln(S)}{S}}.$$

X_1 и X_2 — требуемые нормально распределенные случайные величины (доказательство, см. пример 7.9).

Специалисты рекомендуют использовать для сравнения разные источники случайных чисел — это будет указывать на стабильность результатов. По высказыванию Дж. Морсалья «генератор случайных чисел похож на sex. Если он хорош, то это прекрасно. Если плох, то все равно приятно».

10.3. Вероятностный подход к понятию информации

Что наша жизнь?
Игра!
Добро и зло — одни мечты!
Труд, честность — сказки для бабья.
Кто прав, кто счастлив здесь, друзья!
Сегодня ты, а завтра я!

(Ария Германа. Опера
П. И. Чайковского «Пиковая дама»)

Известна игра, когда один из участников отгадывает, что загадал второй. Вопрос допускает только два ответа «да» и «нет» (да — 1, нет — 0). Игра называется Бар-Кохба. Согласно легенде, Бар-Кохба («сын звезды») был предводителем восстания 135 г. в Иудее против владычества римлян. Превосходящее по силам войско осадило крепость, которую оборонял гарнизон под командованием Бар-Кохбы. Согласно легенде, он послал в стан врага лазутчика, которого римляне схватили и отрезали язык. Несчастный бежал из стана врагов, пришел к Бар-Кохбе, но не смог рассказать, что увидел у противника. Тогда Бар-Кохба стал задавать ему вопросы, на которые можно было ответить только либо «да», либо «нет», и получил необходимую информацию. Таким образом, в известной мере игру Бар-Кохба нужно считать предтечей теории информации. Видимо, еще в древности была известна возможность закодировать любую информацию в виде последовательности двух символов «0» и «1».

Записав последовательность полученных ответов в виде «0» и «1», мы получим число в двоичной системе исчисления. 1 бит (*bit — binary digit* — двоичная единица) — единица информации, которую можно закодировать «0» и «1».

С примерами кодирования и декодирования информации мы встречаемся постоянно. Например, пишем ⟶ кодируем, читаем ⟶ декодируем, радио-сигнал — кодирование ⟶ антенна ⟶ изображение на экране телевизора — декодирование; пластинки — механическое кодирование, кассеты — магнитное кодирование, лазерные диски — оптическое кодирование.

Рассмотрим количество информации на примере игры Бар-Кохба. В списке 32 человека. Сколько вопросов нужно задать, чтобы отгадать человека, которого задумали? Разделим список на две части по 16 человек, зададим вопрос, в какой части загаданный человек, и так далее 5 раз:

$$32 \longrightarrow 16 \longrightarrow 8 \longrightarrow 4 \longrightarrow 2 \longrightarrow 1.$$
$$1в \quad 2в \quad 3в \quad 4в \quad 5в$$

Таким образом, необходимо задать 5 вопросов или получить информацию 5 бит. А если список из N = 48 человек, то одного из них отгадать можно $[log_2 N] + 1$ вопросами. Действительно, если N = 64, то нужно 6 бит информации (6 вопросов).

Имеем при N = 48:

$$log_2 32 < log_2 N < log_2 64,$$
$$5 < log_2 N < 6.$$

Определение 1. Если в заданном множестве H, содержащем N элементов, выделен элемент X ($X \in H$), то чтобы найти X, необходимо получить количество информации, равное

$$H(X) = log_2 N \text{ битам.} \qquad (10.5)$$

Это формула Хартли.

Определение 2. Если не все элементы из N равновероятны:

x_i	x_1	x_2	...	x_N					
p_i	p_1	p_2	...	p_N					

то количество информации, необходимое для поиска элемента X, равно

$$H(X) = p_1 log_2 \frac{1}{p_1} + p_2 log_2 \frac{1}{p_2} + \cdots + p_N log_2 \frac{1}{p_N} — \qquad (10.6)$$

это *формула Шеннона* (партнер загадывает x_i с вероятностями p_i и сыграно много партий).

При $p_1 = p_2 = \cdots = p_N = \frac{1}{N}$, $H(X) = log_2 N$ — *формула Хартли*.

Например, бросаем две монеты. Случайная величина X — число орлов, $X = \{0, 1, 2\}$.

x_i	0	1	2
p_i	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Среднее число знаков для составления кода одного бросания (кодовое слово) по формуле Шеннона:

$$H(X) = \frac{1}{4} log_2 \frac{1}{\frac{1}{4}} + \frac{1}{2} log_2 \frac{1}{\frac{1}{2}} + \frac{1}{4} log_2 \frac{1}{\frac{1}{4}} = \frac{1}{4} \cdot 2 + \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 = 1{,}5.$$

Действительно: $X = 0$: 00_2 (0 цифр),
$X = 1$: 1_2 (1 цифра),
$X = 2$: 10_2 (2 цифры).

$M(X) = 2 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 1{,}5$ — математическое ожидание длины кодового слова. Так как для значения длины кодового слова, соответствующей исходу одного бросания, 1 вероятность равна 0,5; для значения (длины кодового слова) 2 вероятность равна 0,5.

В переводе на язык игры Бар-Кохба нам следует спросить у партнера, не выпал ли у него один герб. В случае утвердительного ответа мы будем знать все, в противном случае следует задать вопрос, не выпало ли у него 2 герба. Таким образом, в игре Бар-Кохба по возможности следует задавать вопросы так, чтобы утвердительные и отрицательные ответы были равновероятны (или близки к ним).

Формулу, названную впоследствии формулой Шеннона, в XIX веке вывел Больцман для решения другой задачи. Он показал, что если в газе, состоящем из множества молекул, вероятности состояний соседних из них $p_1, p_2, \dots, p_n$, то энтропия (Клаузиус, 1865) системы определяется так

$$H = c(p_1 log_2 \frac{1}{p_1} + p_2 log_2 \frac{1}{p_2} + \cdots + p_n log_2 \frac{1}{p_n}), \qquad (10.7)$$

где $c = const$.

Энтропия служит мерой неупорядоченности системы или, иначе, это *мера неопределенности* системы. Таким образом, неопределенность — это недостача информации или отрицательная информация.

Обычно понятие энтропии используется для описания процессов передачи информации в технических или природных системах, кроме того, в прикладной статистике существует ряд информационных характеристик качества моделей, оперирующих понятием энтропии.

Системный подход постулирует наличие двух видов систем: замкнутых и открытых. Деление весьма условное и зависит от целей решаемых задач.

В открытых системах, особенно при наличии внешних воздействий (катализаторов) последовательно возникают все более устойчивые диссипативные структуры, что характерно для процессов самоорганизации. В качестве примеров подобных систем приводят процессы циркуляции в атмосфере, биологическую жизнь.

«Закон Людвига фон Берталанфи (в открытых системах противоположен второму началу термодинамики) утверждает возможность возникновения негэнтропийных тенденций за счет обмена со средой массой, энергией, ИНФОРМАЦИЕЙ. А позднее понято, что не только за счет открытости, но и за счет АКТИВНЫХ элементов системы!» (В. Н. Волкова).

Для закрытых систем постулируется: первое начало термодинамики — закон сохранения энергии; второе начало термодинамики — закон возрастания энтропии. Закон возрастания энтропии приводит к возможному объяснению необратимости процессов (в том числе старения, разрушения), наличии стрелы времени и объясняет невозможность путешествия во времени из-за бесконечного барьера энтропии.

Если $f(x)$ — функция плотности вероятности непрерывной случайной величины X, где $x \in D \subseteq R$, то энтропия определяется как

$$H(X) = -M(\log_2 f(x)) = -\int f(x)\log_2 f(x)dx. \quad (10.8)$$

На промежутках $(0,1); (0,+\infty); (-\infty,+\infty)$ *законы распределения с наибольшей энтропией соответственно: равномерный, показательный* $(M(X)=1)$ *и нормальный* $(M(X)=0, D(X)=1)$ [69].

Указанный факт отражает *принцип максимума энтропии.* При заданной информации о «поведении» среды, состояния неопределенной среды описываются распределениями, доставляющими максимум энтропии.

Для доказательства используем неравенство Йенсена, выражающее свойство выпуклой вниз функции — секущая выше графика.

Неравенство Йенсена в вероятностной формулировке.

Пусть функция φ выпукла вниз на D, математическое ожидание X и $\varphi(x)$ конечны $(|M(X)| < \infty, (|M(\varphi(x))| < \infty)$, тогда верно неравенство Йенсена:

$$f\big(M(X)\big) \le M\big(f(x)\big). \quad (10.9)$$

Доказательство опирается на тот факт, что выпуклая вниз функция всегда выше любой своей касательной, то есть для любых точек из области определения выполняется неравенство

$$f(x) \ge f(y) + f'(x)(x-y). \quad (10.10)$$

Пусть $y = M(X)$, тогда неравенство (10.21) перпишется в виде

$$f(x) \ge f(M(X)) + f'(x)(x - M(X)). \quad (10.11)$$

Возьмем математическое ожидание от обеих частей неравенства (10.11), используя три легко выводимых свойства математического ожидания:

1) если $X \ge Y$, то $M(X) \ge M(Y)$,

2) $M(x - M(X)) = 0$,

3) $M\left(f\big(M(X)\big)\right) = f\big(M(X)\big)$.

Имеем

$$M\big(f(x)\big) \ge M\left(f\big(M(X)\big)\right) + M\left(f'(x)\big(x - M(X)\big)\right)$$

или

$$f\big(M(X)\big) \le M\big(f(x)\big).$$

Из неравенства (10.20) можно получить, например, что

$$M(e^x) \ge e^{M(X)}, M(X^2) \ge \big(M(X)\big)^2, M(\ln(x)) \le \ln\big(M(X)\big), (X > 0). \quad (10.12)$$

Опираясь на неравенство Йенсена, рассмотрим несколько утверждений.

Утверждение 1. Максимум энтропии H, равный $\log_2 N$, достигается при

$$p_1 = p_2 = \cdots = p_N = \frac{1}{N}.$$

Доказательство. Рассмотрим функцию $f(x) = x\log_2 x$.

$f'(x) = \log_2 x + \frac{1}{\ln(2)}$, $f''(x) = \frac{1}{x} > 0$, при $x > 0$, следовательно, функция f выпукла вниз, как и энтропия, вычисленная по формуле Шеннона (10.6). Запишем неравенство Йенсена для случайной величины X, принимающей значения p_i с вероятностями $\frac{1}{N}$:

$$f\left(\frac{1}{N}\sum_{i=1}^{N} p_i\right) \le \sum_{i=1}^{N} \frac{1}{N}\log_2 f(p_i),$$
$$f\left(\frac{1}{N}\right) \le \frac{1}{N}\sum_{i=1}^{N} \log_2 f(p_i),$$
$$-\frac{1}{N}\log_2 N \le \frac{1}{N}\sum_{i=1}^{N} p_i \log_2 p_i,$$

следовательно,

$$H(X) = -\sum_{i=1}^{N} p_i \log_2 p_i \le \log_2 N,$$

что равносильно утверждению 1.

Утверждение 2. Для произвольных функций плотности вероятности $h(x)$ и $g(x)$ верно неравенство

$$\int h(x)\, ln\big(h(x)\big)\, dx \ge \int h(x)\, ln\big(g(x)\big)\, dx. \qquad (10.13)$$

Доказательство. Преобразуем неравенство (10.24) и приведем к виду

$$\int_{-\infty}^{+\infty} h(x) ln\left(\frac{g(x)}{h(x)}\right) dx = M\left(ln\left(\frac{g(x)}{h(x)}\right)\right) \le 0, \qquad (10.14)$$

где $h(x)$ — плотность вероятности случайной величины X.

Как легко убедиться, взяв вторую производную, функция $ln(x)$ выпукла вниз ($X > 0$). Поэтому, применив неравенство Йенсена, получим

$$M\left(ln\left(\frac{g(x)}{h(x)}\right)\right) \ge ln\left(M\left(\frac{g(x)}{h(x)}\right)\right),$$

$$ln\left(M\left(\frac{g(x)}{h(x)}\right)\right) = ln\left(\int_{-\infty}^{+\infty} h(x)\frac{g(x)}{h(x)}dx\right) = ln\big(\int_{-\infty}^{+\infty} g(x)dx\big) = \ln(1) = 0,$$

что доказывает неравенство (10.13).

Утверждение 3. Максимум энтропии H для случайной величины $X \in (0, 1)$ достигается для равномерного закона.

Доказательство. В неравенстве (10.14) в качестве $g(x)$ рассмотрим равномерный закон на (0, 1), $g(x) = 1$.

Преобразуем неравенство (10.14) и приведем к виду

$$\int_0^1 h(x) ln\left(\frac{1}{h(x)}\right) dx = M\left(ln\left(\frac{1}{h(x)}\right)\right) \le 0. \qquad (10.15)$$

Очевидно, что верхняя граница достигается при $h(x) = 1$, что и требовалось доказать.

Утверждение 4. Максимум энтропии H для случайной величины $X \in \in (0, +\infty)$ достигается для показательного закона.

Доказательство. В неравенстве (10.14) в качестве $g(x)$ рассмотрим показательный закон на (0, $+\infty$), $g(x) = e^{-x}$.

Преобразуем неравенство (10.13) и приведем к виду

$$\int_0^{+\infty} h(x) ln\big(h(x)\big) dx \ge \int_0^{+\infty} h(x) ln(e^{-x}) dx,$$
$$-\int_0^{+\infty} h(x) ln\big(h(x)\big) dx \le -\int_0^{+\infty} h(x)(-x) dx, \qquad (10.16)$$

итак,

$$-\int_0^{+\infty} h(x) ln\big(h(x)\big) dx \le M(X). \qquad (10.17)$$

Равенство в (10.17) достигается при $h(x) = e^{-x}$, что и требовалось доказать.

Утверждение 5. Максимум энтропии H для случайной величины $X \in (-\infty, +\infty)$ достигается для нормального закона $N(0,1)$.

Доказательство. В неравенстве (10.13) в качестве $g(x)$ рассмотрим нормальный закон на $(-\infty, +\infty)$, $g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

Преобразуем неравенство (10.13) и приведем к виду

$$-\int_{-\infty}^{+\infty} h(x) ln\big(h(x)\big) dx \le -\int_{-\infty}^{+\infty} h(x) ln\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right) dx, \qquad (10.18)$$

учитывая, что $M(X^2) = D(X) + (M(X))^2$, имеем

$$-\int_{-\infty}^{+\infty} h(x)\left(-ln\sqrt{2\pi} - \frac{x^2}{2}\right) dx = ln\sqrt{2\pi}\int_{-\infty}^{+\infty} h(x)dx + \frac{1}{2}\int_{-\infty}^{+\infty} x^2 h(x)dx =$$
$$= ln\sqrt{2\pi}\cdot 1 + \frac{1}{2} = ln\sqrt{2\pi e}.$$

Верхняя граница достигается при $h(x) = g(x)$, что и требовалось доказать.

Замечание. 1. Можно оценить среднее количество вопросов, необходимых для ответа. Пусть каждый человек обладает словарным запасом из *n* слов, тогда он сможет разбить свой словарный запас на *m* групп так, что вероятности нахождения нужного (ключевого слова) для каждой группы одинаковы и равны

$$p = \frac{1}{m},$$

то вероятность неудачи

$$q = 1 - p = \frac{m-1}{m}.$$

Таким образом, процесс поиска можно описать биноминальным распределением.

I случай — Хеширование — безуспешный поиск. Итак, если *X* — число вопросов для нахождения ответа, то

$min\, x = 0$ — сразу отгадал, не задавая вопросы,

$max\, x = n$ — практически $n \to \infty$.

$$M(X) = \frac{n}{m}\,, D(X) = \frac{n(m-1)}{m^2}\,, \sigma(X) = \sqrt{n/m}\,.$$

II случай — Хеширование — успешный поиск.

В «базе данных» студента есть информация о ключе (или иначе о заданном вопросе), тогда $min\, x = 0$, $max\, x = n$.

Математическое ожидание числа вопросов $M(r) = \frac{n-1}{2m} + 1$. С ростом n, число вопросов стремится к $0{,}5 \ln n$, а среднее квадратическое отклонение к $(\ln n)/\sqrt{12}$, что в какой-то степени ускоряет процесс поиска ответа.

1. *Задача для исследования*. Оцените возможное значение *n* и приведите численную оценку среднего количества вопросов и времени, если современные исследования психологов показали, что человек может держать в сознании (памяти) 7 ± 2 чанка (бита) информации. ■

10.4. Байесовские сети

... это графические структуры для представления вероятностных отношений между большим количеством переменных и для осуществления вероятностного вывода на основе этих переменных.

(А. Л. Тулупьев, С. И. Николенко, А. В. Сироткин. Байесовские сети: Логико-вероятностный подход, 2006)

Теорема умножения предполагает, что

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

Отсюда

$$P(A/B) = \frac{P\,(B/A)P(A)}{P(B)} —$$

это известная формула Байеса.

Используя теоретико-множественный смысл события, можно записать

$$B = (B \cap A) \cup (B \cap \bar{A}\,),$$

тогда в силу независимости

$$P(B) = P(B \cap A) + P\,(B\ \cap \bar{A}\,)$$

или

$$P(B) = P(B/A)P(A) + P(B/\bar{A}\,)P(\bar{A}\,).$$

Теперь формулу Байеса можно переписать в виде

$$P(A/B) = \frac{P(B/A)P(A)}{P(B/A)P(A)+P(B/\bar{A}\,)P(\bar{A}\,)}.$$

В современных информационных технологиях эта формула используется как основа для управления неопределенностью — «делать выводы вперед и назад».

Основой современного логического вывода считается «если..., то...» правила. В нашем случае, «если событие *H* истинно, то событие *E* будет наблюдаться с вероятностью *P*», а если событие *E* уже произошло, то какова вероятность истинности *H*?

Пусть *H* — событие, заключающееся в том, что данная гипотеза верна, *E* — событие, заключающееся в том, что наступило определенное доказательство (свидетельство), которое может подтвердить правильность указанной гипотезы. Тогда формула Байеса примет вид

$$P(H/E) = \frac{P(E/H)P(H)}{P(E/H)P(H)+P(E/\bar{H})P(\bar{H})}.$$

Она устанавливает связь гипотезы *H* и свидетельства *E* и в то же время свидетельства с неподтвержденной гипотезой. *P*(*H*) — априорная вероятность гипотезы, известная до наступления события *E*. Предполагается, что *P*(*H*) и *P*(*E*/*H*) находятся опытным или экспериментальным путем. Формулу Байеса обобщают на случай множества гипотез ($H_1, H_2, ..., H_m$) и множества свидетельств ($E_1, E_2, ..., E_n$).

Вероятности каждой из гипотез можно определить по формуле

$$P(H_i \,/\, E_1E_2 \ldots E_n) = \frac{P(E_1E_2 \ldots E_n/\, H_i)P(H_i)}{\sum_{k=1}^{m} P(E_1E_2 \ldots E_n/\, H_k)P(H_k)}, \qquad (10.19)$$

где $i = 1, 2, \ldots, m$.

Сложность формулы заключается в необходимости знать все условные вероятности знаменателя, поэтому часто делается довольно сильное предположение о *независимости свидетельств* (подход называют наивный Байес — *naive Bayes*). Тогда формула приобретает вид

$$P(H_i \,/\, E_1E_2 \ldots E_n) = \frac{P(E_1E_2 \ldots E_n/\, H_i)P(H_i)}{\sum_{k=1}^{m} P(E_1/\, H_k)P(E_2/\, H_k) \ldots P(E_n/\, H_k)P(H_k)}. \qquad (10.20)$$

Если $\{E_1,\ E_2, \ldots\ , E_n\} =: D$, то формулу 10.2 можно переписать как

$$P(H_i \,/\, D) = \frac{P(D/\, H_i)P(H_i)}{P(D)} \qquad (10.21)$$

или

$$P(H_i \,/\, D) \propto P(D/\, H_i)P(H_i) — \qquad (10.22)$$

апостериорная вероятность гипотезы H_i *пропорциональна* ($\propto$) *произведению правдоподобия полученной информации на априорную оценку вероятности*, где $1/P(D) = const$ — коэффициент пропорциональности (масштабный множитель), который не зависит от H_i и обеспечивает равенство суммы всех апостериорных вероятностей единице [116, 118].

Пример 10.6. Имеется три взаимно независимых состояния фирмы.

Вероятность	$i = 1$	$i = 2$	$i = 3$
$P(H_i)$	0,6	0,3	0,1
$P(E_1/H_i)$	0,3	0,7	0,2
$P(E_2/H_i)$	0,6	0,8	0,0

Гипотезы:

H_1 — «средняя надежность фирмы»,

H_2 — «высокая надежность фирмы»,

H_3 — «низкая надежность фирмы».

Имеется два условно независимых свидетельства, подтверждающих в разной степени исходные гипотезы:

E_1 — «наличие прибыли у фирмы»,

E_2 — «своевременный расчет с бюджетом».

Новые факты, получаемые в процессе сбора, будут повышать или понижать вероятности гипотез. Пусть с вероятностью 1 наступило событие E_2, тогда апостериорные вероятности для гипотез согласно формуле Байеса, для одного свидетельства:

$$P(H_i \,/\, E_2) = \frac{P(E_2/\, H_i)P(H_i)}{\sum_{k=1}^{3} P(E_2/\, H_k)P(H_k)},$$

$i = 1, 2, 3$.

Имеем

$$P(H_1/\, E_2) = \frac{0{,}6 \cdot 0{,}6}{0{,}6 \cdot 0{,}6 + 0{,}3 \cdot 0{,}8 + 0{,}1 \cdot 0{,}0} = \frac{0{,}36}{0{,}6} = 0{,}6.$$

$$P(H_2 \,/\, E_2) = \frac{0{,}3 \cdot 0{,}8}{0{,}6 \cdot 0{,}6 + 0{,}3 \cdot 0{,}8 + 0{,}1 \cdot 0{,}0} = \frac{0{,}24}{0{,}6} = 0{,}4.$$

$$P(H_3 \,/\, E_2) = \frac{0{,}1 \cdot 0{,}0}{0{,}6 \cdot 0{,}6 + 0{,}4 \cdot 0{,}8 + 0{,}1 \cdot 0{,}0} = 0.$$

После того как событие E_2 произошло, доверие к гипотезе H_3 понизилось, а доверие к H_2 возросло. Если есть факты, подтверждающие и событие E_1, и событие E_2, то при условии их независимости формула Байеса будет выглядеть в следующем виде:

$$P(H_i / E_1E_2) = \frac{P(E_1/ H_i)P(E_2/ H_i)P(H_i)}{\sum_{k=1}^{3} P(E_1/ H_k)P(E_2/ H_k)P(H_k)},$$

где $i = 1, 2, 3$.

Таким образом:

$$P(H_1 /E_1E_2) = \frac{0{,}3\cdot 0{,}6\cdot 0{,}6}{0{,}3\cdot 0{,}6\cdot 0{,}6+0{,}7\cdot 0{,}8\cdot 0{,}3+0{,}2\cdot 0{,}0\cdot 0{,}1} = \frac{0{,}108}{0{,}276} = 0{,}391,$$

$$P(H_2 /E_1E_2) = \frac{0{,}7\cdot 0{,}8\cdot 0{,}3}{0{,}3\cdot 0{,}6\cdot 0{,}6+0{,}7\cdot 0{,}8\cdot 0{,}3+0{,}2\cdot 0{,}0\cdot 0{,}1} = \frac{0{,}168}{0{,}276} = 0{,}609,$$

$$P(H_3 / E_1E_2) = \frac{0{,}2\cdot 0{,}0\cdot 0{,}1}{0{,}3\cdot 0{,}6\cdot 0{,}6+0{,}7\cdot 0{,}8\cdot 0{,}3+0{,}2\cdot 0{,}0\cdot 0{,}1} = 0.$$

После получения свидетельств E_1 и E_2 осталось только две гипотезы: H_1 и H_2. При этом H_2 более вероятно, чем H_1.

Данный пример иллюстрирует процесс распространения вероятностей по элементам экспертной системы (ЭС), основанной на байесовских сетях, при поступлении в нее новых свидетельств. Можно показать, что последовательное поступление свидетельств приводит к результатам, аналогичным применению формулы Байеса для одновременно поступающих свидетельств.

Байесовская сеть доверия (БСД) — это направленный ациклический граф со следующими свойствами [116]:

– каждая вершина — событие, описывающееся случайной величиной,

– вершины, связанные с «родительскими», определяются таблицей или функцией условных вероятностей,

– вероятности состояний вершин без «родителей» являются безусловными.

Таким образом, в байесовских сетях доверия вершины — случайные величины, дуги — вероятностные зависимости, определяющиеся таблицей или функцией условных вероятностей.

Пример 10.7. (*Построение байесовской сети доверия*.) Фирма обанкротилась. Основные факторы: надежность фирмы (обобщенный внутренний фактор), экономический кризис (обобщенный внешний фактор) (рис. 10.5). Рассмотрим ситуацию, в которой первая вершина «надежность» — *reliability* имеет два состояния «высокая» — *high*, «невысокая» — *not high*, а вторая «кризис» — *crisis*, два состояния кризис «повлиял» — *affected*, «не повлиял» — *not affected*. Вершины «надежность» и «экономический кризис» не имеют родительских вершин, поэтому они являются маргинальными, то есть ни от чего не зависят (рис. 10.5, 10.6).

Рис. 10.5 — Основные факторы

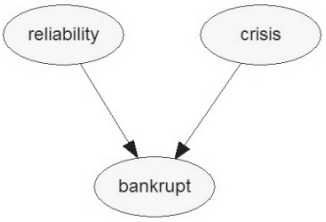


Рис. 10.6 — Main factors

Приведенные таблицы показывают вероятности пребывания вершины «обанкротилась» — *bankrupt* в определенном состоянии, обусловленном состоянием родительских вершин (табл. 10.1, 10.2).

Таблица 10.1

Априорные вероятности

Априорная вероятность *P*(*reliability*)		Априорная вероятность *P*(*crisis*)	
affected	0.1	high	0.1
not affected	0.9	not high	0.9

Таблица 10.2

Условные вероятности *P*(*bankrupt*/*crisis*, *reliability*)

crisis	affected		not affected	
reliability	high	not high	high	not high
bankrupot	0.95	0.85	0.9	0.02
not bankrot	0.05	0.15	0.1	0.98

Современные программные средства (*Netica*, *Hugin* и др.) позволяют строить более сложные байесовские сети доверия, вводить новые свидетельства и получать решения (выводы) за счет пересчета новых вероятностей в вершинах, соответствующих новым свидетельствам, например *P*(фирма = «*bankrupt*») = = 0,1832 (рис. 10.7).

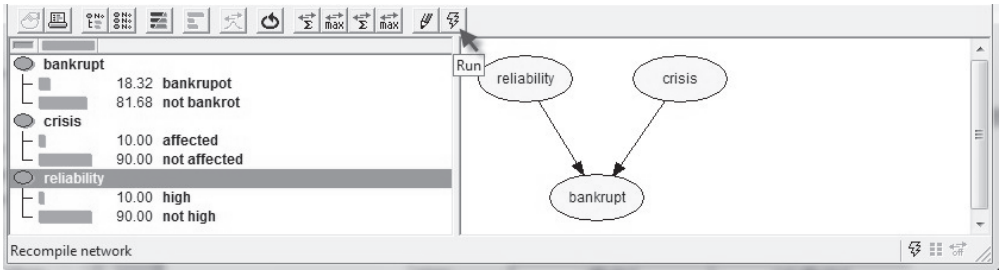


Рис. 10.7 — Пересчет новых вероятностей

Пусть известно, что фирма обанкротилась (с *вероятностью 1 или на 100%*). После этого можно узнать вероятность того, как повлияет экономический кризис. Для приведенных выше данных, опираясь на результаты вывода путем распространения сумм по БСД, можно показать, что кризис повлиял с вероятностью 0,469, а надежность фирмы с вероятностью 0,494 (рис. 10.8).

Можно предположить, что на банкротство фирмы с вероятностью 0,531 не повлияет кризис и с вероятностью 0,506 не повлияет надежность фирмы. Однако этот вывод является преждевременным. Для нахождения наиболее вероятной комбинации состояния вершин в программе *Netica* вместо распространения сумм нужно использовать распространение максимумов.

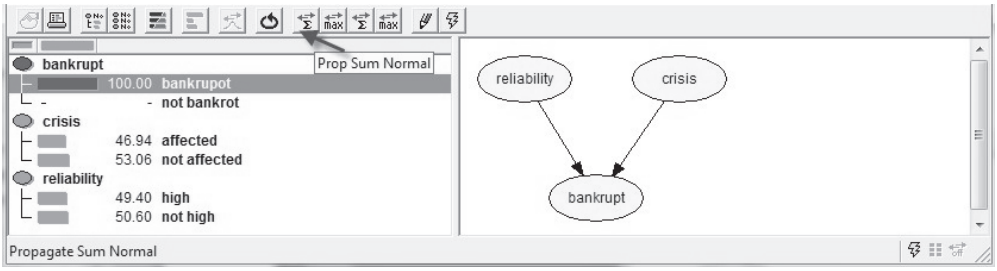


Рис. 10.8 — Вероятность влияния факторов

Каждое из состояний вершин, имеющее значение 100.00, будет принадлежать к наиболее вероятной комбинации состояний. В рассматриваемом примере наиболее вероятно, что кризис не повлияет, а надежность фирмы будет высокая (рис. 10.9).

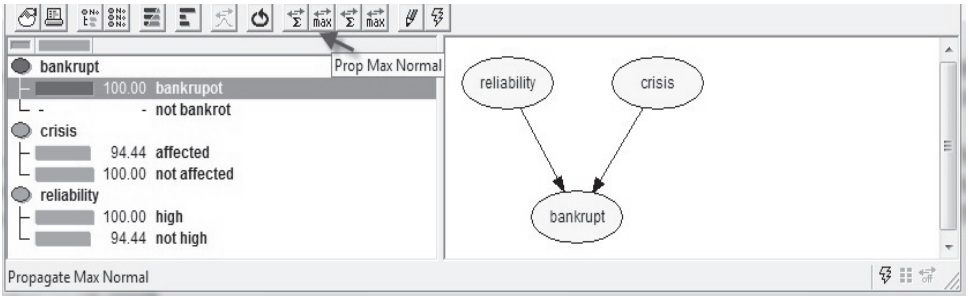


Рис. 10.9 — Определение наиболее вероятной комбинации состояний

Рассмотрим наиболее вероятную комбинацию состояний, полученных при наличии факта, что фирма обанкротилась.

P(*reliability = «high», crisis = «not affected» / bankrupt= «bankrupt»*)

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

В наших обозначениях:

P(*reliability = «high», crisis = «not affected» / bankrupt*)= =*P*(*reliability = «high», crisis = «not affected», bankrupt= «bankrupt»*)/*P*(*bankrupt= «bankrupt»*)=*0,081/0,1832=0,442*.

Итак, наиболее вероятная ситуация, что кризис не повлиял и надежность фирмы была высокой, имеет значение вероятности 0,442 (рис. 10.10, 10.11). При разработке систем принятия решений и экспертных систем используют диаграммы влияния, к БСД добавляют вершины решения — *decisions* (прямоугольники) и вершины пользы — *utility* (ромбы) (рис. 10.12).

Предположим, что мы для повышения надежности фирмы инвестируем 8000 д. е. или 0 д. е. и таблицы условных вероятностей для дополнительных переменных *reliability`* и *crisis`* (рис. 10.12), полученные на основе обработки знаний экспертов, имеют вид, представленный на рисунке 10.13.

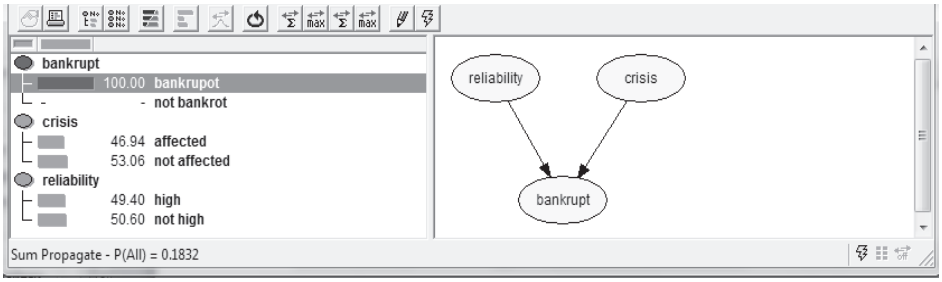


Рис. 10.10 — Определение вероятности

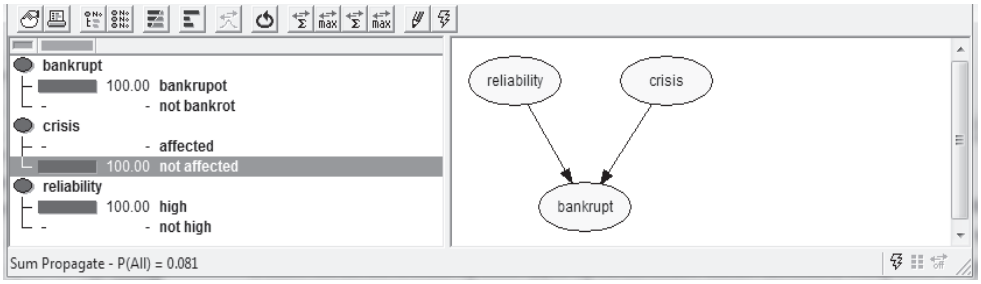


Рис. 10.11 — Наиболее вероятная ситуация

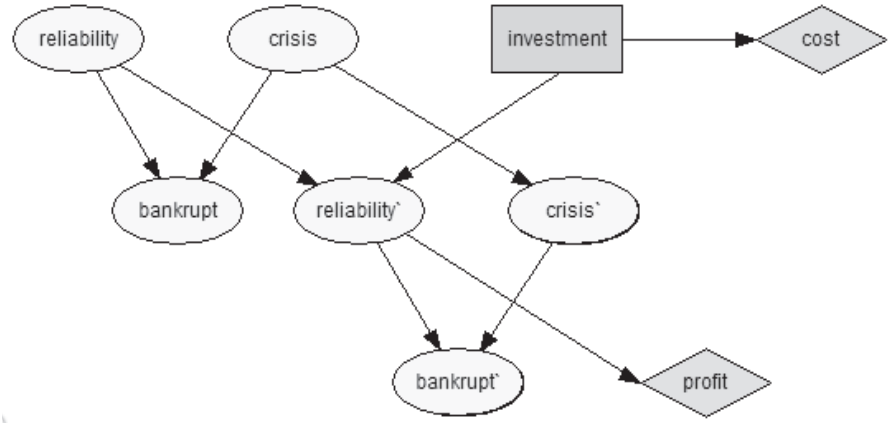


Рис. 10.12 — Диаграмма влияния

investment	investment		not	
reliability	high	not high	high	not high
high	0.2	0.99	0.99	0.02
not high	0.8	0.01	0.01	0.98

crisis	affected	not affected
affected	0.6	0.05
not affected	0.4	0.95

Рис. 10.13 — Инвестирование и условные вероятности

Тогда полезность от инвестиций составляет 10 487 д. е. и 4989 д. е. в противном случае (рис. 10.14). То есть инвестиции более предпочтительны. Если получено свидетельство о том, что фирма банкрот, то результативность инвестиций 11 805,7 д. е., а их отсутствие 11 979 д. е., то есть инвестировать нет смысла (рис. 10.15).

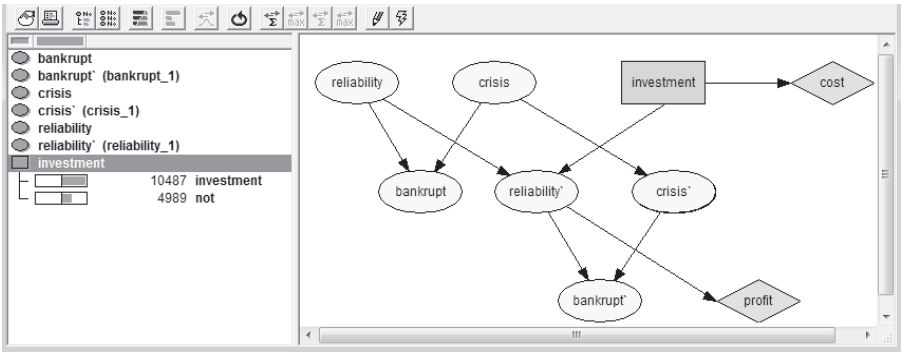


Рис. 10.14 — Полезность инвестиций

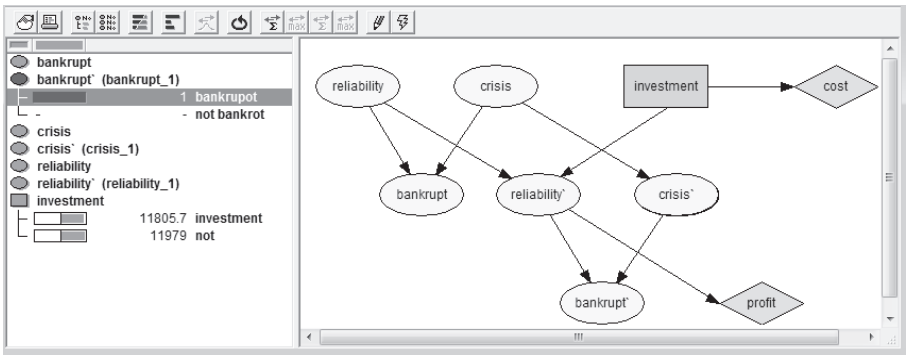


Рис. 10.15 — Результативность инвестиций

Байесовские сети доверия имеют широкое применение при разработке экспертных систем и систем принятия решений для оценки рисков в различных областях деятельности: медицине, финансах, коммерции и т. д.

Темы (вопросы) для самоконтроля

1. Алгоритм и сложность алгоритма.
2. Алгоритм нахождения максимума и его вероятностный анализ.
3. Хеширование и его вероятностный анализ.
4. Случайные числа и их применение.
5. Генераторы случайных чисел.
6. Энтропия. Информация (формулы Хартли и Шеннона).
7. Неравенство Йенсена в вероятностной формулировке.
8. Равномерный, показательный и нормальный законы распределения вероятностей как результат достижения максимума энтропии на отрезке, луче и прямой соответственно.
9. Байесовская сеть (*naive Bayes*).

«Дело не в цифрах, ..., —
а в том, что вы с ними делаете».

(*М. Дж. Кендалл, А. Стьюарт*
Статистические выводы и связи)

Часть II

Математическая статистика

Математическая статистика — раздел математики, посвященный математическим методам систематизации, обработки и использования статистических данных для научных и практических выводов. При этом статистическими данными называются сведения о числе объектов в какой-либо более или менее обширной совокупности, обладающих теми или иными признаками.

А. Н. Колмогоров, Ю. В. Прохоров.
Математическая энциклопедия, 1982

Статистика — это математическая теория, позволяющая познать мир через опыт.

В. Томпсон

Математическая статистика изучает задачи, связанные с правилами индуктивного поведения.

Ю. Нейман

Основная идеология методов многомерного статистического анализа сводится к использованию теории алгебраических инвариантов, не изменяющихся при линейных преобразованиях (например, собственные значения, собственные вектора, определители, декомпозиция матриц, корреляция между переменными и т. д.).

Справочная система Statistica

...Э. Резерфорд, известный физик и директор лаборатории им. Кавендиша, набирая на работу новых сотрудников, задавал им два прямых вопроса: «Вы окончили университет с отличием?» и «Вы умеете считать?». Ответ «Да» на оба вопроса был необходимым условием поступления на работу.

М. Я. Кельберт, Ю. М. Сухов.
Основные понятия теории вероятностей и математической статистики

Его (Аристотеля) величайшим и в то же время чреватым наиболее опасными последствиями вкладом в науку была идея классификации, которая проходит через все его работы ... Аристотель ввел способ классификации предметов, основанный на сходстве и различии ...

Дж. Бернал

Введение

Математическая статистика является разделом математики, в котором рассматриваются методы и приемы получения, систематизации, обработки и представления статистических данных об изучаемых объектах, процессах и явлениях в целях принятия обоснованных научных и практических решений. Математическая статистика позволяет обосновать ответы на вопросы: можно ли судить о явлении по ограниченной информации о нем (спрос и предложение на рынке товаров); случайно или закономерно изучаемое явление; как зависит результативный признак от факторного (как зависит урожайность от дозы внесения удобрений, при прочих равных условиях?); сколько необходимо провести наблюдений для объективного суждения об изучаемом явлении; какой фактор сильнее влияет на результат (влияние вида удобрений на урожайность) и т. д.

«Окружающий нас мир насыщен информацией — разнообразные потоки данных окружают нас, захватывая в поле своего действия, лишая правильного восприятия действительности. Не будет преувеличением сказать, что информация становится частью действительности и нашего сознания.

Без адекватных технологий анализа данных человек оказывается беспомощным в жестокой информационной среде и скорее напоминает броуновскую частицу, испытывающую жестокие удары со стороны и не имеющую возможности рационально принять решение.

Статистика позволяет компактно оценить данные, понять их структуру, провести классификации, увидеть закономерности в хаосе случайных явлений. Удивительно, что даже простейшие методы визуального и разведочного анализа данных позволяют существенно прояснить информацию, первоначально порождающую нагромождением цифр» (В. Боровиков — научный директор *StatSoft Russia. Statistica*. Искусство анализа данных на компьютере).

Методы математической статистики можно разделить на описательные (дескриптивные) и аналитические. *Описательные методы* позволяют судить о совокупности реальных наблюдений с помощью таблиц, графиков, характеристик положения (среднее арифметическое значение, мода, медиана), характеристик рассеяния (среднее линейное отклонение, среднее квадратическое отклонение, дисперсия, коэффициент вариации), характеристик вида распределения (асимметрия и эксцесс) и т. п. Эти результаты (описательной статистики), а также ряд методов изучения зависимостей — корреляционно-регрессионный

анализ и элементы теории субъективного (байесовского) вывода (элементы аналитической статистики) были получены научной школой Френсиса Гальтона, Карла Пирсона до 1900 г.

Под шкалой обычно понимается кортеж $\langle U,\ \varphi,\ V\rangle$, где U — множество реальных объектов, V — множество элементов некоторой знаковой системы с допустимым множеством операций, φ — способ отображения U на V ($U \overset{\varphi.}{\rightarrow} V$).

Различают шкалы количественные (конечные, бесконечные, дискретные и непрерывные — по множеству значений) и качественные. Качественные и количественные шкалы можно классифицировать по множеству допустимых операций (табл. I).

Таблица I

Классификация шкал

Шкала		Допустимые операции	Комментарии	Показатели центральной тенденции
Качественная	номинальная	$X = Y,$ $X \neq Y$ (сравнения)	Числа заменяют названия или имена, например, переменная пол $= \begin{cases} 1, \text{если пол мужской,} \\ 0, \text{если пол женский} \end{cases}$	*Мо — мода*
	порядковая	$X = Y,$ $X < Y, X > Y$ (сравнения и порядка)	Числа позволяют устанавливать порядок между объектами, например: минералы по шкале твердости Мооса: 1 — тальк, 2 — гипс, …, 10 — алмаз; разряд рабочего; оценка на экзамене	*Ме — медиана,* *Мо — мода*
Количественная	интервальная	$X = Y,$ $X < Y, X > Y,$ $X - Y, X + Y$	В этой шкале измеряется календарное время, t^0C, t^0F	$\bar{x}$ — средняя арифметическая, *Ме — медиана,* *Мо — мода*
	отношений	$X = Y,$ $X < Y, X > Y,$ $X - Y, X + Y,$ $X/Y, X * Y$	В этой шкале измеряются рост, вес, время, цены товаров, доходы и т. д.	$x_{\text{ср.геом.}} =$ $= \sqrt[n]{x_1 x_2 \ldots x_n},$ $\bar{x}$ — средняя арифметическая, *Ме — медиана,* *Мо — мода*

Следует отметить, что *статистические методы — это методы анализа чисел как таковых, а не истинных значений некоторого признака*. «Математика может избавить нас от мучительной необходимости размышлять, но мы должны платить за эту привилегию, испытывая муки раздумий как до того, как математика вступает в действие, так и после» (Каплан, 1964).

Одной из задач статистики является установление взаимосвязи между признаками, характеризующимися сопряженностью — когда изменению одного признака в среднем соответствует изменение другого.

Однако, прежде чем говорить о зависимости или независимости признаков, их нужно измерить. «*Статистические выводы могут быть адекватны реальности только тогда, когда они не зависят от того, какую единицу измерения предпочитает исследователь, т. е. когда они инвариантны относительно допустимого преобразования шкалы*» [88].

Наиболее общими мерами взаимосвязи признаков являются информационные меры связи, оперирующие понятием энтропии как количественной меры неопределенности (формулы Шеннона, Хартли и др., см. раздел 10.4). Одна из наиболее известных — *семантическая мера информации Харкевича* (I_{ij}), которая характеризует наличие причинно-следственных связей между факторами:

$$I_{ij} = Log_2 \frac{p_{ij}}{p_j}, \qquad \text{(II. 1)}$$

где p_{ij} — вероятность перехода объекта управления в j −е состояние в условиях действия *i*-го фактора, p_j — вероятность самопроизвольного перехода в j −е состояние или в среднем.

Замечание. 1. А. А. Харкевич — основатель института проблем передачи информации, носящего сегодня его имя. Приведенная выше *формула* (II. 1) предложена А. А. Харкевичем как *мера оценки целесообразности информации* (*знания*), где p_j — вероятность достижения цели до получения дополнительной информации о действии i − го фактора, p_{ij} — вероятность достижения цели после получения и использования информации.

2. Формула Харкевича и ее обобщение положены Е. В Луценко (КубГАУ) в основу теории системно-когнитивного анализа, реализованной в интеллектуальной информационной системе «Эйдос», обеспечивающей изучение предметной области и осуществление процесса поддержки принятия решений на основании ретроспективных данных [76]. ■

Взаимосвязи качественных признаков более подробно изучаются специалистами по социологии и психологии (педагогике) ввиду того, что они больше всех имеют дело с анализом социальных явлений, которые появляются в результате обработки анкет, работы с экспертами и т. д. В таблице II классификация шкал продолжается описанием наиболее часто встречающихся статистик и показателей связи. Шкалы могут понижаться (количественные шкалы к порядковым или номинальным, порядковые к номинальным), обратное преобразование считается некорректным. Все алгоритмы классификации приводят описание объектов к номинальной или порядковой шкале.

Аналитические методы позволяют на основании выборочных наблюдений сделать статистически значимые выводы о наличии закономерностей для всей совокупности. Они обычно основываются на соответствующих вероятностных моделях, предполагающих нормальное (или другое известное) распределение совокупности единиц изучаемого признака (методы *параметрической статистики*). Основателем теории статистического вывода, аргументирующей воз-

можность дедуктивного изучения генеральной совокупности на основании выборочных оценок, полученных с точки зрения наибольшего правдоподобия («*происходит то, что должно было произойти*», то есть имеет наибольшую вероятность для дискретного случая или максимум функции плотности вероятности — для непрерывного) и имеющих определенные свойства (несмещенность, состоятельность, эффективность) был Рональд Фишер [20, 114].

Он же заложил основы математической теории активного эксперимента, позволяющего целенаправленно изучать сложные системы в отличие от пассивного эксперимента (сплошное или выборочное наблюдение).

Таблица II

Классификация шкал (продолжение)

Шкала		Статистики	Показатели вариации	Показатели связи
Качественная	номинальная	$w_i = \frac{n_i}{n}$ — относительные частоты	$D(w) = w_i(1 - w_i)$, $D(\overline{w}) = \sum_i \frac{w_i(1-w_i)n_i}{n}$	χ^2 — Пирсона, V — Крамера, $\varphi^* = (\varphi_1 - \varphi_2)\sqrt{\frac{n_1 n_2}{n_1+n_2}}$, $\varphi = 2arcsin\sqrt{w_i}$
	порядковая	$Q_1 = Q_{0,25}$ — нижняя квартиль: $P(X < Q_1) = 0{,}25$; $Q_2 = Q_{0,50}$ — медиана: $P(X < Q_2) = 0{,}50$; $Q_3 = Q_{0,75}$ — верхняя квартиль: $P(X < Q_3) = 0{,}75$; q_τ — квантиль порядка τ: $P(X < q_\tau) = \tau$, например: P_τ — процентиль при $\tau = 0{,}01; 0{,}02 \ldots; 0{,}99$; q_τ — дециль при $\tau = 0{,}1; 0{,}2 \ldots; 0,9$	$I_{0,50} = Q_{0,75} - Q_{0,25}$ интерквартильный размах; $I_{0,80} = q_{0,90} - q_{0,10}$ интердецильный размах	r_s — коэффициент корреляции Спирмена, r_K — коэффициент корреляции Кендалла
Количественные	интервальная	Начальные моменты: $M_s = \frac{\sum x_i^s n_i}{n}$; центральные моменты: $m_s = \frac{\sum (x_i - \overline{x})^s n_i}{n}$; основные моменты: $r_s = \frac{\sum (x_i - \bar{x})^s n_i}{n\sigma_x^s}$	$D(X)$ — дисперсия, σ_x — стандартное отклонение; $R = x_{max} - x_{min}$ — размах вариации; $V = \frac{\sigma(x)}{\bar{x}} 100\%$ — коэффициент вариации	r — коэффициент корреляции Пирсона $r = cos\,\varphi = \frac{(a,b)}{\lvert a\rvert\lvert b\rvert}$, где φ — угол между векторами наблюдений: $a(x_1, \ldots, x_n)$, $b(y_1, \ldots, y_n)$
	отношений			

Перечисленные ранее описательные и аналитические методы относятся к *параметрической статистике*, позволяющей решать задачи оценки параметров и проверки гипотез в рамках распределений известного вида (К. Пирсон, Р. Фишер, Е. Пирсон, Ю. Нейман, 1900–1933). Включение в регрессионную модель многомерных зависимых переменных приводит к *общей линейной модели*. Возможность учета нелинейных взаимодействий и расширение множества зависимых переменных от непрерывных (в том числе и нормально распределенных) до дискретных называется *обобщенными линейными.* Они включают линейные и логистические модели регрессии (Дж. Нельдер, Р. Веддербурн, 1970-е гг.).

Развитие компьютерной техники в 1980–1990-е годы послужило новому витку в обработке данных. Так, идеи компьютерного «обучения математических моделей» по данным (прецедентам), восходящие к работам по нейронным сетям МакКалока — Питтса (1943), персептрона Ф. Розенблатта (1957), Л. Дж. Фогеля (эволюционное программирование, 1960), А. Г. Ивахненко (МГУА — метод группового учета аргументов, 1969), сегодня нашли себя в ряде областей науки, имеющих разные названия, но выдвигающих очень сходные идеи: *машинное обучение, статистическое обучение, распознавание образов* и т. д. Так, развитие компьютерной техники в 1980-е годы послужило развитию *нелинейных методов решения задач классификации и регрессии,* опирающиеся на введение деревьев регрессии и классификации (Л. Брейман и др.). В 1986 г. были введены *обобщенные аддитивные модели* (Т. Хасти, Р. Тибшириани). Получила толчок Байесовская статистика, опирающаяся на идею размножения выборочных данных и т. д.

Другим направлением являются методы *непараметрической статистики*, которые не опираются на нормальное распределение (или любое другое) и не используют его свойства. Различают несколько этапов ее развития.

1) Начало этому направлению послужила работа А. Н. Колмогорова (1933), посвященная изучению отклонения эмпирической функции распределения от теоретической $F(x)$, начатому В. И. Гливенко (1933). Оказалось, что результаты не зависят от вида функции распределения (статистика Колмогорова). Работа в этом направлении была продолжена Н. В. Смирновым (статистика Смирнова, 1939).

2) *Ранговые критерии* (Вилкоксон, 1945).

3) Использование ранговых критериев для оценки параметров (Ходжес, Леман, 1963).

4) *Робастные методы*, предполагающие малую чувствительность к отклонениям, неоднородностям и загрязненности выборки, простейший метод — статистическая группировка (П. Хьюбер, 1964).

5) *Ядерные оценки* для оценки плотности распределения, решения задач оптимизации, идентификации, регрессии, распознавания и т. д. (Розенблат, 1956; Парзен, 1962).

6) Одно из значимых приложений — *обработка экспертных данных* (например, с использованием медианы Кемени). Известны вероятностные, логические, геометрические и другие подходы. Интенсивное развитие шло параллельно с теорией исследования операций в годы Второй мировой войны и послевоенные годы, хотя отдельные методы типа мозгового штурма были разработаны ранее. В СССР в 1960–1970-е годы шло освоение этого направления, с 1973 г.

работает семинар «Экспертные оценки и анализ данных»[15], которым в настоящее время руководят Ф. Т. Алескеров и Д. А. Новиков (ИПУ РАН).

Последние десятилетия интенсивно развивается *нечисловая статистика*, характерные черты которой — выборка из элементов произвольных пространств, использование показателей различия и расстояния (А. И. Орлов, с 1979).

Основная цель математической статистики — это получение и обработка данных для статистически значимой поддержки процесса принятия решения при решении задач планирования, управления, прогнозирования, включающая:

1) упорядочение исходной совокупности единиц наблюдения по соответствующим признакам и правилам;

2) нахождение числовых характеристик (параметров) предполагаемых законов распределения случайных величин по известной совокупности единиц;

3) проверку статистических гипотез о виде неизвестного распределения или о параметрах известных распределений;

4) вероятностную оценку статистической (стохастической) зависимости между различными объектами или (и) их свойствами.

Математическая статистика позволяет обосновать статистические выводы и связи, опираясь на данные [46, 47].

В математической статистике предполагается вероятностная природа данных, то есть неявно (а зачастую и явно, опираясь на центральную предельную теорему) полагается, что наблюдения $X_1, X_2, X_3, \dots, X_n$ генерируются некоторым случайным процессом с функцией распределения $F(x)$ и, таким образом, являются выборкой объема n из генеральной совокупности с распределением $F(x)$.

В большинстве практических ситуаций вид теоретической функции распределения предполагается заранее, например опираясь на теоретические предпосылки или проверку соответствующей гипотезы. Итак, если речь идет о том, что изучаемая функция распределения принадлежит к определенному классу функций и зависит от одного или нескольких параметров $F(x, \theta_1, \theta_2, \dots, \theta_k)$, то определение функции сводится к оценке выборочных параметров. Эти условия решения поставленных выше задач приводят к понятию *параметрической статистики.*

Если при решении не используется предположение о распределении известного вида, то говорят о *непараметрической статистике*, в которой наиболее развиты:

– задачи проверки гипотез по наблюдениям независимых и одинаково распределенных величин (*согласованности распределений* — подчиненности одному закону распределения; *независимости* признаков одного объекта по наблюдениям над совокупностью аналогичных объектов; *случайности*);

– *ранговые критерии* (табл. III);

– задачи оценки неизвестных распределений и параметров, неизвестная функция распределения может определяться как эмпирическая $F_n(x)$;

– *робастные методы* оценки параметров распределений и моделей.

[15] URL: https://www.ipu.ru/conference/workshops/local/expert%20assessments%20and%20data%20analysis.

Таблица III

Критерии статистического вывода о сравнении выборок

Критерий	Две выборки ($k = 2$)		Несколько выборок ($k > 2$)	
	независимые	зависимые	независимые	зависимые
Ранговый	Непараметрические			
	U — Манна — Уитни, критерий серий	T — Вилкоксона, критерий знаков	H — Краскалла – Уоллеса (ранговый дисперсионный анализ)	χ^2 — Фридмана (ранговый дисперсионный анализ)
Количественный	Параметрические			
	t –распределение Стьюдента (независимые выборки)	t –распределение Стьюдента (зависимые выборки)	F — Фишера — Снедекора (дисперсионный анализ без повторных измерений)	F — Фишера — Снедекора (дисперсионный анализ с повторными измерениями)

Асимптотические методы. Пусть имеется последовательность независимых одинаково распределенных случайных величин $\{X_i\}$, тогда свойства ее выборочной реализации $\{x_i\}$ зависят от распределения генеральной совокупности. Однако при увеличении n ($n \to \infty$) характер поведения числовых характеристик $\{x_i\}$ практически не зависит от вида распределения. Поэтому отдельно изучают числовые характеристики при фиксированном n и соответствующем законе распределения и при $n \to \infty$ (асимптотические свойства).

Последовательность $\{X_n\}$ называется асимптотически нормальной, если существуют такие последовательности $\{a_n\}$ и $\{b_n\}$ ($b_n > 0$), что

$$\frac{X_n - a_n}{\sqrt{b_n}} \xrightarrow{d} N(0,1), \qquad n \to \infty. \tag{II. 2}$$

Утверждение, что последовательность $\{X_n\}$ асимптотически нормальна, записывается как

$$X_n \to N(a_n, b_n), \qquad n \to \infty. \tag{II. 3}$$

Теорему Линденберга — Леви (см. раздел 8.3) с учетом формулы (II. 3) можно записать как

$$\sum_{i=1}^{n} X_i \to N(na, n\sigma^2), \; n \to \infty. \tag{II. 4}$$

Из указанной теоремы следует, что при $n \to \infty$ выборочные числовые характеристики генерального распределения, имеющего конечные начальные моменты первого и второго порядка (т. е. математическое ожидание и дисперсию), асимптотически нормальны:

$$\overline{X} \to N\left(a, \frac{\sigma^2}{n}\right), \tag{II. 5}$$

$$\hat{\sigma}^2, s^2 \to N\left(\sigma^2, \frac{\mu_4 - \sigma^2}{n}\right), \tag{II. 6}$$

$$F_n(x) \to N\left(F(x), \frac{F(x)(1-F(x))}{n}\right). \tag{II. 7}$$

При наличии выбросов (артефактов) предположения о нормальном (и других законах) нарушаются и предлагается применять *робастные*[16] *методы оценивания*, позволяющие значительно снизить вредное влияние больших выбросов на

[16] Robust (*англ.*) — устойчивый.

оценку и получить приемлемую итоговую оценку искомых параметров. Альтернативным вариантом являются процедуры классификации (простой и многомерной группировки), очистки (заполнения пропусков в данных, устранение дубликатов и противоречий, снижения размерности признакового пространства и удаление незначащих факторов), трансформации данных (нормализации и квантования, понижения разнообразия уникальных значений), направленные на формирование однородных совокупностей, а также отдельное изучение причин «выбросов» и «артефактов». Собственно, с этого момента и начинается понимание, что даже небольшие отклонения в данных ведут к тому, что параметрические методы работают неудовлетворительно, а непараметрические зачастую недостаточны для решения практических задач. Появляется идея использования бутстреп-метода — метода размножения выборок, позволяющего увеличить точность измеряемых параметров, уменьшить дисперсию оценок, так как если $D(x) = \sigma^2$, то $D(\overline{x}) = \sigma^2/n$. *В основе бутстреп-метода лежит «метод складного ножа», или джекнайф — оценка, с которой начинается идея разделения выборки на обучающую и проверочную последовательности, что является лейтмотивом современных теорий статистического (машинного) обучения.*

Это приводит к идее *анализа данных* (*прикладной статистики*), а точнее *разведочного анализа данных*, провозглашенного Дж. Тьюки в 1960-е годы, когда наряду с вероятностной природой данных может рассматриваться геометрическая и логическая, а далее когнитивная (то есть опирающаяся на знания), которая реализуется сегодня во всевозможных статистических пакетах и других информационных системах обработки данных, использующих идеологию разведочного анализа данных в информационных технологиях класса *KDD* и *Data Mining*.

В настоящее время применимость методов математической статистики опирается еще на одну идею классификации методов обработки данных, связанную с объемом выборочной совокупности n [92]:

– $0 \le n \le 8$ — экспертные оценки;

– $9 \le n < 30$ — малая выборка — точные формулы выборочного метода;

– $30 \le n \le 100$ — умеренные выборки, сложный случай между «малой выборкой» и «асимптотическим подходом»;

– $n \ge 100 * k,\ (k \in \mathbb{N})$ — асимптотические теоремы теории вероятностей;

– при небольшом числе переменных (до 10–30), относительно небольшой и однородной совокупности данных применяются методы прикладной статистики (анализа данных);

– при большом числе переменных (свыше 30–50), большой, неоднородной совокупности данных применяются *технологии обработки данных KDD* и *Data Mining,* реализующие процесс получения данных, очистки и предобработки, построения моделей для поддержки процесса принятия решений;

– развитие технологий *Big Data* (2008), связанных как с объемом (от 500 млн — 1 млрд записей), так и с появлением потоковых данных, их разнородностью привело к необходимости использования глубокого обучения (*Deep Learning — DL*, 1986), иллюстрируемого работой многослойных нейронных сетей, позволяющих получить приемлемые результаты уже от 5 млн прецедентов (см. раздел 21.3 в [53]).

Глава 11
Вариационные ряды распределения

11.1. Построение и графическое изображение вариационных рядов

Функционирование систем в окружающем мире отображается в виде наблюдаемых и ненаблюдаемых переменных (показателей), отражающих закономерности массовых случайных явлений, которые проявляются в статистических совокупностях. Например, в сельском хозяйстве говорят о совокупности коров определенной породы, совокупностях фермерских хозяйств и т. д. Обработка данных проводится независимо от их природы как числовых значений, отражающих количественные и качественные характеристики реальных совокупностей. Следует отметить, что обрабатываемые данные представляют собой модель, отображающую свойства совокупности и формирующуюся эмпирически. Поэтому даже строго детерминированные наблюдения удобно представлять как реализацию случайных величин $X_1, X_2, X_3, \ldots, X_n$, подчиняющихся одному закону распределения $F(x)$. Реально, конечно, все наблюдения формируются как под воздействием причинно-следственных связей, так и множества случайных факторов, что позволяет использовать теорию вероятностей в качестве обоснования выводов математической статистики и во многих случаях оказывается эффективно при решении задач управления и принятия решений. В экономике особую роль играет тема «Вариационные ряды», в которой предположение о том, что изучается случайная реализация некоторой гипотетической «генеральной совокупности» — выборочная совокупность, отодвигается на второй план, на первый выдвигается эмпирическая функция распределения, ее числовые характеристики и пр., что, по-видимому, объясняется традиционно «большими» объемами изучаемых совокупностей. В реальных социально-экономических системах нельзя проводить активные эксперименты, поэтому данные обычно представляют собой наблюдения за происходящим процессом по одному или нескольким признакам, например: курс валюты на бирже в течение месяца, урожайность сельскохозяйственных культур в хозяйствах, производительность труда рабочих за смену, число детей в семье, объем производства и продаж продукции организаций и т. п. Результаты наблюдений — это, в общем случае, ряд чисел по отдельным единицам совокупности, расположенных в беспорядке, который для изучения и наглядного представления необходимо упорядочить.

Операция, заключающаяся в расположении значений признака по возрастанию или убыванию, называется *ранжированием* данных. После операции ранжирования статистические данные можно сгруппировать так, чтобы в каждой группе признак принимал одно и то же значение, которое называется *вариантом* (x_i). Число единиц в каждой группе называется *частотой* варианта (n_i).

Размахом вариации называется число

$$R = x_{max} - x_{min}, \tag{11.1}$$

где x_{max} — наибольшее значение признака; x_{min} — наименьшее значение признака по данной совокупности единиц.

Сумма всех частот (n_i) называется объемом совокупности (n):

$$\sum_{i=1}^{k} n_i = n_1 + n_2 + \cdots + n_k = n.$$

Отношение частоты данного варианта к объему совокупности называется *относительной частотой* (w_i) или *частостью* этого варианта:

$$w_i = \frac{n_i}{n}, \tag{11.2}$$

$$\sum_{i=1}^{k} w_i = \sum_{i=1}^{k} \frac{n_i}{n} = \frac{\sum_{i=1}^{k} n_i}{n} = 1. \tag{11.3}$$

Упорядоченная последовательность вариантов изучаемого признака и соответствующих им частот (или частостей) называется *статистическим рядом.* Различают *атрибутивные* (построенные по качественным признакам) и *вариационные ряды* (построенные по количественным признакам). Вариационные ряды строятся по дискретным и непрерывным признакам. Дискретный — это признак, принимающий отдельные целочисленные значения, например разряд рабочего, оценка на экзамене, число комнат в квартире, поголовье скота и т. п. Непрерывным называется признак, значения которого заполняют какой-то промежуток, например месячная зарплата работника, себестоимость единицы продукции, цена реализации, вес животного, затраты на производство и т. п.

Дискретным вариационным рядом называется ранжированная последовательность вариант с соответствующими частотами и (или) частостями. Ряд представляется в виде таблицы.

Значение признака (x_i)	x_1	x_2	$\cdots$	x_i	$\cdots$	x_k
Частота (n_i)	n_1	n_2	$\cdots$	n_i	$\cdots$	n_k
Частость (w_i)	w_1	w_2	$\cdots$	w_i	$\cdots$	w_k

Пример 11.1. В результате тестирования группа из 24 человек набрала баллы: 4, 0, 3, 4, 1, 0, 3, 1, 0, 4, 0, 0, 3, 1, 0, 1, 1, 3, 2, 3, 1, 2, 1, 2.

Построить дискретный вариационный ряд.

Решение. Проранжируем исходный ряд, подсчитаем частоту и частость вариант: 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4.

В результате получим дискретный вариационный ряд (табл. 11.1).

Если признак непрерывный или число значений дискретного признака велико, то в этом случае строится *интервальный вариационный ряд.* Для построения такого ряда весь промежуток изменения признака разбивается на ряд отдельных интервалов и подсчитывается количество значений в каждом из них. Интервальный вариационный ряд может быть построен как с равными, так и не равными по длине интервалами (обычно прогрессивно возрастающими или убывающими).

Таблица 11.1

Распределение студентов по баллам

Балл, x_i	Число студентов, n_i	Относительная частота, w_i
0	6	$6/24 = 0{,}250$
1	7	$7/24 = 0{,}292$
2	3	$3/24 = 0{,}125$
3	5	$5/24 = 0{,}208$
4	3	$3/24 = 0{,}125$
$\sum$	24	1,000

Таким образом, с целью сжатого описания данных вариационный ряд группируют и представляют в виде дискретного или интервального вариационного ряда.

Будем считать, что отдельные (частичные) интервалы имеют одну и ту же длину. Число интервалов (k) в случае нормально распределенной или близкой к ней совокупности можно определить по формуле Стерджесса:

$$k = 1 + 3{,}322 \lg(n). \quad (11.4)$$

Длина частичного интервала определяется по формуле

$$h = \frac{R}{k} = \frac{x_{max} - x_{min}}{k}. \quad (11.5)$$

Если имеются аномальные, т. е. резко отклоняющиеся значения признака, и они не отбрасываются, то строится вариационный ряд с открытыми крайними интервалами, а в расчетах длины интервала эти аномальные значения не учитываются.

Пример 11.2. Имеются следующие данные по числу работников на 100 га сельскохозяйственных угодий ($n = 60$).

4,45	5,03	4,74	4,02	4,69	3,51	7,10	5,47	4,77	6,03
4,36	3,02	4,50	4,65	3,72	3,00	4,79	3,70	3,50	3,58
2,44	4,26	9,75	6,20	4,54	4,14	6,07	4,49	6,13	3,75
6,20	7,14	6,97	5,34	8,70	5,53	6,93	8,32	3,23	7,60
5,39	5,06	6,37	9,52	6,47	3,95	8,26	4,05	4,71	6,57
3,75	7,11	6,13	7,85	5,07	7,89	5,03	3,89	6,44	4,44

Необходимо построить интервальный вариационный ряд с равными интервалами, найти относительные частоты и накопленные частоты.

Решение. Для определения числа групп подставим значение $n = 60$ в формулу Стерджесса: $k = 1 + 3{,}322 lg 60 \approx 6{,}907; k = 7$.

Найдем длину частичного интервала: $h = \frac{9{,}75-2{,}44}{7} = \frac{7{,}31}{7} \approx 1{,}0$.

Построим интервальный вариационный ряд, для этого в качестве начального значения используем x_{min}. Разобьем интервал изменения признака X на $k = 7$ частичных интервалов с шагом $h = 1{,}0$ и подсчитаем число хозяйств в каждом частичном интервале (табл. 11.2). Так как длина интервала *h* была округлена до 1,0, то значение 9,75 оказалось вне интервала 8,40–9,40, поэтому последний интервал открытый.

Накопленная частота определяется путем последовательного суммирования частот вариационного ряда. В примере 11.2: 4=4; 4+14=18; 18+17=35 и т. д. Она показывает, сколько единиц совокупности имеет значения признака до верхней границы данного интервала.

В вариационных рядах с неравными интервалами определяется плотность распределения как отношение частот или частостей к величине соответствующего интервала.

Накопленная частость показывает, какая доля единиц совокупности не превышает данного значения.

Вариационные ряды изображают графически с помощью полигона, гистограммы, кумуляты и огивы. *Полигон частот* — это ломаная, отрезки которой соединяют точки $(x_1; n_1)$, $(x_2; n_2)$, …, $(x_k; n_k)$. *Полигон относительных частот* — это ломаная, отрезки которой соединяют точки: $(x_1; \frac{n_1}{n})$, $(x_2; \frac{n_2}{n})$,…, $\left(x_k; \frac{n_k}{n}\right)$. По данным примера 11.1 построим полигон частот и относительных частот (рис. 11.1, 11.2).

Таблица 11.2

Распределение хозяйств по численности работников на 100 га сельскохозяйственных угодий

Группы хозяйств по численности работников на 100 га сельскохозяйственных угодий, чел.	Число хозяйств в группе (n_i)	Накопленное число хозяйств (S_i)	Относительная частота (w_i)	Относительные накопленные частоты $\left(\frac{n_i}{n}\right)$
2,40–3,40	4	4	0,067	0,067
3,40–4,40	14	18	0,233	0,300
4,40–5,40	17	35	0,284	0,584
5,40–6,40	9	44	0,150	0,734
6,40–7,40	8	52	0,133	0,867
7,40–8,40	5	57	0,083	0,950
Свыше 8,40	3	60	0,050	1,000
Итого:	60	–	1,000	–

Интервальный вариационный ряд с равными интервалами графически изображается в виде *гистограммы* (Пирсона), причем открытые крайние интервалы предварительно закрываются. В примере 10.2: 8,4+1,0 = 9,4. *Гистограммой частот* называется фигура, состоящая из прямоугольников с основанием h и высотами n_i, ординаты соответствующей *Кумуляты распределения — накопленные частоты*. Для *гистограммы относительных частот* в качестве высоты рассматривают n_i/n. Гистограмма относительных частот — графический аналог функции плотности вероятности, с ростом объема совокупности ее площадь для произвольного интервала стремится к площади под функцией плотности вероятности над этим же интервалом. Огива строится в прямоугольной системе координат. По оси абсцисс откладываются номера единиц наблюдения признака, а по оси абсцисс — значения признака в порядке возрастания значений.

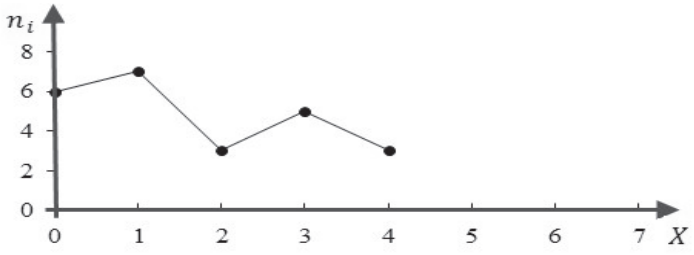


Рис. 11.1 — Полигон распределения студентов по баллам

Полигон относительных частот будет иметь следующий вид (рис. 11.3).

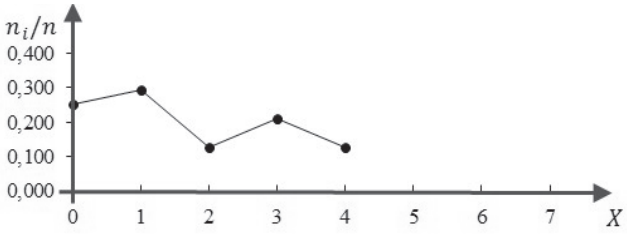


Рис. 11.2 — Полигон относительных частот

Замечание. «Я до сих пор живо помню, когда я был еще ребенком, мой отец привел меня на край города, где на берегу стояли ивы, и велел мне сорвать наугад сотню ивовых листочков. После отбора листьев с поврежденными кончиками у нас осталось 89 целых листиков.

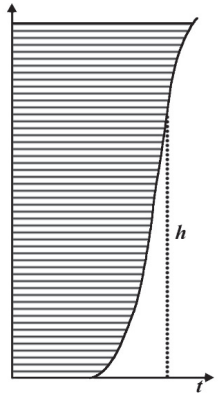


Рис. 11.3

Вернувшись домой, мы расположили их в ряд по росту, как солдат. Затем мой отец через кончики листьев провел кривую и сказал: “Это и есть кривая Кетле (*огива*). Глядя на нее, ты видишь, что посредственности всегда составляют подавляющее большинство и лишь немногие поднимаются выше или так и остаются внизу”. Если эту кривую расположить вертикально (рис. 11.3) и в качестве единицы масштаба на оси ординат выбрать отрезок, длина которого равна высоте всей фигуры, то ордината h, соответствующая абсциссе t, будет, очевидно, представлять собой частоту (или долю) тех ивовых листьев, длина которых меньше t. И так как частота h приближенно равна вероятности p, то наша кривая приближенно представляет $p = F(t)$ — функцию распределения длины листьев» (Б. Л. ван дер Варден, Математическая статистика, [14]). ■

Построим гистограмму распределения частот и кумуляту по данным примера 11.2 (рис. 11.4, 11.5).

График гистограммы относительных частот можно получить из графика (рис. 11.4) сжатием в 60 раз вдоль оси ординат.

Наряду с гистограммой и полигоном относительных частот, для оценки $f(x)$ — теоретической плотности распределения по эмпирическим данным, в современных исследованиях рассматривают ядерные оценки плотности, например, на рисунке 11.6, выполненном в эконометрическом пакете $gretl$ по данным примера 11.2. Суть метода в равномерном разбиении диапазона данных на ряд контрольных точек и вычисления оценочной плотности в каждой по формуле

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right),$$

где n — количество точек данных; h — параметр пропускной способности («ширина окна»); $K(t)$ — функция ядра (непрерывной ограниченной четной функции с интегралом, равным единице), чаще всего рассматривают ядро Гаусса, прямоугольное или Епанечникова; $f(x)$ называют оценкой Розенблатта — Парзена [69].

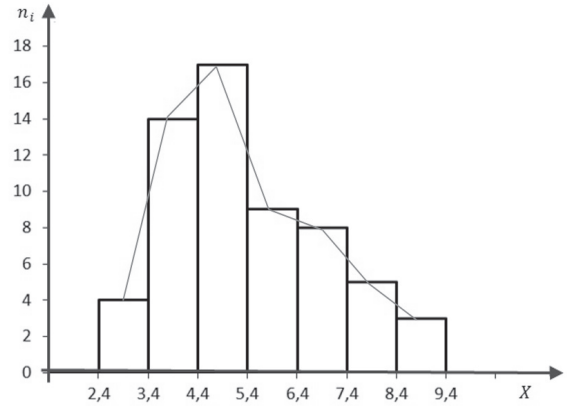


Рис. 11.4 — Распределение хозяйств по численности работников на 100 га сельскохозяйственных угодий, чел.

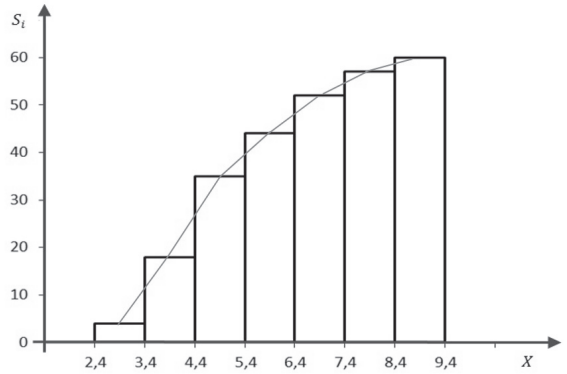


Рис. 11.5 — Кумулята распределения хозяйств по численности работников на 100 га сельхозугодий

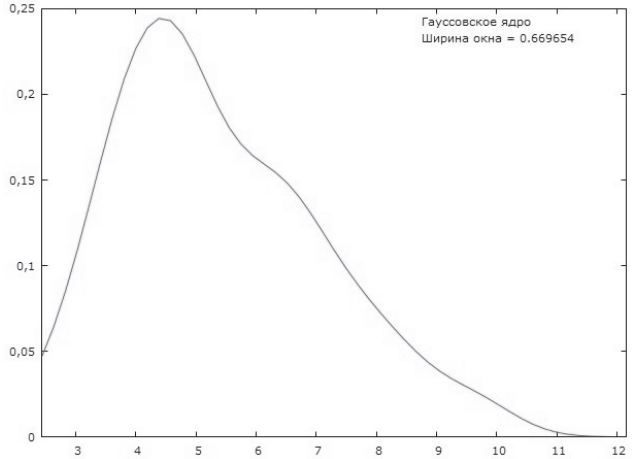


Рис. 11.6 — Ядерная оценка плотности хозяйств по численности работников на 100 га сельскохозяйственных угодий, чел.

11.2. Меры центральной тенденции

Вариационный ряд содержит достаточно полную информацию о величине и колеблемости (вариации) значений изучаемого признака и позволяет получить первое представление о фактически наблюдаемом распределении. В большинстве практических задач требуется определить числовые характеристики вариационного ряда (аналогичные характеристикам распределения в теории вероятностей): положения (средняя арифметическая, мода, медиана); рассеяния (дисперсия, среднее квадратическое отклонение, коэффициент вариации); скошенности (коэффициент асимметрии) и островершинности (эксцесс) распределения.

Характеристики положения вариационного ряда.

Средней арифметической ($\overline{X}$) дискретного вариационного ряда называется отношение суммы произведений вариант на соответствующие частоты к сумме частот:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} = \frac{\sum x_i n_i}{n}. \qquad (11.6)$$

В интервальном вариационном ряду x_i — среднее значение *i*-го интервала, как полусумма его границ.

Если имеются только значения признака по отдельным единицам совокупности, то средняя арифметическая равна отношению суммы значений признака на объем совокупности:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}. \qquad (11.7)$$

Средняя арифметическая имеет те же единицы измерения, что и варианты признака. Она отражает центральную тенденцию распределения.

Свойства средней арифметической.

1. Средняя арифметическая постоянной величины равна самой постоянной:

$$\overline{C} = C.$$

2. Если все варианты признака умножить или разделить на одно и то же число *h*, то средняя арифметическая соответственно изменится в *h* раз:

$$\overline{hx} = \frac{\sum_{i=1}^{k} (hx_i) n_i}{n} = \frac{h\sum_{i=1}^{k} x_i n_i}{n} = h\bar{x}. \qquad (11.8)$$

3. Сумма отклонений вариант признака от их средней арифметической, взвешенная соответствующими частотами, равна нулю:

$$\sum (x_i - \bar{x}) n_i = \sum (x_i n_i - \bar{x} n_i) = \sum x_i n_i - \bar{x} \sum n_i = \\ = \sum x_i n_i - \sum x_i n_i = 0.$$

4. Если все варианты признака увеличить (уменьшить) на одно и то же число, то средняя арифметическая увеличится (уменьшится) на то же число, т. е.:

$$\overline{x \pm c} = \frac{\sum (x_i \pm c) n_i}{n} = \frac{\sum x_i n_i \pm c \sum n_i}{n} = \bar{x} \pm c. \qquad (11.9)$$

5. Если все частоты вариант признака умножить на одно и то же число, то средняя арифметическая не изменится.

$$\frac{\sum x_i n_i h}{\sum n_i h} = \frac{h \sum x_i n_i}{h \sum n_i} = \frac{\sum x_i n_i}{\sum n_i} = \overline{x}. \qquad (11.10)$$

6. Если вариационный ряд разбит на несколько непересекающихся групп, то общая средняя равна средней арифметической из групповых средних, взвешенных по объемам групп.

$$\overline{x} = \frac{\sum_{j=1}^{l} \overline{x}_j n_j}{\sum_{j=1}^{l} n_j}, \overline{x}_j = \frac{\sum_{i=1}^{k} x_{ij} n_{ij}}{n_j}, \quad (11.11)$$

где $\overline{x}_j$ — групповые средние, n_j — численность $j-$ой группы, x_{ij} — значение i-го варианта признака в $j-$ой группе, n_{ij} — частота i-го варианта признака в $j-$ой группе,

$$j = 1, 2, \dots, l; i = 1, 2, \dots, k.$$

Модой (M_o) дискретного вариационного ряда называется значение признака, имеющее наибольшую частоту.

Медианой (Me) дискретного вариационного ряда называется значение признака, которое делит вариационный ряд на две равные части.

Если дискретный вариационный ряд имеет $2n$ членов в ранжированной совокупности: $x_1, x_2, \dots, x_n, x_{n+1}, \dots, x_{2n}$, то

$$Me = \frac{x_n + x_{n+1}}{2}. \quad (11.12)$$

Если дискретный вариационный ряд в ранжированной совокупности имеет $2n+1$ членов: $x_1, x_2, \dots, x_n, x_{n+1}, \dots, x_{2n+1}$, то

$$Me = x_{n+1}. \quad (11.13)$$

В примере 11.1:

$$\bar{X} = \frac{0 \cdot 6 + 1 \cdot 7 + 2 \cdot 3 + 3 \cdot 5 + 4 \cdot 3}{24} = \frac{40}{24} = 1{,}67.$$

$$Mo = 1;$$

$$Me = \frac{1+1}{2} = 1.$$

Значит, средний балл теста составил 1,67, наиболее часто студенты набирали один балл из четырех, половина студентов набрала до одного балла включительно, а половина — один балл и выше.

Для интервальных вариационных рядов имеют место формулы:

а) медианы:

$$Me = x_{\text{Me}} + h \frac{0{,}5n - S_{Me-1}}{n_{Me}}, \quad (11.14)$$

где x_{Me} — нижняя граница медианного интервала; h — длина медианного интервала; n — объем совокупности; S_{Me-1} — накопленная частота интервала, предшествующего медианному интервалу; n_{Me} — частота медианного интервала;

б) моды:

$$Mo = x_{Mo} + h \frac{(n_{Mo} - n_{Mo-1})}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})}, \quad (11.15)$$

где x_{Mo} — нижняя граница модального интервала; h — длина модального интервала; n_{Mo} — частота модального интервала; n_{Mo-1} — частота интервала перед модальным интервалом; n_{Mo+1} — частота интервала после модального интервала.

По данным примера 11.2:

$$Me = 4{,}40 + 1{,}0 \frac{0{,}5 \cdot 60 - 18}{17} = 5{,}10;$$

$$Mo = 4{,}40 + 1{,}0 \frac{(17-14)}{(17-14)+(17-9)} = 4{,}67.$$

Половина предприятий имеет численность работающих на 100 га сельскохозяйственных угодий до 5,1 чел., а половина — больше 5,1. Наиболее часто встречаются хозяйства с численностью работающих 4,7 человека на 100 га сельскохозяйственных угодий.

Мода и медиана используются в качестве характеристики среднего положения в случае, если границы ряда нечеткие или если ряд не симметричен.

Замечание. Известно, что множество результатов наблюдений, расположенных на числовой прямой, имеет свойство группироваться относительно некоторого центра. Этот центр часто рассматривают как «свертку» всех наблюдений, которая используется в анализе. Центр может быть описан различными статистиками или, иначе, мерами центральной тенденции: модой, медианой, средней арифметической. При выборе меры центральной тенденции необходимо иметь в виду следующее.

1. В малой группе мода не стабильна — при небольших изменениях она может сильно измениться: $M_0(1,1,1, 3, 5, 7, 7, 8) = 1$, $M_0(0, 1,2, 3, 5, 7, 7, 8) =7$.

2. На медиану не влияет величина «больших» и «малых» значений, например, если последнее значение утроить, то медиана не изменится.

3. На величину среднего влияет каждое значение. Если одно из значений изменится на C единиц, то $\bar{X}$ изменится в том же направлении на C/n единиц.

4. Медиана — лучшая характеристика центральной тенденции, когда гистограмма исходных данных унимодальна (имеет одну моду). Поэтому, например, для адекватного описания дохода или заработной платы часто используют медиану.

5. Некоторые множества не имеют центральной тенденции, это имеют в виду, когда говорят о «средней температуре по больнице», «средней высоте дома на улице, где расположены одноэтажные и пятиэтажные дома» и т. д.

Дж. Гласс, Дж. Стенли в книге «Статистические методы в педагогике и психологии» [26] приводят следующий анекдот, обобщающий множество проблем, возникающих в процессе применения разных мер центральной тенденции. «Однажды пятеро мужчин сидели рядом на скамейке парка. Двое были бродягами, имущество которых выражалось в 25 центах. Третий был рабочим, чей счет в банке и имущество составляли 2000 долларов. Четвертый владел 15 000 долларов в различных формах. Пятый же был мультимиллионером с чистым доходом 5 000 000 долларов. Поэтому модальный актив группы составил 25 центов. Эта цифра точно характеризует двоих, но является чрезвычайно некорректной для трех других. Медиана, составляющая 2000 долларов, несколько меняет дело для всех, кроме рабочего. Среднее, 1 003 400,10 долларов, не является вполне удовлетворительным даже для мультимиллионера. Если мы должны выбрать одну меру тенденции, возможно, это была бы мода, которая точно описывает 40 процентов группы. Однако, если сказать, что “модальный актив пяти лиц, сидящих на скамье парка, равен 25 центам”, то нам пришлось бы сделать вывод о том, что общий актив группы приблизительно составляет 1,25 доллара, что меньше фактического более чем на пять миллионов долларов. Очевидно, нет меры, адекватной этим “странным соседям по скамейке”, которые просто не имеют “центральной тенденции”».

Отсутствием меры центральной тенденции часто иллюстрируется ограниченность методов математической статистики и предлагается расширить ее возможности в анализе данных. ■

11.3. Показатели вариации

Показатели центральной тенденции (M_0, M_e, $\overline{X}$) не исчерпывают всех свойств распределения. В одних случаях значения признака концентрируются тесно около среднего значения, а в других наблюдается значительное рассеяние. Под воздействием множества причин и факторов изучаемый признак принимает различные значения по отдельным единицам совокупности. Вариация характеризуется и измеряется системой показателей.

Для изучения степени изменчивости признака используются следующие показатели вариации:

1) размах вариации $R = x_{max} - x_{min}$;

2) среднее линейное отклонение;

3) дисперсия и среднее квадратическое отклонение.

Из приведенных показателей наиболее часто применяется среднее квадратическое отклонение и дисперсия, вследствие их математических свойств и возможности разложения дисперсии на составные части, а также четкой интерпретации результатов.

Среднее линейное отклонение есть средняя из абсолютных отклонений вариант признака от средней арифметической:

простое $$L = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n}, \qquad (11.16)$$

взвешенное $$L = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|n_i}{n}. \qquad (11.17)$$

Дисперсия дискретного ряда распределения представляет среднюю из квадратов отклонений вариант признака от средней арифметической:

$$\sigma^2 = \frac{\sum(x_i - \overline{x})^2 n_i}{n}, \qquad (11.18)$$

или

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

Среднее квадратическое отклонение дискретного ряда распределения есть корень квадратный из дисперсии:

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2 n_i}{n}}, \qquad (11.19)$$

выражается в тех же единицах, что и признак, по которому оно рассчитывается.

Коэффициент вариации:

$$V = \frac{L}{\bar{x}} \cdot 100\% \quad \text{или} \quad V = \frac{\sigma}{\bar{x}} \cdot 100\%. \qquad (11.20)$$

Коэффициент вариации выражается в процентах, характеризует относительную вариацию признака от среднего значения и обычно служит для сравнения колеблемости разных признаков по одной совокупности или одного признака по разным совокупностям. Коэффициент вариации не рассчитывается, если среднее значение признака близко к нулю, а среднее квадратическое отклонение значительно отличается от нуля.

Замечание. Пусть имеется дискретный вариационный ряд (x_i; n_i), где $\bar{x} = \frac{\sum x_i n_i}{n}, \sigma^2(x) = \frac{\sum (x_i - \bar{x})^2 n_i}{n}$, $i = \overline{1,k}$. Рассмотрим соответствующий ряд относительных величин (y_i; n_i), где $y_i = \frac{x_i}{\bar{x}}$. Средняя арифметическая нового ряда равна 1:

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

Дисперсия нового ряда:

$$\sigma^2(y) = \frac{\sum (y_i - \bar{y})^2 n_i}{n} = \frac{\sum \left(\frac{x_i}{\bar{x}} - 1\right)^2 n_i}{n} = \frac{\sum (x_i - \bar{x})^2 n_i}{(\bar{x})^2 n} = \frac{\sigma^2(x)}{(\bar{x})^2},$$

отсюда

$$\sigma(y) = \sigma\left(\frac{x}{\bar{x}}\right) = \frac{\sigma(x)}{\bar{x}} =: V. \blacksquare$$

Свойства дисперсии.

1. Дисперсия постоянной величины равна нулю:

$$\sigma^2(C) = \frac{\sum (C - \bar{C})^2}{n} = 0. \qquad (11.21)$$

2. Если все варианты признака увеличить (уменьшить) на одно и то же число *C*, то дисперсия и среднее квадратическое отклонение не изменятся, т. е.

$$D(X \pm C) = \frac{\sum \left((x_i \pm c) - \overline{(x \pm c)}\right)^2 n_i}{n} = \frac{\sum (x_i \pm c - \bar{x} \mp c)^2 n_i}{n} = \frac{\sum (x_i - \bar{x})^2 n_i}{n} = \sigma^2, \qquad (11.22)$$

$$\sigma(X \pm C) = \sigma(X). \qquad (11.23)$$

3. Если все варианты признака умножить на одно и то же число, то дисперсия изменится в квадрат этого числа, а среднее квадратическое отклонение изменится в это число раз (по модулю):

$$D(hX) = \frac{\sum (x_i h - \bar{x} h)^2 n_i}{n} = \frac{h^2 \sum (x_i - \bar{x})^2 n_i}{n} = h^2 \sigma^2, \qquad (11.24)$$

$$\sigma(hX) = |h| \sigma.$$

4. Если все частоты вариант умножить на одно и то же число, то дисперсия и среднее квадратическое отклонение не изменятся.

$$\frac{\sum (x_i - \bar{x})^2 n_i h}{nh} = \frac{\sum (x_i - \bar{x})^2 n_i}{n} = \sigma^2(X).$$

5. Свойство минимальности дисперсии.

$$S(a) = \frac{\sum (x_i - a)^2 n_i}{n} \to min \quad \text{при } a = \bar{x}. \qquad (11.25)$$

Минимум (11.25) достигается при условии, что $\frac{dS(a)}{da} = 0$, откуда

$$-2 \sum (x_i - a) n_i = 0, \ \sum (x_i - a) n_i = 0,$$

$$a = \frac{\sum x_i n_i}{\sum n_i} =: \bar{x}.$$

Следствие 1. Средний квадрат отклонений значений x_i от их средней арифметической равен среднему квадрату отклонений x_i от произвольной постоянной *a* минус квадрат разности между средней арифметической ($\bar{x}$) и этой произвольной постоянной.

$$\text{Пусть } \sigma_x^2 = \frac{\sum(x_i-\bar{x})^2 n_i}{n}, \quad \sigma_a^2 = \frac{\sum(x_i-a)^2 n_i}{n},$$

тогда

$$\sigma_x^2 = \sigma_a^2 - (\bar{x}-a)^2. \tag{11.26}$$

Рассмотрим тождество

$$\sum(x_i-a)^2 n_i = \sum(x_i-\bar{x}+\bar{x}-a)^2 n_i =$$
$$= \sum(x_i-\bar{x})^2 n_i + \sum(\bar{x}-a)^2 n_i + 2\sum(x_i-\bar{x})(\bar{x}-a)n_i.$$

Так как

$$\sum(\bar{x}-a)^2 n_i = (\bar{x}-a)^2 \sum n_i = (\bar{x}-a)^2 n,$$
$$\sum(x_i-\bar{x})n_i = 0,$$

имеем

$$\sum(x_i-a)^2 n_i = \sum(x_i-\bar{x})^2 n_i + (\bar{x}-a)^2 n,$$
$$\frac{\sum(x_i-a)^2 n_i}{n} = \frac{\sum(x_i-\bar{x})^2 n_i}{n} + (\bar{x}-a)^2$$

или

$$\sigma_a^2 = \sigma_x^2 + (\bar{x}-a)^2,$$

откуда следует формула (11.26).

Следствие 2. Дисперсия равна средней арифметической из квадратов значений признака минус квадрат средней арифметической:

$$\sigma^2 = \overline{x^2} - (\bar{x})^2, \tag{11.27}$$

где $$\overline{x^2} = \frac{\sum x_i^2 n_i}{n}.$$

Формула получается из следствия 1 свойства 5 при $a = 0$.

6. *Правило сложения дисперсий*. Если объединяются несколько распределений в одно, то общая дисперсия σ_0^2 нового распределения равна сумме средней арифметической из дисперсий объединяемых распределений с дисперсией частных средних относительно общей средней нового распределения.

Варианты (*i*)	Частное распределение (*j*)						
	1	2	⋯	*j*	⋯	*m*	Σ
x_1	n_{11}	n_{12}	⋯	n_{1j}	⋯	n_{1m}	n_1
x_2	n_{21}	n_{22}	⋯	n_{2j}	⋯	n_{2m}	n_2
x_3	n_{31}	n_{32}	⋯	n_{3j}	⋯	n_{3m}	n_3
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
x_i	n_{i1}	n_{i2}	⋯	n_{ij}	⋯	n_{1i}	n_i
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
x_k	n_{k1}	n_{k2}	⋯	n_{kj}	⋯	n_{km}	n_k
Σ	N_1	N_2	⋯	N_j	⋯	N_m	N

Или, иначе говоря, общая дисперсия равна сумме внутригрупповой и межгрупповой дисперсий:

$$\sigma_o^2 = \overline{\sigma^2} + \delta^2, \tag{11.28}$$

или

$$\sigma_o^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x}_o)^2 n_i}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_{ij} - \overline{x}_o)^2 n_{ij}}{N}, \tag{11.29}$$

где n_{ij} — частота i-го варианта j-го частного распределения (j = 1, 2, …, m; i = 1, 2, …, k);

x_{ij} — значение i-го варианта j-го частного распределения;

$N_j = \sum_{i=1}^{k} n_{ij}$ — объем (частота) j-го частного распределения;

$n_i = \sum_{j=1}^{m} n_{ij}$ — частота i-го варианта нового распределения;

$N = \sum_{i=1}^{k} n_i = \sum_{j=1}^{m} N_j = \sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}$ — объем нового распределения;

$\overline{x}_j = \frac{\sum_{i=1}^{k} x_i n_{ij}}{N_j}$ — средняя арифметическая j-го частного распределения;

$\overline{x}_o = \frac{\sum_{i=1}^{m} x_i n_i}{N}$ — средняя арифметическая нового распределения;

$\sigma_j^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x}_j)^2 n_{ij}}{N_j}$ — дисперсия j-го частного распределения;

$\overline{\sigma^2} = \frac{\sum_{j=1}^{m} \sigma_j^2 N_j}{N}$ — внутригрупповая дисперсия как средняя из групповых дисперсий;

$\delta^2 = \frac{\sum_{j=1}^{m}(\overline{x}_j - \overline{x}_o)^2 N_j}{N} = \frac{\sum_{j=1}^{m} \overline{x}_j{}^2 N_j}{N} - (\overline{x}_o)^2$ — межгрупповая дисперсия.

Рассмотрим вывод формулы (11.28). Запишем формулу для частной дисперсии j-ой группы наблюдений:

$$\sigma_j^2 = \frac{\sum_{i=1}^{k} x_i^2 n_{ij}}{N_j} - \left(\bar{x}_j\right)^2. \tag{11.30}$$

Перепишем формулу (11.30) в виде

$$\sigma_j^2 N_j = \sum_{i=1}^{k} x_i^2 n_{ij} - \left(\bar{x}_j\right)^2 N_j$$

или

$$\sum_{i=1}^{k} x_i^2 n_{ij} = \sigma_j^2 N_j + \left(\bar{x}_j\right)^2 N_j. \tag{11.31}$$

Просуммируем все уравнения (11.31) по индексу j:

$$\sum_{j=1}^{m}\sum_{i=1}^{k} x_i^2 n_{ij} = \sum_{j=1}^{m} \sigma_j^2 N_j + \sum_{j=1}^{m}\left(\bar{x}_j\right)^2 N_j, \tag{11.32}$$

где $\sum_{j=1}^{m}\sum_{i=1}^{k} x_i^2 n_{ij} = \sum_{i=1}^{k} x_i^2 n_i$.

Поделив (11.32) на общую численность совокупности и вычитая квадрат общей средней, получим

$$\left(\frac{\sum_{i=1}^{k} x_i^2 n_i}{N} - (\overline{x}_o)^2\right) = \frac{\sum_{j=1}^{m} \sigma_j^2 N_j}{N} + \left(\frac{\sum_{j=1}^{m}\left(\bar{x}_j\right)^2 N_j}{N} - (\overline{x}_o)^2\right), \tag{11.33}$$

или

$$\sigma_o^2 = \overline{\sigma^2} + \delta^2,$$

что и требовалось доказать.

Общая дисперсия характеризует совместное влияние всех факторов на изменение изучаемого признака.

Дисперсия частного распределения — это средний квадрат отклонений значений признака в группе от групповой средней. Характеризует вариацию признака, обусловленную действием на него всех прочих факторов, кроме признака, положенного в основу разбиения общей совокупности на части.

Межгрупповая дисперсия — средний квадрат отклонений групповых средних от общей средней по всей совокупности. Характеризует влияние признака, положенного в основу группировки на изменение изучаемого признака.

Формула (11.28) и ее обобщение при расчленении многомерной статистической совокупности по нескольким признакам одновременно, используется в схемах факторного и дисперсионного анализа для разложения общей вариации признака по ее источникам (например, для p признаков:

$$\sigma_o^2 = \delta_1^2 + \delta_2^2 + \cdots + \delta_p^2 + \overline{\sigma^2}).$$

11.4. Моменты вариационного ряда. Асимметрия и эксцесс

Средняя арифметическая и дисперсия являются частными случаями более общего понятия моментов вариационного ряда.

Моменты вариационных рядов в математической статистике находятся по формулам, аналогичным формулам (4.13)–(4.18).

Моментом *s*-го порядка называется средняя из *s*-х степеней отклонений вариант признака от некоторого числа *C*:

$$\mu_s = \overline{(x - C)^s} = \frac{\sum (x_i - C)^s n_i}{n}. \qquad (11.34)$$

Обычно рассматриваются моменты до четвертого порядка включительно, $s = 0, 1, 2, 3, 4$.

Если постоянная *C* равна нулю, то моменты называются начальными

$M_s = \dfrac{\sum x_i^s n_i}{n}$ — начальный момент $s-$ го порядка.

Начальный момент:

нулевого порядка $M_0 = \dfrac{\sum x_i^0 n_i}{n} = \dfrac{\sum n_i}{n} = 1$;

первого порядка $M_1 = \dfrac{\sum x_i^1 n_i}{n} = \overline{x}$ — средняя арифметическая;

второго порядка $M_2 = \dfrac{\sum x_i^2 n_i}{n} = \overline{x^2}$ — средняя квадратическая;

третьего порядка $M_3 = \dfrac{\sum x_i^3 n_i}{n}$;

четвертого порядка $M_4 = \dfrac{\sum x_i^4 n_i}{n}$.

Если постоянная *C* равна средней арифметической, то моменты называются центральными

$m_s = \overline{(x - \overline{x})^s} = \dfrac{\sum (x_i - \overline{x})^s n_i}{n}$ — центральный момент $s-$ го порядка.

Центральный момент:

нулевого порядка $m_0 = \frac{\sum(x_i-\overline{x})^0 n_i}{n} = \frac{\sum n_i}{n} = 1$;

первого порядка $m_1 = \frac{\sum(x_i-\overline{x})^1 n_i}{n} = 0$;

второго порядка $m_2 = \frac{\sum(x_i-\overline{x})^2 n_i}{n} = \sigma^2$ — дисперсия вариационного ряда;

третьего порядка $m_3 = \frac{\sum(x_i-\overline{x})^3 n_i}{n}$;

четвертого порядка $m_4 = \frac{\sum(x_i-\overline{x})^4 n_i}{n}$.

Центральные моменты выражаются через начальные моменты, используя формулу

$$m_s = M_s - C_s^1 M_{s-1} M_1 + C_s^2 M_{s-2} M_1^2 - C_s^3 M_{s-3} M_1^3 + \ldots \pm M_1^s. \quad (11.35)$$

$$m_2 = M_2 - M_1^2;\ m_3 = M_3 - 3M_1 M_2 + 2M_1^3;$$

$$m_4 = M_4 - 4M_1 M_3 + 6M_1^2 M_2 - 3M_1^4.$$

$r_s = \frac{\sum(x_i-\bar{x})^s n_i}{n\sigma_x^s}$ — основной центральный момент, s-го порядка,

$r_{s,h} = \frac{\sum(x_i-\bar{x})^s (y_i-\bar{y})^h n_i}{n\sigma_x^s \sigma_y^h}$ — основной момент порядка s, h.

Соотношения между начальными и центральными моментами в математической статистике соответствуют формулам (4.19–4.20).

Для характеристики симметричности изменения частот вариационного ряда применяется коэффициент асимметрии:

$$Ka = \frac{m_3}{\sigma^3} = \frac{\sum(x_i-\bar{x})^3 n_i}{n\sigma^3}. \quad (11.36)$$

Коэффициент асимметрии характеризует скошенность вариационного ряда, когда частоты значений признака, равноотстоящих от средней, отличаются друг от друга. Для симметричных вариационных рядов $Ka = 0$. При положительном значении коэффициента асимметрии преобладают варианты, большие средней арифметической, а при отрицательном значении — варианты, меньшие средней арифметической. Вариационный ряд имеет правостороннюю асимметрию, если $Ka > 0$, или левостороннюю асимметрию, если $Ka < 0$.

Мерой крутости распределения частот вариационного ряда служит эксцесс:

$$Ex = \frac{m_4}{\sigma^4} - 3 = \frac{\sum(x_i-\bar{x})^4 n_i}{n\sigma^4} - 3. \quad (11.37)$$

Для нормального распределения эксцесс равен нулю. Если значение $Ex \approx 0$, то распределение частот имеет нормальное распределение. Положительное значение эксцесса свидетельствует об островершинности распределения по сравнению с нормальным, а отрицательное значение — о плосковершинности распределения по сравнению с нормальным. Большое значение эксцесса может указывать на «тяжелые хвосты» распределения ($F(-x)$ и $1 - F(x)$ при $x \to \infty$) и выбросы.

Рассчитаем среднюю арифметическую, дисперсию, коэффициенты асимметрии и эксцесса по данным примера 11.2. Построим вспомогательную таблицу 11.3.

Таблица 11.3

Вспомогательная таблица для расчета
числовых характеристик ряда распределения

Группы хозяйств по численности работников на 100 га сельхозугодий, чел.	Среднее значение интервала (x_i)	Число хозяйств в группе (n_i)	$x_i n_i$	$x_i - \overline{x}$	$(x_i - \overline{x})^2 n_i$	$\frac{x_i - \overline{x}}{\sigma}$	$\left(\frac{x_i - \overline{x}}{\sigma}\right)^3 n_i$	$\left(\frac{x_i - \overline{x}}{\sigma}\right)^4 n_i$
2,40–3,40	2,9	4	11,6	–2,5	25,0	–1,582	–15,837	25,054
3,40–4,40	3,9	14	54,6	–1,5	31,5	–0,949	–11,965	11,355
4,40–5,40	4,9	17	83,3	–0,5	4,25	–0,316	–0,536	0,169
5,40–6,40	5,9	9	53,1	0,5	2,25	0,316	0,284	0,090
6,40–7,40	6,9	8	55,2	1,5	18,0	0,949	6,837	6,488
7,40–8,40	7,9	5	39,5	2,5	31,25	1,582	19,796	31,317
Свыше 8,40	8,9	3	26,7	3,5	36,75	2,215	14,719	32,602
Итого	–	60	324,0		149,0	2,215	13,298	107,075

Среднее значение признака составит

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{324{,}0}{60} = 5{,}4.$$

Дисперсия и среднее квадратическое отклонение:

$$\sigma^2 = \frac{\sum (x_i - \overline{x})^2 n_i}{n} = \frac{149{,}0}{60} = 2{,}483,$$

$$\sigma = \sqrt{\frac{\sum (x_i - \overline{x})^2 n_i}{n}} = \sqrt{\frac{149{,}0}{60}} = \sqrt{2{,}483} = 1{,}58.$$

Коэффициент вариации:

$$V = \frac{\sigma}{\bar{x}} \cdot 100\% = 29{,}2\%.$$

Таким образом, средняя численность работников на 100 га сельскохозяйственных угодий по исследуемой совокупности хозяйств составила 5,4 человека. Численность работников по хозяйствам в среднем колебалась в промежутке $\bar{x} \pm \sigma = 5{,}4 \pm 1{,}58$, то есть от 3,82 до 6,98 чел. на 100 га сельскохозяйственных угодий. Этот интервал, а также коэффициент вариации показывают, что имеются незначительные различия в обеспеченности хозяйств рабочей силой.

Коэффициент асимметрии:

$$Ka = \frac{\mu_3}{\sigma^3} = \frac{\sum (x_i - \bar{x})^3 n_i}{n\sigma^3} = \frac{13{,}298}{60} = 0{,}221.$$

Эксцесс:

$$E_x = \frac{\sum (x_i - \bar{x})^4 n_i}{n\sigma^4} - 3 = \frac{107{,}075}{60} - 3 = -1{,}215.$$

Найденное значение коэффициента асимметрии (недостаточно близкое к нулю) указывает, что распределение имеет небольшую правостороннюю асимметрию. Эксцесс значительно отличен от нуля, что говорит о плосковершинном распределении и возможном отличии распределения вариационного ряда от нормального распределения.

Темы (вопросы) для самоконтроля

1. Дискретный вариационный ряд и его графическое изображение (полигон частот, относительных частот), числовые характеристики (средняя арифметическая, мода, медиана, характеристики вариации).
2. Средняя арифметическая и ее свойства.
3. Дисперсия вариационного ряда и ее свойства.
4. Правило сложения дисперсий.
5. Интервальный вариационный ряд и его графическое изображение (гистограмма, кумулята) и числовые характеристики.
6. Ядерная оценка плотности вариационного ряда.
7. Моменты вариационного ряда.
8. Асимметрия и эксцесс вариационного ряда.

Глава 12
Выборочный метод

12.1. Понятие о выборочном методе

Для того чтобы исследовать какой-либо процесс или явление, необходимо собрать соответствующие данные по единицам совокупности. При сплошном наблюдении отбираются и обследуются все единицы изучаемой совокупности. Примером сплошного наблюдения служит перепись населения, статистическая отчетность организаций, учет рождаемости и смертности населения и т. п.

В реальных условиях обычно бывает трудно или экономически нецелесообразно, а иногда и невозможно, исследовать всю совокупность, характеризующую изучаемый процесс или явление *(генеральную совокупность)*. Формально генеральная совокупность представляет собой множество результатов всех мыслимых наблюдений, которые могли быть получены при данном комплексе условий. Её можно рассматривать как случайную величину, заданную на пространстве элементарных событий с заданным полем событий. Генеральная совокупность может быть как конечной, так и бесконечной.

Рис. 12.1 — Принципиальная схема выборочного исследования

Свойства (закон распределения и его параметры) генеральной совокупности неизвестны, поэтому возникает задача их оценки по части единиц совокупности. На практике широко применяется выборочное наблюдение, когда отбирается и исследуется часть генеральной совокупности. *Выборочная совокупность* — это совокупность единиц, отобранная по определенным правилам из генеральной совокупности и подлежащая исследованию.

Современная наука трактует выборочное обследование как необходимый этап в изучении сложных систем, характеризующихся большими объемами информации. Можно представить использование выборки при проведении исследований (для формирования статистического вывода) в виде итеративного процесса (рис. 12.1): генеральная совокупность — случайный отбор — выборка — свертка информации (представление данных несколькими числами) для получения выборочных оценок параметров генеральной совокупности (средней арифметической, дисперсии) или скрытых зависимостей в данных (уравнения регрессии) — выводы о генеральной совокупности (распространение выборочных оценок на генеральную совокупность) и их верификация (проверка соответствия данным генеральной совокупности) — при несоответствии новая итерация.

Замечание. 1. *Sample* — образец, *sampling metods* — выборочный метод (*англ.*). Сегодня в статистике рассматривают термин «выборочный метод», а в машинном обучении используют понятие «сэмплинг» (сэмплирование — осуществление выборки), имея в виду современное развитие выборочного метода в IT, которое может опираться как на имеющиеся данные, так и на известные законы распределения (см. гл. 7, 19).

2. По ряду причин в социально-экономических исследованиях итеративный подход ограничен (рис. 12.1). Однако в связи развитием IT следует отметить, что подобный итеративный процесс лежит в основе целого ряда успешных статистических процедур, применяемых в алгоритмах машинного обучения (*machine learning*), позволяющих получать множество моделей на одних выборках и проверять их на других (кросс-проверка), реализуя таким образом принцип многообразия математических моделей, для получения наиболее устойчивых результатов. В основе лежит так называемый метод бутстреп или бутстрап (*bootstrap*), основанный на многократном извлечении выборки (не обязательно из генеральной совокупности) и получения численных характеристик и математических моделей по каждой из них. Идея бутстреп-метода получила реализацию только после интенсивного развития IT, она позволяет формировать множество альтернативных моделей (*ансамбль моделей*) и увеличить точность прогноза. ■

Так как выборочная совокупность является частью генеральной совокупности, то при проведении выборочного обследования возникают ошибки, которые подразделяются на ошибки регистрации и ошибки репрезентативности.

Ошибки регистрации — это неточности между зафиксированными значениями признака в процессе проведения статистического наблюдения и действительными его значениями, например ошибки в установлении фактов, в расчете значений признаков по единицам совокупности, приписки и т. п. Ошибки регистрации возникают при сплошном и выборочном наблюдении.

Ошибки репрезентативности (*представительности*) — это разность между значениями характеристик генеральной и выборочной совокупности. Ошибки репрезентативности бывают систематическими и случайными. Систематические ошибки возникают вследствие нарушения правил отбора единиц из генеральной в выборочную совокупность, а случайные — вследствие изучения только части единиц генеральной совокупности, а не всех единиц.

Для получения хороших оценок характеристик генеральной совокупности необходимо, чтобы выборка была *репрезентативной* (представительной). Репрезентативность, в силу закона больших чисел, достигается случайностью отбора единиц из генеральной в выборочную совокупность. Ошибка репрезентативности $\varepsilon = \left|\theta - \tilde{\theta}\right|$ представляет разность между параметром генеральной совокупности θ и параметром выборочной совокупности $\tilde{\theta}$, который называется выборочной статистической оценкой параметра θ.

Задача выборочного метода заключается в том, чтобы на основе выборочной совокупности получить такие статистические оценки $\tilde{\theta}$, которые наиболее точно характеризовали значения параметров генеральной совокупности θ.

Различают следующие основные виды выборок:

1) *собственно-случайная,* когда отбор единиц из генеральной в выборочную совокупность производится способом жеребьевки или с использованием таблицы случайных чисел;

2) *типическая* — генеральная совокупность предварительно разбивается на группы по одному или нескольким типическим признакам, а затем отбор осуществляется из каждой выделенной группы обычно случайным или механическим способом. Различают:

а) *равномерные* выборки (при равенстве объемов исходных групп в генеральной совокупности выбирается одинаковое количество элементов из каждой);

б) *пропорциональные* (численность выборок формируют пропорционально численностям или средним квадратическим отклонениям групп генеральной совокупности);

в) *комбинированные* (численность выборок пропорциональна средним квадратическим отклонениям и численностям групп генеральной совокупности);

3) *механическая* — отбор единиц проводится через определенный интервал;

4) *серийная* — отбор проводится не по отдельным единицам, а сериями или группами единиц, которые обследуются сплошным способом;

5) *комбинированная* — используются различные комбинации вышеуказанных методов, например типическая выборка сочетается с механической и собственно случайной.

Различают два способа отбора единиц — повторный и бесповторный. При повторном отборе единицы после отбора возвращаются обратно в генеральную совокупность, т. е. каждая единица может быть отобрана более одного раза, а при бесповторном — выбранные единицы не возвращаются в генеральную совокупность. Обычно используется бесповторный способ отбора. Как уже отмечалось в части I, посвященной теории вероятностей, повторный отбор соответствует биномиальному закону распределения, а бесповторный отбор соответствует гипергеометрическому закону распределения.

12.2. Статистические оценки параметров генеральной совокупности

После осуществления выборки возникает задача оценки числовых характеристик генеральной совокупности по характеристикам выборочной совокупности.

Различают точечные и интервальные оценки.

Точечная оценка характеристики генеральной совокупности — это число, определяемое по выборке.

Пусть $\tilde{\theta} = \tilde{\theta}_n$ — выборочная статистическая оценка, вычисленная по результатам *n* наблюдений случайной величины *X*, используемая для оценки θ — характеристики генеральной совокупности (в качестве θ может быть *M*(*X*), *D*(*X*), *Mo* и др.).

Оценка параметров генеральной совокупности опирается на выборочные точечные оценки — *статистики*, являющиеся функциями выборочных данных:

$$\theta \approx \tilde{\theta}(X_1, X_2, \dots, X_n). \qquad (12.1)$$

В левой части приближенного равенства (12.1) находится параметр генеральной совокупности, а в правой — случайная величина, являющаяся функцией выборочных значений. Например: точечная оценка математического ожидания $M(X) = a$ определяется как выборочная средняя арифметическая:

$$\bar{x} = \frac{\sum x_i n_i}{n}, \qquad (12.2)$$

точечная оценка вероятности p_i определяется как относительная частота:

$$w_i = \frac{n_i}{n}. \qquad (12.3)$$

Качество оценки $\tilde{\theta}$ устанавливается по трем свойствам: несмещенность, эффективность, состоятельность.

1. *Несмещенность.* Оценка $\tilde{\theta}_n$ генеральной характеристики θ называется несмещенной, если математическое ожидание оценки $\tilde{\theta}_n$ равно оцениваемому параметру генеральной совокупности θ, при любом объеме выборки. То есть $M(\tilde{\theta}_n) = \theta$. Если $M(\tilde{\theta}_n) \neq \theta$, то такая оценка называется смещенной. Смещение — $Bias$[17] $=: M(\tilde{\theta}_n) - \theta$.

Оценка $\tilde{\theta}_n$ называется *асимптотически несмещенной*, если свойство $M(\tilde{\theta}_n) \to \theta$ выполняется при $n \to \infty$.

2. *Эффективность*. Несмещенная оценка $\tilde{\theta}_n$ генеральной характеристики θ называется эффективной, если среди всех подобных оценок той же характеристики она имеет наименьшую дисперсию:

$$D(\tilde{\theta}_n) \to min. \qquad (12.4)$$

3. *Состоятельность.* Оценка $\tilde{\theta}_n$ является состоятельной оценкой характеристики генеральной совокупности θ, если для любого $\varepsilon > 0$ выполняется следующее равенство:

$$\lim_{n\to\infty} P\left(\left|\tilde{\theta}_n - \theta\right| < \varepsilon\right) = 1. \qquad (12.5)$$

[17] Bias (*англ.*) — смещение, популярное понятие в анализе данных и машинном обучении.

Это означает, что при увеличении объема выборки n выборочная характеристика $\tilde{\theta}_n$ стремится по вероятности к генеральной характеристике θ:

$$\tilde{\theta}_n \xrightarrow{P} \theta.$$

Теорема 1. Если статистическая оценка является несмещенной, а дисперсия ее стремится к нулю при $n \to \infty$, то такая оценка является и состоятельной.

Доказательство. Воспользуемся неравенством Чебышёва:

$$P(|X - M(X)| < \varepsilon) \geq 1 - \frac{D(X)}{\varepsilon^2}.$$

Так как $\tilde{\theta}$ является случайной величиной, то в неравенстве Чебышёва случайную величину X можно заменить на $\tilde{\theta}$:

$$P\left(\left|\tilde{\theta} - M\left(\tilde{\theta}\right)\right| < \varepsilon\right) \geq 1 - \frac{D\left(\tilde{\theta}\right)}{\varepsilon^2}.$$

При $n \to \infty, D(\tilde{\theta}) \to 0$, поэтому второй член неравенства будет равен нулю, и так как вероятность не может быть больше единицы, то

$$P\left(\left|\tilde{\theta} - M\left(\tilde{\theta}\right)\right| < \varepsilon\right) = 1. \qquad (12.6)$$

По условию $\tilde{\theta}$ является несмещенной оценкой θ, поэтому

$$\lim_{n\to\infty} P\left(\left|\tilde{\theta} - \theta\right| < \varepsilon\right) = 1,$$

что и требовалось доказать.

Теорема во многих случаях позволяет доказать состоятельность оценок $\tilde{\theta}$.

Статистика $\tilde{\theta}$, определяемая по формуле (12.1), называется *достаточной статистикой* для параметра θ, если она использует всю информацию относительно оцениваемого параметра, содержащуюся в выборке $X_1, X_2, \ldots, X_n$.

В математической статистике обычно неявно полагается, что оценки параметров изучаемых распределений, а также параметров моделей, опирающихся на идеологию статистического вывода, в идеальном случае должны удовлетворять перечисленным ранее свойствам несмещенности, эффективности, состоятельности. При этом *точные значения наблюдений сопровождаются наложением* (аддитивным и (или) мультипликативным) двух типов *ошибок*: *случайных* (вызванных влиянием внешних факторов, агрегированием ошибок измерений, вычислений, метода моделирования и других естественных погрешностей); *систематических* (вызванных тремя основными источниками: методикой наблюдений или проведения опыта, смещением в показаниях приборов или способе измерения данных, ошибками наблюдателя). Изучение случайных ошибок составляет предмет теории статистического вывода (выборочного метода, проверки статистических гипотез) и не должно противоречить «здравому смыслу», в предположении, что систематические ошибки устранены (или минимизированы). Таким образом, можно положить, что проводится наблюдение n различных величин X_1, $X_2, \ldots, X_n$, которые представляются в виде $x_1, x_2, \ldots, x_n$. Причем можно считать, что каждое наблюдаемое значение представляет собой результат суммы (или произведения) точного значения и некоторой погрешности, вызванной как случайными, так и систематическими ошибками:

$$x_{i\text{н}} = x_{i\text{т}} + \varepsilon_i.$$

В математической статистике полагается отсутствие систематических ошибок, поэтому считается, что $M(\varepsilon_i) = 0$, а $D(\varepsilon_i) = const$.

Замечание. Физически систематические ошибки (ошибки смещения) можно проиллюстрировать с помощью доски Гальтона (рис. 5.5) — если установить ее под наклоном, то «колоколообразная» форма распределения будет отклоняться от первоначального центра. Теперь коэффициент асимметрии будет либо больше, либо меньше нуля (увеличится влияние хвостов распределения), изменятся и характеристики центральной тенденции (мода, медиана, математическое ожидание), которые совпадали для симметричного случая, что и соответствует идее смещения из-за внешнего воздействия (внесения асимметрии). Проиллюстрировать эффект смещения можно с использованием биномиального закона с параметрами (рис. 3.2): $Bin(10; 0{,}2)$, $Bin(10; 0{,}5)$, $Bin(10; 0{,}8)$. ■

Большая дисперсия указывает на большое рассеяние данных относительно центра и, следовательно, не эффективность.

Кажущаяся теоретичность введенных выше свойств точечных оценок имеет достаточно много практических примеров, иллюстрирующих проблему *систематических ошибок* (*ошибок смещения*).

1. В книге «Возможно да, возможно нет. Фишер. Статистический вывод» [20] приводится несколько вариантов попадания стрелками в цель (или игры в дартс), центр соответствует началу координат, которые можно интерпретировать с точки зрения эффективности и несмещенности (рис. 12.2).

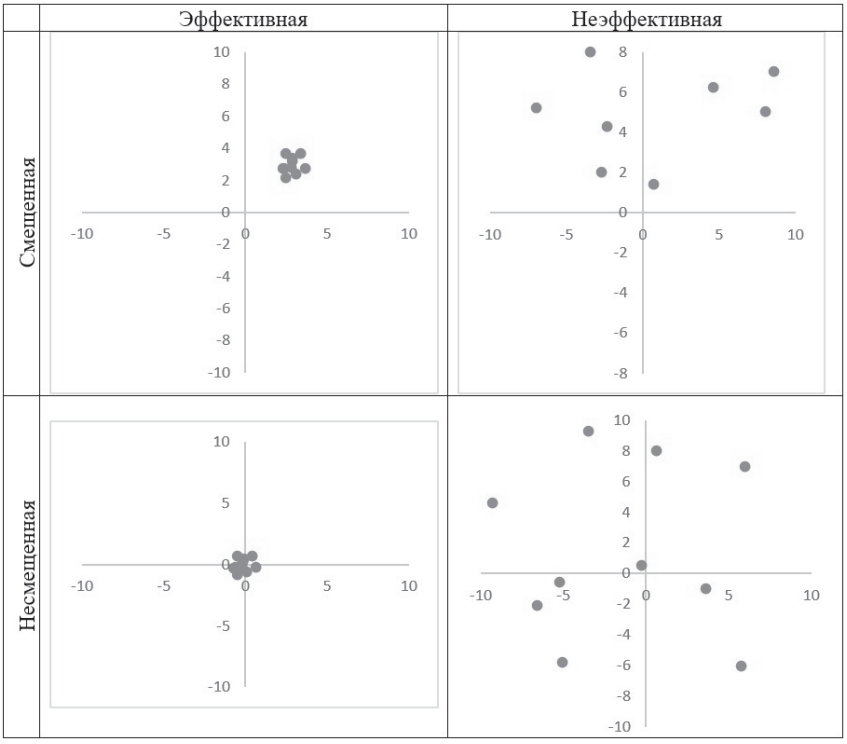


Рис. 12.2 — Пример смещения и эффективности оценки

2. Выборочные обследования мнения клиентов розничных сетей способны значительно увеличить продажи. В то же время предвзятые оценки экспертов (специалистов узкой предметной области) приводят к тому, что, например, смещенные оценки объема выпуска продукции сильно влияют на результат. Так, в книге Г. Кимбла [58] приводится следующий пример. Одна компания, занимающаяся выпуском продукции из пластмасс, обратилась в консалтинговую фирму с просьбой выяснить пути похищения части готовой продукции. Анализ показал, что обследования в компании проводились в начале каждого месяца после наладки оборудования, что вызывало смещение оценки результатов производства продукции (изделий из пластмассы) в большую сторону, так как со временем расход сырья и материалов увеличивался, соответственно, выход продукции был ниже ожидаемого.

3. С описания решения задачи несмещенности начинается книга Дж. Шуровьески [132]. «В один из осенних дней 1906 года британский ученый Фрэнсис Гальтон оставил свой дом в городе Плимуте и отправился на сельскую ярмарку. Гальтону было восемьдесят пять лет. Он вполне ощущал свой возраст, однако его все еще переполняла любознательность, благодаря которой было написано немало научных трудов по статистике и теориям наследования, принесших ему известность (включая скандальную). Гальтон ехал на ежегодную выставку животноводства и птицеводства Западной Англии — региональное мероприятие, на которое собирались местные фермеры и горожане, чтобы оценить достоинства домашнего скота и птицы — коров, овец, лошадей, свиней, кур. ...Прогуливаясь по выставке, Гальтон наткнулся на стенд, около которого проводились соревнования по угадыванию веса. На всеобщее обозрение был выставлен откормленный бык, и собравшаяся толпа должна была на глазок определить вес животного. (А точнее, они должны были угадать вес этого быка после того, как его “забьют и освежуют”.) За шесть пенсов вы могли купить проштампованный и пронумерованный билет, в который надо было внести ваше имя, адрес и прогноз. За самые точные ответы были обещаны призы. Счастье попытали примерно восемьсот человек. Это была разношерстная публика — как мясники и фермеры, явно искушенные в оценке веса скота, так и люди, наверняка далекие от животноводства.

“Участие приняли множество непрофессионалов, — писал впоследствии Гальтон в научном журнале *Nature*, — клерки и прочие из тех, кто, не имея специальных знаний о лошадях, делают ставки на бегах, опираясь на мнение газет, друзей или собственное разумение”. Гальтону тут же пришла на ум аналогия с демократией, когда люди с радикально различающимися способностями и интересами получают каждый по одному голосу. “Средний участник конкурса был экипирован знаниями для точной оценки веса забитого и освежеванного быка не лучше, чем средний избиратель — для оценки качеств того или иного претендента или особенностей большинства политических вопросов, по которым он голосует”, — сетовал он.

Гальтон хотел установить, на что способен “средний избиратель”, поскольку намеревался доказать, что его возможности очень малы. Поэтому он превратил конкурс в импровизированный эксперимент. Когда соревнование закончилось и призы были розданы, Гальтон позаимствовал у его организаторов

билеты и подверг их ряду статистических тестов. Гальтон рассортировал билеты с прогнозами (всего 787 — ему пришлось исключить тринадцать билетов, ибо они были заполнены неразборчиво) в порядке убывания точности и выстроил график, чтобы убедиться, что он будет представлять собой колоколообразную, гауссову кривую. Затем он сложил все оценки участников и вывел усредненный прогноз группы. Эта цифра представляла собой, можно сказать, коллективную мудрость плимутской толпы. Если бы толпа была одним человеком, именно так этот человек оценил бы вес быка. Гальтон, несомненно, полагал, что средний прогноз группы будет очень далек от истины. Казалось очевидным, что коллективное решение толпы, состоящей как из мудрецов, так и из людей посредственных и недалеких, скорее всего, окажется неудачным. Но Гальтон ошибся. Толпа предположила, что вес быка, после того как его забьют и освежуют, составит 1197 фунтов. После того как его действительно забили и освежевали, оказалось, что бык весил 1198 фунтов. Иными словами, оценка толпы оказалась очень точной. Возможно, в конечном итоге селекция не так уж много значила. Позднее Гальтон писал: “Результат был в большей степени в пользу надежности демократических суждений, чем того можно было ожидать”. Это было явное преуменьшение».

В цитированной книге [132] приводятся многочисленные примеры того, что экспертные опросы специалистов приводят к смещению, обусловленному их узкой специализацией, и наоборот — внесение элемента случайности в перечень опрашиваемых может увеличить точность результатов. Примеры систематических ошибок (завышения или занижения оценки по сравнению генеральной характеристикой) часто вызваны асимметрией информации. Например, известный статистик Абрахам Вальд в годы Второй мировой войны рекомендовал защищать броней части «выживших» самолетов, которые остались целыми, ибо по «не выжившим» самолетам информации нет, и вполне вероятно, что причиной этому служили соответствующие повреждения.

Один из самых известных примеров смещенной выборки — опрос около *десяти миллионов респондентов* (без учета статистической методики), проведенный американским журналом «Литературный обзор» в 1936 г. среди своих подписчиков, который предсказал победу в президентских выборах А. Лэндона над Ф. Рузвельтом. Основатель института общественного мнения Дж. Гэллап, проводя регулярный стратифицированный опрос, учитывая мнение всего *двух тысяч респондентов,* предсказал победу Ф. Рузвельта [153].

Фактически различным аспектам систематических ошибок посвящены работы известного финансового аналитика Н. Талеба («Черный лебедь» и др.). Поэтому рекомендуется использовать мнение специалистов различных областей, используя случайный выбор или, как это назвал Р. Фишер, рандомизацию.

Правильный выбор единиц совокупности наблюдений (методика, надежность измерений, точность наблюдателя), случайность гарантируют репрезентативность (представительность) выборки и надежность статистических выводов. Для улучшения результатов обработки (уменьшения систематических ошибок) рекомендуется, например, метод «складного ножа» и его развитие — бутстреп-метод.

12.3. Методы нахождения точечных оценок неизвестных параметров

Задача поиска точечной оценки параметра θ формулируется следующим образом: имеется генеральная совокупность с функцией распределения $F(x,\ \theta)$, (или функцией плотности вероятности $f(x,\ \theta)$), необходимо оценить параметр θ (в общем случае вектор θ) по результатам выборки $x_1, x_2, \dots, x_n$.

Поиск оценок параметров распределения по выборочным данным можно осуществить несколькими методами. Один из первых — *метод моментов* (К. Пирсон, 1894). Предлагается эмпирические моменты, введенные по аналогии с моментами случайных величин, принимать за оценки соответствующих теоретических моментов, а неизвестные параметры распределения выражать через эти моменты.

Эмпирические и соответствующие им теоретические (начальные, центральные и основные) моменты (разделы 6.3, 11.4) определяются по следующим формулам.

Случайные величины	Моменты	
	Эмпирические	Теоретические
Дискретные	$M_s = \overline{x^s} = \frac{\sum x_i^s n_i}{n}$, $m_s = \frac{\sum (x_i - \overline{x})^s n_i}{n}$, $r_s = \frac{\sum (x_i - \overline{x})^s n_i}{n\sigma_x^s}$, $r_{s,h} = \frac{\sum (x_i - \overline{x})^s (y_i - \overline{y})^h n_i}{n\sigma_x^s \sigma_y^h}$	$\alpha_s = \sum x_i^s p_i\,(x,\theta)$, $\mu_s = \sum \big(x_i - M(x)\big)^s p_i\,(x,\theta)$, $r_s = \frac{\sum \big(x_i - M(x)\big)^s p_i(x,\theta)}{n\sigma_x^s}$, $r_{s,h} = \frac{\sum (x_i - M(X))^s (y_i - M(Y))^h p_i\big((x_i, y_i), \theta\big)}{\sigma_x^s \sigma_y^h}$
Непрерывные	$M_s = \overline{x^s} = \frac{\sum x_i^s n_i}{n}$, $m_s = \frac{\sum (x_i - \overline{x})^s n_i}{n}$, $r_s = \frac{\sum (x_i - \overline{x})^s n_i}{n\sigma_x^s}$, $r_{s,h} = \frac{\sum (x_i - \overline{x})^s (y_i - \overline{y})^h n_i}{n\sigma_x^s \sigma_y^h}$	$\alpha_s = \int_{-\infty}^{+\infty} x_i^s f(x,\theta)dx$, $\mu_s = \int_{-\infty}^{+\infty} (x_i - M(x))^s f(x,\theta)dx$, $r_s = \frac{\int_{-\infty}^{+\infty} (x_i - M(x))^s f(x,\theta)dx}{\sigma_x^s}$, $r_{s,h} = \frac{\iint (x_i - M(x))^s (y_i - M(y))^h f(x,\theta)dxdy}{\sigma_x^s \sigma_y^h}$

Приравнивая теоретические и эмпирические моменты для любых законов, мы получим, что оценкой математического ожидания является средняя арифметическая

$$\widehat{M}(X) = \overline{x} = \frac{\sum x_i n_i}{n},$$

а оценка дисперсии определяется как

$$\widehat{D}(X) = \frac{\sum (x_i - \overline{x})^2 n_i}{n}.$$

Метод моментов позволяет получить эффективные и состоятельные оценки лишь для нормального закона распределения ($\overline{x}$ и s^2 — состоятельные

и эффективные оценки математического ожидания a и дисперсии σ^2). Для других законов распределения оценки параметров, полученные методом моментов, оказываются смещенными и малоэффективными. Однако асимптотическая нормальность больших выборок обеспечивает применимость метода моментов в этом случае.

Вторым методом поиска точечных оценок, наиболее важным как с теоретической, так и с практической точки зрения, является метод максимального правдоподобия.

Метод максимального правдоподобия[18] (ММП) (систематически разработан Р. Фишером в 1912 г., хотя идеологически он восходит к Д. Бернулли и К. Гауссу) является наиболее распространенным методом оценивания выборочных характеристик во многих задачах математической статистики и ее приложений, что обусловлено получением состоятельных и асимптотически нормальных оценок с наименьшей дисперсией, при этом наилучшим образом используется вся информация, содержащаяся в выборке.

Пусть имеется выборка из n независимых и одинаково распределенных случайных величин $X_1, X_2, \dots, X_n$, подчиняющихся закону распределения вида $f(X, \theta)$, зависящему от параметра θ, тогда для любой реализации указанных случайных величин $x_1, x_2, \dots, x_n$ их совместную плотность вероятности (вероятность в дискретном случае) можно представить в виде функции правдоподобия

$$L(x, \theta) = \prod_{i=1}^{n} f(x_i, \theta), \qquad (12.7)$$

опираясь на следствие из теоремы 3 о произведении n независимых событий (см. раздел 1.4).

Полагается, что чем больше значение функции правдоподобия $L(\theta)$, тем более правдоподобно (вероятно), при заданном значении параметра θ, появление выборки $x = (x_1, x_2, \dots, x_n)$. В силу монотонности функций $L(x, \theta)$ и $l(x, \theta) = ln\left(L(x, \theta)\right) = \sum_{i=1}^{n} ln\big(f(x_i, \theta)\big)$, они имеют максимум в общей точке, поэтому для упрощения преобразований обычно рассматривают *логарифмическую функцию правдоподобия* $l(x, \theta)$.

Измерение расстояний между распределениями (12.7) при двух различных значениях параметра θ часто осуществляется с использованием специальной характеристики — *информации Фишера*[19].

Пусть на логарифмическую функцию правдоподобия $l(x, \theta)$ наложено условие *регуляризации* (существование непрерывной производной в любой точке области определения):

$$M_\theta \left[\frac{\partial l(x_1, x_2, \dots, x_n, \theta)}{\partial \theta}\right] = 0, \qquad (12.8)$$

тогда *информацией Фишера* $I(\theta)$ для данной статистической модели, при n независимых испытаниях, *называется дисперсия функции вклада выборки*. Для

[18] MLE — Maximum Likelihood Estimation (*англ.*) — метод максимального правдоподобия.

[19] В отличие от I_{ij} — семантической меры информации (II.1) А. А. Харкевича, информация $I_n(\theta)$ Р. Фишера исключает из рассмотрения содержательную сторону и связана с ожиданием разрешения неопределенности — чем выше ожидание определенного результата, тем меньше информации.

случайного вектора $x = (x_1, x_2, \dots, x_n)$ с функцией плотности вероятности $f(x, \theta)$ эта величина равна математическому ожиданию, при заданном θ, квадрата производной логарифмической функции правдоподобия по θ (или отрицательному значению ее второй производной):

$$I_n(\theta) = M_\theta \left[\frac{\partial l(x_1, x_2, \dots, x_n, \theta)}{\partial \theta}\right]^2 = -M_\theta \left[\frac{\partial^2 l(x_1, x_2, \dots, x_n, \theta)}{\partial \theta^2}\right]. \quad (12.9)$$

$I(\theta)$ характеризует математическое ожидание квадрата относительной скорости изменения условной плотности вероятности $f(x, \theta)$ в точке x [61].

Количество информации в одном наблюдении равно

$$I_i(\theta) = M_\theta \left[\frac{\partial l(x_i, \theta)}{\partial \theta}\right]^2. \quad (12.10)$$

Для регулярных моделей все $I_i(\theta)$ равны между собой, кроме того, так как x_i независимы, то дисперсия суммы независимых случайных величин равна сумме дисперсий этих величин. Поэтому для n независимых случайных величин $I_n(\theta) = nI_i(\theta)$, оценка параметра может быть получена по каждому из n испытаний.

Теорема 2 (неравенство Рао — Крамера). Для дисперсии произвольной несмещенной оценки параметра θ, удовлетворяющей условиям регулярности:

1) множество x_i $\{f(x_i, \theta) > 0\}$ не зависит от θ;

2) существует первая и вторая производная по θ под знаком интеграла:

$$\int_{-\infty}^{+\infty} f(x_i, \theta)\, dx_i = 1; \quad (12.11)$$

3) существует первая производная по θ под знаком интеграла:

$$\int_{-\infty}^{+\infty} \tilde{\theta}(x_i) f(x_i, \theta)\, dx_i = \theta, \quad (12.12)$$

выполняется неравенство Рао — Крамера:

$$D_\theta(\tilde{\theta}) \geq \frac{1}{I_n(\theta)}, \quad (12.13)$$

где $I_n(\theta) = nI_i(\theta)$ — информация Фишера.

Оценка $\tilde{\theta}$, для которой в неравенстве Рао — Крамера (12.13) достигается знак равенства, называется *эффективной,* так как выполняется формула (12.4).

Доказательство опирается на возможность дифференцирования под знаком интеграла (12.11), (12.12). Имеем, соответственно, так как

$$\frac{df(x_i)}{d} = \frac{dlnf(x_i)}{dx_i} f(x_i),$$

то

$$\int_{-\infty}^{+\infty} \frac{dlnf(x_i)}{dx_i} f(x_i)\ dx_i = 0, \quad (12.14)$$

$$\int_{-\infty}^{+\infty} \tilde{\theta}(x_i) \frac{dlnf(x_i)}{dx_i} f(x_i)\ dx_i = 1. \quad (12.15)$$

Умножив (12.14) на θ и вычитая результат из (12.15), получим

$$\int_{-\infty}^{+\infty} (\tilde{\theta}(x_i) - \theta) \frac{dlnf(x_i)}{dx_i} f(x_i), dx_i = 1, \quad (12.16)$$

применив к (12.16) неравенство Коши — Шварца — Буняковского (6.34), получим

$$1 = \left(\int_{-\infty}^{+\infty} (\tilde{\theta}(x_i) - \theta) \frac{dlnf(x_i)}{dx_i} f(x_i)\ dx_i\right)^2 \leq$$

$$\leq \int_{-\infty}^{+\infty} (\tilde{\theta}(x_i) - \theta)^2 f(x_i) dx_i \int_{-\infty}^{+\infty} \left(\frac{dlnf(x_i)}{dx_i}\right)^2 f(x_i) dx_i.$$

Следовательно,

$$1 \le D_\theta \tilde{\theta}(x_i) M_\theta \left[\frac{\partial l(x_i,\theta)}{\partial \theta}\right]^2.$$

Откуда имеем

$$D_\theta \tilde{\theta}(x_i) \ge \frac{1}{M_\theta\left[\frac{\partial l(x_i,\theta)}{\partial \theta}\right]^2}. \qquad (12.17)$$

Учитывая независимость и одинаковое распределение компонент вектора $x = (x_1, x_2, \dots, x_n)$, получим, что

$$f(x,\theta) = f(x_1,\theta) \dots f(x_n,\theta).$$

Пусть $Z_i = \frac{\partial L(x_i,\theta)}{\partial \theta}$, тогда, учитывая (12.14), $M_\theta(Z_i) = 0$, следовательно,

$$M_\theta \left[\frac{\partial l(x,\theta)}{\partial \theta}\right]^2 = D_\theta\left(\sum_{i=1}^n Z_i\right) = nI_i(\theta) = I_n(\theta),$$

в силу определения количества информации (12.10), подставив результат в формулу (12.17), получим неравенство Рао — Крамера (12.13), позволяющее дать нижнюю оценку дисперсии оцениваемого параметра θ, что обосновывает ее эффективность.

Замечание. 1) Пусть в неизменных условиях проводятся n опытов по схеме Бернулли и в каждом может произойти или не произойти событие A, $P(A) = p$, $P\left(\overline{A}\right) = 1 - p = q$. Тогда относительная частота $\frac{k}{n}$ появления события A является состоятельной, несмещенной и эффективной оценкой вероятности p.

Состоятельность относительной частоты как оценки вероятности в испытаниях по схеме Бернулли следует из теоремы Бернулли (см. раздел 8.2):

$$\lim_{n\to\infty} P\left(\left|\frac{k}{n} - p\right| < \varepsilon\right) = 1 \qquad (\varepsilon > 0).$$

Для нахождения математического ожидания относительной частоты положим $k = X$, где случайная величина X распределена по биномиальному закону

$$M\left(\frac{k}{n}\right) = M\left(\frac{X}{n}\right) = \frac{1}{n} M(X) = \frac{1}{n} np = p.$$

Таким образом, так как $M\left(\frac{k}{n}\right) = p$, то относительная частота $\frac{k}{n}$ является несмещенной оценкой p.

Найдем дисперсию $\frac{k}{n}$:

$$D\left(\frac{k}{n}\right) = D\left(\frac{X}{n}\right) = \frac{1}{n^2} D(X) = \frac{1}{n^2} npq = \frac{pq}{n}.$$

Полагая, что $X = \sum_{i=1}^n x_i$ — результат суммы индикаторных случайных величин, распределенных по закону Бернулли:

$$f(x_i,p) = \begin{cases} p, & \text{при } x_i = 1, \\ 1-p, & \text{при } x_i = 0. \end{cases}$$

Тогда функция правдоподобия равна

$$L(p) = \prod_{i=1}^n f(x_i,p) = p^x (1-p)^{n-x}.$$

Логарифмическая функция правдоподобия:

$$l(p) = x ln p + (n-x) ln(1-p).$$

Производные логарифмической функции правдоподобия:

$$\frac{\partial l(x,p)}{\partial p} = \frac{x}{p} - \frac{(n-x)}{1-p},$$

$$\frac{\partial^2 l(x,p)}{\partial p^2} = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}.$$

Следовательно,

$$M\left(\frac{\partial^2 l(x,p)}{\partial p^2}\right) = M\left(-\frac{x}{p^2} - \frac{n-x}{(1-p)^2}\right) =$$

$$= -M\left(\frac{x}{p^2}\right) - M\left(\frac{n-x}{(1-p)^2}\right) = -\frac{M(x)}{p^2} - \frac{n-M(x)}{(1-p)^2} =$$

$$= -\frac{np}{p^2} - \frac{n-np}{(1-p)^2} = -\frac{n}{p} - \frac{n}{1-p} = -n\left(\frac{1}{p} + \frac{1}{q}\right) = -\frac{n}{pq} = -I_n(\theta).$$

Полагая, что для оценки вероятности сделано n экспериментов в неравенстве Рао — Крамера, нижняя оценка будет равна

$$\frac{1}{I_n(\theta)} = \frac{pq}{n},$$

это значение совпадает с дисперсией относительной частоты, значит, относительная частота $\frac{k}{n}$ является эффективной оценкой вероятности.

2) Случайная величина X распределена по нормальному закону. Требуется по результатам наблюденных значений $x = (x_1, x_2, \dots, x_n)$ этой величины оценить параметры a и σ нормального закона:

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

Следовательно, функция правдоподобия имеет вид

$$L(x, a, \sigma) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left(-\frac{\sum(x_i-a)^2}{2\sigma^2}\right),$$

а логарифмическая функция правдоподобия

$$l(x, a, \sigma) = -\frac{n}{2} ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}\sum(x_i - a)^2.$$

Для нахождения оценок параметров необходимо получить частные производные по a и σ и приравнять их к нулю. Продифференцировав $l(x, a, \sigma)$ по a и σ, имеем

$$\begin{cases} \frac{\partial l(x,a,\sigma)}{\partial a} = \frac{1}{\sigma^2}\sum(x_i - a), \\ \frac{\partial l(x,a,\sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum(x_i - a)^2, \end{cases}$$

$$\begin{cases} \frac{1}{\sigma^2}\sum(x_i - a) = 0, \\ -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum(x_i - a)^2 = 0. \end{cases}$$

Решая систему, получим

$$\hat{a} = \overline{x} = \frac{\sum x_i}{n},$$

$$\widehat{\sigma^2} = \frac{\sum(x_i - \overline{x})^2}{n}.$$

Эти оценки с учетом повторностей совпадают с оценками, полученными методом моментов.

3) Оценки, полученные ММП, иногда являются смещенными, поэтому приходится вводить поправки, однако с увеличением объема выборки смещение уменьшается (асимптотическая несмещенность). Проблемы, связанные с решением сложных систем уравнений, благодаря распространению компьютеров и соответствующих программ обработки данных, сегодня решены. Отметим некоторые важные свойства оценок, полученных ММП:

– оценки состоятельны;

– ММП дает эффективную и достаточную оценку, если она существует;

– оценки ММП асимптотически эффективны;

– если $\tilde{\theta}$ — оценка параметра генеральной совокупности θ, полученная ММП, то при $n \to \infty$, она асимптотически стремится к нормальному закону распределения с $M(\tilde{\theta}) = \theta$,

$$D(\tilde{\theta}) = \frac{1}{M_\theta\left[\frac{\partial l(x_1,x_2,\dots,x_n,\theta)}{\partial\theta}\right]^2} = -\frac{1}{M_\theta\left[\frac{\partial^2 l(x_1,x_2,\dots,x_n,\theta)}{\partial\theta^2}\right]} = \frac{1}{I_n(\theta)} : \tilde{\theta} \to N\left(\theta, \frac{1}{I_n(\theta)}\right).$$

4) В статистике кроме средней арифметической (простой и взвешенной) для оценки центральной тенденции используют: моду, медиану; средние степенные, среднюю гармоническую, среднюю геометрическую — для нахождения средних, соответствующих принятым единицам измерения. Для оценки вариации признаков используют: размах вариации, среднее линейное отклонение, дисперсию, среднее квадратическое отклонение, коэффициент вариации. ■

Некоторые альтернативные методы поиска выборочных оценок.

1) *Вероятностная бумага.* На основании выборочных наблюдений $x_1, x_2, \dots, x_n$ можно построить эмпирическую функцию распределения. Пусть известен вид функции распределения $F(x) = P(X < x)$, тогда эмпирическая функция является приближением к ее некоторому варианту, полученному сдвигом и масштабированием: $F\left(\frac{x-m}{\sigma}\right)$. Если для функции распределения выполняются (что обычно верно по ее определению) свойства непрерывности и монотонности, то существует обратная функция F^{-1}. Отсюда следует, что

$$(x, y) \to \left(x, F\left(\frac{x-m}{\sigma}\right)\right) \to \left(x, F^{-1}(F\left(\frac{x-m}{\sigma}\right)\right) \to \left(x, \frac{x-m}{\sigma}\right).$$

Значит, график функции распределения переходит в график прямой:

$$y = \frac{x-m}{\sigma} = \frac{1}{\sigma}x - \frac{m}{\sigma}.$$

То есть m — точка пересечения с прямой $y = 0$, а σ — котангенс угла наклона этой прямой к оси абсцисс в системе координат $(x, F(x))$.

При построении графиков обычно данные упорядочиваются, а оси абсцисс и ординат могут быть различны, и в соответствии с этим графики имеют названия:

$(F(x), F_n(x))$ — «вероятность — вероятность»,

$(x, F(x))$ — «нормальный график».

Рассмотрим иллюстративный пример. Для этого с использованием генератора случайных чисел получим 15 чисел $X \in N(3, 1)$: 3,046; 2,600; 1,232; 3,404; 3,113; 3,292; 3,880; 3,492; 1,827; 3,642; 2,473; 2,710; 2,432; 2,306; 2,971.

Полученные графики в *Statistica 10 Ru* представлены на рисунках 12.3, 12.4.

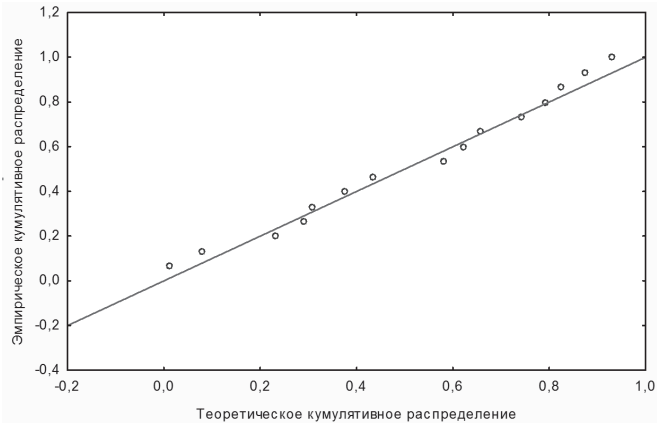
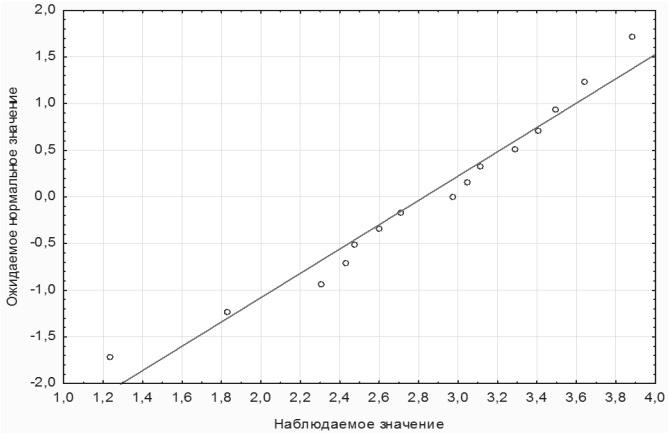
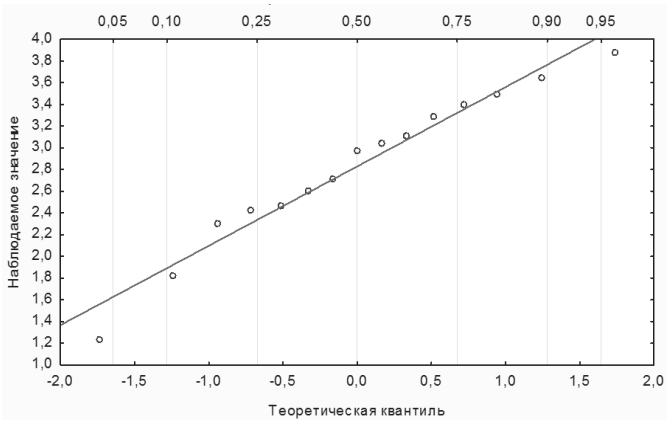


Рис. 12.3 — График типа «вероятностная бумага» (вероятность — вероятность)



а) нормальный



б) квантиль — квантиль

Рис. 12.4 — Графики типа «вероятностная бумага»

В частности, из рисунка 12.4*а* оценка математического ожидания $m \approx 2{,}9$, а $\sigma \approx 0{,}8$, что можно объяснить малой выборкой.

Квантильные оценки можно получить в системе координат с осями: X — теоретические квантили, Y — наблюдаемые квантили.

Близость данных к прямой линии может визуально подтверждать гипотезу о виде закона распределения, которые, например, в системе *Statistica* можно итеративно подбирать.

2) *Метод наименьших квадратов* (МНК) часто называется в качестве основного метода математической статистики и вообще прикладной математики. Он может быть получен с помощью метода максимального правдоподобия, исходя из предположения о нормальном законе распределения (см. пункт 2 в предыдущем замечании). Более подробно он рассмотрен в разделе регрессионного анализа.

3) *Робастные оценки*, устойчивые к выбросам, разрабатывались с начала 1960-х годов, когда Дж. Тьюки (1960) показал неустойчивость оценки метода максимального правдоподобия к небольшому отклонению эмпирической плотности распределения от теоретической. Это демонстрировало ограниченную прикладную ценность классической статистики и послужило началом развития *прикладной статистики* (или *анализа данных*), у истоков которой стоял Дж. Тьюки. В ней рассматривались различные варианты отказа от классических рекомендаций статистики, в том числе отказ от законов распределения, ориентирующихся на вероятностную природу генерации данных, и предлагалось изучение логической, геометрической и когнитивной (основанной на знаниях) природы данных.

Различные варианты оптимального решения задач поиска параметров распределений и моделей в условиях загрязнения предлагали: П. Хьюбер (1964–1984), Л. Мешалкин (1971), Д. Эндрюс (1972), А. Ершов (1978), А. Гуда (1997), А. Шурыгин (2000) и многие другие [133, 185]. Робастные оценки широко используются в статистических и эконометрических пакетах. Например, в свободно распространяемом эконометрическом пакете *gretl* предлагаются следующие робастные оценки при построении непараметрических моделей регрессии:

– *Least Absolute Deviation* (метод наименьших модулей или медианная регрессия),

– *Quantile regression* (квантильная регрессия),

– *Nadaraya-Watson* (ядерная оценка Надарая — Ватсона),

– *Loess* (локально-взвешенная полиномиальная регрессия).

4) *Байесовские оценки.* Байесовская статистика позволяет практически всем статистическим процедурам поставить в соответствие байесовские, например, это проявляется в интенсивно развивающемся и свободно распространяемом пакете *JASP*, поддерживаемым университетом Амстердама. Основная идея в том, что появление новых наблюдений позволяет пересматривать оценки статистических параметров, используя средневзвешенное значение прошлой оценки и новых наблюдений с весами, зависящими от дисперсии прошлых данных на некотором промежутке времени (см. гл. 19, 20 в [53]).

12.4. Оценка генеральной средней и дисперсии по выборочной средней и дисперсии

Пусть изучаемый признак является дискретным и в генеральной совокупности задан следующей таблицей.

x_i	x_1	x_2	…	x_k
N_i	N_1	N_2	…	N_k

$$N_1 + N_2 + \cdots + N_k = N.$$

Отбор единиц из генеральной в выборочную совокупность проводится случайным повторным способом.

Возможные значения изучаемого признака при отборе первой, второй, …, n-ой единицы выборочной совокупности являются дискретными случайными величинами $X_1, X_2, \ldots, X_n$. Так как отобранная единица возвращается назад в генеральную совокупность, то случайные величины $X_1, X_2, \ldots, X_n$ будут иметь один и тот же закон распределения. Поэтому случайные величины $X_1, X_2, \ldots, X_n$ являются одинаково распределенными и независимыми, у которых совпадут числовые характеристики.

Рассмотрим случайную величину X_i. Эта единица может принимать значения $x_1, x_2, \ldots, x_k$ с вероятностями $\frac{N_1}{N}, \frac{N_2}{N}, \ldots, \frac{N_k}{N}$ соответственно. Следовательно, закон распределения случайной величины X_i будет иметь следующий вид.

x_i	x_1	x_2	…	x_k
p_i	$\frac{N_1}{N}$	$\frac{N_2}{N}$	…	$\frac{N_k}{N}$

Найдем математическое ожидание случайной величины X_i:

$$M(X_i) = x_1\frac{N_1}{N} + x_2\frac{N_2}{N} + \cdots + x_k\frac{N_k}{N} = \frac{x_1N_1 + x_2N_2 + \cdots + x_kN_k}{N},$$

$$M(X_i) = \bar{x}_\text{г}. \qquad (12.18)$$

Значит, все случайные величины $X_1, X_2, \ldots, X_n$ будут иметь математические ожидания, равные среднему значению признака в генеральной совокупности. Выборочная средняя ($\bar{X}_\text{в}$) есть средняя арифметическая возможных результатов единичных выборок. Так как *выборочная средняя* также является случайной величиной, найдем ее математическое ожидание:

$$\text{M}(\bar{X}_\text{в}) = M\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{M(X_1) + M(X_2) + \cdots + M(X_n)}{n},$$

$$\text{M}(\bar{X}_\text{в}) = \frac{n\bar{x}_\text{г}}{n} = \bar{x}_\text{г}\,. \qquad (12.19)$$

Таким образом, выборочная средняя является несмещенной оценкой генеральной средней при любом объеме выборки. Можно показать также, что выборочная средняя является состоятельной оценкой генеральной средней.

Определим дисперсию случайной величины X_i.

$$D(X_i) = M(x_i - \bar{x}_\text{г})^2 p_i = D_\text{г} = D.$$

Тогда дисперсия среднего арифметического независимых случайных величин $X_1, X_2 \ldots, X_n$ будет определяться по формуле

$$D(\bar{X}_{\text{в}}) = D\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) = \frac{D(X_1)+D(X_2)+\cdots+D(X_n)}{n^2},$$

$$D(\bar{X}_{\text{в}}) = \frac{nD}{n^2} = \frac{D}{n}. \qquad (12.20)$$

Значит дисперсия среднего арифметического случайных величин $X_1, X_2, \ldots, X_n$ в n раз меньше дисперсии признака в генеральной совокупности.

Выборочная дисперсия $D_{\text{в}}$ (далее по тексту σ^2) не обладает свойством несмещенности.

Пусть случайные величины X_i независимы и имеют один закон распределения, то есть $M(X_i) = M(X), D(X_i) = D(X)$. Из свойств дисперсии известно, что изменение всех вариант на величину a не изменяет дисперсию. Согласно формуле (11.26) при $a = 0$ имеем

$$\sigma_x^2 = \frac{\sum(X_i)^2}{n} - (\bar{x})^2. \qquad (12.21)$$

Найдем математическое ожидание (12.21):

$$M(\sigma_x^2) = M\left(\frac{\sum_{i=1}^n X_i^2}{n}\right) - M\left(\left(\frac{\sum_{i=1}^n X_i}{n}\right)^2\right) = \qquad (12.22)$$

$$= M\left(\frac{\sum_{i=1}^n X_i^2}{n}\right) - M\left(\frac{\sum_{i=1}^n X_i^2}{n^2}\right) - 2M\left(\left(\frac{\sum_{i<j} X_i X_j}{n^2}\right)^2\right) =$$

$$= \frac{n-1}{n^2} M\left(\frac{\sum_{i=1}^n X_i^2}{n}\right) - 2M\left(\left(\frac{\sum_{i<j} X_i X_j}{n^2}\right)^2\right) =$$

$$= \frac{n-1}{n^2}\sum_{i=1}^n M\left(X_i^2\right) - \frac{2}{n^2}\sum_{i<j} M(X_i X_j).$$

Дисперсия не зависит от положения начала координат, выберем его так, чтобы $M(X) = 0$. Тогда если $X - M(X) =: \dot{X}$, то

$$M\left(X_i^2\right) = M\left(\dot{X}^2\right) = D(X), \sum_{i=1}^n M\left(X_i^2\right) = nD(X).$$

В силу независимости переменных X_i и X_j

$$M\left(X_i X_j\right) = M\left(\dot{X}_i \dot{X}_j\right) = M(\dot{X}_i) M\left(\dot{X}_j\right) = 0.$$

Имеем

$$M(\sigma_x^2) = \frac{n-1}{n^2} nD(X) - 0 = \frac{n-1}{n} D(X).$$

Следовательно,

$$M(\sigma_x^2) = \frac{n-1}{n}\sigma^2. \qquad (12.23)$$

Формула (12.23) показывает, что *выборочная дисперсия* не обладает свойством несмещенности.

На практике используют *исправленную выборочную дисперсию* s^2, которая является несмещенной оценкой дисперсии генеральной совокупности:

$$s^2 = \frac{n}{n-1} D(X) = \frac{\sum(x_i - \bar{x}_{\text{в}})^2 n_i}{n-1}, \qquad (12.24)$$

s — «исправленное» среднее квадратическое отклонение.

Более кратко, учитывая, что случайные величины x_i независимы и имеют один закон распределения («генеральной» совокупности X), мы будем иметь:

$$M(X_i - a)^2 = \sigma_X^2, M(\bar{x} - a)^2 = \frac{\sigma_X^2}{n}, \ (a = const).$$

Отсюда, опираясь на формулу (11.26) ($\sigma_x^2 = \sigma_a^2 - (\bar{x} - a)^2$), получим

$$M(\sigma_x^2) = \sigma_X^2 - \frac{\sigma_X^2}{n} = \frac{n-1}{n}\sigma_X^2.$$

Объяснить формулу (12.24) можно, опираясь на одно из свойств средней арифметической — $\sum(x_i - \bar{x})n_i = 0$, следовательно, n разностей линейно зависимы, поэтому в расчетах делят сумму квадратов разностей на $(n - 1)$.

Кроме того, в расчетах используют среднюю ошибку выборки (стандартное отклонение), вытекающую из формулы (12.20):

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}. \qquad (12.25)$$

Для оценки точности приближенного равенства (12.1) используется метод доверительных интервалов (систематически разработанный американским статистиком Ю. Нейманом в 1940-е годы).

Интервальной называют оценку, которая определяется двумя числами — границами интервала. Она позволяет ответить на вопрос: внутри какого интервала и с какой вероятностью находится неизвестное значение оцениваемого параметра θ генеральной совокупности.

Пусть $\tilde{\theta}$ — точечная оценка параметра генеральной совокупности θ. Чем меньше разность $|\tilde{\theta} - \theta|$, тем точнее и лучше оценка, т. е. в неравенстве $|\tilde{\theta} - \theta| \leq \Delta$, чем меньше Δ, тем точнее оценка.

Обычно говорят о *доверительной вероятности* (надежности оценки) $\gamma = 1 - \alpha$, с которой θ будет находиться в интервале

$$(\tilde{\theta} - \Delta \leq \theta \leq \tilde{\theta} + \Delta),$$

где Δ — предельная ошибка выборки, которая может быть либо задана наперед, либо вычислена; α — риск или уровень значимости (вероятность того, что неравенство будет неверным).

Доверительной вероятностью γ называется вероятность, с которой осуществляется неравенство: $\gamma = P(|\tilde{\theta} - \theta| \leq \Delta)$.

Интервал $(\tilde{\theta} - \Delta \leq \theta \leq \tilde{\theta} + \Delta)$, который покрывает неизвестный параметр генеральной совокупности θ с доверительной вероятностью γ, называется *доверительным интервалом*.

Оценка указанного доверительного интервала может быть получена с помощью неравенства Чебышёва (при $\varepsilon = \Delta$). В качестве γ принимают значения 0,90; 0,95; 0,99; 0,999. Доверительная вероятность показывает, что в $(1 - \alpha)100\%$ случаев оценка θ будет накрываться указанным интервалом.

12.5. Доверительные интервалы характеристик генеральной совокупности

Как следует из выводов асимптотической теории, опирающейся на центральную предельную теорему, при достаточно большом объеме совокупности, статистики, отраженные в формулах (II.2)–(II.7), асимптотически распределены нормально, что обуславливает актуальность рассмотрения нормального закона.

В случае нормально распределенных величин, *точные распределения выборочных характеристик выражаются через несколько распределений* (χ^2 Пирсона, t −распределение Стьюдента и F Фишера — Снедекора), что

и составляет содержание нижеследующих теорем 3 и 4. Большое теоретическое и практическое значение имеет следующая теорема о совместном распределении выборочного среднего $\bar{x}$ и выборочной дисперсии S^2, которая является точной, а не асимптотической.

Теорема 3 (*лемма Р. Фишера*[20]). Если имеется независимая выборка $x_1, x_2, \dots, x_n$ из нормального распределения с математическим ожиданием a и дисперсией σ^2 $(x_i \in N(a, \sigma^2), i = 1, 2, \dots, n)$, то:

1) $\overline{x}$ и S^2 — независимы:

$$(\bar{x} = \frac{\sum_i x_i}{n};\ S^2 = \frac{\sum_i (x_i - \overline{x})^2}{n}),$$

2) $\bar{x} \to N\left(a, \frac{\sigma}{\sqrt{n}}\right)$, (12.26)

3) $S^2 \xrightarrow{d} \frac{\sigma^2}{n} \chi^2_{n-1}$, (12.27)

где χ^2_{n-1} — распределение хи-квадрат Пирсона с $\nu = n - 1$ степенями свободы.

Доказательство. 1) Величина S^2 не изменится при замене случайной величины X на V, где $v_i = x_i - a$ $(a = const)$. Значит, распределение S^2 не зависит от a, и мы можем считать, что $a = 0$ и x_i имеют нормальное распределение $N(0, \sigma^2)$. Перейдем теперь от случайных величин x_i к системе случайных величин y_i:

$$\begin{cases} y_1 = \frac{1}{\sqrt{1\cdot 2}}(x_1 - x_2), \\ y_2 = \frac{1}{\sqrt{2\cdot 3}}(x_1 + x_2 - 2x_3), \\ y_3 = \frac{1}{\sqrt{3\cdot 4}}(x_1 + x_2 + x_3 - 3x_4), \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ y_{n-1} = \frac{1}{\sqrt{(n-1)\cdot n}}(x_1 + x_2 + \cdots + x_{n-1} - (n-1)x_n), \\ y_n = \frac{1}{\sqrt{n}}(x_1 + x_2 + \cdots + x_n). \end{cases} \quad (12.28)$$

Выпишем коэффициенты системы (12.28) в виде матрицы.

$$A_{nn} = \begin{pmatrix} \frac{1}{\sqrt{1\cdot 2}} & -\frac{1}{\sqrt{1\cdot 2}} & 0 & 0 & \dots & 0 \\ \frac{1}{\sqrt{2\cdot 3}} & \frac{1}{\sqrt{2\cdot 3}} & -\frac{2}{\sqrt{2\cdot 3}} & \dots & \dots & \dots \\ \frac{1}{\sqrt{3\cdot 4}} & \frac{1}{\sqrt{3\cdot 4}} & \frac{1}{\sqrt{3\cdot 4}} & -\frac{3}{\sqrt{3\cdot 4}} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \frac{1}{\sqrt{(n-1)\cdot n}} & \frac{1}{\sqrt{(n-1)\cdot n}} & \frac{1}{\sqrt{(n-1)\cdot n}} & \frac{1}{\sqrt{(n-1)\cdot n}} & \dots & -\frac{1}{\sqrt{(n-1)\cdot n}} \\ \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \dots & \frac{1}{\sqrt{n}} \end{pmatrix} \quad (12.29)$$

[20] 1. Лемма (*греч.*) — предположение, доказанное утверждение, использующееся для доказательства теорем.

2. Лемма Фишера является основой доказательства ряда положений теории статистического вывода (выборочного метода и проверки статистических гипотез), что объясняет название «лемма».

Вспомним некоторые факты из курса линейной алгебры. Система векторов $x_1, \dots, x_n$ ортогональна, если их скалярное произведение $(x_i, x_j) = 0$, при $i \neq j$, а $(x_i, x_i) = 1$. Заметим, что строки и столбцы матрицы A_{nn} удовлетворяют условиям ортогональности, то есть она (матрица A_{nn}) является ортогональной, значит, обратная матрица совпадает с транспонированной $A^{-1} = A^T$ и произведение обратной матрицы и исходной равно произведению исходной и транспонированной матриц $A^{-1}A = A^TA = E$. Отсюда следует, что

$$|A| = \pm 1.$$

Определим некоторые свойства ортогональных векторов.

Взаимно ортогональные ненулевые вектора линейно независимы.

Если $x_1, \dots, x_n$ — ортогональная система векторов, то

$$(x_1 + \dots + x_n)^2 = x_1^2 + \dots + x_n^2.$$

Соотношения ортогональности отражают вращение исходной системы координат при переходе к новой системе и представляют собой характеристики линейного преобразования координат некоторой точки, равные косинусам углов наклона новых осей относительно старых, они связаны соотношениями ортогональности старых и новых осей с единичными ортами. Вращение можно записать в матричной форме:

$$Y = AX,$$

где $A = A_{nn}$, $Y = (y_1, \dots, y_n)$ — вектор-строка, $X = (x_1, \dots, x_n)$ — вектор-столбец. Учитывая ортогональность матрицы A, получим

$$Y^TY = (AX)^TAX = X^TA^TAX = X^TA^{-1}AX = X^TEX = X^TX \text{ или}$$

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n x_i^2.$$

То есть сумма квадратов координат при преобразовании с использованием матрицы A сохраняется, оправдывая термин «вращение системы координат».

Учитывая свойства числовых характеристик, получим

$$M(y_i) = 0, \quad D(y_i) = \sigma^2, \quad M(y_iy_j) = 0, (i \neq j).$$

Значит, $y_i \in N(0, \sigma^2)$ как линейная комбинация $x_i \in N(0, \sigma^2)$ при всех $i = 1, \dots, n$; так как $M(y_iy_j) = 0$, $(i \neq j)$, то величины не коррелированы, а в случае нормального закона — они и независимы при любом числе переменных.

Согласно (12.28)

$$y_n = \bar{x}\sqrt{n}.$$

Рассмотрим преобразование суммы

$$\sum_i x_i^2 = \sum_i (x_i - \bar{x} + \bar{x})^2 = \sum_i (x_i - \bar{x})^2 + n(\bar{x})^2,$$

так как $\sum_i (x_i - \bar{x}) = 0$, где $\bar{x} = \frac{\sum_i x_i}{n}$.

$\sum_i (x_i - \bar{x})^2$ и $n(\bar{x})^2$ — ортогональные *статистики* (симметрические функции от n независимых случайных величин, образующих выборку, по терминологии Р. Фишера).

Отсюда имеем

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n x_i^2 = nS^2 + y_n^2$$

или

$$nS^2 = \sum_{i=1}^{n-1} y_i^2. \qquad (12.30)$$

Значит,

$$\bar{x} = \frac{y_n}{\sqrt{n}},\ S^2 = \frac{1}{n}\sum_{i=1}^{n-1} y_i^2,$$

следовательно, $\overline{x}$ и S^2 — независимы.

2) $\bar{x} \to N\left(a, \frac{\sigma}{\sqrt{n}}\right)$. Это утверждение следует из асимптотических свойств средней арифметической.

3) Из равенства (12.30) получим

$$\frac{nS^2}{\sigma^2} = \sum_{i=1}^{n-1} \frac{y_i^2}{\sigma^2} \to \chi^2_{n-1}, \qquad (12.31)$$

следовательно, $S^2 \xrightarrow{d} \frac{\sigma^2}{n}\chi^2_{n-1}$ — по определению распределения хи-квадрат Пирсона с $\nu = n - 1$ степенями свободы.

Числом степеней свободы квадратичной формы от наблюдений называется ее ранг — число линейно независимых строк или столбцов симметричной матрицы квадратичной формы.

Из теоремы 3 следует еще два утверждения.

Теорема 4.

4) Если имеется независимая выборка $x_1, x_2, \dots, x_n$ из нормального распределения с математическим ожиданием a и дисперсией σ^2

$(x_i \in N(a, \sigma^2), i = 1, 2, \dots, n)$, то

$$\frac{\bar{x}-a}{s}\sqrt{n} \xrightarrow{d} t_{n-1}, \qquad (12.32)$$

где t_{n-1}— t-распределение Стьюдента с $k = n - 1$ степенями свободы.

Действительно, $\frac{\frac{\bar{x}-a}{\sigma}\sqrt{n}}{\frac{s}{\sigma}} = \frac{Z}{\sqrt{\frac{s^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{1}{n-1}\sum_i\left(\frac{x_i-\bar{x}}{\sigma}\right)^2}} = \frac{Z}{\sqrt{\frac{1}{n-1}\chi^2_{n-1}}} \xrightarrow{d} t_{n-1},$

где $Z = \frac{\bar{x}-a}{\sigma}\sqrt{n} \xrightarrow{d} N(0,1)$.

5) Пусть $x_1, x_2, \dots, x_n$ и $y_1, y_2, \dots, y_m$ две независимые выборки из $(n + m)$ независимых нормально распределенных величин с параметрами $N(a, \sigma^2)$, тогда, если

$$\bar{x} = \frac{\sum_i x_i}{n}, S_1^2 = \frac{\sum_i (x_i-\bar{x})^2}{n}, \bar{y} = \frac{\sum_j y_j}{m}, S_2^2 = \frac{\sum_j \left(y_j-\bar{y}\right)^2}{m}, \text{ то}$$

$$\frac{\frac{nS_1^2}{n-1}}{\frac{mS_2^2}{m-1}} = \frac{\sum_i (x_i-\bar{x})^2/(n-1)}{\sum_j \left(y_j-\bar{y}\right)^2/(m-1)} \xrightarrow{d} F(k_1 = n-1, k_2 = m-1), \qquad (12.33)$$

где $F(k_1 = n - 1, k_2 = m - 1)$ — распределение Фишера — Снедекора с числом степеней свободы числителя $k_1 = n - 1$ и знаменателя $k_2 = m - 1$. Что соответствует определению распределения Фишера — Снедекора.

Итак, из предположения, что наблюдения подчиняются нормальному закону распределения с параметрами a и σ, получены следующие основные результаты.

Параметры	Статистика
a — известно	$\sum_{i=1}^{n}\left(\frac{x_i-a}{\sigma}\right)^2 \xrightarrow{d} \chi_n^2$
σ — известно	$\frac{\bar{x}-a}{\sigma}\sqrt{n} \xrightarrow{d} N(0,1)$
a — неизвестно	$\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2 \xrightarrow{d} \chi_{n-1}^2$
σ — неизвестно	$\frac{\bar{x}-a}{s}\sqrt{n} \xrightarrow{d} t_{n-1}$
a_1, a_1 — неизвестно	$\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2(m-1)}{\sum_{i=1}^{m}(y_i-\overline{y})^2(n-1)} \xrightarrow{d} F(k_1=n-1, k_2=m-1)$

Рассмотрим построение доверительных интервалов для параметров нормально распределенной генеральной совокупности.

1. *Построение доверительного интервала для математического ожидания нормально распределенной генеральной совокупности*.

1.1. Пусть выборка $x_1, x_2, \dots, x_n$ состоит из независимых нормально распределенных случайных величин с параметрами a и σ, причем σ *известно*, а величину a оцениваем по выборке

$$a \approx \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,. \qquad (12.34)$$

Поставим перед собой задачу оценить точность этого приближенного равенства, т. е. указать границы (доверительные пределы), в которых практически достоверно лежит неизвестное число a.

Для нормального закона распределения, вероятность отклонения нормально распределенной случайной величины X с параметрами a и σ от математического ожидания $(M(X) = a)$ не более, чем на величину $\Delta > 0$, определяется по формуле

$$P(|X-a|<\Delta) = 2\Phi\left(\frac{\Delta}{\sigma}\right).$$

Можно показать, что для независимых нормально распределенных случайных величин $X_1, X_2, \dots, X_n \in N(a, \sigma^2)$ любая линейная комбинация

$\sum_{i=1}^{n} c_i X_i = c_1X_1 + c_2X_2 + + \cdots + c_nX_n$ распределена нормально:

$$\sum_{i=1}^{n} c_i X_i \to N\left(\sum_{i=1}^{n} c_i a\,;\ \sum_{i=1}^{n}(c_i\sigma)^2\right).$$

В частности, средняя арифметическая $X_1, X_2, \dots, X_n$ асимптотически нормальна, в силу формулы (II. 5):

$$\frac{\bar{x}-a}{\sigma(\bar{x})} \to N(0,1).$$

Итак, величина $\bar{x}$, определенная по формуле (12.12), распределена нормально с параметрами $\overline{x} = a$ и $\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}$. Поэтому вероятность того, что $(\bar{x}-a)$ не превзойдет по абсолютной величине некоторого наперед заданного числа $\Delta = \Delta_{\bar{X}} > 0$, равна

$$P(|\bar{x}-a|<\Delta_{\bar{X}}) = 2\Phi\left(\frac{\Delta\sqrt{n}}{\sigma}\right), \qquad (12.35)$$

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

Пусть α вероятность ошибки предположения о том, что $(\bar{x} - a)$ не превзойдет по абсолютной величине некоторого наперед заданного числа $\Delta_{\bar{x}} > 0$. Тогда зададим коэффициент доверия $\gamma = 1 - \alpha$ таким, чтобы рассматриваемое событие с вероятностью γ можно было считать практически достоверным, и пусть t_γ — корень уравнения $2\Phi\left(t_\gamma\right) = \gamma$ — *квантиль стандартного нормального распределения*, которую можно найти по таблицам функции Лапласа или нормальной функции распределения, например:

$\gamma = 1 - \alpha$	α	t_γ
0,9	0,1	1,65
0,95	0,05	1,96
0,99	0,01	2,58
0,9973	0,0027	3,00
0,999	0,001	3,29

Определим из условия

$$\frac{\Delta_{\bar{x}}\sqrt{n}}{\sigma} = t_\gamma$$

число $\Delta_{\bar{X}}$:

$$\Delta_{\bar{x}} = t_\gamma \frac{\sigma}{\sqrt{n}}.$$

Для данного $\Delta_{\bar{X}}$:

$$P\left(|\bar{x} - a| < t_\gamma \frac{\sigma}{\sqrt{n}}\right) = 2\Phi\left(t_\gamma\right) = \gamma.$$

Таким образом, с вероятностью γ

$$|\bar{x} - a| < t_\gamma \frac{\sigma}{\sqrt{n}},$$

где $2\Phi\left(t_\gamma\right) = \gamma$.

Последнее неравенство запишем в виде

$$\bar{x} - t_\gamma \frac{\sigma}{\sqrt{n}} < a < \bar{x} + t_\gamma \frac{\sigma}{\sqrt{n}}\,. \qquad (12.36)$$

Значит, интервал со случайными концами

$$\bar{x} - t_\gamma \frac{\sigma}{\sqrt{n}} \quad \text{и} \quad \bar{x} + t_\gamma \frac{\sigma}{\sqrt{n}}$$

с вероятностью γ покрывает неизвестное значение $a = M(X)$. Этот интервал является *доверительным интервалом для a,* соответствующим *коэффициенту доверия* γ. Доверительные границы в этом случае таковы:

$$\bar{x} - t_\gamma \frac{\sigma}{\sqrt{n}} \;\text{ и }\; \bar{x} + t_\gamma \frac{\sigma}{\sqrt{n}}.$$

Оценка (12.36) предполагает известным среднее квадратичное отклонение (*стандартное отклонение*) σ, которое на практике чаще всего бывает неизвестно. Если величину σ в неравенстве (12.36) заменить ее приближенным значением

$$s = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}, \qquad (2.37)$$

то коэффициент доверия оценки (12.36) уменьшится.

1.2. Если величина σ *неизвестна*, пользуются другим способом построения доверительного интервала для математического ожидания, основанном на использовании *t*-распределения Стьюдента, где s^2 — исправленное выборочное

среднее квадратическое отклонение, а t_γ находится из решения уравнения с функцией плотности распределения вероятности *t*-распределения Стьюдента для двухсторонней области с $k = (n-1)$ степенями свободы и заданной надежности $\gamma = 1 - \alpha$:

$$\left|\frac{\bar{x}-a}{s}\sqrt{n}\right| < t_{n-1,\alpha}$$

или

$$\int_{-t}^{t} f(t)dt = \gamma, \qquad (12.38)$$

решения которого представлены в таблице приложения 3.

Каков бы ни был закон распределения независимых одинаково распределенных величин, имеющих конечную дисперсию, их сумма распределена приближенно нормально при больших значениях n (согласно центральной предельной теореме). Поэтому доверительный интервал для математического ожидания в общем случае можно представить в виде

$$\bar{x}_в - t_\gamma \frac{s}{\sqrt{n}} < \bar{x}_г < \bar{x}_в + t_\gamma \frac{s}{\sqrt{n}}, \qquad (12.39)$$

где $\Delta = t_\gamma \frac{s}{\sqrt{n}}$ — предельная ошибка выборочной средней (*ошибка выборки*), s — «исправленное» среднее квадратическое отклонение, $a = \bar{x}_г$ и $\bar{x}_в = \bar{x}$ — генеральная и выборочная средние соответственно.

В более общем случае выбор статистики и формулы предельной погрешности зависит от вида и объема выборки (табл. 12.1). Доверительные интервалы для математического ожидания (генеральной средней $\bar{x}_в \to a$) — формула (12.40), для вероятности (доли $w \to p$) — формула (12.41):

$$\bar{x}_в - \Delta_{\bar{x}} < \bar{x}_г < \bar{x}_в + \Delta_{\bar{x}}, \qquad (12.40)$$

$$w - \Delta_w < p < w + \Delta_w. \qquad (12.41)$$

В таблице 12.1 использованы обозначения:

1) $t_\gamma = t$ — квантиль распределения, соответствующая уровню значимости α (или доверительной вероятности $\gamma = 1 - \alpha$):

а) при $n \geq 30$, $t = u_{\alpha/2}$ — квантиль нормального закона распределения (приложение 1);

б) при $n < 30, t$ — квантиль распределения Стьюдента с $k = n - 1$ степенями свободы для двусторонней области (приложение 3);

2) σ^2 — выборочная дисперсия:

а) при $n \geq 30$, $\sigma^2 = \frac{\sum(x_i-\bar{x})^2 n_i}{n}$;

б) при $n < 30$ вместо σ^2 берут «исправленную» выборочную дисперсию $s^2 = \frac{\sum(x_i-\bar{x})^2 n_i}{n-1}$;

3) $pq = w(1-w)$ — дисперсия относительной частоты в схеме повторных независимых испытаний;

4) N — объем генеральной совокупности;

5) n — объем выборки;

6) $\overline{\sigma^2}$ — средняя арифметическая групповых дисперсий (внутригрупповая дисперсия);

7) $\overline{pq}$ — средняя арифметическая дисперсий групповых долей;

8) $\delta^2_{\text{м.с.}}$ — межсерийная дисперсия;
9) $pq_{\text{м.с.}}$ — межсерийная дисперсия доли;
10) N_c — число серий в генеральной совокупности;
11) n_c — число отобранных серий (объем выборки);
12) Δ — предельная ошибка выборки.

Таблица 12.1

Определение предельной ошибки и необходимого объема выборки для различных способов отбора

		Выборка					
		Собственно-случайная		Типическая		Серийная	
		повторная	бесповторная	повторная	бесповторная	повторная	бесповторная
Предельная ошибка, Δ	средней	$t\cdot\sqrt{\frac{\sigma^2}{n}}$	$t\cdot\sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}$	$t\cdot\sqrt{\frac{\overline{\sigma^2}}{n}}$	$t\cdot\sqrt{\frac{\overline{\sigma^2}}{n}\left(1-\frac{n}{N}\right)}$	$t\cdot\sqrt{\frac{\delta^2_{м.с}}{n_c}}$	$t\cdot\sqrt{\frac{\delta^2_{м.с}}{n_c}\left(1-\frac{n_c}{N_c}\right)}$
	доли	$t\cdot\sqrt{\frac{pq}{n}}$	$t\cdot\sqrt{\frac{pq}{n}\left(1-\frac{n}{N}\right)}$	$t\cdot\sqrt{\frac{\overline{pq}}{n}}$	$t\cdot\sqrt{\frac{\overline{pq}}{n}\left(1-\frac{n}{N}\right)}$	$t\cdot\sqrt{\frac{pq_{мс.}}{n_c}}$	$t\cdot\sqrt{\frac{pq_{мс.}}{n_c}\left(1-\frac{n_c}{N_c}\right)}$
Необходимая численность, n	средней	$\frac{t^2\sigma^2}{\Delta^2}$	$\frac{t^2\sigma^2 N}{t^2\sigma^2+\Delta^2 N}$	$\frac{t^2\overline{\sigma^2}}{\Delta^2}$	$\frac{t^2\overline{\sigma^2}N}{t^2\sigma^2+\Delta^2 N}$	$\frac{t^2\delta^2_{м.с}}{\Delta^2}$	$\frac{t^2\delta^2_{м.с}N_c}{t^2\delta^2_{м.с}+\Delta^2 N_c}$
	доли	$\frac{t^2pq}{\Delta^2}$	$\frac{t^2Npq}{t^2pq+\Delta^2 N}$	$\frac{t^2\overline{pq}}{\Delta^2}$	$\frac{t^2N\overline{pq}}{t^2\overline{pq}+\Delta^2 N}$	$\frac{t^2pq_{м.с}}{\Delta^2}$	$\frac{t^2pq_{м.с}N_c}{t^2pq_{м.с}+\Delta^2 N_c}$

Замечание. 1) «Средняя ошибка типической выборки» ≤ «средняя ошибка механической выборки» ≤ «средняя ошибка собственно случайной выборки», поэтому на практике при расчете оценок механической выборки используют формулы собственно-случайной повторной выборки.

2) В случае повторного отбора, описывающегося биномиальным законом распределения, дисперсия средней арифметической:

$$D(\bar{x})=\frac{\sigma^2}{n}.$$

Для бесповторного отбора, учитывая формулу дисперсии для гипергеометрического закона распределения (3.29) и то, что при большом объеме генеральной совокупности

$$\frac{N-n}{N-1}\to\frac{N-n}{N}=1-\frac{n}{N},$$

получим формулу дисперсии

$$D(\bar{x})=\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right). \blacksquare$$

2. Рассмотрим *построение доверительного интервала для дисперсии нормально распределенной генеральной совокупности*.

2.1. Пусть выборка $x, x_2, \dots, x_n$ состоит из независимых нормально распределенных случайных величин с параметрами a_0 и σ, причем a_0 *известно, а величину* σ *оцениваем по выборке* с помощью статистики:

$$S_0^2 = \frac{\sum_i (x_i - \bar{x})^2}{n},$$

перепишем ее в виде

$$S_0^2 = \frac{\sigma^2}{n} \sum_i \left(\frac{x_i - \bar{x}}{\sigma} \right)^2 = \frac{\sigma^2}{n} \chi_n^2,$$

χ_n^2 — распределение хи-квадрат Пирсона с n степенями свободы.

Отсюда $\chi_n^2 \xrightarrow{d} \frac{nS_0^2}{\sigma^2}$.

Следовательно, можно определить границы доверительного интервала для $\frac{nS_0^2}{\sigma^2}$, опираясь на значения квантилей распределения хи-квадрат c_1 и c_2:

$$P(\chi_n^2 \le c_1) = 1 - \frac{\alpha}{2},$$
$$P(\chi_n^2 \ge c_2) = \frac{\alpha}{2}.$$

Имеем

$$\frac{nS_0^2}{c_2} < \sigma^2 < \frac{nS_0^2}{c_1}. \qquad (12.42)$$

Это и есть искомый доверительный интервал.

2.2. Пусть выборка $x, x_2, \dots, x_n$ состоит из независимых нормально распределенных случайных величин *с неизвестными параметрами* a и σ. Из теоремы 3 следует, что

$$\frac{nS^2}{\sigma^2} \xrightarrow{d} \chi_{n-1}^2.$$

Проводя рассуждения аналогичные выводу (12.42), с доверительной вероятностью $\gamma = 1 - \alpha$ получим доверительный интервал

$$\frac{nS_0^2}{c_2} < \sigma^2 < \frac{nS_0^2}{c_1}, \qquad (12.43)$$

где $P(\chi_{n-1}^2 \le c_1) = 1 - \frac{\alpha}{2}$, $P(\chi_{n-1}^2 \ge c_2) = \frac{\alpha}{2}$.

Пример 12.1. При уровне доверительной вероятности 0,95 определить доверительный интервал для средней численности работников на 100 га сельскохозяйственных угодий в примере 11.2, учитывая, что проводилась 10%-ная случайная бесповторная выборка. При тех же условиях найти необходимый объём выборки для уменьшения предельной ошибки в два раза.

Решение. Средняя численность работников на 100 га сельскохозяйственных угодий составляет 5,4. Доверительный интервал для средней определяется по формуле

$$\bar{x}_в - \Delta_{\bar{x}_в} < \bar{x}_г < \bar{x}_в + \Delta_{\bar{x}_в},$$

где $\bar{x}_в$ — выборочная средняя; $\bar{x}_г$ — средняя генеральной совокупности; $\Delta_{\bar{x}}$ — предельная ошибка выборки для средней.

Предельная ошибка выборки при случайном бесповторном отборе определяется по формуле (табл. 12.1):

$$\Delta_{\bar{x}} = t\sigma(\bar{x}) = t\sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)},$$

где t — квантиль нормального закона распределения, при $\gamma = 0{,}95$, $t = 1{,}96$ (прил. 1); N — объем генеральной совокупности, где $n/N = 0{,}1, n = 60$, следовательно, $N = 600$; σ^2 — выборочная оценка дисперсии генеральной совокупности; так как объем выборочной совокупности $n = 60 > 30$, то в качестве выборочной дисперсии используем $\sigma^2 = 2{,}483$.

Имеем $\Delta_{\bar{x}} = 1{,}96\sqrt{\frac{2{,}483}{60}\left(1 - \frac{60}{600}\right)} = 0{,}38.$

Значит, с доверительной вероятностью γ=0,95 можно утверждать, что средняя численность работников на 100 га сельскохозяйственных угодий во всей совокупности хозяйств находится в границах

$$\bar{x} \pm \Delta_{\bar{x}} = 5{,}4 \pm 0{,}38,$$ то есть от 5,02 до 5,78.

Необходимый объем выборки, чтобы предельная ошибка не превышала $0{,}5\Delta_{\bar{x}} = 0{,}19$, при заданном уровне доверительной вероятности в случае случайного бесповторного отбора определяется по формуле (табл. 12.1):

$$n = \frac{t^2\sigma^2 N}{t^2\sigma^2 + \Delta^2 N}.$$

Следовательно,

$$n = \frac{1{,}96^2 \cdot 2{,}483^2 \cdot 600}{1{,}96^2 \cdot 2.483^2 + 0{,}19^2 \cdot 600} = 184.$$

Значит, для уменьшения предельной ошибки в два раза объем совокупности необходимо увеличить в три раза.

Темы (вопросы) для самоконтроля

1. Сущность выборочного метода.
2. Генеральная совокупность.
3. Способы случайного отбора.
4. Выборочная совокупность и ее свойства.
5. Точечные оценки параметров генеральной совокупности.
6. Метод моментов.
7. Метод максимального правдоподобия.
8. Вероятностная бумага, квантильные оценки, метод наименьших квадратов, робастные методы оценивания.
9. Выборочная средняя и выборочная дисперсия.
10. Интервальные оценки параметров генеральной совокупности.
11. Доверительный интервал для средней арифметической.
12. Доверительный интервал для доли.
13. Доверительный интервал для дисперсии.
14. Необходимая численность выборки.

Глава 13
Проверка статистических гипотез

Выбор. Проблемой
является выбор.
(кф. «Матрица»)

13.1. Понятие и виды статистических гипотез

Статистической гипотезой называется всякое высказывание о генеральной совокупности, проверяемое по выборке. Она может касаться вида неизвестного распределения, отдельных параметров распределений, связей между случайными величинами и т. п. Например, совокупность малых предприятий по выручке от реализации продукции распределяется по нормальному закону, совокупность семей по среднемесячному доходу на члена семьи распределена по логарифмически нормальному закону. Известно, что совокупность безработных территориального образования распределяется по нормальному закону, выдвигается гипотеза, что средний стаж работы безработного составляет 10 лет. Если сравниваются две или более генеральных совокупностей, имеющих один и тот же закон распределения, то могут быть выдвинуты гипотезы о равенстве средних значений или дисперсий этих совокупностей.

Статистические гипотезы подразделяются на:

а) *параметрические* — это гипотезы, сформулированные относительно параметров распределения известного вида (среднего значения, дисперсии и т. д.);

б) *непараметрические* — это гипотезы, сформулированные относительно вида распределения и использующие только *частоты* и *ранги* (например, определение по выборке нормальности генеральной совокупности).

Процесс использования выборки для проверки гипотезы называется *статистическим доказательством*. Выдвигаемая гипотеза называется *нулевой или основной* H_0. Наряду с нулевой гипотезой рассматривается ей противоположная гипотеза, которая называется *альтернативной*, или *конкурирующей*, и обозначается H_1.

Например, H_0: $M(X) = 1$, математическое ожидание случайной величины в генеральной совокупности, распределенной по показательному закону, равно единице. Тогда конкурирующая гипотеза может иметь вид H_1: $M(X) > 1$, или $M(X) < 1$, или $M(X) \neq 1$ (математическое ожидание больше 1, или меньше 1, или не равно 1). Выдвигаемая гипотеза может содержать одно или несколько предположений. Если *параметрическая гипотеза* содержит одно утверждение, то она называется *простой*, а если множество утверждений — то *сложной*. Простой будет гипотеза, что среднее время безотказной работы холодильника определенной марки составляет 45 тыс. часов, а сложной — что среднее время безотказной работы составит менее 45 тыс. часов. Обычно рассматривают следующие сочетания нулевой и альтернативной гипотез: простая, простая; простая, сложная; сложная, сложная.

Так как проверка статистических гипотез осуществляется по выборочным данным, то нельзя быть уверенным об истинности или ложности выдвинутой гипотезы. Выбор между гипотезами H_0 и H_1 может сопровождаться ошибками двух родов. Ошибка первого рода заключается в том, что будет отвергнута верная нулевая гипотеза. Ошибка *первого рода* α означает вероятность принятия H_1, если верна гипотеза H_0, т. е. $\alpha = P(H_1/H_0)$. Ошибка второго рода состоит в том, что будет принята неправильная гипотеза. Ошибка *второго рода* означает вероятность принятия H_0, если верна гипотеза H_1: $\beta = P(H_0/H_1)$. Существует правильное решение двух родов: $P(H_0/H_0) = 1 - \alpha$ и $P(H_1/H_1) = 1 - \beta$ (табл. 13.1).

Таблица 13.1

Ошибки первого и второго рода

Принятая гипотеза	H_0	H_1
H_0 — верна	$P(H_0/H_0) = 1 - \alpha$	$P(H_1/H_0) = \alpha$
H_0 — не верна	$P(H_0/H_1) = \beta$	$P(H_1/H_1) = 1 - \beta$

Правило, по которому принимается решение о том, верна или не верна гипотеза H_0, называется *критерием*, где:

$\alpha = P(H_1/H_0)$ — *уровень значимости критерия*;

$M = 1 - \beta = P(H_1/H_1)$ — *мощность критерия*.

Статистическим критерием (*тестом, решающим правилом*) называют случайную величину K, с помощью которой принимают решение о принятии или отклонении H_0.

Для проверки параметрических гипотез используют *критерии значимости*, основанные на статистиках: u, χ^2, t, F (прил. 5–7). Непараметрические гипотезы проверяют с помощью *критериев согласия*, использующих статистики: χ^2 Пирсона, λ Колмогорова, Смирнова, ω^2 — Мизеса и т. д. Так как критерий K является случайной величиной, то он принимает множество возможных значений соответствующего распределения.

После выбора статистического критерия, все множество его значений разбивается на два непересекающихся подмножества. Одно подмножество содержит значения критерия, при которых нулевая гипотеза принимается. Это подмножество называется областью допустимых значений критерия или областью принятия гипотезы H_0. Второе подмножество содержит значения критерия, при которых нулевая гипотеза отвергается и называется критической областью. Точка, разделяющая эти подмножества, называется критической точкой, которая обозначается $K_{\text{кр}}$. По данным выборки находится наблюдаемое значение критерия. Затем наблюдаемое значение критерия сравнивается с критическим значением. Если наблюдаемое значение критерия принадлежит критической области, то нулевую гипотезу отвергают, а если принадлежит области допустимых значений — то гипотезу принимают.

Например, $H_0: M(X) = 10$. В зависимости от вида альтернативной гипотезы рассматриваются три случая.

1. Если конкурирующая гипотеза имеет вид $H_1: M(X) \neq 10$, то в этом случае рассматривают двустороннюю критическую область и используют функцию плотности вероятности $f\ (K/H_0)$, для определения соответствующих квантилей (границ области принятия гипотезы — левой $(K_{1-\alpha/2})$ и правой $(K_{\alpha/2})$). Площадь под криволинейной трапецией плотности распределения слева от $K_{1-\alpha/2}$ и справа от $K_{\alpha/2}$ равна $\alpha/2$.

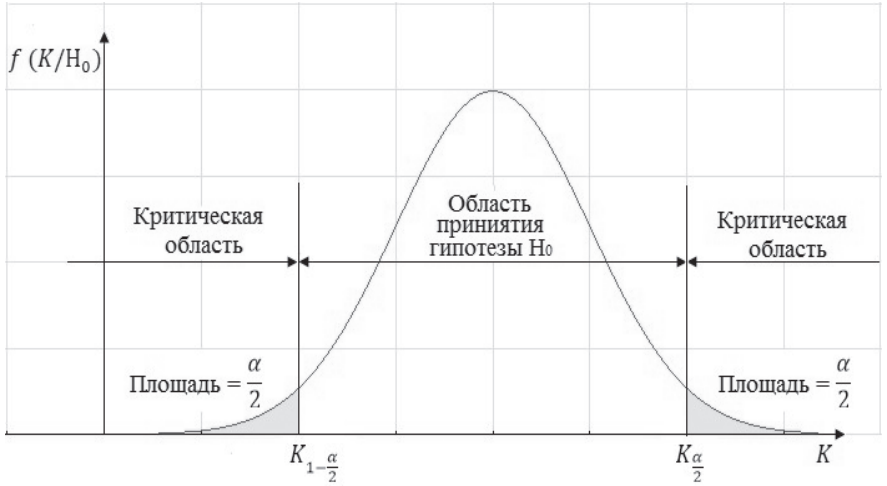


Рис. 13.1 — Двусторонняя критическая область

Общая площадь, ограниченная криволинейной трапецией плотности распределения, квантилями и осью абсцисс, равна $(1\ -\ \alpha)$ (рис. 13.1):

$$P\left(K_{1-\frac{\alpha}{2}} < K < K_{\frac{\alpha}{2}}\right) = 1 - \alpha. \qquad (13.1)$$

2. Если конкурирующая гипотеза, $H_1: M(X) > 10,$ то рассматривается правосторонняя критическая область (площадь под криволинейной трапецией справа от K_α) равна α (рис. 13.2):

Рис. 13.2 — Правосторонняя критическая область

$$P(K > K_\alpha) = \int_{K_\alpha}^{+\infty} f(K/H_0)dK = \alpha. \quad (13.2)$$

3. Если $H_1: M(X) < 10$, то рассматривается левосторонняя критическая область (площадь под криволинейной трапецией слева от $K_{1-\alpha}$ равна α) (рис. 13.3):

$$P(K < K_\alpha) = \int_{-\infty}^{K_{1-\alpha}} f(K/H_0)dK = \alpha. \quad (13.3)$$

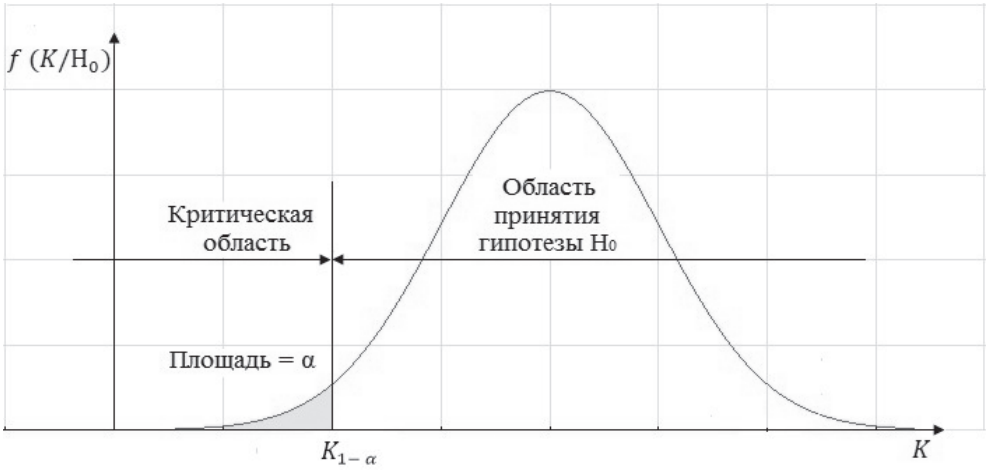


Рис. 13.3 — Левосторонняя критическая область

Алгоритм проверки статистических гипотез сводится к следующим пунктам.

1. Располагая выборочными данными $(x_1, x_2, \dots, x_n)$, формулируется нулевая гипотеза H_0 и конкурирующая гипотеза H_1.

2. Задаются уровнем значимости α (обычно принимают $\alpha = 0{,}1$; 0,05; 0,01; 0,001).

3. Рассматривается выборочная статистика критерия K, обычно одна из перечисленных ниже:

u — нормальное распределение;

χ^2 — распределение Пирсона (хи-квадрат);

t — распределение Стьюдента;

F — распределение Фишера — Снедекора.

4. На основании выборки $(x_1, x_2, \dots, x_n)$ определяется наблюдаемое значение критерия (статистики) K (например, прил. 5–7).

5. В зависимости от вида альтернативной гипотезы выбирается по соответствующей таблице квантиль критерия для двусторонней $\left(K_{1-\frac{\alpha}{2}}$ и $K_{\frac{\alpha}{2}}\right)$ или односторонней области ($K_{1-\alpha}$ или K_α) (прил. 1–4).

6. Сравнивается наблюдаемое значение критерия с критическим значением и формулируется вывод. Если наблюдаемое значение критерия попадает в критическую область, то гипотеза H_0 отвергается. В противном случае принимается гипотеза H_0 и считается, что H_0 не противоречит выборочным данным (при этом существует возможность ошибки с вероятностью равной α).

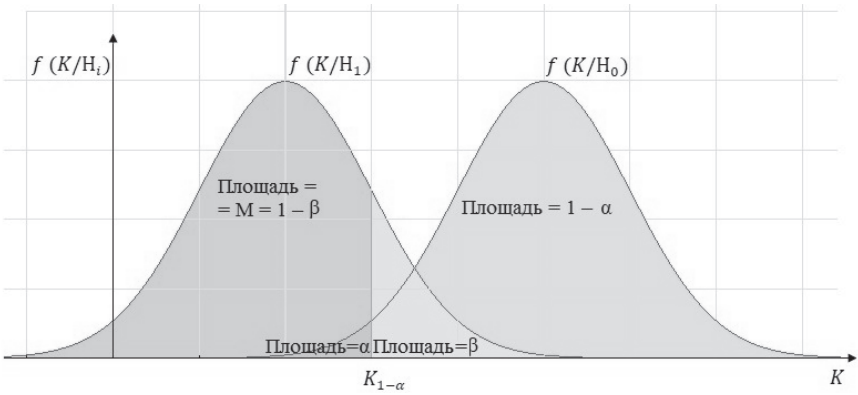


Рис. 13.4 — Геометрическая интерпретация вероятностей ошибок первого, второго рода и мощности критерия K для левосторонней критической области

Пусть рассматривается левосторонняя критическая область (рис. 13.4), тогда, передвигая квантиль $K_{1-\alpha}$, влево мы будем увеличивать вероятность ошибки II рода (β). Ошибка второго рода зависит от числа наблюдений (n), уровня значимости (α), вида альтернативной гипотезы (H_1), критерия (K). Практическая достоверность доказательства истинности гипотез доставляется бесконечным объемом изучаемой выборки. Таким образом, увеличение объема выборки ведет к уменьшению ошибок первого и второго рода.

Чем меньше уровень значимости, тем меньше будет вероятность отклонить верную нулевую гипотезу. Уровень значимости обычно задается заранее. Чем больше мощность критерия, тем меньше вероятность ошибки второго рода. Желательно минимизировать вероятности ошибок первого и второго рода, что невозможно при заданном объеме выборки. Обычно задают границу вероятности отклонения нулевой гипотезы, когда она верна, и при этом условии минимизируют вероятность ошибки второго рода (β), обращая внимание на мощность критерия.

Рассматриваются два варианта статистических критериев значимости.

1) *Критерии* (*тесты*) *с заранее определенным уровнем значимости* (вероятностью ошибки первого рода), которые позволяют принять решение «отклонить гипотезу H_0», либо «нет оснований, опираясь на результаты выборки, для отклонения гипотезы H_0». Вероятность ошибки второго рода неизвестна (вероятность принятия ложной гипотезы H_0), поэтому нельзя принять решение о том, что «гипотеза H_0 верна».

2) *Критерии с наименьшим уровнем значимости p-значение* (*p-value*). В статистических (эконометрических) пакетах при оценке значимости параметров распределений или статистических моделей (например, $H_0: \theta = 0$) выдается $p-$значение, указывающее наименьший уровень значимости, при котором нулевая гипотеза отвергается (в предположении, что нулевая гипотеза верна), при больших значениях нулевая гипотеза принимается. Этот подход соответствует идеологии Р. Фишера о том, что уровень значимости нельзя фиксировать заранее. Вероятность *p-value* соответствует вероятности получить эмпирические данные при условии того, что гипотеза H_0 верна:

$$p - value\ = P(x_1, x_2, \dots, x_n\ /\ H_0).$$

При проверке статистических гипотез часто решаются следующие *задачи проверки гипотез: сравнения* (средних, дисперсий), *согласия* (согласованность эмпирических данных и теоретического распределения), *однородности* нескольких выборок, *независимости* двух и более выборок, *случайности* (например, при проверке качества датчика случайных чисел).

Замечание. 1. $p-$ начение *является вероятностной мерой доказательства* того, что нулевая гипотеза не верна, но не является вероятностной мерой ошибки при принятии нулевой гипотезы. Р. Фишер предлагал рассматривать следующую шкалу измерений доказательства против гипотезы H_0.

$p-$ значение	Уровень вероятности доказательства против гипотезы H_0
0,1	*Borderline* — граничный
0,05	*Moderate* — умеренный
0,025	*Substantial* — существенный
0,01	*Strong* — сильный
0,001	*Overwhelming* — очень сильный

2. Следует отметить, что возможность принятия гипотезы происходит из принципа невозможности наступления маловероятных событий. Те события, вероятность которых близка к 1, принимаются как достоверные. Возникает проблема выбора уровня риска (уровня значимости α). В одних случаях возможно пренебрегать событиями, если $p < 0{,}05$, в других нельзя пренебрегать событиями, которые могут появиться с $p = 0{,}001$ (разрушение сооружений, транспортных средств и т. д.).

3. Критерии значимости строятся на основании идеологии построения доверительных интервалов для выборочных статистик.

4. Для оценки во сколько раз вероятнее получить имеющиеся данные при наличии различий (гипотеза H_1), *чем при их отсутствии* (гипотеза H_0), предлагается использовать байесовский фактор

$$BF_{10} = \frac{P(x_1, x_2, \dots, x_n \,/\, H_1)}{P(x_1, x_2, \dots, x_n \,/\, H_0)},$$

позволяющий построить шкалу свидетельств в пользу H_1, против H_0 или наоборот $BF_{01} = 1/BF_{10}$ (см. гл. 19–20 в [53]). ■

13.2. Проверка гипотезы о среднем значении нормально распределенной генеральной совокупности

Пусть имеется некоторая генеральная совокупность, которая распределена по нормальному закону в отношении признака *X*. Математическое ожидание непосредственно неизвестно, но можно предположить, что оно равно некоторому значению. Другими словами, предполагается, что *генеральная средняя a, заранее неизвестная, равна некоторому значению a_0. Дисперсия генеральной совокупности известна и равна σ^2*. Для проверки данной гипотезы из генеральной

совокупности взята случайная выборка объема n, по которой найдена выборочная средняя $\bar{X}$. Так как выборочная средняя является несмещенной оценкой генеральной средней, то выдвинем нулевую гипотезу H_0: $\bar{X} = a_0$. Если нулевая гипотеза неверна, то выборочная средняя может быть больше или меньше значения генеральной средней.

а) Проверим нулевую гипотезу H_0: $\bar{X} = a_0$, при H_1: $\bar{X} \neq a_0$.

В качестве критерия проверки нулевой гипотезы используется стандартное нормальное распределение, так как в силу асимптотической теории из центральной предельной теоремы следует формула (II.5) $\bar{X} \to N(a_0, \sigma^2(\bar{X}))$, где

$$\sigma(\bar{X}) = \sqrt{\frac{\sigma^2}{n}},$$

поэтому стандартизированная случайная величина U подчиняется стандартному нормальному закону распределения:

$$u = \frac{\bar{X} - a_0}{\frac{\sigma}{\sqrt{n}}} \to N(0, 1).$$

Наблюдаемое значение критерия находится по формуле

$$u_{\text{н.}} = \frac{\bar{X} - a_0}{\frac{\sigma}{\sqrt{n}}} = \frac{(\bar{X} - a_0)\sqrt{n}}{\sigma}. \qquad (13.4)$$

Критическое значение критерия находится по таблице распределения Ф(x) (прил. 1), для двухсторонней критической области, исходя из равенства

$$\Phi\left(u_{\text{кр.}}\right) = \frac{1-\alpha}{2}. \qquad (13.5)$$

Затем сравнивается наблюдаемое значение критерия с критическим.

Если $|u_{\text{н.}}| > u_{\text{кр.}}$, то нулевая гипотеза отвергается, выборочная средняя значимо отличается от генеральной средней.

Если $|u_{\text{н.}}| < u_{\text{кр.}}$, то нулевая гипотеза не отвергается, нет оснований отвергнуть нулевую гипотезу.

Пример 13.1. Урожайность озимой пшеницы определенного сорта по совокупности крестьянских хозяйств распределяется по нормальному закону с известным средним квадратическим отклонением σ = 6,4 ц/га и генеральной средней $\bar{X}_{\text{г}} = 60{,}0$ ц/га.

По выборочной совокупности 50 крестьянских хозяйств найдена выборочная средняя урожайность, составившая 63 ц/га. При уровне значимости α = 0,05 проверить нулевую гипотезу H_0: $\bar{X} = \bar{X}_{\text{г}} = 60$, при конкурирующей гипотезе H_1: $\bar{X} \neq 60$.

Решение. Необходимо рассмотреть критерий $K = u$, где

$$u_{\text{н.}} = \frac{(\bar{X} - \bar{X}_{\text{г}})\sqrt{n}}{\sigma} = \frac{(63{,}0 - 60{,}0)\sqrt{50}}{6{,}4} = 3{,}31.$$

а) По условию конкурирующая гипотеза имеет вид $\bar{X} \neq 60{,}0$, поэтому критическая область двусторонняя. Найдем критическую точку из равенства $\Phi(u_{\text{кр},\alpha/2}) = (1-\alpha)/2 = (1-0{,}05)/2 = 0{,}475$, используя приложение 1: $u_{\text{кр.}} = 1{,}96$.

Так как $|u_{\text{н.}}| > u_{\text{кр.}}$, следует отклонить нулевую гипотезу, то есть выборочная средняя и гипотетическая генеральная средняя различаются значимо.

б) Проверим нулевую гипотезу H_0: $\bar{X} = a_0$, при H_1: $\bar{X} > a_0$.

При конкурирующей гипотезе H_1: $\bar{X} > a_0$ критическая область является правосторонней. Критическую точку находят из равенства

$$\Phi(u_{\text{кр}}) = \frac{1-2\alpha}{2}. \quad (13.6)$$

Если $u_{\text{н.}} > u_{\text{кр.}}$, то выдвинутая нулевая гипотеза отвергается, выборочная средняя значимо отличается от генеральной средней.

Если $u_{\text{н.}} < u_{\text{кр.}}$, то нулевая гипотеза принимается, нет оснований отвергнуть нулевую гипотезу.

По условию примера 13.1 конкурирующая гипотеза имеет вид H_1: $\bar{X} > 60{,}0$, поэтому критическая область правосторонняя. Найдем критическую точку из равенства:

$$\Phi(u_{\text{кр},\alpha/2}) = (1–2\alpha)/2 = (1–2\cdot 0{,}05)/2 = 0{,}45.$$

Согласно приложению 1, $u_{\text{кр}} = 1{,}645$. Так как $u_{\text{н}} > u_{кр}$, то следует отклонить нулевую гипотезу.

в) Проверим нулевую гипотезу H_0: $\bar{X} = a_0$, при H_1: $\bar{X} < a_0$.

Наблюдаемое значение находится по формуле (13.4). Критическую точку находят из равенства (13.6), учитывая, что критическая точка левосторонняя.

Если $u_{\text{н.}} < u_{\text{кр.}}$, то выдвинутая нулевая гипотеза отвергается, выборочная средняя значимо отличается от генеральной средней.

Если $u_{\text{н.}} > u_{\text{кр.}}$, то нулевая гипотеза принимается, нет оснований отвергнуть нулевую гипотезу.

Учитывая, что $u_{\text{кр. левост.}} = - u_{\text{кр. правост.}}$, то вывод можно формулировать, как и в пункте б).

Если по условию примера 13.1 средняя выборочная урожайность составила 58 ц/га, то конкурирующая гипотеза имеет вид H_1:
$\bar{X} < 60$, критическая область левосторонняя.
При $\alpha = 0{,}05$, $u_{\text{кр.}} = -1{,}64$.

$$u_{\text{н.}} = \frac{(\bar{X}-\bar{X}_{\text{г}})\sqrt{n}}{\sigma} = \frac{(58{,}0-60{,}0)\sqrt{50}}{6{,}4} = -2{,}21.$$

Так как $u_{\text{н.}} < u_{\text{кр.}}$, то выдвинутая нулевая гипотеза отвергается, выборочная средняя значимо отличается от генеральной средней.

Пусть *генеральная совокупность распределена по нормальному закону, но числовые характеристики непосредственно неизвестны,* несмещенными их оценками служат выборочная средняя и «исправленная» выборочная дисперсия s^2 (разд. 12.4).

Тогда, согласно теореме 4, в качестве критерия проверки нулевой гипотезы используется t-распределение Стьюдента. Наблюдаемое значение критерия находится по формуле

$$t_{\text{н.}} = \frac{\bar{X}-a_0}{s/\sqrt{n}} = \frac{(\bar{X}-a_0)\sqrt{n}}{s}. \quad (13.7)$$

Критическая область строится в зависимости от вида конкурирующей гипотезы, как рассмотрено выше. При заданном уровне значимости α и числе степеней свободы $k = n - 1$ по таблице распределения Стьюдента (прил. 3) находится критическое значение критерия для односторонней или двухсторонней

критической области. Сравнивая наблюдаемое значение критерия с критическим значением, формулируется вывод.

Пример 13.2. Ожидаемая урожайность подсолнечника может составить 30 ц с 1 га. В результате посева на 6 делянках одинаковой площади была получена урожайность с 1 га: 31,8; 28,8; 29,4; 30,2; 32,2; 30,6. При уровне значимости 0,05 проверить гипотезу о равенстве средней урожайности подсолнечника 30 ц/га.

Решение. Считая распределение урожайности по участкам нормально распределенной случайной величиной, проверим нулевую гипотезу H_0: $\bar{X} = a_0$, при H_1: $\bar{X} \neq a_0$. По условию $a_0 = 30$. Дисперсия генеральной совокупности неизвестна, выборка малая по объему, найдем «исправленную» выборочную дисперсию, для чего составим таблицу 13.2.

Средняя урожайность подсолнечника составила 30,5 ц/га. По участкам урожайность в среднем колебалась в границах 30,5±1,3 ц/га, т. е. от 29,2 до 31,8 ц/га.

$$t_{\text{н.}} = \frac{(\bar{X}-a_0)\sqrt{n}}{s} = \frac{(30{,}5-30{,}0)\cdot\sqrt{6}}{1{,}325} = 0{,}92.$$

По таблице *t*-распределения Стьюдента при уровне значимости α = 0,05 и числе степеней свободы $k = n - 1 = 6 - 1 = 5$, $t_{\text{кр.}}$ = 2,57.

Таблица 13.2

Урожайность подсолнечника с 1 га, ц

№ п/п	Урожайность, ц/га	$x_i - \bar{X}$	$(x_i - \bar{X})^2$
1	31,8	1,3	1,69
2	28,8	−1,7	2,89
3	29,4	−1,1	1,21
4	30,2	−0,3	0,09
5	32,2	1,7	2,89
6	30,6	0,1	0,01
Итого	183,0	0,0	8,78

$$\bar{X} = \frac{\sum x_i}{n} = \frac{183{,}0}{6} = 30{,}5;\ s^2 = \frac{\sum(x_i-\bar{X})^2}{n-1} = \frac{8{,}78}{6-1} = 1{,}756;$$

$$s = \sqrt{1{,}756} = 1{,}325.$$

Сравниваем наблюдаемое значение критерия с критическим значением. Так как $t_{\text{н.}} < t_{\text{кр.}}$, то нулевая гипотеза принимается, средняя урожайность подсолнечника на всех участках может составить 30 ц/га.

Гипотеза может быть проверена с использованием доверительного интервала средней арифметической.

$$\bar{X} - t_{кр}\frac{s}{\sqrt{n}} \le a_0 \le \bar{X} + t_{кр}\frac{s}{\sqrt{n}},$$

$$30{,}5 - 2{,}57\cdot\frac{1{,}325}{\sqrt{6}} \le a_0 \le 30{,}5 + 2{,}57\cdot\frac{1{,}325}{\sqrt{6}},\ 29{,}1 \le a_0 \le 31{,}9.$$

С доверительной вероятностью 0,95 можно утверждать, что средняя урожайность на всех участках может находиться в пределах от 29,1 до 31,9 ц/га. Так как этот интервал покрывает значение a_0 = 30, то нулевая гипотеза принимается.

Проверим нулевую гипотезу H_0: $\bar{X} = a_0$, при H_1: $\bar{X} > a_0$.

Критическая область является правосторонней. При уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $k = n - 1 = 6 - 1 = 5$, $t_{\text{кр.}} = 2{,}01$. Так как $t_{\text{н.}} < t_{\text{кр}}$, то нулевая гипотеза принимается.

13.3. Проверка гипотезы о числовом значении генеральной доли

Производится n независимых испытаний, вероятность появления события A в каждом из n независимых испытаний постоянна и равна p, которая заранее неизвестна. Предполагается, что число испытаний достаточно велико. Необходимо проверить гипотезу, что неизвестная вероятность p равна значению p_0. Долю единиц, обладающих данным признаком в выборочной совокупности, т. е. частость, обозначим w. Причем $w = \frac{k}{n}$, где k — число единиц, обладающих данным признаком в выборочной совокупности. Требуется определить, значимо или нет различаются относительная частота и генеральная доля.

В силу локальной и интегральной теорем Муавра — Лапласа известно, что при достаточно большом числе повторных независимых испытаний по схеме Бернулли ($n \to \infty$)

$$u = \frac{k - np}{\sqrt{npq}} \to N(0, 1),$$

следовательно,

$$u = \frac{\frac{k}{n} - p}{\sqrt{\frac{pq}{n}}} \to N(0, 1),$$

поэтому относительная частота асимптотически имеет нормальный закон распределения с математическим ожиданием $M\left(\frac{k}{n}\right) = p$ и средним квадратическим отклонением $\sigma\left(\frac{k}{n}\right) = \sqrt{\frac{pq}{n}}$. Следовательно, в качестве критерия проверки гипотезы можно использовать нормально распределенную случайную величину. Если выдвигается нулевая гипотеза H_0: $p = p_0$, при H_1: $p \neq p_0$, то наблюдаемое значение критерия u находится по формуле

$$u_{\text{н.}} = \frac{w - p_0}{\sqrt{\frac{pq}{n}}}. \tag{13.8}$$

Критическое значение критерия находится по таблице распределения $\Phi(x)$, представленной в приложении 1, для двухсторонней критической области, исходя из равенства

$$\Phi\left(u_{\text{кр.}}\right) = \frac{1 - \alpha}{2}. \tag{13.9}$$

Затем сравнивается наблюдаемое значение критерия с критическим.

Если $|u_{\text{н.}}| > u_{\text{кр.}}$, то нулевая гипотеза отвергается, выборочная доля значимо отличается от генеральной доли.

Если $|u_{\text{н.}}| < u_{\text{кр.}}$, то нулевая гипотеза принимается, нет оснований отвергнуть нулевую гипотезу.

Если конкурирующая гипотеза имеет вид H_1: $p > p_0$ или H_1: $p < p_0$, то критическую точку находят для правосторонней или левосторонней критической области исходя из равенства

$$\Phi\left(u_{\text{кр.}}\right) = \frac{1-2\alpha}{2}. \qquad (13.10)$$

Пример 13.3. Руководство некоторой политической партии утверждает, что на предстоящих выборах в парламент за кандидатов партии проголосует 20,0% избирателей. Независимая социологическая служба провела опрос 600 случайно выбранных будущих избирателей, из которых за кандидатов этой партии отдадут свои голоса 100. Можно ли при уровне значимости 0,05 согласиться с мнением руководства партии?

Решение. По условию p_0 = 0,2; $w = \frac{k}{n} = \frac{100}{600}$ = 0,1667. Гипотеза H_0: $p = p_0 = 0{,}2$, при альтернативной H_1: $p < 0{,}2$. Найдем наблюдаемое значение критерия по формуле (13.8).

$$u_{\text{н.}} = \frac{w-p_0}{\sqrt{\frac{pq}{n}}} = \frac{0{,}1667-0{,}2}{\sqrt{\frac{0{,}2\cdot 0{,}8}{600}}} = \ -\ 2{,}02.$$

При α = 0,05, $\Phi\left(u_{\text{кр}}\right) = \frac{1-2\alpha}{2} = \frac{1-2\cdot 0{,}05}{2} = 0{,}45$, $u_{\text{кр.}}$ = 1,645. Так как $|u_{\text{н.}}| > u_{\text{кр.}}$, то нулевая гипотеза отвергается, нет оснований утверждать, что 20% избирателей отдадут голоса за кандидатов партии.

К этому же выводу можно прийти при конкурирующей гипотезе H_1: $p \neq 0{,}2$, так как при α =0,05, $u_{\text{кр}}$ = 1,96.

13.4. Проверка гипотезы о дисперсиях двух независимых нормально распределенных генеральных совокупностей

Пусть имеются две независимые нормально распределенные генеральные совокупности (случайные величины) X_1 и X_2 , дисперсии которых σ_1^2 и σ_2^2.

Необходимо проверить нулевую гипотезу о равенстве дисперсий H_0: $\sigma_1^2 = \sigma_2^2$, при альтернативной гипотезе H_1: $\sigma_1^2 > \sigma_2^2$ или H_1: $\sigma_1^2 \neq \sigma_2^2$. Из первой совокупности образована случайная выборка объема n_1, а из второй — объема n_2. В качестве оценок σ_1^2 и σ_2^2 используются «исправленные» выборочные дисперсии

$$s_1^2 = \sigma_{1\text{выб.}}^2 \frac{n_1}{n_1-1} \text{ и } s_2^2 = \sigma_{2\text{выб.}}^2 \frac{n_2}{n_2-1}.$$

Задача сводится к проверке нулевой гипотезы о равенстве дисперсий s_1^2 и s_2^2 при конкурирующей гипотезе H_1: $s_1^2 > s_2^2$.

Отношение исправленных выборочных дисперсий в силу теоремы 4 подчиняется распределению *F* Фишера — Снедекора, поэтому гипотеза проверяется с помощью соответствующего критерия.

Наблюдаемое значение критерия находится по формуле

$$F_{\text{н.}} = \frac{s_1^2}{s_2^2}. \qquad (13.11)$$

Критическое значение F-критерия находится по таблице (прил. 4) при уровне значимости α и числе степеней свободы числителя $k_1 = n_1 - 1$ и знаменателя $k_2 = n_2 - 1$, соответственно большей и меньшей дисперсий.

Если $F_{\text{н.}} > F_{\text{кр.}}$, то нулевая гипотеза отвергается, дисперсия признака по первой совокупности значимо больше дисперсии признака по второй совокупности.

Если $F_{\text{н.}} < F_{\text{кр.}}$, то нулевая гипотеза принимается, дисперсии признака по двум совокупностям различаются не значимо.

Пример 13.4. Испытывались два сорта риса. Первый сорт высевался на 8 делянках одинаковой площади, а второй на 10 делянках. По первому сорту получена средняя урожайность 75 ц/га при среднем квадратическом отклонении 6,0 ц/га, а по второму 70 и 4,0 ц/га соответственно. При уровне значимости 0,05 проверить нулевую гипотезу H_0: $s_1^2 = s_2^2$ при H_1: $s_1^2 > s_2^2$.

Решение. По условию задачи: $\bar{X}_1$=75,0; $\sigma_1^2 = 6^2 = 36$; $\bar{X}_2$ = 70,0; σ_2^2 = 16; n_1 = 8; n_2 = 10.

$$s_1^2 = \sigma_{1\text{выб.}}^2 \frac{n_1}{n_1 - 1} = 36 \cdot \frac{8}{7} = 41{,}1; \; s_2^2 = \sigma_{2\text{выб.}}^2 \frac{n_2}{n_2 - 1} = 16 \cdot \frac{10}{9} = 17{,}8.$$

$$F_{\text{н.}} = \frac{s_1^2}{s_2^2} = \frac{41{,}1}{17{,}8} = 2{,}31.$$

По таблице F −критерия Фишера — Снедекора, $F_{\text{кр}} = 3{,}29$ при уровне значимости α = 0,05 и числе степеней свободы $k_1 = n_1 - 1 = 7$ и $k_2 = n_2 - 1 = 9$. Так как $F_{\text{н.}} < F_{\text{кр.}}$, то нулевая гипотеза принимается, т. е. различия в колеблемости урожайности двух сортов риса статистически не значимы.

Если проверяется нулевая гипотеза о равенстве дисперсий H_0: $\sigma_1^2 = \sigma_2^2$, при альтернативной гипотезе H_1: $\sigma_1^2 \neq \sigma_2^2$, то наблюдаемое значение критерия находится также по формуле (13.11). Критическое значение находится по таблице F-критерия Фишера — Снедекора, при числе степеней свободы $k_1 = n_1 - 1$, $k_2 = n_2 - 1$ и уровне значимости $\alpha/2$.

13.5. Проверка гипотезы о равенстве двух средних независимых нормально распределенных генеральных совокупностей

Пусть имеются *две независимые нормально распределенные генеральные совокупности* X_1 *и* X_2 *, дисперсии которых известны* и равны σ_1^2 и σ_2^2. Из первой генеральной совокупности образована случайная выборка объема n_1, а из второй — объема n_2. По этим выборкам найдены выборочные средние $\bar{x}_1$ и $\bar{x}_2$. Выборочные средние будем рассматривать как нормально распределенные независимые случайные величины $\bar{X}_1$ и $\bar{X}_2$, у которых $D(\bar{X}_1) = \frac{\sigma_1^2}{n_1}$, $D(\bar{X}_2) = \frac{\sigma_2^2}{n_2}$. При уровне значимости α необходимо проверить нулевую гипотезу о равенстве математических ожиданий случайных величин X_1 и X_2:

$$H_0: M(X_1) = M(X_2).$$

Выборочные средние являются несмещенными оценками генеральных средних, поэтому нулевая гипотеза будет иметь вид H_0: $M(\bar{X}_1) = M(\bar{X}_2)$ или $H_0: \bar{X}_1 = \bar{X}_2$. Выборки независимы, поэтому и средние значения независимы, при условии справедливости нулевой гипотезы:

$$M(\bar{X}_1 - \bar{X}_2) = M(\bar{X}_1) - M(\bar{X}_2) = 0, \quad (13.12)$$

$$D(\bar{X}_1 - \bar{X}_2) = D(\bar{X}_1) + D(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \quad (13.13)$$

То есть случайная величина $(\bar{X}_1 - \bar{X}_2) \to N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$.

Следовательно,

$$u = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \to N(0,1).$$

Наблюдаемое значение критерия u определяется по следующей формуле:

$$u_{\text{н.}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} . \quad (13.14)$$

Критическое значение критерия находится по таблице (прил. 1). Если конкурирующая гипотеза имеет вид $H_1: \bar{X}_1 > \bar{X}_2$ или H_1: $\bar{X}_1 < \bar{X}_2$, то критическую точку находят для правосторонней или левосторонней критической области, исходя из равенства

$$\Phi(u_{\text{кр.}}) = \frac{1-2\alpha}{2}.$$

Если конкурирующая гипотеза имеет вид $H_1: \bar{X}_1 \neq \bar{X}_2$, то критическая точка находится для двухсторонней критической области, исходя из равенства

$$\Phi(u_{\text{кр.}}) = \frac{1-\alpha}{2}.$$

Затем сравнивается наблюдаемое значение критерия с критическим.

Если $|u_{\text{н.}}| > u_{\text{кр.}}$, то нулевая гипотеза отвергается, выборочные средние значения различаются значимо.

Если $|u_{\text{н.}}| < u_{\text{кр.}}$, то нулевая гипотеза принимается, нет оснований отвергнуть нулевую гипотезу.

В большинстве практических задач дисперсии генеральных совокупностей неизвестны. Если объемы обеих выборок большие, то выборочные дисперсии мало отличаются от дисперсий генеральных совокупностей. Проверка нулевой гипотезы проводится аналогично изложенному ранее. В формуле (13.14) в качестве дисперсий σ_1^2 и σ_2^2 необходимо взять выборочные дисперсии.

Пусть имеются две независимые нормально распределенные генеральные совокупности X_1 и X_2, дисперсии которых неизвестны. Из первой генеральной совокупности образована случайная выборка объема n_1, а из второй — объема n_2. По этим выборкам найдены выборочные средние $\bar{x}_1$ и $\bar{x}_2$. Выборочные средние будем рассматривать как нормально распределенные независимые случайные величины $\bar{X}_1$ и $\bar{X}_2$. Они являются несмещенными оценками генеральных средних, поэтому нулевая гипотеза H_0 будет иметь вид:

$$M(\bar{X}_1) = M(\bar{X}_2) \text{ или } H_0: M(\bar{X}_1) - M(\bar{X}_2) = 0.$$

Если предположить, что обе выборки имеют одинаковые дисперсии, то для проверки нулевой гипотезы можно использовать критерий Стьюдента.

Если дисперсии не равны, то нулевую гипотезу об их равенстве необходимо проверить, используя критерий Фишера — Снедекора.

При условии справедливости нулевой гипотезы о равенстве генеральных дисперсий рассмотрим стандартизированную случайную величину, которая, как и исходные генеральные совокупности, будет подчиняться нормальному закону распределения (но уже стандартному):

$$u = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{D(\bar{X}_1 - \bar{X}_2)}}.$$

В силу независимости выборок

$$D(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_{x1}^2}{n_1} + \frac{\sigma_{x2}^2}{n_2},$$

причем *дисперсии выборок равны неизвестной общей дисперсии*

$$\sigma^2 = \sigma_{x_1}^2 = \sigma_{x_2}^2,$$

следовательно,

$$D(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_{x1}^2}{n_1} + \frac{\sigma_{x2}^2}{n_2} = \sigma^2 \frac{(n_1+n_2)}{n_1 n_2}.$$

Имеем

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{D(\bar{X}_1 - \bar{X}_2)}} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \to N(0,1). \qquad (13.15)$$

Для оценки неизвестной дисперсии σ^2 рассмотрим исправленные выборочные дисперсии

$$s_1^2 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2}{n_1 - 1}; \quad s_2^2 = \frac{\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2}{n_2 - 1}.$$

В силу леммы Фишера:

$$\frac{(n_1-1)s_1^2}{\sigma^2} = \sum_{i=1}^{n_1}\left(\frac{x_{1i} - \bar{x}_{1i}}{\sigma}\right)^2 \xrightarrow{d} \chi^2_{(n_1-1)};$$

$$\frac{(n_2-1)s_2^2}{\sigma^2} = \sum_{i=1}^{n_1}\left(\frac{x_{2i} - \bar{x}_{2i}}{\sigma}\right)^2 \xrightarrow{d} \chi^2_{(n_2-1)}.$$

По свойству воспроизводимости χ^2-распределения, их сумма распределена по закону χ^2 с $(n_1 + n_2 - 2)$ степенями свободы

$$\frac{(n_1-1)s_1^2}{\sigma^2} + \frac{(n_2-1)s_2^2}{\sigma^2} \xrightarrow{d} \chi^2_{(n_1+n_2-2)}. \tag{13.16}$$

Величина

$$s^2 = \frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{(n_1+n_2-2)}$$

является несмещенной оценкой для σ^2.

Действительно,

$$M\left(\frac{s_1^2(n_1-1)}{(n_1+n_2-2)} + \frac{s_2^2(n_2-1)}{(n_1+n_2-2)}\right) = \frac{(n_1-1)\sigma^2}{(n_1+n_2-2)} + \frac{(n_2-1)\sigma^2}{(n_1+n_2-2)} = \sigma^2.$$

Кроме того,

$$\frac{s^2}{\sigma^2} = \frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{\sigma^2(n_1+n_2-2)} \xrightarrow{d} \frac{\chi^2_{(n_1+n_2-2)}}{(n_1+n_2-2)}.$$

Рассмотрим величину

$$t = \frac{(\bar{X}_1-\bar{X}_2)\sqrt{\frac{n_1n_2}{n_1+n_2}}}{s} = \frac{\frac{(\bar{X}_1-\bar{X}_2)}{\sigma}\sqrt{\frac{n_1n_2}{n_1+n_2}}}{\frac{s}{\sigma}} =$$

$$= \frac{\frac{(\bar{X}_1-\bar{X}_2)}{\sigma}\sqrt{\frac{n_1n_2}{n_1+n_2}}}{\sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{\sigma^2(n_1+n_2-2)}}} = \frac{Z}{\sqrt{\frac{\chi^2_{(n_1+n_2-2)}}{(n_1+n_2-2)}}} \xrightarrow{d} t_{(n_1+n_2-2)}, \tag{13.17}$$

где $Z = \frac{(\bar{X}_1-\bar{X}_2)}{\sigma}\sqrt{\frac{n_1n_2}{n_1+n_2}} \to N(0,1)$.

Следовательно, в силу теоремы 4 и формул (13.15)–(13.17) величина

$$t = \frac{\bar{X}_1-\bar{X}_2}{\sqrt{s_1^2(n_1-1)+s_2^2(n_2-1)}}\sqrt{\frac{(n_1+n_2-2)n_1n_2}{n_1+n_2}} \tag{13.18}$$

не зависит от $\bar{X}_1, \bar{X}_2$, σ^2 и подчиняется распределению Стьюдента с числом степеней свободы $k = n_1 + n_2 - 2$.

Если $n_1 = n_2 = n$, то формулу (13.18) можно представить в виде

$$t = \frac{\bar{X}_1-\bar{X}_2}{\sqrt{s_1^2+s_2^2}}\sqrt{n}. \tag{13.19}$$

В зависимости от вида конкурирующей гипотезы строится критическая область.

Если конкурирующая гипотеза имеет вид H_1: $\bar{X}_1 \neq \bar{X}_2$, то критическая точка находится для двухсторонней критической области, чтобы вероятность попадания критерия в каждый интервал была равна $\alpha/2$:

$$\mathrm{P}(t < t_{\text{лев.}}) = \frac{\alpha}{2}, \mathrm{P}(t > t_{\text{прав.}}) = \frac{\alpha}{2}.$$

Так как критерий Стьюдента симметричен относительно нуля, то $t_{\text{прав.}}(\alpha, k) = -t_{\text{лев.}}(\alpha, k)$, поэтому достаточно найти одну критическую точку. Если $|t_{\text{н.}}| < t_{\text{кр.}}$, то нулевая гипотеза принимается. Уровень значимости берется по верхней строчке таблицы значений критерия, при уровне значимости α и числе степеней свободы

$$k = n_1 + n_2 - 2.$$

Если $|t_{\text{н.}}| > t_{\text{кр.}}$, то нулевая гипотеза отвергается, средние значения значимо различаются по двум совокупностям.

Если конкурирующая гипотеза имеет вид $H_1: \bar{X}_1 > \bar{X}_2$ или $H_1: \bar{X}_1 < \bar{X}_2$, то критическую точку находят для правосторонней или левосторонней критической области, чтобы, при условии справедливости нулевой гипотезы,

$$P(t > t_{\text{прав.}}) = \alpha.$$

Значение $t_{\text{кр}}$ находится по таблице при заданном уровне значимости и числе степеней свободы $k = n_1 + n_2 - 2$. Сравнивая наблюдаемое значение критерия с критическим, формулируется вывод.

Пример 13.5. По результатам испытаний двух сортов риса (пример 13.4), можно ли утверждать, что эти сорта существенно различаются по уровню урожайности?

Решение. По условию задачи: $\bar{X}_1$=75,0, s_1^2 = 41,1, $\bar{X}_2$=70,0, s_2^2 = 17,8, n_1=8, n_2= 10. $H_0: \bar{X}_1 = \bar{X}_2$, при $H_1: \bar{X}_1 \neq \bar{X}_2$. Выборки независимые, малые по объему, причем имеют статистически не значимо отличающиеся дисперсии, что доказано в примере 13.4, поэтому необходимо использовать критерий Стьюдента. Найдем наблюдаемое значение критерия:

$$t_{\text{н.}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2(n_1-1)+s_2^2(n_2-1)}} \cdot \sqrt{\frac{(n_1+n_2-2)n_1n_2}{n_1+n_2}} =$$

$$= \frac{75{,}0-70{,}0}{\sqrt{41{,}1\cdot 7+17{,}8\cdot 9}} \cdot \sqrt{\frac{(8+10-2)\cdot 8\cdot 10}{8+10}} = 1{,}99.$$

По таблице для двухсторонней критической области при уровне значимости α = 0,05 и числе степеней свободы $k = 8 + 10 - 2 = 16$, $t_{\text{кр.}}$ = 2,12. Так как $|t_{\text{н.}}| < t_{\text{кр.}}$, то нулевая гипотеза принимается и нет оснований утверждать, что два сорта риса значимо различаются по урожайности.

Если $H_0: \bar{X}_1 = \bar{X}_2$, при $H_1: \bar{X}_1 > \bar{X}_2$, то $t_{\text{н.}}$ =1,99. При α = 0,05 и числе степеней свободы $k = 8 + 10 - 2 = 16$ по таблице для односторонней критической области $t_{\text{кр.}}$ =1,75. Так как $|t_{\text{н.}}| > t_{\text{кр.}}$, то нулевая гипотеза отвергается и на данном уровне значимости можно утверждать, что первый сорт значимо превосходит второй по средней урожайности. Вывод оказался противоречивым возможно вследствие малых объемов выборок. На практике обычно используется более строгий критерий с использованием двухсторонней критической области.

Замечание. 1. Задача проверки гипотезы о равенстве средних двух нормальных распределений в случае, что дисперсии неизвестны и не равны (гипотеза о равенстве дисперсий, проверяемая по критерию Фишера — Снедекора, отвергается), не имеет удовлетворительного решения и называется «*проблема Беренса — Фишера*». Приближенное решение предлагает использовать статистику распределения Стьюдента:

$$t_{\text{н.}} = |\bar{X}_1 - \bar{X}_2| / \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

с числом степеней свободы

$$k = \left[\frac{s_1^2/n_1 + s_2^2/n_2}{\frac{s_1^2/n_1}{n_1 - 1} + \frac{s_2^2/n_2}{n_2 - 1}}\right],$$

где $[x]$ — целая часть числа x.

2. Для выборок разного объема рекомендуется применять критерий Крамера — Уэлча [88]. ■

Пример 13.6. Оценить существенность различий в средней урожайности двух сортов озимой пшеницы. Для 1-го сорта средняя урожайность $\bar{X}_1 = 65{,}6$ ц/га и выборочная дисперсия $S_1^2 = 8{,}05$, для 2-го сорта средняя урожайность $\bar{X}_2 = 78{,}4$ ц/га и выборочная дисперсия $S_2^2 = 14{,}31$. Объемы выборок $n_1 = 5$ и $n_2 = 5$ соответственно.

Решение. Выдвигаем нулевую гипотезу о том, что средние урожайности двух сортов пшеницы не отличаются друг от друга, т. е. $H_0: \bar{X}_1 = \bar{X}_2$, при альтернативной гипотезе $H_1: \bar{X}_1 \neq \bar{X}_2$ — урожайности существенно различны. Примем уровень значимости $\alpha = 0{,}05$.

Так как выборки независимые, причем $n_1 = n_2$, то применим критерий Стьюдента с $k = n_1 + n_2 - 2$ степенями свободы.

$$t_{\text{н.}} = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{78{,}4 - 65{,}6}{\sqrt{\frac{8{,}05}{5} + \frac{14{,}31}{5}}} = \frac{12{,}8}{2{,}11} = 6{,}07.$$

По данным приложения 3 определим критическое значение *t*-распределения при уровне значимости 0,05 и числе степеней свободы $k = 5 + 5 - 2 = 10$:

$$t_{\text{кр.}} = t_{0{,}05;10} = 2{,}31.$$

Так как $t_{\text{н.}} > t_{\text{кр.}}$, то нулевую гипотезу следует отклонить. Следовательно, два сорта озимой пшеницы отличаются статистически значимо по величине средней урожайности.

13.6. Проверка гипотезы о значимости средней разности двух зависимых нормально распределенных генеральных совокупностей

Пусть имеются две нормально распределенные генеральные совокупности с одинаковой дисперсией. Единицы наблюдения одной генеральной совокупности попарно связаны каким-то общим условием с единицами наблюдения другой генеральной совокупности, то есть может существовать корреляция между парами наблюдений из двух выборок. Из этих совокупностей образованы выборки объема *n* единиц, взятых попарно. Обозначим через x_{1i} — значения признака по *i*-той единице первой совокупности, x_{2i} — значения признака по *i*-той единице второй совокупности. Разность значений $d_i = x_{1i} - x_{2i}$. Находится средняя разность значений:

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \overline{x_1 - x_2} = \bar{x}_1 - \bar{x}_2. \qquad (13.20)$$

Так как x_{1i} и x_{2i} извлечены из нормальных генеральных совокупностей с одинаковой дисперсией, то их разность также будет подчиняться нормальному закону $\bar{d} \to N\left(M\left(\frac{1}{n}\sum_{i=1}^{n} d_i\right), D\left(\frac{1}{n}\sum_{i=1}^{n} d_i\right)\right)$.

В силу зависимости, $\bar{x}_1$ и $\bar{x}_2$

$$D(\bar{d}) = D(\bar{x}_1 - \bar{x}_2) = D(\bar{x}_1) + D(\bar{x}_2) - 2cov(\bar{x}_1, \bar{x}_2),$$

но в силу леммы Фишера $\bar{d}$ и исправленная выборочная дисперсия

$$s_d^2 = \frac{1}{n-1}\sum\left((x_{1i} - x_{2i}) - \bar{d}\right)^2 = \frac{\sum(d_i - \bar{d})^2}{n-1}$$ линейно не зависимы.

Выдвигается нулевая гипотеза H_0: $\bar{d} = 0$, при альтернативной гипотезе H_0: $\bar{d} \neq 0$.

Рассмотрим величину

$$t = \frac{\bar{d}}{s_d} = \frac{\frac{\bar{d}}{\sigma}}{\frac{s_d}{\sigma}} = \frac{\frac{\sum_{i=1}^{n} d_i}{n\sigma}}{\sqrt{\frac{1}{n-1}\sum\left(\frac{(x_{1i} - x_{2i}) - \bar{d}}{\sigma}\right)^2}} \to t_{n-1}, \qquad (13.21)$$

так как, полагая, что выполняется нулевая гипотеза, получим

$$\frac{\sum_{i=1}^{n} d_i}{n\sigma} \to N(0,1), \quad \frac{1}{n-1}\sum\left(\frac{(x_{1i} - x_{2i}) - \bar{d}}{\sigma}\right)^2 \to \frac{\chi^2_{n-1}}{n-1}. \qquad (13.22)$$

Итак, рассмотрим алгоритм проверки гипотезы о значимости средней разности для зависимых выборок. Нулевая гипотеза проверяется с использованием *t*-критерия Стьюдента.

Наблюдаемое значение критерия определяется по формуле

$$t_{\text{н.}} = \frac{\bar{d}}{s_{\bar{d}}}, \qquad (13.23)$$

где

$$s_{\bar{d}} = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n(n-1)}} = \sqrt{\frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n(n-1)}}, \qquad (13.24)$$

n — число наблюдений.

По таблице критических точек распределения Стьюдента при уровне значимости α для двухсторонней критической области и числе степеней свободы $k = n - 1$, находится критическое значение t-критерия. Сравнивая наблюдаемое значение критерия с критическим, формулируется соответствующий вывод.

Если $|t_{\text{н.}}| > t_{\text{кр.}}$, то нулевая гипотеза отвергается, средняя разность значений значимо отличается от нуля по двум совокупностям.

Если $|t_{\text{н.}}| < t_{\text{кр.}}$, то нулевая гипотеза принимается, средняя разность значений не значимо отличается от нуля по двум совокупностям.

Пример 13.7. Два сорта озимой пшеницы испытывались на одинаковом числе участков на протяжении семи лет (табл. 13.3).

При уровне значимости $\alpha = 0{,}05$ проверить нулевую гипотезу о значимости различий в урожайности двух сортов озимой пшеницы.

Решение. Так как имеются две зависимости выборки, т. е. существует определенная корреляция между урожайностью сортов по годам, то необходимо оценить значимость не разности двух выборочных средних, а средней разности.

Выдвигаем нулевую гипотезу: средняя величина различий в урожайности пшеницы равна нулю, $H_0 : \overline{X_1 - X_2} = 0$ при $H_1 : \overline{X_1 - X_2} \neq 0$.

По данным таблицы 13.3 найдем среднюю разность $\bar{d}$ и ошибку средней разности $s_{\bar{d}}$:

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{35}{7} = 5;$$

$$s_{\bar{d}} = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n(n-1)}} = \sqrt{\frac{46}{7(7-1)}} = 1{,}046; t_{\text{н.}} = \frac{5}{1{,}046} = 4{,}78.$$

Таблица 13.3

Вспомогательная таблица для расчета ошибки средней разности

Годы	Урожайность, ц/га		Разность $d_i = x_{1i} - x_{2i}$	$(d_i - \bar{d})$	$(d_i - \bar{d})^2$
	x_{2i}	x_{1i}			
2009	46	52	6	1	1
2010	43	48	5	0	0
2011	46	45	–1	–6	36
2012	51	56	5	0	0
2013	52	58	6	1	1
2014	48	55	7	2	4
2015	52	59	7	2	4
Сумма	–	–	35	0	46

При $\alpha = 0{,}05;\ k = n - 1 = 7 - 1 = 6, t_{\text{кр.}} = 2{,}45.$

Сопоставив наблюдаемое значение t с критическим значением, можно сделать вывод, что два сорта озимой пшеницы существенно различаются по уровню средней урожайности.

Если дисперсии не равны, то нулевую гипотезу об их равенстве необходимо проверить, используя критерий Фишера — Снедекора.

13.7. Проверка гипотезы о равенстве долей двух независимых нормально распределенных генеральных совокупностей

Пусть имеется две генеральные совокупности. Доля единиц, обладающих данным признаком в первой совокупности, равна p_1, а по второй совокупности — p_2. Необходимо при заданном уровне значимости проверить гипотезу о равенстве генеральных долей:

$$H_0: p_1 = p_2.$$

Конкурирующая гипотеза:

а) $H_1: p_1 \neq p_2$; б) $H_1: p_1 > p_2$; в) $H_1: p_1 < p_2$.

Из генеральных совокупностей образованы две независимые выборки: из первой совокупности объема n_1, а из второй — объема n_2. Предполагается, что обе выборки достаточно большие по объему. Обозначим m_1 — число единиц, обладающих данным признаком по первой совокупности, а m_2 — число единиц,

обладающих данным признаком по второй совокупности. Тогда выборочные доли w_1 и w_2 по каждой совокупности составят:

$$w_1 = \frac{m_1}{n_1}, \quad w_2 = \frac{m_2}{n_2}.$$

Так как объемы выборок достаточно велики, то выборочные доли приближенно распределяются по нормальному закону с математическими ожиданиями p_1 и p_2, дисперсиями $\frac{p_1(1-p_1)}{n_1}$, $\frac{p_2(1-p_2)}{n_2}$.

В качестве критерия проверки нулевой гипотезы принимается случайная величина U. При условии справедливости нулевой гипотезы эта величина распределяется нормально с параметрами $M(U) = 0$, $\sigma(U) = 1$. Наблюдаемое значение критерия находится по формуле

$$U_{\text{н.}} = \frac{\frac{m_1}{n_1} - \frac{m_2}{n_2}}{\sqrt{\frac{m_1+m_2}{n_1+n_2}\left(1-\frac{m_1+m_2}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}. \qquad (13.25)$$

Если конкурирующая гипотеза имеет вид $H_1: p_1 \neq p_2$, то критическая точка находится для двухсторонней критической области, исходя из равенства

$$\Phi(u_{\text{кр.}}) = \frac{1-\alpha}{2}.$$

Если конкурирующая гипотеза имеет вид $H_1: p_1 > p_2$ или $H_1: p_1 < p_2$, то критическую точку находят, исходя из равенства

$$\Phi(u_{\text{кр.}}) = \frac{1-2\alpha}{2}.$$

Если $|U_{\text{н.}}| > U_{\text{кр.}}$, то нулевая гипотеза отвергается, выборочные доли различаются значимо.

Если $|U_{\text{н.}}| < U_{\text{кр.}}$, то нулевая гипотеза принимается, нет оснований отвергнуть нулевую гипотезу.

Пример 13.8. В торговую сеть поступает однотипный товар от двух производителей. Проведен устный опрос случайно взятых покупателей. По продукции первого производителя положительную оценку качества продукции высказал 91 покупатель из 100 опрошенных, а по продукции второго производителя 99 из 120 опрошенных. При уровне значимости 0,05 проверить нулевую гипотезу о равенстве долей покупателей, ответивших положительно в отношении качества продукции двух сравниваемых производителей.

Решение. По условию задачи $n_1 = 100, m_1 = 91, n_2 = 120, \quad m_2 = 99$. Необходимо, при уровне значимости $\alpha = 0{,}05$, проверить гипотезу оравенстве генеральных долей $H_0: p_1 = p_2$, при конкурирующей гипотезе $H_1: p_1 \neq p_2$. Оценками p_1 и p_2 служат относительные частоты w_1 и w_2. Тогда нулевую гипотезу можно записать как $H_0: w_1 = w_2$, при $H_1: w_1 \neq w_2$.

Найдем наблюдаемое значение критерия U.

$$w_1 = \frac{m_1}{n_1} = \frac{91}{100} = 0{,}91, \quad w_2 = \frac{m_2}{n_2} = \frac{99}{120} = 0{,}825.$$

$$U_{\text{н.}} = \frac{0{,}91-0{,}825}{\sqrt{\frac{91+99}{100+120}\left(1-\frac{91+99}{100+120}\right)\left(\frac{1}{100}+\frac{1}{120}\right)}} = \frac{0{,}085}{0{,}0464} = 1{,}83.$$

При уровне значимости $\alpha = 0{,}05$, $U_{\text{кр.}} = 1{,}96$. Так как $U_{\text{н.}} < U_{\text{кр.}}$, то нулевая гипотеза принимается, нет значимых различий в доле покупателей, высказавшихся положительно по качеству продукции двух производителей.

Если конкурирующая гипотеза имеет вид $H_1: p_1 > p_2$, то при уровне значимости $\alpha = 0{,}05$, $U_{\text{кр.}} = 1{,}65$, следовательно, нулевая гипотеза отвергается и верна альтернативная гипотеза.

13.8. Проверка гипотезы о виде распределения

В самом общем виде математическая статистика изучает эмпирические данные, которые являются реализацией непрерывных случайных величин. Рассмотрим различные случайные величины X_1, X_2, $X_3, \dots, X_n$, которые взаимно независимы и одинаково распределены с непрерывной функцией распределения $F(x)$. Расположим значения случайных величин в порядке неубывания, получим упорядоченную совокупность

$$X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \cdots \leq X_{(n)}, \qquad (13.26)$$

которая называется *вариационным рядом*, где $X_{(i)}$ — варианты или порядковые статистики, которые имеют определенные названия, например $X_{(\tau)} = q_\tau$ — квантиль порядка τ означает, что $P(X < q_\tau) = \tau$. В математической и экономической статистике применяют квартили и децили. Нижняя квартиль $Q_{0,25}$, верхняя квартиль $Q_{0,75}$, медиана $Me(X) = Q_{0,5}$. Децили: $q_{0,1}$, $q_{0,2}$, $q_{0,3}$, $q_{0,4}$, $q_{0,5}$, $q_{0,6}$, $q_{0,7}$, $q_{0,8}$, $q_{0,9}$. Например, если рассматривать население по уровню дохода, то нижняя и верхняя децили отделяют 10% наиболее бедных и 10% наиболее богатых от остального населения. Для характеристики рассеяния (разброса) членов вариационного ряда можно использовать интерквартильный и интердецильный размах, соответственно:

$$R_{0,5} = \left|Q_{0,25} - Q_{0,75}\right|, \quad R_{0,8} = \left|q_{0,9} - q_{0,1}\right|.$$

Рассмотрим индикаторную случайную величину

$$I(X_i < x) = \begin{cases} 1, & \text{при } X_i < x, \\ 0, & \text{при } X_i \geq x. \end{cases} \qquad (13.27)$$

Среднее арифметическое индикаторных случайных величин $I(X_i < x)$ называют эмпирической функцией распределения:

$$F_n(x) = \frac{\text{количество } X_i \in (-\infty; x)}{n} = \frac{1}{n}\sum_{i=1}^{n} I(X_i < x). \qquad (13.28)$$

Эмпирическая функция распределения по вариационному ряду имеет скачки в точке X_i, равные $\frac{k}{n}$, где k — число членов вариационного ряда, совпадающих с X_i. Функция $F_n(x)$ имеет вид[21]:

[21] Неравенство $X_{(k)} < x \leq X_{(k+1)}$ означает непрерывность слева. Аналогично можно было ввести непрерывность справа: $X_{(k)} \leq x < X_{(k+1)}$.

$$F_n(x) = \begin{cases} 0, & \text{при } x \le X_{(1)}, \\ \dots & \\ \frac{k}{n}, & \text{при } X_{(k)} < x \le X_{(k+1)}, \\ \dots & \\ 1, & \text{при } x > X_{(n)}, \end{cases} \qquad (13.29)$$

где $k = 1,\ 2, \dots, n-1$.

Альтернативной формой задания эмпирической функции является таблица: для дискретного вариационного ряда, представляющая вариационный ряд как совокупность вариантов, их частот и (или) соответствующих частостей; для интервального вариационного ряда, представляющего значения непрерывного или дискретного признака, сгруппированного по интервалам. Графически таблица может быть представлена полигоном частот, гистограммой частот и относительных частот, кумулятой (см. ниже).

Согласно результатам раздела 3.3, индикатор $I(X_i < x)$ при каждом x подчиняется распределению Бернулли, причем $p = P(X_i < x) = F(x)$, математическое ожидание индикаторной случайной величины $I(X_i < x)$ равно вероятности

$$p = P(X_i < x),$$

дисперсия равна

$$p(1-p) = P(X_i < x)(1 - P(X_i < x)).$$

Таким образом, $\sum_{i=1}^{n} I(X_i < x) = nF_n(x)$ *подчиняется биномиальному закону распределения с числом испытаний* n *и вероятностью успеха в одном опыте* p:

$$P\left(F_n(x) = \frac{k}{n}\right) = C_n^k F^k(x)(1 - F(x))^{n-k}, \quad k = 0, 1, \dots, n.$$

Следовательно, свойства эмпирической функции распределения аналогичны свойствам средней арифметической независимых, одинаково распределенных по биномиальному закону случайных величин:

1) $M(F_n(x)) = \frac{1}{n}\sum_{i=1}^{n} M\big(I(X_i < x)\big) = \frac{1}{n}\sum_{i=1}^{n} P(X_i < x) = F(x)$.

2) $D(F_n(x)) = \frac{1}{n}\sum_{i=1}^{n} D\big(I(X_i < x)\big) =$

$= \frac{1}{n}\sum_{i=1}^{n} P(X_i < x)\big(1 - P(X_i < x)\big) = \frac{1}{n}F(x)\big(1 - F(x)\big)$.

3) $\sqrt{n}(F_n(x) - F(x)) \to u \in N(0, F(x)\big(1 - F(x)\big)$.

Действительно, рассмотрим случайную величину

$$\sqrt{n}(F_n(x) - F(x)) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i < x) - F(x)\right) =$$

$$= \left(\frac{\sum_{i=1}^{n} I(X_i < x) - nF(x)}{\sqrt{n}}\right) = \left(\frac{\sum_{i=1}^{n} I(X_i < x) - nM(F_n(x)}{\sqrt{n}}\right).$$

Согласно теореме Линденберга — Леви (раздел 8.3):

$$\frac{\sum_{i=1}^{n} I(X_i < x) - M(F_n(x))}{\sqrt{n}} \to N(0, F(x)\big(1 - F(x)\big).$$

Эмпирическая функция сходится по вероятности (и почти наверное) к теоретической при любых $x \in R$:

$$F_n(x) \xrightarrow{p} F(x).$$

Как показывает *теорема Гливенко — Кантелли* (1933), для непрерывной функции распределения $F(x)$ и эмпирической функции $F_n(x)$, распределение функции

$$D_n = \sup_{x \in R} |F_n(x) - F(x)|$$

не зависит от вида функции $F(x)$ (sup означает точную верхнюю грань).

В 1940 г. А. Н. Колмогоров уточнил теорему Гливенко, введя специальное обозначение для функции распределения случайной величины $\sqrt{n}D_n$:

$$K_n(\lambda) = P\left(\sqrt{n}D_n < \lambda\right).$$

$K(\lambda)$ — функция Колмогорова, значения которой затабулированы.

Теорема (А. Н. Колмогоров). Для любого $\lambda > 0$ при $n \to \infty$

$$K_n(\lambda) \to K(\lambda) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda^2}.$$

К важнейшим задачам математической статистики относится оценка степени расхождений между наблюдаемыми или эмпирическими частотами вариационного ряда и теоретическими частотами, найденными в предположении, что изучаемый признак распределяется по определенному закону, например нормальному, равномерному, показательному и т. п. Если между наблюдаемыми и теоретическим частотами имеются небольшие расхождения, значит, высокая согласованность частот или частостей, то можно заключить, что в основе эмпирического распределения лежат те же причины, что и теоретического распределения. Проверка гипотезы о виде распределения осуществляется с помощью критериев согласия, к важнейшим из которых относятся хи-квадрат Пирсона, Колмогорова, Романовского, ω^2 — Мизеса и другие. *Критерий согласия предназначен для проверки нулевой гипотезы, что случайная выборка X_1, X_2, X_3,..., X_n образована из генеральной совокупности X с функцией распределения F(X),* вид функции, а соответственно и закон распределения, считается известным, а параметры — неизвестными.

При проверке гипотезы о нормальном распределении генеральной совокупности сравниваются эмпирические (наблюдаемые) и теоретические (вычисленные в предположении нормальности распределения) частоты. Для этого наиболее часто используется статистика хи-квадрат (χ^2) с $k = l - r - 1$ степенями свободы (l — число групп или классов, на которые разбит вариационный ряд, r — число оцениваемых параметров предполагаемого распределения). При нормальном распределении оценивается математическое ожидание и среднее квадратическое отклонение, следовательно, $r = 2$, значит $k = l - 3$. Если $\chi^2_{\text{н}} \geq \chi^2_{\text{кр}}$, то нулевая гипотеза отвергается и считается, что предположение о нормальности распределения не согласуется с имеющимися данными. В противном случае ($\chi^2_{\text{н}} < \chi^2_{\text{кр}}$) нулевая гипотеза принимается.

Проверка гипотезы о нормальном распределении может осуществляться в следующей последовательности.

1) Из генеральной совокупности извлекается случайная выборка, по которой составляется вариационный ряд с равными интервалами (интервалы могут быть неравными). Определяется выборочная средняя, дисперсия и среднее квадратическое отклонение. Эти характеристики служат статистическими оценками непосредственно неизвестных математического ожидания и дисперсии нормального распределения.

2) Переходят к стандартизированной случайной величине $t = (X - \bar{x})/\sigma$, по которой определяют концы интервалов:

$$t_i = \frac{x_i - \bar{x}}{\sigma}, t_{i+1} = \frac{x_{i+1} - \bar{x}}{\sigma}. \quad (13.30)$$

3) Находят теоретические вероятности попадания нормально распределенной случайной величины в частичный интервал:

$$p_i = \Phi(t_{i+1}) - \Phi(t_i), \quad (13.31)$$

где

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-\frac{x^2}{2}}\, dx. \quad (13.32)$$

4) Вычисляют теоретические частоты в предположении справедливости нулевой гипотезы о нормальности распределения:

$$n_i' = np_i. \quad (13.33)$$

5) Выбирается критерий, по которому будет проверяться нулевая гипотеза, в данном случае критерий хи-квадрат Пирсона. Наблюдаемое значение критерия находится по формуле

$$\chi^2_{\text{н}} = \sum_{i=1}^{l} \frac{(n_i - n_i')^2}{n_i'} = \sum_{i=1}^{l} \frac{(n_i)^2}{n_i'} - n. \quad (13.34)$$

6) По таблице критических точек χ^2-распределения Пирсона при заданном уровне значимости α и числе степеней свободы $k = l - 3$, находится критическое значение критерия для правосторонней критической области.

7) Сравнивая наблюдаемое значение критерия с критическим, формулируется вывод о степени различий частот, то есть о законе распределения.

Если параметры распределения полагаются известными (простая гипотеза), то приведенный ранее критерий (13.24) называется *хи-квадрат Пирсона*, его обоснование дает *теорема Пирсона*.

Если вспомнить схему полиномиального распределения, то события A_i соответствуют попаданию в i-й интервал ($i = 1, \dots, l$) с вероятностью p_i, соответственно с математическим ожиданием $n_i' = np_i$, и дисперсией np_iq_i. В результате n опытов полагается, что число успехов для события A_i (попаданий в i-ый интервал) равно n_i, $\sum p_i = 1$, $\sum n_i = n$.

В качестве меры расхождения «теоретических» частот ($n_i' = np_i$) и эмпирических частот используется величина χ^2. В частности, верна следующая теорема.

Теорема (К. Пирсон, 1900). Если нулевая гипотеза верна, то при фиксированном l и $n \to \infty$

$$\chi^2_{\text{н}} = \sum_{i=1}^{l} \frac{(n_i - n_i')^2}{n_i'} \to \chi^2_{l-1}. \quad (13.35)$$

В противном случае $\chi^2_{\text{н}} \to \infty$.

Доказательство. Пусть $l = 2$, тогда $n = n_1 + n_2$, $1 = p_1 + p_2$, следовательно,

$$\chi^2_{\text{н}} = \frac{(n_1 - np_1)^2}{np_1} + \frac{(n_2 - np_2)^2}{np_2} = \frac{(n_1 - np_1)^2}{np_1} + \frac{(n - n_1 - n(1-p_1))^2}{n(1-p_1)} =$$
$$= \frac{(n_1 - np_1)^2}{np_1} + \frac{(-n_1 + np_1)^2}{n(1-p_1)} = \frac{(n_1 - np_1)^2}{n}\left(\frac{1}{p_1} + \frac{1}{(1-p_1)}\right) = \frac{(n_1 - np_1)^2}{np_1(1-p_1)}.$$

Положим, что n_1 — число успехов в схеме Бернулли из n испытаний (или сумма n случайных величин, распределенных по закону Бернулли), поэтому, согласно центральной предельной теореме (Линденберга — Леви):

$$\frac{(n_1-np_1)}{\sqrt{np_1(1-p_1)}} =: Z \to N(0,1).$$

Соответственно, согласно определению распределения χ^2 Пирсона:

$$Z^2 = \left(\frac{(n_1-np_1)}{\sqrt{np_1(1-p_1)}}\right)^2 \to \chi_1^2.$$

Общий случай можно получить аналогично лемме Фишера:

$$\frac{nS^2}{\sigma^2} = \sum_{i=1}^{l}\left(\frac{x_i-\overline{x}}{\sigma}\right)^2 \to \chi_{l-1}^2,$$

введя соответствующие переобозначения.

Если параметры распределения не известны (сложная гипотеза), критерий часто называется *хи-квадрат Фишера*, с $k = l-r-1$ степенями свободы, где r — число оцениваемых параметров распределения.

Замечание [109]. 1. Для дальнейшего изложения рассмотрим некоторые факты, опираясь на сведения из главы 6. Для k-мерной нормально распределенной случайной величины $X(x_1, x_2, \dots, x_k), X \in R^k$, плотность распределения имеет вид

$$f(X) = [(2\pi)^k|\Sigma|]^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(X-M)^T\Sigma^{-1}(X-M)\right], \qquad (13.36)$$

где:

квадратичная форма:

$$(X-M)^T\Sigma^{-1}(X-M) = \sum_{i,j}(x_i-\mu_i)\,\Sigma^{-1}\left(x_j-\mu_j\right), \qquad (13.37)$$

вектор математических ожиданий: $M = \{\mu_1, \mu_2, \dots, \mu_k\}$,

ковариационная матрица Σ:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kk} \end{bmatrix}, \qquad (13.38)$$

элементы которой $\sigma_{ij} = cov\left(x_i, x_j\right) = M\left[\left(x_i - M(x_i)\right)\left(x_j - M\left(x_j\right)\right)\right]$,

обратная ковариационная матрица Σ^{-1} и корреляционная матрица:

$$R = \begin{pmatrix} 1 & \cdots & \rho_{1k} \\ \vdots & \ddots & \vdots \\ \rho_{k1} & \cdots & 1 \end{pmatrix},$$

элементы которой $\rho_{ij} = \frac{cov(x_i, x_j)}{\sqrt{\sigma_{ii}^2\sigma_{jj}^2}}$.

2. Если матрица Σ является диагональной

$$\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{kk} \end{bmatrix},$$

то квадратичная форма (13.37) может быть записана как

$$(X-M)^T\Sigma^{-1}(X-M) = \sum_{i,j}(x_i-\mu_i)\,\Sigma^{-1}\left(x_j-\mu_j\right) = \sum_i\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2,$$

таким образом, многомерная функция плотности вероятности может быть представлена как произведение одномерных функций плотности вероятности:

$$f(x_1, x_2, \dots, x_k) = f(x_1)f(x_2)\dots f(x_k),$$

что означает их независимость. Диагональность ковариационной матрицы означает, что координаты вектора X не коррелированы.

Итак, *если вектор* $X(x_1, x_2, \dots, x_k)$ *имеет нормальное распределение, то из независимости координат вектора X следует их некоррелированность, и наоборот, из некоррелированности следует независимость*.

3. Имеет место *многомерная центральная предельная* теорема. ■

Теорема. Если случайные вектора $X_1, X, \dots, X_n$ независимы, одинаково распределены и имеют вектор математических ожиданий $M(\mu_1, \mu_2, \dots, \mu_k)$, невырожденную ковариационную матрицу Σ, то при $n \to \infty$:

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n - na}{\sqrt{n}} \to Z, \qquad (13.39)$$

где $Z \in N(0, \Sigma)$.

4. Рассмотрим несколько утверждений линейной алгебры, связанных некоторыми фактами многомерных законов распределения (полагается, что дисперсии компонент существуют), которые не сложно доказать.

Утверждение 1. Пусть (a, z) — скалярное произведение неслучайного вектора $a = \{a_1, \dots, a_l\}$ на случайный вектор $Z = \{z_1, z_2, \dots, z_l\}$, тогда дисперсия (a, Z) определяется по формуле

$$D(a, Z) = (\Sigma_z a, a). \qquad (13.40)$$

Действительно,

$$D(a, Z) = D\left(\sum_{i=1}^{l}(a_i, z_i)\right) = \sum_{i,j=1}^{l} a_i\, a_j cov\left(z_i, z_j\right) = (\Sigma_Z a, a).$$

Утверждение 2. Если распределение вектора $Z = \{z_1, z_2, \dots, z_l\}$ сосредоточено в некоторой гиперплоскости, то ковариационная матрица Σ_Z вырождена.

Уравнение гиперплоскости L с нормальным вектором $a = \{a_1, a_2, \dots, a_l\}$: $(a, Z) = r$, где r — действительное число. Сосредоточенность распределения вектора Z в плоскости L достигается, если при некотором значении r выполняется достоверное событие $(a, Z) = r$, вероятность которого равна единице:

$$P\big((a, Z) = r\big) = 1,$$

или иначе $D(a, Z) = 0$. Опираясь на утверждение 1, получим $D(a, Z) = (\Sigma_Z a, a) = 0$, что и означает вырожденность матрицы Σ_Z.

Утверждение 3. Пусть $U = AV + a$, где A — неслучайная матрица, a — неслучайный вектор, $V = (v_1, v_2, \dots, v_l)$ — случайный вектор. Тогда ковариационная матрица случайного вектора U определяется как

$$\Sigma_U = A\Sigma_V A^T, \qquad (13.41)$$

где Σ_V — ковариационная матрица вектора V, A^T — транспонированная матрица A.

Найдем дисперсию

$$D(V, U) = D(V, AV + a) = D(V, AV) = D(A^T V, A^T AV) = D(A^T V, V) = \\ = (\Sigma_V A^T V, A^T V) = (A^T \Sigma_V A^T V, V).$$

В силу формулы (13.40)

$$D(U, V) = (\Sigma_U V, V).$$

Значит формула (13.41) верна.

Возвращаясь к полиномиальному распределению, рассматриваемому в качестве трактовки теоремы Пирсона, рассмотрим вектор числа «успехов»

$X = (n_1, n_2, \dots, n_l)$, где $\sum n_i = n$. Так как проводится n испытаний, то случайную величину X можно представить как сумму индикаторных случайных величин, элементы которых распределены по закону Бернулли, для этого введем случайный вектор

$$Y_m = (y_{m1}, y_{m2}, \dots, y_{mi}, \dots, y_{mk}),$$

где $y_{mi} = 1$, если событие A_i произошло на шаге m, $y_{mj} = 0$ $(i \neq j)$. Итак, $X = Y_1 + Y_2 + \cdots + Y_n$. Случайные векторы $Y_1, Y_2, \dots, Y_n$ независимы и одинаково распределены. Если Σ_Y — матрица ковариаций любого из векторов Y_m, в силу их одинаковой распределенности, то из равенства

$$X = Y_1 + Y_2 + \cdots + Y_n$$

следует, что

$$\Sigma_X = \Sigma_{Y_1} + \Sigma_{Y_2} + \cdots + \Sigma_{Y_n} = n\Sigma_Y.$$

Как уже было сказано ранее, в основе реализации случайного вектора Y_m лежит полиномиальное распределение с вектором вероятностей $p = (p_1, p_2, \dots, p_l)$. Поэтому каждая компонента вектора распределена по закону Бернулли, следовательно: дисперсии

$$D(y_{mi}) = p_i q_i, (1 - p_i = q_i);$$

ковариации

$$cov(y_{mi}, y_{mj}) = M(y_{mi}\, y_{mj}) - M(y_{mi})M(y_{mj}) = -p_i p_j,$$

так как $y_{mi}\, y_{mj} = 0$, при $i \neq j$.

Таким образом, ковариационная матрица Σ_Y имеет вид

$$\Sigma_Y = \begin{pmatrix} p_1q_1 & -p_1p_2 & \dots & -p_1p_l \\ -p_2p_1 & p_2q_2 & \dots & -p_2p_l \\ \dots & \dots & \dots & \dots \\ -p_lp_1 & -p_lp_2 & \dots & p_lq_l \end{pmatrix}. \qquad (13.42)$$

При большом числе испытаний, в силу многомерной центральной теоремы, согласно (13.39), вектор

$$\frac{X - M(X)}{\sqrt{n}} = \left(\frac{n_1 - np_1}{\sqrt{n}}, \frac{n_2 - np_2}{\sqrt{n}}, \dots, \frac{n_l - np_l}{\sqrt{n}}\right), \qquad (13.43)$$

будет стремиться к $N(0,\ \Sigma_Y)$ — многомерному нормальному распределению. Так как $n_1 + n_2 + \dots + n_l = n$, то вектор X расположен в этой гиперплоскости, поэтому, в силу утверждения 1 (см. замечание выше) матрица Σ_Y вырождена. Открытие К. Пирсона было в том, что удобно рассматривать не величины

$$t_i = \frac{n_i - np_i}{\sqrt{n}}, \text{ а}$$
$$z_i = \frac{n_i - np_i}{\sqrt{np_i}} = \frac{t_i}{\sqrt{p_i}}, \qquad (13.44)$$

где $i = 1, \dots, l$.

Соответствующее распределение величин z_i сосредоточено в гиперплоскости L:

$$z_1\sqrt{p_1} + z_2\sqrt{p_2} + \cdots + z_l\sqrt{p_l} = 0. \qquad (13.45)$$

Этот факт можно представить в виде следующей теоремы [109].

Теорема. В гиперплоскости L, при $n \to \infty$, распределение вектора $Z = \{z_1, z_2, \dots, z_l\}$ сходится в L к нормальному распределению $N(0, E)$, где E — единичная матрица.

Доказательство. Покажем, что в гиперплоскости L ковариационная матрица вектора Z будет единичной, то есть $\Sigma_Z = E$.

Пусть вектор $a = \{a_1, a_2, \dots, a_l\}$ лежит в L и его длина равна 1:

$$\sum_{i=1}^{l} a_i \sqrt{p_i} = 0, \ \sum_{i=1}^{l} a_i^2 = 1.$$

Вычислим скалярное произведение $(\Sigma_Z a, a) = D(a, Z)$, где Σ_Z — ковариационная матрица вектора Z. Запишем вектор a в матричной форме, для этого умножим его на единичную матрицу $a = aE =: A$.

В силу утверждения 3, по формуле (13.41) имеем $\Sigma_Z = A\Sigma_V A^T$, следовательно,

$$\Sigma_Z = \begin{pmatrix} \frac{1}{\sqrt{p_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{p_2}} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{\sqrt{p_l}} \end{pmatrix} \begin{pmatrix} p_1 q_1 & -p_1 p_2 & \dots & -p_1 p_l \\ -p_2 p_1 & p_2 q_2 & \dots & -p_2 p_l \\ \dots & \dots & \dots & \dots \\ -p_l p_1 & -p_l p_2 & \dots & p_l q_l \end{pmatrix}$$

$$\begin{pmatrix} \frac{1}{\sqrt{p_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{p_2}} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{\sqrt{p_l}} \end{pmatrix} = \begin{pmatrix} q_1 & -\sqrt{p_1 p_2} & \dots & -\sqrt{p_1 p_l} \\ -\sqrt{p_2 p_1} & q_2 & \dots & -\sqrt{p_2 p_l} \\ \dots & \dots & \dots & \dots \\ -\sqrt{p_l p_1} & -\sqrt{p_l p_2} & \dots & q_l \end{pmatrix}.$$

Отсюда

$$\begin{gathered} (\Sigma_Z a, a) = (1 - p_1) a_1^2 - \sqrt{p_1 p_2} a_1 a_2 - \dots - \sqrt{p_1 p_l} a_1 a_l + \\ + (1 - p_2) a_2^2 - \sqrt{p_2 p_1} a_2 a_1 - \dots - \sqrt{p_2 p_l} a_2 a_l + \dots \\ \dots + (1 - p_l) a_2^2 - \sqrt{p_l p_1} a_l a_1 - \dots - \sqrt{p_l p_{l-1}} a_l a_{l-1} = \\ = a_1^2 + \dots + a_l^2 - a_1 \sqrt{p_1} \sum_{i=1}^{l} a_i \sqrt{p_i} - a_2 \sqrt{p_2} \sum_{i=1}^{l} a_i \sqrt{p_i} - \\ - \dots - a_l \sqrt{p_l} \sum_{i=1}^{l} a_i \sqrt{p_i} = a_1^2 + \dots + a_l^2 = 1, \end{gathered}$$

так как по условию вектор $a = (a_1, a_2, \dots, a_l)$ лежит в гиперплоскости L, то $\sum_{i=1}^{l} a_i \sqrt{p_i} = 0$, в силу формулы (13.45).

Следовательно, ковариационная матрица вектора z будет единичной, что и требовалось доказать.

Следствие. Из леммы Фишера согласно (12.31) при $n \to \infty$ получим, что распределение величины

$$\sum_{i=1}^{l} \left(\frac{n_i - np_i}{\sqrt{np_i}} \right)^2 = \sum_{i=1}^{l} z_i^2 \to \chi_{l-1}^2$$

стремится к χ_{l-1}^2 — распределению хи-квадрат Пирсона с $\nu = l - 1$ степенями свободы (полагается, что математическое ожидание заранее не известно).

Пример 13.9. По данным группировки личных подсобных хозяйств населения (ЛПХ) региона по размеру земельной площади (табл. 13.4) проверить гипотезу, что данное распределение можно описать нормальным законом распределения.

Решение. По данным таблицы 13.4 рассчитаем выборочные характеристики: средний размер земельной площади на личное подсобное хозяйство, дисперсию и среднее квадратическое отклонение.

Из условия следует, что точные параметры гипотетического и нормального закона нам неизвестны, поэтому нулевую гипотезу (H_0) словесно можно сформулировать следующим образом: $F(x)$ является функцией нормального распределения с параметрами

$$M(X) = a \text{ и } D(X) = \sigma^2.$$

Для проверки этой нулевой гипотезы найдем точечные оценки математического ожидания и среднего квадратического отклонения нормально распределенной случайной величины:

$$\bar{x} = \frac{\sum x_i^* n_i}{n} = \frac{102{,}72}{715} = 0{,}144\,; \qquad \sigma^2 = \frac{\sum x_i^{*2} n_i}{n} - \bar{x}^2 =$$

$$= \frac{17{,}2768}{715} - 0{,}144^2 = 0{,}003427; \sigma = \sqrt{0{,}003427} = 0{,}0585 \approx 0{,}058.$$

По выборочной совокупности средний размер земельной площади на одно личное подсобное хозяйство населения составил 14,4 соток. По ЛПХ размер земельной площади в среднем колебался в границах от 8,6 до 20,2 соток.

Вычислим теоретические вероятности (p_i) попадания случайной величины $X \to N$ (0,144; 0,058) в частичные интервалы (x_i; x_{i+1}) по формулам (13.30)–(13.32), теоретические частоты (13.33) и наблюдаемое значение критерия хи-квадрат Пирсона (13.34) (табл. 13.5).

Таблица 13.4

Группировка личных подсобных хозяйств населения по размеру земельной площади

Группы личных подсобных хозяйств по размеру земельной площади, га $x_i - x_{i+1}$	Число хозяйств в группе n_i	Среднее значение интервала x_i^*	$x_i^* n_i$	$x_i^{*2} n_i$
До 0,06	65	0,04	2,60	0,1040
0,06–0,10	112	0,08	8,96	0,7168
0,10–0,14	167	0,12	20,04	2,4048
0,14–0,18	170	0,16	27,20	4,3520
0,18–0,22	115	0,20	23,00	4,6000
0,22–0,26	79	0,24	18,96	4,5504
Свыше 0,26	7	0,28	1,96	0,5488
Итого	715	–	102,72	17,2768

Таблица 13.5

Расчет теоретических частот вариационного ряда

Интервалы $x_i - x_{i+1}$	n_i	t_i	t_{i+1}	$\Phi(t_i)$	$\Phi(t_{i+1})$	p_i	n_i'	$\frac{(n_i - n_i')^2}{n_i'}$
До 0,06	65	–∞	–1,45	–0,5	–0,4265	0,0735	52,6	2,92
0,06–0,10	112	–1,45	–0,76	–0,4265	–0,2764	0,1501	107,3	0,20
0,10–0,14	167	–0,76	–0,07	–0,2764	–0,0279	0,2485	177,7	0,64
0,14–0,18	170	–0,07	0,62	–0,0279	0,2324	0,2603	186,1	1,39
0,18–0,22	115	0,62	1,31	0,2324	0,4049	0,1725	123,3	0,56
0,22–0,26	79	1,31	2,00	0,4049	0,4772	0,0723	51,7	14,42
Свыше 0,26	7	2,00	∞	0,4772	0,5	0,0228	16,3	5,31
Итого	715	–	–	–	–	1,0000	715,0	25,44

В результате вычислений получили $\chi^2_{\text{н}} = 25{,}44$. Найдем по таблице квантилей χ^2-распределения (приложение 2), при заданном уровне значимости α=0,01 и числе степеней свободы $\nu = k - r - 1 = 7 - 2 - 1 = 4$, критическое значение:

$$\chi^2_{\text{кр.}} = \chi^2_{0{,}01;\,4} = 13{,}28.$$

Так как $\chi^2_{\text{н.}} > \chi^2_{\text{кр.}}$ (25,44>13,28), то имеются основания для отклонения нулевой гипотезы о нормальном законе распределения личных подсобных хозяйств по размеру земельной площади с имеющимися параметрами $a = 0{,}144$ и $\sigma = 0{,}058$.

Если принять уровень значимости α = 0,05 при числе степеней свободы $\nu = k - r - 1 = 7 - 2 - 1 = 4$, то критическое значение:

$$\chi^2_{\text{кр.}} = \chi^2_{0{,}05;\,4} = 9{,}49.$$

При этом уровне значимости наблюдаемое значение критерия также больше критического, поэтому нулевая гипотеза отвергается и нет оснований утверждать, что распределение ЛПХ по земельной площади описывается нормальным законом.

Замечание.

1. Так как случайная величина *X*, распределенная по нормальному закону, определена на $(-\infty; +\infty)$, то в примере 13.9 наименьшее значение стандартизованной переменной заменено на –∞, а наибольшее значение заменено на $+\infty$.

2. Применение критерия χ^2 для проверки гипотезы о нормальности распределения предполагает наличие в каждом частичном интервале не менее пяти единиц, в противном случае желательно объединять эти интервалы с соседними.

3. Проверка гипотезы о принадлежности случайной величины показательному, биномиальному, пуассоновскому или другому распределению основывается на применении в описанном алгоритме соответствующих функций распределения или плотностей распределения.

4. Вид альтернативной гипотезы существенно влияет на статистический вывод, поэтому ее нужно выдвигать, исходя из реального смысла рассматриваемой задачи. Практически в качестве альтернативной гипотезы следует выбрать такую, которая побуждает лицо, принимающее решение, к действиям. Например, при проверке гипотезы о том, что средний вес взятого продукта соответствует объявленному номиналу (H_0), в качестве альтернативной гипотезы (H_1) следует принять гипотезу — средний вес меньше номинала (покупателя обвешивают).

5. Отклонение нулевой гипотезы не есть доказательство верности альтернативной. Если необходимо проверить H_1, то для этого нужно взять уже другую выборку.

6. Принимая H_0, мы отнюдь не доказываем ее абсолютную истинность. Можно лишь утверждать при принятии H_0 то, что она не противоречит имеющимся данным с вероятностью $(1 - \alpha)$ (например, при $\alpha = 0{,}05$ нет противоречия в 95 случаях из 100). Ведь достаточно всего одного прецедента, чтобы опровергнуть принимаемую гипотезу. Например, наше предположение по виду и поведению человека о том, что он бомж, может быть опровергнуто наличием у него счета в банке на большую сумму. ■

13.9. Проверка гипотезы об однородности выборок

Если имеются две независимые выборки, то проверяется гипотеза, что эти выборки образованы из одной и той же генеральной совокупности. Гипотеза может быть проверена с помощью критерия Колмогорова — Смирнова.

Пусть имеются две независимые генеральные совокупности с неизвестными функциями распределения $F_1(x)$ и $F_2(x)$. Из каждой генеральной совокупности взята выборка объемами n_1 и n_2 соответственно. Проверяется нулевая гипотеза $H_0: F_1(x) = F_2(x)$ — о равенстве функций распределения, при конкурирующей гипотезе $H_1: F_1(x) \neq F_2(x)$. Предполагается, что функции распределения непрерывны, а проверка нулевой гипотезы проводится с помощью критерия Колмогорова — Смирнова. Расчет критерия основан на сравнении двух эмпирических функций распределения.

Наблюдаемое значение статистики Колмогорова — Смирнова определяется по формуле

$$\lambda_{\text{н.}} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \max\left|F_{n_1}(x) - F_{n_2}(x)\right|, \qquad (13.46)$$

где $F_{n_1}(x)$ — эмпирическая функция распределения по первой выборке объема n_1; $F_{n_2}(x)$ — эмпирическая функция распределения по второй выборке объема n_2.

Замечание. Н. В. Смирнов доказал, что если $F_{n_1}(x) = F_{n_2}(x)$ и непрерывны, то при $n_1, n_2 \to \infty, (n_1/n_2) = t, 0 < t < \infty$ величина

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \max\left|F_{n_1}(x) - F_{n_2}(x)\right| \to \lambda, \qquad (13.47)$$

где λ имеет закон распределения, определенный функцией Колмогорова $K(\lambda)$.

Поэтому соответствующий критерий проверки гипотезы о том, что две выборки взяты из генеральной совокупности с одним законом распределения называют статистикой Колмогорова — Смирнова.

Эмпирическая функция распределения — это накопленные частости распределения значений признака по вариационному ряду распределения.

Нулевая гипотеза принимается, если наблюдаемое значение критерия λ меньше критического значения при уровне значимости α. Нулевая гипотеза отвергается, если наблюдаемое значение критерия больше критического значения. Объем выборок должен быть большой ($n_1 \geq 50;\ n_2 \geq 50$). ■

Пример 13.10. Выборочным методом изучались среднедушевые месячные доходы семей по двум регионам. Можно ли утверждать, что среднемесячные доходы семей описываются одной и той же функцией распределения при уровне значимости 0,05.

Решение. Сравнивая модули разности функций видно, что

$$max\left| F_{n_1}(x) - F_{n_2}(x)\right| = 0{,}108.$$

Тогда наблюдаемое значение критерия Колмогорова — Смирнова равно

$$\lambda_{\text{н.}} = \sqrt{\frac{650 \cdot 750}{650+750}} \cdot 0{,}108 = 2{,}015.$$

При уровне значимости α=0,05 критическое значение критерия Колмогорова — Смирнова $\lambda_{\text{кр.}}$=1,36. Так как $\lambda_{\text{н.}} > \lambda_{\text{кр.}}$, то нулевая гипотеза отвергается, значит, выборочные совокупности семей описываются разными функциями распределения. Эти совокупности семей по величине среднемесячного дохода на одного члена семьи являются не однородными.

Для проверки нулевой гипотезы об однородности двух независимых выборок, если данные представлены в виде интервальных вариационных рядов можно воспользоваться критерием χ^2.

Таблица 13.6

Распределение семей по величине среднемесячных доходов

Группы семей по среднемесячному доходу на члена семьи, тыс. руб.	Число семей по региону		Накопленное число семей		$F_{n_1}(x) = \frac{S_{1i}}{n_1}$	$F_{n_2}(x) = \frac{S_{2i}}{n_2}$	$\left\lvert F_{n_1}(x) - F_{n_2}(x)\right\rvert$
	n_{1i}	n_{2i}	S_{1i}	S_{2i}			
До 10,0	108	76	108	76	0,166	0,101	0,065
10,0–20,0	231	235	339	311	0,522	0,414	0,108
20,0–30,0	127	158	466	469	0,718	0,625	0,093
30,0–40,0	72	99	538	568	0,828	0,757	0,071
40,0–50,0	41	61	579	629	0,891	0,839	0,052
50,0–60,0	25	39	604	668	0,929	0,891	0,038
60,0–70,0	20	25	624	693	0,960	0,924	0,036
Свыше 70,0	26	57	650	750	1,000	1,000	
Сумма	650	750					

Наблюдаемое значение критерия находится по формуле

$$\chi^2_{\text{н.}} = n\left(\sum_{i=1}^{m}\sum_{j=1}^{l}\frac{n_{ij}^2}{n_{i*}n_{*j}} - 1\right),$$

где $n_{i*} = \sum_{j=1}^{l} n_{ij};\ n_{*j} = \sum_{i=1}^{m} n_{ij};\ n = \sum_{i=1}^{m} n_{i*} = \sum_{j=1}^{l} n_{*j}.$

Критическое значения критерия находится по таблице значений χ^2— критерия Пирсона при уровне значимости α и числе степеней свободы $k = (m-1)(l-1)$. Если наблюдаемое значение критерия меньше критического значения, то нулевая гипотеза об однородности двух совокупностей принимается, а в противном случае она отвергается.

Пример 13.11. По данным предыдущего примера проверить нулевую гипотезу об однородности двух выборочных совокупностей семей по среднемесячному доходу на одного члена семьи.

Решение. Исходные данные представим в таблице 13.7.

Таблица 13.7

Расчет теоретических частот

Интервалы		0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	$n_{*j} = \sum_{i=1}^{9} n_{ij}$
Частоты	n_{1i}	108	231	127	72	41	25	20	26	650
	n_{2i}	76	235	158	99	61	39	25	57	750
$n_{i*} = \sum_{j=1}^{2} n_{ij}$		184	466	285	171	102	64	45	83	n=1400
$n_{1i}^2/(n_{i*}n_{*j})$		0,098	0,176	0,087	0,047	0,025	0,015	0,014	0,012	0,474
$n_{2i}^2/(n_{i*}n_{*j})$		0,042	0,158	0,117	0,076	0,049	0,032	0,018	0,052	0,544

$\frac{108^2}{650\cdot184} = 0{,}098; \frac{231^2}{650\cdot466} = 0{,}176$ и т. д.

$$\chi^2_{\text{н.}} = 1400(0{,}474 + 0{,}544 - 1) = 25{,}2.$$

При уровне значимости $\alpha = 0{,}05$ и числе степеней свободы

$$k = (2-1)(8-1) = 7,\ \chi^2_{\text{кр.}} = 14{,}1.$$

Так как $\chi^2_{\text{н.}} > \chi^2_{\text{кр.}}$, то нулевая гипотеза об однородности двух совокупностей отвергается.

13.10. Проверка гипотезы о независимости выборок

Для проверки гипотезы о независимости пары признаков A и B, представленных таблицей сопряженности, может использоваться модернизированный критерий хи-квадрат. Таблица представляет собой частоты встречаемости уровней изучаемых признаков. По столбцам расположены градации фактора B, по строкам — градации фактора A, n_{ij} — частота встречаемости наблюдений, обладающих признаками $\mathrm{A}_i B_j$, суммы по строкам и столбцам — маргинальные частоты (n_i — для i −ой строки, N_j — для j-го столбца), N — общее число наблюдений.

Для проверки нулевой гипотезы

$$H_0: P\left(A_i \mathrm{B}_j\right) = P(A_i)P\left(\mathrm{B}_j\right),$$

рассмотрим статистику

$$\chi^2 = \sum_{j=1}^{m} \sum_{i=1}^{k} \frac{(n_{ji} - N\hat{P}(\mathrm{A}_i)\hat{P}(\mathrm{B}_j))^2}{N\hat{P}(\mathrm{A}_i)\hat{P}(\mathrm{B}_j)}. \qquad (13.48)$$

Признак A	Признак B						Σ
	B_1	B_2	$\cdots$	B_j	$\cdots$	B_m	
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1m}	n_1
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2m}	n_2
A_3	n_{31}	n_{32}	$\cdots$	n_{3j}	$\cdots$	n_{3m}	n_3
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{im}	n_i
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{km}	n_k
Σ	N_1	N_2	$\cdots$	N_j	$\cdots$	N_m	N

Принятие статистического решения опирается на следующую теорему.

Теорема. Если гипотеза о независимости признаков A и B, представленных таблицей сопряженности, верна, то при $n \to \infty$

$$\chi^2 \to \chi^2_{(m-1)(k-1)}. \qquad (13.49)$$

Доказательство этой теоремы основывается на использовании критерия обобщенного отношения правдоподобия (см. 13.12).

Пример 13.12. Рассмотрим данные Ф. Гальтона, приводимые В. Романовским и Н. Дружининым [40]. A_1 и B_1 — голубой цвет глаз у отца и сына, A_2 и B_2 — зеленый и серый, A_3 и B_3 — темно-серый, A_4 и B_4 — карий и черный.

i \ j	B_1	B_2	B_3	B_4	Σ
A_1	194	83	25	56	358
A_2	70	124	34	36	264
A_3	41	41	55	43	180
A_4	30	36	23	109	198
Σ	335	284	137	244	1000

Решение. Применение критерия независимости опирается на сопоставление эмпирических наблюдений и теоретически ожидаемых численностей. Положим, что признаки A и B независимы, тогда расслоение параллельных групп будет пропорционально, то есть относительное распределение отцов и сыновей по группам (признакам A и B).

Последовательно умножая итоги по строкам и столбцам и разделив произведения на общую численность (например, $\frac{335 \cdot 358}{1000} = 119{,}930$; $\frac{284 \cdot 358}{1000} = 101{,}672$; …; $\frac{244 \cdot 198}{1000} = 48{,}312$), получим:

i \ j	B_1	B_2	B_3	B_4	Σ
A_1	119,930	101,672	49,046	87,352	358,000
A_2	88,440	74,976	36,168	64,416	264,000
A_3	60,300	51,120	24,660	43,920	180,000
A_4	66,330	56,232	27,126	48,312	198,000
Σ	335,000	284,000	137,000	244,000	1000,000

Рассчитаем величину хи-квадрат по формуле (13.48). Используя MS Excel, получим следующее.

45,74639	3,429101	11,78914	11,25272
3,844794	32,05496	0,129955	12,53523
6,177280	2,003412	37,32829	0,019271
19,89852	7,279375	0,627585	76,23434

Отсюда

$$\chi^2 = \left(\frac{(194-119{,}93)^2}{119{,}93} + \frac{(70-88{,}44)^2}{88{,}44} + \cdots + \frac{(109-48{,}312)^2}{48{,}312}\right) =$$
$$= 45{,}74639 + 3{,}844797 + \cdots + 76{,}23434 = 270{,}350.$$

Критическое значение распределения хи-квадрат Пирсона имеет $k = (m-1)(k-1)$ степеней свободы, то есть $k = (4-1)(4-1) = 9$. Используя таблицы или доступные программы можно найти, что при уровне значимости $\alpha = 0{,}05$ и $\alpha = 0{,}01$ критические значения соответственно равны 16,919 и 21,666. Сравним $\chi^2_{\text{н.}}$ и $\chi^2_{\text{кр.}}$, получим, что $\chi^2_{\text{н.}} > \chi^2_{\text{кр.}}$. Таким образом, гипотезу о независимости следует отклонить — изучаемые признаки связаны. Пользуясь пакетом *gretl*, легко найти, что для $\chi^2 = 270{,}350$ $p-$значение $(p-value)$ равно 4,98934e-053.

Замечание. 1. Величина χ^2 может указать, что отклонения от теоретических частот, вычисленные в предположении независимости факторов, имеют такие значения, что их нельзя назвать случайными. Кроме того, величина χ^2 является основой для вычисления коэффициента сопряженности Чупрова, позволяющего измерить степень связи признаков:

$$K = \frac{\varphi^2}{\sqrt{(m-1)(l-1)}} = \frac{\chi^2/N}{\sqrt{(m-1)(l-1)}}.$$

В примере 13.12, $K = \frac{270{,}350/1000}{\sqrt{(4-1)(4-1)}} = \frac{0{,}270350}{3} = 0{,}0901.$

2. «…открытие критерия “хи-квадрат”, — пишет Н. К. Дружинин, — несомненно, принадлежит к числу таких важных открытий в математической статистике, которые способны раскрыть единство основ многих математико-статистических методов» [40]. ■

13.11. Проверка гипотезы о случайности выборок

Множество примеров применения генераторов случайных чисел (см. часть I глава 10) приводит к мысли о проверке случайности используемых «случайных последовательностей» чисел и, соответственно, о качестве работы генераторов случайных чисел. Обычно говорят о необходимости использования нескольких критериев для проверки гипотезы случайности, большое количество подобных критериев рассматривается во втором томе «Искусства программирования» Д. Кнута [60].

Один из основных критериев проверки, используемый в сочетании с другими критериями, — критерий хи-квадрат Пирсона.

Пример 13.13. Подбрасываются две игральные кости, проверим гипотезу о том, что они «правильные», если имеется таблица следующих результатов для сумм числа очков на верхних гранях *S*: вероятностей (p_S), эмпирических частот (n_S), теоретических частот (np_S).

S	2	3	4	5	6	7	8	9	10	11	12
p_S	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$
n_S	10	16	24	38	48	50	41	38	28	20	7
np_S	8,889	17,778	26,667	35,556	44,444	53,333	44,444	35,556	26,667	17,778	8,889

Решение. Опираясь на исходные данные и формулу (13.48), получим

$$\chi^2_{\text{н.}} = \left(\frac{(10-8{,}889)^2}{8{,}889} + \frac{(16-17{,}778)^2}{17{,}778} + \cdots + \frac{(7-8{,}889)^2}{8{,}889}\right) = 2{,}425.$$

Если принять уровень значимости α=0,05 при числе степеней свободы $\nu = k - 1 = 11 - 1 = 10$, то критическое значение равно

$$\chi^2_{\text{кр.}} = \chi^2_{0{,}05;\,10} = 3{,}94.$$

Так как $\chi^2_{\text{н.}} < \chi^2_{\text{кр.}}$ (2,425 < 3,94), то имеются основания для принятия нулевой гипотезы о «правильности» игральных костей.

13.12. Оптимальные критерии*

Теория построения оптимальных критериев известна давно (начало ее создания относится к 1928–1933 годам, см. замечание далее). Практические приложения появились примерно в 1940–1950-е годы и остаются актуальными, например, в задачах радиотехники [71, 72], для выделения полезного сигнала на фоне помех в различных информационных ситуациях на земле, на поверхности воды и в космосе [34, 39, 103, 115]. В частности, например, при конструировании радиоприёмника или радиолокационной системы (РЛС), минимизирующих не средний квадрат ошибки, а вероятность ошибки (с известными штрафами за различные виды ошибок) при обнаружении сигнала (с известным законом распределения, функцией плотности вероятности) на фоне аддитивного гауссова шума. Статистические аспекты теории сигналов опираются на понятие пространства сигналов (функциональное пространство), поэтому изложение предполагает использование функционального анализа и требует введения более глубоких математических понятий. Актуальность этого подхода в области изучения социально-экономических и других систем стала осознана примерно с 1990-х годов в связи с ростом количества имеющихся данных и возможностей их обработки. Поэтому наше изложение в настоящем разделе преимущественно опирается на теоретические источники, в основном [12, 34, 46, 54, 71–73, 85, 92, 98, 111, 125], и при первом чтении может быть опущено, так как практическая значимость вводимых понятий сразу не очевидна, однако следует понимать, что желающие заниматься анализом данных и далее машинным обучением, встретятся с материалами, основывающимися на введенных понятиях.

Проверка статистических гипотез, опирающаяся только на уровень значимости α, очень часто позволяет удовлетворительно решать задачи. Используемые при этом критерии называются критериями значимости. Они позволяют с фиксированным риском отклонить нулевую гипотезу, в противном случае считается, что нет оснований для отклонения нулевой гипотезы. Вероятность ошибки второго рода (вероятность принятия ложной нулевой гипотезы) остается неизвестной, поэтому критерии значимости не позволяют принять решение о том, что нулевая гипотеза

верна. Эта ситуация обычно удовлетворяет исследователей, например, нулевая гипотеза — сроки службы изделий, изготовленных по новой и старой технологиям, равны, альтернативная — срок службы изделий, изготовленных по новой технологии, увеличился. Критерий значимости позволяет отклонить нулевую гипотезу, и тогда повышение срока службы изделия, благодаря новой технологии, может считаться доказанным, в противном случае у нас нет оснований, по имеющимся статистическим данным, отклонить нулевую гипотезу.

Практическое применение критерия значимости приводит к трем вариантам статистического решения: нулевая гипотеза отвергается, нулевая гипотеза принимается, никакого действия не принимается до получения новых сведений. Будем считать, что мы располагаем всеми доступными данными, поэтому здесь рассматриваем первые два случая.

Полагая, что нам важно не отвергнуть гипотезу H_0, в случае, когда она верна, и не принять H_1, мы задаем *вероятность ошибки первого рода* α (*вероятность* «*фальшивого открытия*»). Тогда, если для нас менее опасно совершение ошибки второго рода при принятии ложной гипотезы H_0, можно при заданном уровне значимости α минимизировать *вероятность ошибки второго рода* (*вероятность* «*упустить открытие*») или иначе максимизировать мощность критерия $M = 1 - \beta,$ при заданном уровне значимости α, соответствующая критическая область называется *наилучшей критической областью* (НКО). Критерий называют критерием размера α.

Ошибка первого рода возникает, когда отвергается верная гипотеза, ошибка второго рода появляется, когда не отвергается ложная гипотеза. Поэтому статистический критерий должен не отвергать верные гипотезы и отвергать ложные, то есть ошибка второго рода должна быть как можно меньше и соответственно мощность критерия как можно больше.

Пусть рассматривается выборка независимых и одинаково распределенных случайных величин $X = (x_1, x_1, \dots, x_n)$, извлеченных из одной генеральной совокупности, закон распределения которой при выполнении нулевой (H_0) и альтернативной (H_1) гипотезы соответственно описывается функциями плотности вероятности $f_0(X/\ H_0)$ и $f_1(X/\ H_1)$.

Тогда соответствующие функции правдоподобия примут вид

$$L_0(X) = \prod_{i=1}^{n} f_0(x_i/H_0) \text{ и } L_1(X) = \prod_{i=1}^{n} f_1(x_i/H_1).$$

Критическая область $C,$ соответствующая уровню значимости α (размер критической области или размер критерия) в непрерывном случае может быть представлена как

$$P_0(C) = \int f_0(X) I(X \in C) dX = \alpha,$$

следовательно, мощность критерия

$$P_1(C) = \int f_1(X) I(X \in C) dX = 1 - \beta \text{ —}$$

величина, которую нужно максимизировать.

Оказывается, что все критерии, опирающиеся на лемму Фишера и предполагающие однородность и нормальность распределения изучаемых случайных величин, являются наиболее мощными (см. 13.2–13.7), поэтому задача проверки гипотез решается эффективно.

Имеется целый ряд ситуаций, когда задача проверки статистических гипотез, учитывающая вероятности ошибок первого и второго рода, решается эффективно, определяется критериями, построенными по типу критерия Неймана — Пирсона, последовательного анализа А. Вальда, использующими теорию статистических решений, и т. д.

Пусть рассматривается две простые гипотезы (например, $H_0: \Theta = \Theta_0, H_1: \Theta = \Theta_1$) и сравнивается два статистических критерия K_1 и K_2 с вероятностями ошибок первого и второго рода соответственно: $\alpha(K_1) = : \alpha_{H_0}, \beta(K_1), \alpha(K_2) =: \alpha_{H_1}, \beta(K_2)$. Может рассматриваться несколько подходов к сравнению критериев, имеющих корни и продолжение в современной теории статистических решений.

Минимаксный подход. Критерий K_1 не хуже критерия K_2 в смысле минимаксного подхода, если

$$max\ \{\alpha(K_1), \beta(K_1)\} \leq max\ \{\alpha(K_2), \beta(K_2)\}. \qquad (13.50)$$

Критерий K_1 называется минимаксным, если он не хуже других критериев в смысле минимаксного подхода.

Байесовский подход. Пусть $P(H_0) = p_0, P(H_1) = 1 - p_0 = p_1$ — вероятности гипотез. Известна функция потерь (может рассматриваться функция полезности):

$$R = c_0 p_0 + c_1 p_1, \qquad (13.51)$$

где c_0 и c_1 — потери (в случае полезности — прибыль), получаемые при принятии соответствующих гипотез.

Критерий K_1 не хуже критерия K_2 в смысле байесовского подхода, если

$$R(K_1) = c_0 p_0(K_1) + c_1 p_1(K_1) \leq R(K_2) = c_0 p_0(K_2) + c_1 p_1(K_2).$$

Критерий называется байесовским, если он не хуже других в смысле минимаксного подхода.

Наиболее мощный критерий (*НМК*). Критерий K_1 называется наиболее мощным, если при заданном уровне значимости α для любого другого критерия K_2 верно неравенство

$$\beta(K_1) \leq \beta(K_2)$$

или

$$M(K_1) = 1 - \beta(K_1) \geq M(K_2) = 1 - \beta(K_2).$$

Рассмотрим отношение правдоподобия

$$\lambda_{H_1, H_0}(X) = \frac{L_1(X)}{L_0(X)}. \qquad (13.52)$$

Если для данной выборки значение $\lambda_{H_1 H_0} =: \lambda$ велико, то нулевая гипотеза H_0 маловероятна, в противном случае мы полагаем, что H_0 следует принять (*принцип отношения правдоподобия*).

Обычно задаются некоторым числом $k(\alpha) =: k$, из множества векторов $X = \{x_1, x_1, \dots, x_n\} \in R^n$ выбирается такая область C, что если $X \subset$ C, то $\lambda \geq k$.

Замечание. Перечисленные выше подходы имеют прямое отношение к исследованию операций, науке, получившей развитие в годы Второй мировой войны в Англии, целью которой было помочь командованию принять научно обоснованные решения по руководству боевыми операциями. В дальнейшем

успешность применения теории позволила использовать ее при управлении сложными социально-экономическими и техническими системами. Сегодня *задача выбора и принятия решений — это одна из задач системного анализа*. С самого начала была понятна субъективность процесса принятия решений, что выражалось в критериях, часть из которых отражена выше. Опираясь на цель задачи принятия решений, выделяют «классические» критерии оптимальности, предполагающие несколько вариантов состояния психики человека.

1. *Минимаксный, отражающий мышление пессимиста* («если неприятность может случиться, то она случится ...»), полагающего выбор решения с минимальными потерями при наихудшей ситуации (состоянии природы), то есть минимальны максимальные потери (отсюда и название критерия).

2. *Критерий оптимизма-пессимизма Гурвица*, предлагающий каждому человеку поставить в соответствие показатель оптимизма λ — число в интервале от 0 до 1. Лучшее решение то, для которого линейная комбинация потерь меньше:

$$R(H_1, H_2) = \lambda\, min\{c_0, c_1\} + (1 - \lambda)\; max\;\{c_0, c_1\}.$$

При $\lambda=0$ критерий превращается в минимаксный (крайний пессимист), при $\lambda=1$ (крайний оптимист) в «миниминный».

3. *Минимаксного сожаления* (*Сэвиджа*), который моделирует ситуацию угрызений совести по поводу принятия неудачного решения (H_i) с ценой c_i, сожаление можно записать как

$$s_i = c_i - min\;\{c_i\}.$$

К сожалениям s_i следует применить минимаксный критерий: из двух решений лучше то, для которого сожаление при наименее благоприятном состоянии меньше

$$max\; s_i \rightarrow min.$$

4. *Критерий Байеса* полагает, что состояния природы — случайные события, причем состояния природы заранее известны (априорное распределение) и могут определяться в результате обработки наблюдений. Качество решения разумно оценивать средними (в смысле математического ожидания) потерями (13.51). Полагается, что из двух решений лучше то, для которого средние потери меньше.

Если априорные вероятности неизвестны, то можно воспользоваться предположением равновероятности событий

$$p_1 = p_2 = \cdots = p_n = \frac{1}{n},$$

которое следует из принципа недостаточности оснований Лапласа: если нет оснований считать, что одни события более вероятны, чем другие, то следует принять, что они равновероятны. В этом случае байесовский критерий называется *критерием Лапласа*.

5. *Критерий Неймана — Пирсона.* Рассматривается игра с двумя состояниями природы, одно из которых (гипотеза H_0) более важно и находится под контролем. Вводится пороговая величина k, все решения, которые при контролируемом состоянии имеют потери большие, чем k, считаются недопустимыми.

Из двух решений лучшим считается то, у которого потери в состоянии H_1 меньше. Требуется обеспечить малую вероятность ошибки первого рода, а затем второго рода.

6. *Рандомизированное* (*random* (*англ.*) — случайный) *решение* определяется, в отличие от рассмотренных ранее, случайно, с учетом полезности принятия решения:

$$U(j) = \sum_{i=1}^{m} c_{ij} x_{ij},$$

где c_{ij} — цена принятия решения x_{ij}.

7. Сегодня наиболее известны три асимптотически эквивалентных подхода к поиску критериев: тест Вальда, множителей Лагранжа, отношения правдоподобия [45].

Определим нерандомизированный *критерий отношения правдоподобия* δ_{C}:

$$\delta_{\mathrm{C}}(X) = \begin{cases} H_0, \text{если } \lambda < k, \\ H_1, \text{если } \lambda \geq k. \end{cases} \tag{13.53}$$

Имеет место следующая теорема [125].

Теорема. *Критерий отношения правдоподобия является:*

1) *минимаксным* при k таком, что

$$\alpha_{H_0}(\delta_{\mathrm{C}}) = \alpha_{H_1}(\delta_{\mathrm{C}}), \tag{13.54}$$

2) *байесовским* при

$$k = \frac{p_0}{p_1}, \tag{13.55}$$

3) (*лемма Неймана — Пирсона*) критерий δ_{C} размера α, для любого значения $k > 0$, с областью $\mathrm{C} = \{X: L_1(X) > kL_0(X)\}$, имеет наибольшую мощность $(P_1(C) \coloneqq 1 - \beta)$.

Решающая функция δ_{C} (правило, согласно которому принимается или отвергается нулевая гипотеза) может принимать два значения $\{0, 1\}$.

Итак, критерий δ_{C} является *наиболее мощным* среди всех критериев (критических областей C^*) размера $\alpha = \alpha(k) = P_0(C)$.

Наметим *доказательство* леммы Неймана — Пирсона в непрерывном случае. При $X \subset C$: $L_1(X) > kL_0(X)$, $I_C(X) \geq I_{C^*}(X)$ [54]. Рассмотрим неравенство

$$\int (I_C(X) - I_{C^*}(X))(L_1(X) - kL_0(X)dX \geq 0, \tag{13.56}$$

так как множители под знаком интеграла имеют одинаковые знаки ($I_C(X)$, $I_{C^*}(X)$ — индикаторные функции для C и C^*). Если ввести обозначения

$$\int (L_1(X) I_C(X) dX \coloneqq P_1(C) = 1 - \beta, \int (L_1(X) I_{C^*}(X) dX \coloneqq P_1(C^*) = \beta,$$
$$\int (L_0(X) I_C(X) dX \coloneqq P_0(C) = \alpha, \int (L_0(X) I_{C^*}(X) dX \coloneqq P_0(C^*) = 1 - \alpha,$$

то неравенство (13.56) можно переписать как

$$P_1(C) - P_1(C^*) \geq k\big(P_0(C) - P_0(C^*)\big).$$

Если выполняется условие $\beta = P_0(\mathrm{C}^*) \leq P_0(C) = \alpha = \alpha(k)$, то

$$P_1(C) \geq P_1(C^*)$$

или

$$1 - \beta \geq P_1(C^*),$$

то есть критерий с критической областью C имеет наибольшую мощность среди всех критериев с критической областью C^* размера:

$$\beta = P_0(\mathrm{C}^*) \leq P_0(C) = \alpha.$$

Фактически на практике решается задача отыскания для заданного уровня значимости α такого критерия $\mathrm{C} = \{X: L_1(X) > kL_0(X)(X)\}$, что

$$P(C) = \alpha \tag{13.57}$$

(знак больше можно заменить на знак больше или равно:

$$\mathrm{C} = \{X: L_1(X) \geq kL_0(X)\}).$$

Итак, наилучшая критическая область при проверке простой гипотезы H_0 против простой альтернативы H_1 критерия размера α определяется леммой Неймана — Пирсона. При заданном уровне значимости α наилучшая критическая область (НКО) состоит из точек выборочного пространства (выборок объема n), удовлетворяющих неравенству

$$\lambda_{H_1,H_0}(X) = \frac{L_1(X)}{L_0(X)} \geq k(\alpha),$$

где λ определяется как мера расхождения имеющихся в нашем распоряжении выборочных данных с проверяемой гипотезой (H_0).

Для дискретных распределений возникает некоторая трудность, заключающаяся в том, что в упорядоченной совокупности включение очередной точки в критерий отношения правдоподобия не позволяет достичь уровня значимости α, а включение следующей приводит значению уровня значимости, превосходящего α. Поэтому переходят к *рандомизированным критериям*, позволяющим «расщепить» очередную точку и получить суммарную вероятность, равную α. На практике уровень значимости стараются фиксировать, увеличивая или уменьшая первоначальное значение.

Определим рандомизированный *критерий отношения правдоподобия* δ_{C}^*:

$$\delta_{\mathrm{C}}^*(X) = \begin{cases} H_0, \text{если } \lambda < k, \\ H_1, \text{если } \lambda > k, \\ k(\alpha), \text{если } \lambda = k. \end{cases} \qquad (13.58)$$

Критерий, максимизирующий мощность критерия при проверке гипотезы H_0, обычно зависит от альтернативы H_1. Если альтернатива одна, то задача сводится к поиску максимума некоторого интеграла при ряде ограничений. Критерии, которые максимизируют мощность всех альтернатив, называют *равномерно наиболее мощными* (РНМ).

Если существует такая функция $T(x)$, что отношение правдоподобия есть неубывающая функция от $T(x)$, то зависящее от параметра семейство плотностей распределения вероятностей имеет *монотонное отношение правдоподобия*.

Для монотонного отношения правдоподобия относительно θ — параметра изучаемого распределения, существует равномерно наиболее мощный критерий проверки гипотезы $H_0: \theta \leq \theta_0$ при $H_1: \theta > \theta_0$.

На практике обычно избегают рандомизированных критериев, варьируя уровнем значимости α.

Пример 13.14. Построим оптимальный критерий для гипергеометрического закона.

Решение. Пусть имеется множество из N изделий, случайным образом без возвращения отбирается n изделий, $n < N$.

Если множество содержит $M \leq N$ бракованных изделий, то число бракованных изделий в выборке имеет вероятность появиться:

$$H_0: P_{n0} = \frac{C_M^x C_{N-M}^{n-x}}{C_N^n},$$

где

$$M \le N, x \le n;\ x = m_0, m_0+1, m_0+2, \dots,\ min(M,n),\ m_0 = max\{0, n-(N-M)\}.$$

$$H_1: P_{n0} = \frac{C_{M+1}^{x} C_{N-M-1}^{n-x}}{C_N^n},$$

где $M+1 \le N, x \le n;\ x = m_0, m_0+1, m_0+2, \dots,\ min(M+1,n)$,

$$m_0 = max\{0, n-(N-M-1)\}.$$

Рассмотрим отношение правдоподобия:

$$\lambda_{H_1,H_0}(X) = \frac{L_1(X)}{L_0(X)} = \frac{M+1}{N-M}\frac{N-M-n+x}{M+1-x} \ge k. \qquad (13.59)$$

Таким образом, изучаемые распределения имеют монотонное отношение правдоподобия с $T(x) = x$, следовательно, существует равномерно наиболее мощный критерий для проверки гипотезы $H_0: M \le M_0$ при $H_1: M > M_0$.

Пример 13.15. Если в условиях предыдущего примера изделие возвращается обратно, то это приводит к повторным независимым испытаниям по схеме Бернулли с вероятностью $p = \frac{M}{N}$. Построим оптимальный критерий.

Решение. Пусть $0 < p_0 < p_1 < 1$.

$$H_0: P_{n0} = C_n^x p_0^x (1-p_0)^{n-x}, x = \overline{0,n.}$$

$$H_1: P_{n1} = C_n^x p_1^x (1-p_1)^{n-x}, x = \overline{0,n.}$$

Рассмотрим отношение правдоподобия

$$\lambda_{H_1,H_0}(X) = \frac{L_1(X)}{L_0(X)} = \left[\frac{p_1(1-p_0)}{p_0(1-p_1)}\right]^x \left[\frac{(1-p_1)}{(1-p_0)}\right]^n \ge k, \qquad (13.60)$$

распределения также имеют монотонное отношение правдоподобия с $T(x) = x$, следовательно, существует равномерно наиболее мощный критерий для проверки указанных гипотез.

Согласно лемме Неймана — Пирсона, существует такое $k(\alpha)$, что

$$\alpha = P(x \ge k(\alpha)/H_0),$$

$$\beta = P(x < k(\alpha)/H_1).$$

Учитывая, согласно интегральной теореме Муавра — Лапласа, что число положительных успехов x асимптотически нормально, то есть

$x \to N(np, np(1-p))$, получим

$$\alpha = P(x \ge k(\alpha)/H_0) = P\left(u_\alpha = \frac{x-np_0}{\sqrt{np_0(1-p_0)}} \ge \frac{k(\alpha)-np_0}{\sqrt{np_0(1-p_0)}}/H_0\right),$$

$$\beta = P(x < k(\alpha)/H_1) = P\left(-u_\beta = \frac{x-np_1}{\sqrt{np_1(1-p_1)}} < \frac{k(\alpha)-np_1}{\sqrt{np_1(1-p_1)}}/H_1\right).$$

$\Phi(u_\alpha) = 1-\alpha = 1-P(x \ge k(\alpha)/H_0)$ и силу того, что $p_0 < p_1$ (для геометрической визуализации, на рис. 13.4 достаточно поменять местами гипотезы H_0 и H_1 и, соответственно, α и β — рис. 13.5):

$$\Phi(-u_\beta) = \beta = P(x < k(\alpha)/H_1),$$

то получим, что при заданных α и β границу $k(\alpha) =: k$ можно определить, как

$$k \approx np_0 + u_\alpha\sqrt{np_0(1-p_0)} \approx np_1 - u_\beta\sqrt{np_1(1-p_1)}\,. \qquad (13.61)$$

Это и есть искомый критерий.

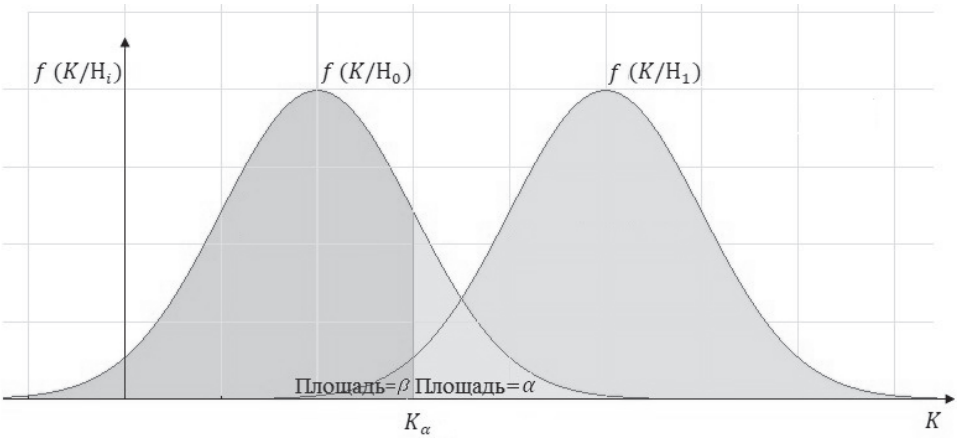


Рис. 13.5 — Геометрическая интерпретация вероятностей ошибок первого, второго рода и мощности критерия K для правосторонней критической области

Отсюда необходимый объем выборки:

$$n \approx \frac{\left(u_\alpha\sqrt{np_0(1-p_0)}+u_\beta\sqrt{np_1(1-p_1)}\right)^2}{(p_0-p_1)^2}. \qquad (13.62)$$

Критерий отношения правдоподобия сложной гипотезы. Для проверки сложных гипотез опираются на *теорему Уилкса*, которая является обобщением теоремы Пирсона (см. раздел 13.8) об асимптотических свойствах статистики χ^2. Смысл указанной теоремы в том, что *статистика*[22]

$$LR = 2\,ln\left(\Lambda_{H_1,H_0}(X)\right) \to \chi^2_\nu \qquad (13.63)$$

асимптотически стремится к распределению χ^2 *Пирсона с* $\nu = p - r$ *степенями свободы*, где p — *размерность вектора параметров* $\Theta(\theta_1, \theta_2, \dots, \theta_r, \theta_{r+1}, \dots, \theta_p)$, r — *размерность вектора оцениваемых параметров* (θ_r), $X = (x_1, x_2, \dots, x_n)$ — выборка из совокупности, с регулярной функцией распределения $F(X, \theta)$ (в смысле существования первых и вторых производных по θ) и функцией плотности вероятности $f(X, \theta)$.

$$\Lambda_{H_1,H_0}(X) = \frac{max\,[L_1(X;\theta):\theta\in\Theta]}{max\,[L_0(X;\theta):\theta\in\Theta_0]} — \qquad (13.64)$$

обобщенное отношение правдоподобия, H_0: $\theta \in \Theta_0$ — нулевая гипотеза, где $\theta \in \Theta_0$; H_1: $\theta \subset \Theta$ — общая альтернатива, $L(X;\theta)$ — функция правдоподобия.

Пример 13.16. Рассмотрим построение критерия, когда имеется n нормально распределенных случайных величин: $X_i \in N(a, \sigma^2)$ при этом рассмотрим два случая: а) дисперсия σ^2 известна, б) дисперсия σ^2 неизвестна [54].

Решение. а) Пусть проверяется гипотеза H_0: $a = a_0$ против H_1: $a \neq a_0$, причем дисперсия σ^2 известна. Тогда

$$max[L_1(X; a; \sigma^2): a = \bar{x} \in \mathrm{R}] = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n}\exp\left(-\frac{\sum_i(x_i-\bar{x})^2}{2\sigma^2}\right),$$

$$max[L_0(X; a_0;\ \sigma^2): a_0 \in \mathrm{R}] = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n}\exp\left(-\frac{\sum_i(x_i-a_0)^2}{2\sigma^2}\right).$$

[22] LR — likelihood ratio test (*англ.*) — тест отношения правдоподобия.

Отсюда

$$LR = 2\,ln\left(\Lambda_{H_1,H_0}(X)\right) = \frac{1}{\sigma^2}\left(\sum_i (x_i - a_0)^2 - \sum_i (x_i - \bar{x})^2\right).$$

В силу формулы (11.26), имеем

$$LR = 2\,ln\left(\Lambda_{H_1,H_0}(X)\right) = \frac{n}{\sigma^2}(\bar{x} - a_0)^2 \to \chi_1^2. \qquad (13.65)$$

Проверка гипотезы будет эквивалентна использованию стандартизированного нормального распределения — статистики $u \in N(0,1)$:

$$u = \sqrt{\chi_1^2} = \frac{|\bar{x} - a_0|\sqrt{n}}{\sigma}. \qquad (13.66)$$

б) Пусть проверяется гипотеза H_0: $a = a_0$ против H_1: $a \neq a_0$, причем дисперсия σ^2 неизвестна и определяется по выборочным данным. Тогда

$$max\left[L_1(X; a; \sigma^2): a = \bar{x} \in \mathrm{R}, \hat{\sigma}^2 = \frac{\sum_i (x_i - \bar{x})^2}{n}\right] =$$

$$= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left(-\frac{\sum_i (x_i - \bar{x})^2}{2\hat{\sigma}^2}\right) = \frac{1}{\left(2\pi\frac{\sum_i (x_i - \bar{x})^2}{n}\right)^{\frac{n}{2}}} \exp\left(-\frac{n}{2}\right),$$

$$max\left[L_0(X; a_0; \sigma^2): a_0 \in \mathrm{R}, \hat{\sigma}^2 = \frac{\sum_i (x_i - a_0)^2}{n}\right] =$$

$$= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left(-\frac{\sum_i (x_i - a_0)^2}{2\hat{\sigma}^2}\right) = \frac{1}{\left(2\pi\frac{\sum_i (x_i - a_0)^2}{n}\right)^{\frac{n}{2}}} \exp\left(-\frac{\sum_i (x_i - a_0)^2}{\sum_i (x_i - \bar{x})^2}\right).$$

Отсюда, учитывая формулу $\sigma_x^2 = \sigma_{a_0}^2 - (\bar{x} - a_0)^2$ (11.26), имеем

$$\Lambda_{H_1,H_0}(X) = \left(\frac{\sum_i (x_i - a_0)^2}{\sum_i (x_i - \bar{x})^2}\right)^{\frac{n}{2}} = \left(1 + \frac{n(\bar{x} - a_0)^2}{\sum_i (x_i - \bar{x})^2}\right)^{\frac{n}{2}} = \left(1 + \frac{n(\bar{x} - a_0)^2}{\sum_i (x_i - \bar{x})^2}\right)^{\frac{n}{2}}.$$

Нулевая гипотеза отклоняется, когда $\frac{n(\bar{x} - a_0)^2}{\sum_i (x_i - \bar{x})^2}$ велико, но величина

$$\frac{n(\bar{x} - a_0)^2}{\frac{\sum_i (x_i - \bar{x})^2}{n-1}} \frac{1}{(n-1)} = t_{n-1}^2 \frac{1}{n-1}, \qquad (13.67)$$

где $\bar{x}$ — выборочная средняя, $\frac{\sum_i (x_i - \bar{x})^2}{n-1} = s^2$ — исправленная выборочная дисперсия, $t = \frac{(\bar{x} - a_0)\sqrt{n}}{s}$ — дробь Стьюдента.

Значит,

$$\Lambda_{H_1,H_0}(X) = \left(1 + \frac{t^2}{n-1}\right)^{\frac{n}{2}}. \qquad (13.68)$$

Таким образом, критерий отношения правдоподобия $\Lambda_{H_1,H_0}(X)$ эквивалентен t-критерию Стьюдента.

Пример 13.17. Пусть опять имеется n нормально распределенных случайных величин: $X_i \in N(a, \sigma^2)$ при этом рассмотрим два случая: а) a известно, б) a неизвестно [54].

Решение. а) Пусть проверяется гипотеза H_0: $\sigma^2 = \sigma_0^2$ против H_1: $\sigma^2 \neq \sigma_0^2$.

Тогда

$$max[L_1(X; a; \sigma^2): a \in R, \sigma^2 = \sigma_0^2] =$$

$$= \frac{1}{\left(2\pi\frac{\sum_i (x_i - a)^2}{n}\right)^{n/2}} \exp\left(-\frac{\sum_i (x_i - a_0)^2}{2\frac{\sum_i (x_i - a_0)^2}{n}}\right) = \left(\frac{n}{2\pi\sum_i (x_i - a)^2}\right)^{n/2} \exp\left(-\frac{n}{2}\right),$$

$$max\left[L_0(X;a;\sigma^2)\colon a\in \mathrm{R},\sigma^2=\frac{\sum_i(x_i-a)^2}{n}\right]=\frac{1}{\left(2\pi\sigma_0^2\right)^{n/2}}\exp\left(-\frac{\sum_i(x_i-a)^2}{2\sigma_0^2}\right),$$

следовательно, обобщенное отношение правдоподобия

$$\Lambda_{H_1,H_0}(X)=\left(\frac{n\sigma_0^2}{\sum_i(x_i-a)^2}\right)^{\frac{n}{2}}\exp\left(-\frac{n}{2}+\frac{\sum_i(x_i-a)^2}{2\sigma_0^2}\right).$$

Значит, статистика

$$LR=2\,ln\left(\Lambda_{H_1,H_0}(X)\right)=n\,ln\left(\frac{n\sigma_0^2}{\sum_i(x_i-a)^2}\right)+\frac{\sum_i(x_i-a)^2-n\sigma_0^2}{\sigma_0^2},$$

$$LR=-n\,ln\left(1+\frac{\sum_i(x_i-a)^2-n\sigma_0^2}{n\sigma_0^2}\right)+\frac{\sum_i(x_i-a)^2-n\sigma_0^2}{\sigma_0^2}.$$

Воспользовавшись асимптотической формулой для логарифма, опирающейся на разложение в ряд Тейлора, при условии, что верна H_0, получим

$$LR\approx\left(\frac{\sum_i(x_i-a)^2-n\sigma_0^2}{\sqrt{2n\sigma_0^2}}\right)^2\to\chi_1^2. \tag{13.69}$$

Замечание. 1. При больших значениях n величина $\frac{\sum_i(x_i-a)^2-n\sigma_0^2}{n\sigma_0^2}\to 0$, следовательно, можно воспользоваться приближенной формулой

$$ln(1+\alpha)=\alpha-\frac{\alpha^2}{2}+O(\alpha^3),$$

где α — бесконечно малая величина ($\alpha\to 0$).

Имеем

$$LR\approx -n\left(\frac{\sum_i(x_i-a)^2-n\sigma_0^2}{n\sigma_0^2}-\frac{1}{2}\left(\frac{\sum_i(x_i-a)^2-n\sigma_0^2}{n\sigma_0^2}\right)^2\right)+\frac{\sum_i(x_i-a)^2-n\sigma_0^2}{\sigma_0^2},$$

$$LR\approx\frac{n}{2}\left(\frac{\sum_i(x_i-a)^2-n\sigma_0^2}{n\sigma_0^2}\right)^2=\left(\frac{\sum_i(x_i-a)^2-n\sigma_0^2}{\sqrt{2n\sigma_0^4}}\right)^2.$$

$$M(x_i-a)^2=\sigma^2, D(x_i-a)^2=2\sigma_0^4,$$

значит,

$$\frac{\sum_i(x_i-a)^2-n\sigma_0^2}{\sqrt{2n\sigma_0^4}}\to N(0,1). \tag{13.70}$$

2. В рассматриваемом примере можно использовать статистику χ^2 — Пирсона, но с другой критической областью, воспользовавшись тем, что

$$\left(\frac{\sum_i(x_i-a)^2}{\sigma_0^2}\right)^2\to\chi_n^2 \tag{13.71}$$

или

$$\left(\frac{\sum_i(x_i-\bar{x})^2}{\sigma_0^2}\right)^2\to\chi_{n-1}^2. \blacksquare \tag{13.72}$$

б) Пусть значение математического ожидания a — неизвестно и проверяется гипотеза $H_0\colon \sigma^2=\sigma_0^2$ против $H_1\colon \sigma^2\neq\sigma_0^2$. Тогда

$$max\left[L_1(X;a;\sigma^2)\colon a=\bar{x},\hat{\sigma}^2=\frac{\sum_i(x_i-\bar{x})^2}{n}\right]=$$

$$=\frac{1}{\left(2\pi\frac{\sum_i(x_i-\bar{x})^2}{n}\right)^{n/2}}\exp\left(-\frac{\sum_i(x_i-\bar{x})^2}{2\frac{\sum_i(x_i-\bar{x})^2}{n}}\right)=\left(\frac{n}{2\pi\sum_i(x_i-\bar{x})^2}\right)^{n/2}\exp\left(-\frac{n}{2}\right),$$

$$max[L_0(X;a): a = \bar{x}, \hat{\sigma}^2 = \sigma_0^2] = \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left(-\frac{\sum_i (x_i - \bar{x})^2}{2\sigma_0^2}\right),$$

следовательно, обобщенное отношение правдоподобия

$$\Lambda_{H_1,H_0}(X) = \left(\frac{n}{e}\right)^{n/2} \left(\frac{\sum_i (x_i - \bar{x})^2}{\sigma_0^2}\right)^{-\frac{n}{2}} \exp\left(\frac{\sum_i (x_i - \bar{x})^2}{2\sigma_0^2}\right).$$

Заметим, что

$$\frac{\sum_i (x_i - \bar{x})^2}{\sigma_0^2} \to \chi^2_{n-1},$$

следовательно, $\Lambda_{H_1,H_0}(X)$ велико, когда χ^2_{n-1} велико или мало, что фактически приводит к использованию в качестве критерия χ^2_{n-1}.

Допустим, что *имеется n независимых наблюдений* x_i и известен закон распределения вида $f(x, \Theta)$, где параметр Θ может быть найден в результате использования ML $(Maximum\ likelihood)$ — метода максимального правдоподобия, предполагающего поиск параметра $\hat{\theta}_{Ml}$ как решения задачи максимизации вероятности получения имеющихся данных:

$$L\big(f(x,\Theta)\big) = \prod_{i=1}^{n} f(x_i, \theta) \to max. \qquad (13.73)$$

Пусть в результате решения (13.73) получена оценка $\hat{\theta}_{Ml}$. Если требуется *проверить простую гипотезу* $H_0: \Theta = \Theta_0$ *против сложной альтернативы* $H_1: \Theta \neq \Theta_0$, то используется *LR* (*likelihood ratio test*) — тест отношения правдоподобия (13.63), утверждающий, что

$$LR = 2ln\left(\Lambda_{H_1,H_0}(X)\right) = 2\big[l\big(\hat{\theta}_{Ml}\big) - l(\theta_0)\big] \to \chi^2_1, \qquad (13.74)$$

где $l\big(\hat{\theta}_{Ml}\big) = ln\,(L\big(\hat{\theta}_{Ml}\big))$ и $l(\theta_0) = ln\,(L(\Theta_0)$ — значения логарифмической функции правдоподобия при $\hat{\theta}_{Ml}$ и θ_0.

Замечание. Принцип отношения правдоподобия была выдвинут в 1928 г., он предполагает, что если отношение $\lambda_{H_1,H_0}(X) = \frac{L_1(X)}{L_0(X)}$ велико, то *простая гипотеза* H_0 маловероятна.

Принцип отношения правдоподобия был доказан авторами в виде леммы Неймана — Пирсона 1933 г. — *критерий отношения правдоподобия* (критерий δ_C размера α, для любого значения $k > 0$, с областью $C = \{X: L_1(X) > kL_0(X)\}$, имеет наибольшую мощность $(P_1(C) := 1 - \beta)$). Заметим, что область с наибольшим значением мощности критерия для заданного уровня значимости α может подбираться на компьютере (см. пример 13.18).

При больших значениях n асимптотические свойства отношения правдоподобия для проверки *сложных гипотез* под названием *критерия обобщенного отношения правдоподобия* (*LR*) были получены Уилксом в 1938 г. в виде теоремы, обобщающей теорему Пирсона (см. 13.8) об асимптотических свойствах статистики χ^2. Как было показано ранее, из *LR* можно вывести большинство статистических тестов, в том числе χ^2 Пирсона, распределение Стьюдента, а также F-критерий дисперсионного анализа (*ANOVA*) и др. На практике, если существует «стандартный» критерий, отношение правдоподобия очень часто позволяет найти его или близкий к нему. ■

Пример 13.18. *Задача о леди, пробующей чай* (задача рассматривалась Р. Фишером (1922) и несколько позже с использованием другого подхода Ю. Нейманом [85]). Некая леди заявляет, что, попробовав чашку чая с молоком, она может определить, что было сначала налито в чашку — молоко или чай. Характерно, что леди не претендует на то, что она может безошибочно определить разницу во вкусе, но утверждает, что, пусть иногда ошибаясь, она чаще отвечает верно, чем неверно. Можно ли утверждать, что леди имеет соответствующие способности, опираясь, например, на данные таблицы 13.8?

Таблица 13.8

Тестирование способностей леди

Мнение леди	Налито	
	чай + молоко (P)	молоко + чай (N)
Чай + молоко (P)	7	2
Молоко + чай (N)	3	8

Решение. Р. Фишер, основатель идеологии статистического вывода и теории статистического эксперимента, привел эту задачу в известной книге *The Design of Experiments*, опираясь на идеи *сравнения, повторности, рандомизации, однородности*. Сравнение предполагает наличие эталона (в случае с леди — это «правильный» чай с молоком, который она пила с детства). Повторность позволяет оценить ошибку эксперимента и приводит к ее уменьшению. Рандомизация (*randomisation, случайное распределение*) позволяет получить правдоподобную несмещенную оценку интересующих исследователя эффектов за счет исключения систематических и несистематических ошибок (трендов), правдоподобную оценку ошибки эксперимента и обеспечить независимость результатов эксперимента (например, за счет использования генератора случайных чисел). Однородность предполагает постоянство условий опыта и наблюдений, например, в случае с чашками чая — постоянство температуры, структуры состава, неизменное состояние рецепторов.

Полагается, что леди тестируется n дней. *Утром ей подают пару чашек чая,* приготовленных по каждому рецепту. Рационально будет считать, что если число X_n правильно классифицированных чашек чая превышает некоторое, заранее известное число, то следует признать способности леди к правильной классификации (тестируемая гипотеза H_0), в противном случае следует отвергнуть ее способности. Если вероятность правильной классификации p известна, то мы имеем дело с простой гипотезой, которая для жюри может выглядеть как $H_0: p = 0{,}5$, то есть претензии дамы на наличие различающих способностей, не обоснованы. Альтернативная гипотеза $H_1: p \neq 0{,}5$.

1) Для проверки гипотезы о том, что «правильность» чая не зависит от диагностических возможностей дамы, используем критерий независимости хи-квадрат, основанный на сопоставлении эмпирических наблюдений и теоретически ожидаемых численностей. Получим теоретические частоты:

$$\frac{10\cdot 9}{20} = 4{,}5;\ \frac{10\cdot 9}{20} = 4{,}5;\ \frac{11\cdot 10}{20} = 5{,}5;\ \frac{11\cdot 10}{20} = 5{,}5.$$

Отсюда

$$\chi^2 = \left(\frac{(7-4{,}5)^2}{4{,}5} + \frac{(3-4{,}5)^2}{4{,}5} + \frac{(2-5{,}5)^2}{5{,}5} + \frac{(8-5{,}5)^2}{5{,}5}\right) =$$
$$= 1{,}38889 + 1{,}13636 + 1{,}38889 + 1{,}13636 = 5{,}0505.$$

Критическое значение хи-квадрат Пирсона имеет $k = (m-1)(l-1)$ степеней свободы, у нас: $k = (2-1)(2-1) = 1$. Пользуясь пакетом *gretl*, легко найти, что для $\chi^2 = 5{,}05051$ $p-$значение $(p - value)$ равно 0,024619 — это вероятность ошибки первого рода, при которой гипотезу о независимости следует отвергнуть — изучаемые признаки связаны.

2) Малая выборка (есть числа в таблице менее 5, и общая сумма частот не превышает 20) снижает достоверность критерия хи-квадрат Пирсона. В этом случае рекомендуется применение точного критерия Фишера (опирающегося на гипергеометрический закон распределения). Итак, если имеется таблица 2×2 вида

Таблица 13.9

Результаты теста

Прогноз	Фактически		Σ
	положительно (H_0)	отрицательно (H_1)	
Положительно (H_0)	a	b	$a+b$
Отрицательно (H_1)	c	d	$c+d$
Σ	$a+c$	$b+d$	$a+b+c+d$

то точный критерий Фишера, опирающийся на гипергеометрический закон распределения, выглядит следующим образом:

$$P = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{a!b!c!d!(a+b+c+d)!}. \qquad (13.75)$$

Таблица 13.10

Значения отношения правдоподобия

N	M	n	x	P	S	$\lambda_{H_1,H_0}(X)$
8	4	4	0	0,014286	0,014286	0,0000
8	4	4	1	0,228571	0,242857	0,3125
8	4	4	2	0,514286	0,757143	0,8333
8	4	4	3	0,228571	0,985714	1,8750
8	4	4	4	0,014286	1	5,0000

Если леди предложили $N = 8, 12$ чашек чая, то по формуле (13.75), получим соответствующие вероятности (табл. 13.10, 13.11), где вероятности (P) получены по формуле (13.75), S — накопленные вероятности, $\lambda_{H_1,H_0}(X)$ — отношение правдоподобия, полученное для гипергеометрического распределения по формуле (13.59).

Из таблиц следует, что при уровне значимости $\alpha = 0{,}05$ для подтверждения способностей леди при $N = 8$ — она не должна ошибиться, при $N = 12$ — она может ошибиться один раз. Таким образом, решение о принятии или отвержении гипотезы о наличии способностей у дамы зависит от количества опытов.

3) С увеличением объема совокупности, как известно, гипергеометрический закон стремится к биномиальному. Предполагается, что дама может совершать ошибки, но относительная частота не сильно отличается от вероятности. Допустим, рассматривается нулевая гипотеза $H_0: p = \frac{1}{2}$, то есть дама совершенно случайно оценивает чай, в качестве альтернативной гипотезы рассмотрим сложную гипотезу $H_1: p > \frac{1}{2}$.

Таблица 13.11

Значения отношения правдоподобия

N	*M*	*n*	*x*	*P*	*S*	$\lambda_{H_1,H_0}(X)$
12	6	6	0	0,001082	0,001082	0,0000
12	6	6	1	0,038961	0,040043	0,1944
12	6	6	2	0,243506	0,283550	0,4667
12	6	6	3	0,432900	0,716450	0,8750
12	6	6	4	0,243506	0,959957	1,5556
12	6	6	5	0,038961	0,998918	2,9167
12	6	6	6	0,001082	1,000000	7,0000

Рассматривается случайная величина $X = \{0, 1, 2, \dots, n\}$ — число правильных ответов дамы в n опытах. Если положить, что $X = k$ и гипотеза H_0 отвергается при некотором t, то критическая область будет состоять из всех точек, для которых $k = t, t+1, \dots, n$. Ошибка первого рода возникает, если гипотеза H_0 верна, но выборочная точка попадает в критическую область, то есть

$$\alpha = P(H_1/H_0) = P(X \geq t/H_0) = \sum_{k=t}^{n} C_n^k \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} = \left(\frac{1}{2}\right)^n \sum_{k=t}^{n} C_n^k. \quad (13.76)$$

Ошибка второго рода возникает, если H_0 — не верна, то есть $p \neq \frac{1}{2}$, но выборочная точка не попадает в критическую область, имеем

$$P(H_0/H_1) = P(X < t/p) = \sum_{k=0}^{t-1} C_n^k (p)^k (1-p)^{n-k} = 1 - \sum_{k=t}^{n} C_n^k (p)^k (1-p)^{n-k}. \quad (13.77)$$

Для оценки вероятности ошибки второго рода при различных значениях $p \neq \frac{1}{2}$ строится график функции мощности критерия:

$$\beta = \beta(p/t, n) = \sum_{k=t}^{n} C_n^k (p)^k (1-p)^{n-k}, \quad (13.78)$$

для правосторонней критической области, так как H_0 отвергается при $k \geq t$ (рис. 13.6).

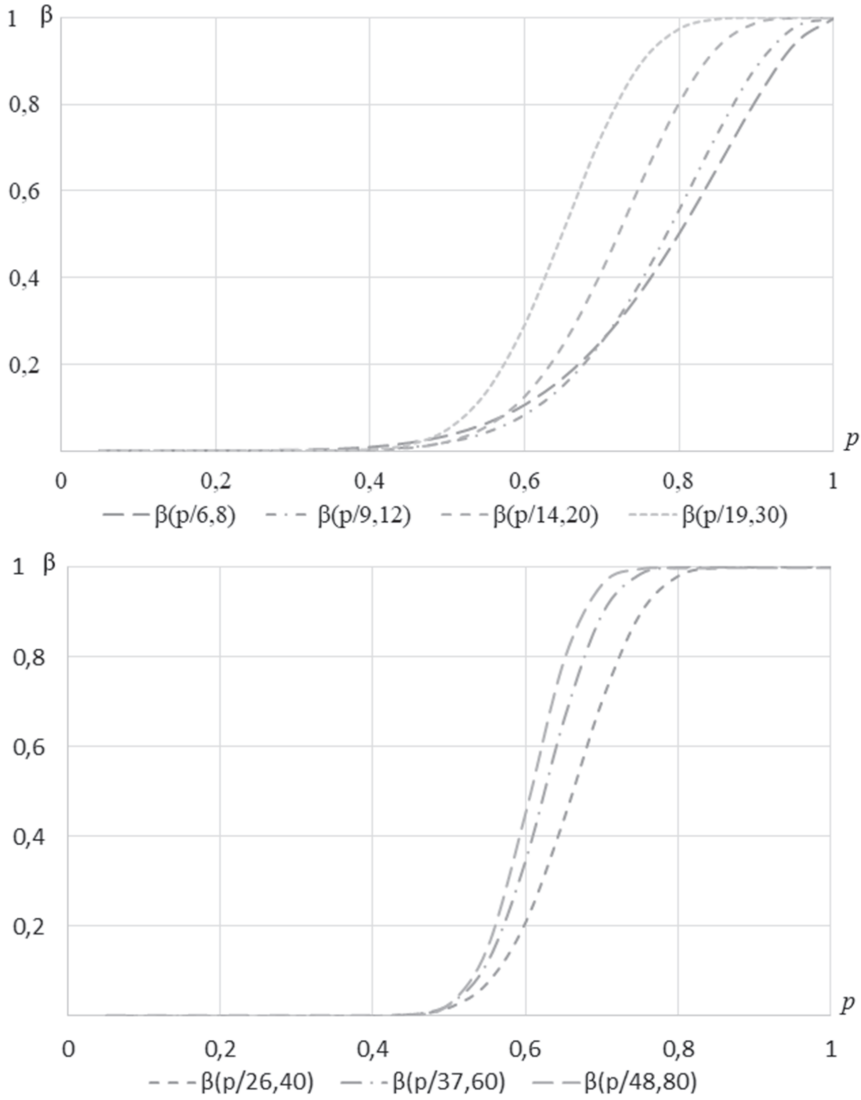


Рис. 13.6 — Функции мощности критерия
(правосторонняя критическая область)

Функция мощности критерия, определенная на множестве простых гипотез — для каждой истинной гипотезы H_0 равна вероятности того, что критерий отвергнет неверную гипотезу H_1.

Перебирая значения $n = 8, 12, 20, 30, 40, 60, 80$, мы для каждого находим (например, используя *MS Excel команда*: «Данные-Подбор параметра») соответственно, $k = 6, 9, 14, 19, 26, 37, 48$ (критические значения, примерно соответствующие значению $\alpha = 0{,}05$ для правосторонней области). Если дама угадывает с вероятностью p, получим значения функции мощности критерия (табл. 13.12), отраженные на рисунке 13.6.

Таким образом, например, при уровне значимости 0,05 способности леди признаются, если она угадает хотя бы 48 пар чашек из 80. Шансы леди доказать свое умение превосходят 0,5, если она угадывает с вероятностью 0,85 при $n = 8$; 0,8 при $n = 12$; 0,75 при $n = 20$; 0,70 при $n = 30, 40$; 0,65 при $n = 60, 80$ (в таблице 3.12 шансы свыше 0,5 выделены цветом). Например, если дама угадывает с вероятностью 0,65, то рисунок 13.5 иллюстрирует факт, что при небольшом количестве опытов (8–12) шансы обнаружить способности леди практически одинаковы. В пределах 20–30 и 40–60 опытов — шансы наглядно растут быстрее, чем в пределах 60–80 опытов.

Таблица 13.12

p	β(p/6,8)	β(p/9,12)	β(p/14,20)	β(p/19,30)	β(p/46,40)	β(p/37,60)	β(p/48,80)
0,50	0,0352	0,0193	0,0207	0,0494	0,0192	0,0259	0,0283
0,55	0,0632	0,0421	0,0553	0,1350	0,0751	0,1210	0,1559
0,60	0,1064	0,0834	0,1256	0,2915	0,2112	0,3493	0,4576
0,65	0,1691	0,1513	0,2454	0,5078	0,4408	0,6616	0,7951
0,70	0,2553	0,2528	0,4164	0,7304	0,7032	0,8959	0,9640
0,75	0,3671	0,3907	0,6172	0,8943	0,8968	0,9846	0,9978
0,80	0,5033	0,5583	0,8042	0,9744	0,9806	0,9992	1,0000
0,85	0,6572	0,7358	0,9327	0,9971	0,9986	1,0000	1,0000
0,90	0,8131	0,8891	0,9887	0,9999	1,0000	1,0000	1
0,95	0,9428	0,9804	0,9997	1,0000	1,0000	1	1
1,00	1	1	1	1	1	1	1

Последнее говорит в пользу з*акона убывающего эффекта*, утверждающего, что *при хороших условиях эксперимента дальнейшее улучшение затруднено*.

4) Рассмотрим в рамках поставленной задачи, полагая, что число отгаданных чашек чая имеет нормальный закон распределения ($N\big(np;\ np(1-p)\big)$, где $n = 30$) три вопроса:

а) *нахождение наилучшей критической области (НКО) для проверки простой гипотезы* $H_0: p = p_0 = 0{,}5$ *против простой альтернативной гипотезы* $H_0: p = p_1 = 0{,}7$ $(p_1 > p_0)$;

б) *определение функции мощности критерия* и вычисление ее значение при $p_1 = 0{,}7$, если $p_0 = 0{,}5$, объем выборки $n = 30$. Уровень значимости $\alpha = 0{,}05$;

в) *проверки простой гипотезы* $H_0: p = p_0 = 0{,}5$ *против сложной альтернативной гипотезы* $H_0: p \neq 0{,}5$ $(p_1 > p_0)$, $\sigma^2 = 0{,}25$.

Решение. а) Граница критической области определяется по формуле (13.61):

$$k \approx np_0 + u_\alpha \sqrt{np_0(1-p_0)}.$$

У нас $k \approx 30 \cdot 0{,}5 + 1{,}645\sqrt{30 \cdot 0{,}5 \cdot (1-0{,}5)} = 19{,}505$, что приближенно соответствует точным результатам (табл. 13.12). Таким образом, граница критической области примерно равна $k_к = 19$, а НКО имеет вид $k > 19$.

б) Функция мощности критерия с использованием функции Лапласа Ф для правосторонней критической области (рис. 13.5, $p_1 > p_0$) имеет вид

$$\beta = \beta(p_1/k, n) = 0{,}5 - \Phi\left(\frac{\frac{k_к}{n} - p_1}{p_1(1-p_1)}\sqrt{n}\right).$$

Мощность критерия:

$$\beta = \beta(p_1 = 0{,}7/k > 19, n = 30) = 0{,}5 - \Phi\left(\frac{\frac{19}{30} - 0{,}7}{0{,}7(1-0{,}7)}\sqrt{30}\right) =$$

$$= 0{,}5 - \Phi(-0{,}7452) = 0{,}5 + 0{,}228 = 0{,}728.$$

При небольшом объеме совокупности различия между биномиальным законом и его приближением в виде нормального закона влияют на результаты вычислений, что мы и наблюдали выше.

в) По формулам (13.65) и (13.74), учитывая, что $\sigma^2 = 0{,}25$, получим

$$LR = 2\,ln\left(\Lambda_{H_1,H_0}(X)\right) = \frac{n}{0{,}25}(\hat{p}_{Ml} - 0{,}5)^2 \to \chi_1^2. \qquad (13.79)$$

При уровне значимости $\alpha = 0{,}05$ с использованием *gretl* получим, что критическое значение критерия хи-квадрат с одной степенью свободы равно 3,84146.

Таким образом, можно получить в зависимости от числа наблюдений n границы $\hat{p}_1 < \hat{p}_{Ml} < \hat{p}_2$, попадание $\hat{p}_{Ml}$ в которые позволяет принять гипотезу H_0 (рис. 13.7).

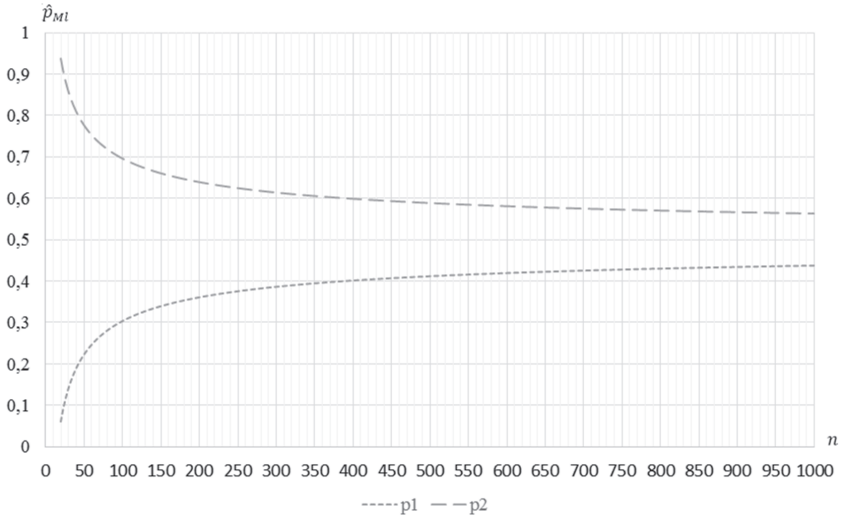


Рис. 13.7 — Доверительные границы $\hat{p}_{Ml}$ принятия нулевой гипотезы (выбор дамы случаен)

Из формулы (13.79) получим

$$(\hat{p}_{Ml} - 0{,}5)^2 = \frac{3{,}84146}{4n},$$

$$\hat{p}_{Ml} = 0{,}5 \pm \sqrt{\frac{3{,}84146}{4n}}.$$

Таким образом, $\hat{p}_1 = 0{,}5 - \sqrt{\frac{3{,}84146}{4n}}$, $\hat{p}_2 = 0{,}5 + \sqrt{\frac{3{,}84146}{4n}}$.

Замечание. 1. Известно, что после окончания Кембриджского университета Р. Фишер, который продолжил развитие идей К. Пирсона и практически в одиночку заложил основы теории статистического вывода (выборочный метод, про-

верку статистических гипотез и др.), после учебы в колледже отказался от работы на кафедре К. Пирсона и поступил на работу в Ротемстедскую биологическую лабораторию для разработки методов обоснования решения проблем в сельском хозяйстве. Результаты опытов фиксировались раз в году, и остальное время было посвящено разработке теории статистического вывода и планирования эксперимента, мысли о котором появились в процессе регулярного чаепития молодежного коллектива лаборатории. Однажды Мюриэль Бристол отказалась пить чай, ибо в аристократических кругах считалось, что необходимо добавлять чай в молоко, а не наоборот. Согласно плану Р. Фишера, учитывающего идеи сравнения, повторности, рандомизации, однородности, был проведен эксперимент, подтвердивший способность леди отличать последовательность добавления чая и молока в чашку (из четырех пар предложенных чашек она угадала все!).

2. Классическое и статистическое определения вероятности предполагают воспроизводимость опытов, массовость, однородность и т. д. На практике неопределенность и непредсказуемость носят принципиальный характер, а результаты деятельности субъекта зачастую приводят к конкретным результатам (наблюдений, экспериментов), по которым требуется оценить последствия принятия решений для субъекта. Если они отрицательны, то говорят о рисках, если положительны, то говорят о шансах. Результаты наблюдений позволяют уменьшить неопределенность будущих рисков и шансов благодаря договоренности о точках отсечения. Например, рассмотрим вероятность как интервал $0 < A < B < 1$. Если частота положительного события меньше A, то следует говорить о риске, если больше B, то говорят о шансе. А между ними область неопределенности. Хотя на практике чаще используется одна точка отсечения, которая используется при принятии решения о наступлении положительного события. ■

Например, для оценки, качества бинарной классификации используется инструмент *ROC*-анализа (*Receiver Operator Characteristic Curve*).

Пусть имеется база данных службы занятости безработных. Для оценки социальной активности условно будем считать, что «положительное событие, $A = 1$» — в течение 60 дней безработный не нашел работу, а «отрицательное, $A = 0$» — нашел. Предположим, что есть тест — некоторая модель принятия решений о наступлении положительных и отрицательных исходов, опирающаяся на определенные свойства изучаемых объектов (например, модель бинарной регрессии, где факторные признаки — социально-экономические характеристики безработных, а результативные принимают два значения $A = \{0, 1\}$). Результаты подсчета частот положительных и отрицательных примеров представляются в виде таблицы классификации (табл. 13.13), где учитываются:

– *TP* (*True Positives*) — верно классифицированные положительные примеры (истинно положительные случаи);

– *TN* (*True Negatives*) — верно классифицированные отрицательные примеры (истинно отрицательные случаи);

– *FN* (*False Negatives*) — положительные примеры, классифицированные как отрицательные (ошибка I рода). Это так называемый «ложный пропуск» — когда интересующее нас событие ошибочно не обнаруживается (ложно отрицательные примеры);

– *FP* (*False Positives*) — отрицательные примеры, классифицированные как положительные (ошибка II рода). Это ложное обнаружение, так как при отсутствии события ошибочно выносится решение о его присутствии (ложно положительные случаи).

Таблица 13.13

Классификации и сопутствующие статистики

Прогноз	Фактически		Σ	Прогностическая ценность	Шансы
	Положительно (H_0)	Отрицательно (H_1)			
Положительно (H_0)	$TP=a$	$FP = b$	$a + b$	$PV_+ = \frac{a}{a+b}$	$\frac{a}{b}$
Отрицательно (H_1)	$FN = c$	$TN = d$	$c + d$	$PV_- = \frac{c}{c+d}$	$\frac{c}{d}$
Σ	$a + c$	$b + d$	$a+b+$ $+c+d$		
Качество классификации	$Se = TPR =$ $= \frac{TP}{TP+FN} \cdot$ $\cdot 100\%$ $Se =$ $= \frac{a}{a+c} \cdot 100\%$	$Sp = 100 - FPR =$ $= \frac{TN}{TN+FP} \cdot 100\%$ $Sp = \frac{d}{b+d} \cdot 100\%$		$P =$ $= \frac{a+c}{a+b+c+d}$ *распространенность (prevalence)*	
Отношение правдоподобия	$LR_+ = \frac{Se}{1-Sp} =$ $= \frac{a/(a+c)}{b/(b+d)}$	$LR_- = \frac{1-Se}{Sp} =$ $= \frac{c/(a+c)}{d/(b+d)}$			
Отношение рисков	$RR = \frac{PV_+}{PV_-} -$ *Risk Ratio*	$RR = \frac{(1-PV_+)}{(1-PV_-)} -$ *Risk Ratio*			
Отношение шансов	$OR = \frac{ad}{bc} = \frac{PV_+}{PV_-}\frac{(1-PV_-)}{(1-PV_+)} -$ *Odds Ratio*				

Знание результатов классификации позволяет дать основные понятия *ROC*-анализа (табл. 13.14).

Таблица 13.14

Основные понятия *ROC*-анализа

Формулы	Понятия
$TPR = \frac{TP}{TP+FN} \cdot 100\%$	*Доля истинно положительных примеров* (*True Positives Rate*)
$FPR = \frac{FP}{TN+FP} \cdot 100\%$	*Доля ложно положительных примеров* (*False Positives Rate*)
$Se = TPR =$ $= \frac{TP}{TP+FN} \cdot 100\%$	*Чувствительность* (*Sensitivity*) — доля истинно положительных случаев
$Sp = 100 - FPR =$ $= \frac{TN}{TN+FP} \cdot 100\%$	*Специфичность* (*Specificity*) — доля истинно отрицательных случаев, которые были правильно идентифицированы моделью

Формула Байеса позволяет связать прогностическую ценность положительного и отрицательного результатов:

$$PV_{+} = \frac{Se \cdot P}{Se \cdot P + (1-Se) \cdot (1-P)}, \quad (13.80)$$

$$PV_{-} = \frac{(1-Se) \cdot (1-P)}{Se \cdot P + (1-Se) \cdot (1-P)}. \quad (13.81)$$

Модель с высокой чувствительностью часто дает истинный результат при наличии положительного исхода (обнаруживает положительные примеры). Наоборот, модель с высокой специфичностью чаще дает истинный результат при наличии отрицательного исхода (обнаруживает отрицательные примеры).

Пусть таблица 13.15, описывающая результаты эксперимента по проверке способностей дамы, получена после проведения серии экспериментов с убывающей субъективной уверенностью леди (субъективной вероятностью).

Таблица 13.15

Возможные результаты проверки способностей дамы

№ п/п	Прогноз	x_i	Субъективная степень уверенности леди	Фактически	y_i	Индикатор
1	P	1	0,95	P	1	1
2	P	1	0,90	P	1	1
3	P	1	0,85	P	1	1
4	N	0	0,80	N	0	1
5	P	1	0,80	P	1	1
6	P	1	0,75	P	1	1
7	N	0	0,70	P	1	0
8	P	1	0,65	P	1	1
9	P	1	0,60	N	0	0
10	P	1	0,55	P	1	1
11	N	0	0,50	P	1	0
12	N	0	0,45	N	0	1
13	P	1	0,40	N	0	0
14	N	0	0,35	N	0	1
15	N	0	0,30	N	0	1
16	N	0	0,25	N	0	1
17	P	1	0,20	N	0	0
18	N	0	0,15	N	0	1
19	N	0	0,10	N	0	1
20	N	0	0,05	N	0	1
accur	–		–	–	0,75	

В системе координат с абсциссой (*FPR=100% — Sp*) и ординатой *Se* строится *ROC*-кривая — множество пар точек (*Sp, Se*), полученных для порога отсечения (*optimalcut-offvalue*) с определенным шагом (например, 0,01). Чем ближе *ROC*-кривая к диагонали (*y=x*), тем она хуже, чем ближе к левому углу (идеальный классификатор) — тем лучше. Сравнение моделей между собой можно проводить с использованием показателя площади под кривой — *AUC* (*Area Under Curve*), чем площадь больше, тем модель классификации лучше.

У нас сбалансированная выборка, так как число опытов равно 20 и насчитывает по 10 чашек «правильного» чая и «неправильного» чая соответственно: *P — positive, N — negative.*

ROC-кривая строится следующим образом. Единичный квадрат $m \times n$ разбивается на 100 частей с шагом 0,1 ($m = 10$ число «правильных чашек чая», $n = 10$ — «число неправильных чашек чая») (рис. 13.8). *ROC*-кривая лучшего классификатора проходит через точку (0, 1).

Начало — точка (0; 0). Если «алгоритм дамы» дает прогноз *N — negative* — делаем шаг вправо, если нет — делаем шаг вверх. Таким образом мы попадем в точку (1; 1). Следует отметить, что пример, основанный на таблице 13.15, носит иллюстративный характер, реальные примеры опираются даже не на сотни, а на тысячи прецедентов.

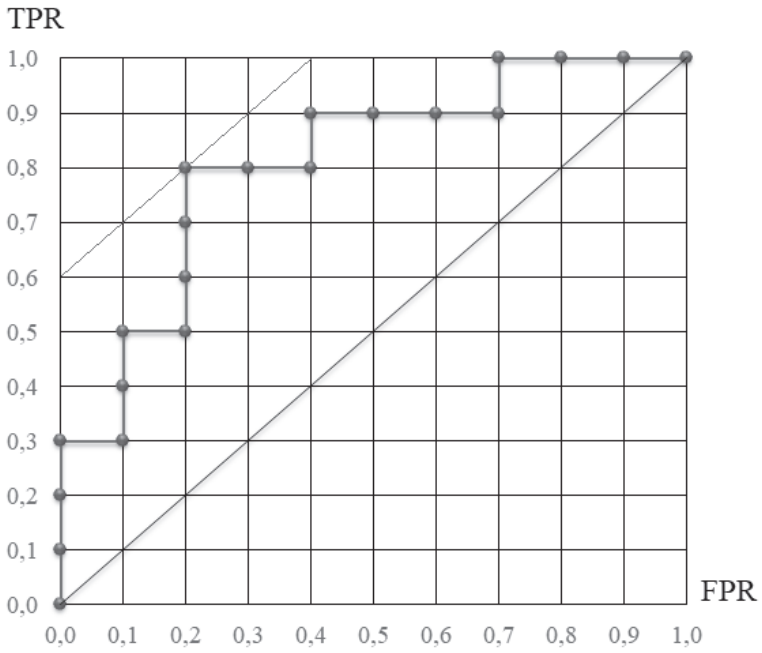


Рис. 13.8 — *ROC*-кривая

Индикатор показывает верную классификацию: если «прогноз *P*» = «фактически *P*» или «прогноз *N*» = «фактически *N*», то 1; иначе — 0. Он позволяет рассчитать

$$accuracy = \frac{a+d}{a+b+c+d} = 0{,}75 — \qquad (13.82)$$

долю верной классификации (табл. 13.15).

Значение порога отсечения, влияющего на соотношение *Se* и *Sp*, соответствует стратегии исследования:

$ocov_1$ — максимальная специфичность (чувствительность) предполагает обеспечить долю истинно отрицательных случаев не ниже определенной границы (например, 90%);

$ocov_2$ — максимальная суммарная чувствительность и специфичность модели, $C = \max_k (Se_k + Sp_k)$;

$ocov_3$ — баланс между чувствительностью и специфичностью, т. е. когда $Se \approx Sp$;

учитывая соотношение позитивных и негативных примеров, строится прямая вида $y = \frac{b+d}{a+c}x + 1$, в точке отсечения проходит касательная к *ROC*-кривой параллельно указанной прямой (если $\frac{b+d}{a+c} > 1$, то модель лучше классифицирует негативные примеры, если $\frac{b+d}{a+c} < 1$ — позитивные, в случае равенства $\frac{b+d}{a+c} = 1$ — возможности классификации равны). Поэтому у нас в качестве точки отсечения нужно выбрать точку (0,2; 0,8), в которой проходит касательная, параллельная прямой $y = x$. То есть если дама уверена в «правильности чая» не менее чем на 80%, то чай «правильный».

Возвращаясь к условию примера 13.16, видим, что рассматривалось $a + c = 7 + 3$ чашек «правильного» чая и $b + d = 2 + 8$ — «неправильного». Дама выявила $a = 7$ чашек с «правильным» чаем (из $a + c$) и $d = 8$ чашек с «неправильным» чаем (из $b + d$).

Таким образом, способности дамы имеют *чувствительность* (*sensitivity*):

$$Se = \frac{a}{a+c} = \frac{7}{7+3} = \frac{7}{10},$$

характеризующую способность диагностировать «правильный чай» как «правильный»;

специфичность (*specificity*):

$$Sp = \frac{d}{b+d} = \frac{8}{2+8} = \frac{8}{10},$$

характеризующую способность диагностировать «неправильный» чай, как «неправильный».

Итак, если очередной тест на проверку способностей дамы обнаруживает высокую чувствительность, то дама выбирает чаще «правильный» чай верно, хотя если она ошибается, то тест более информативен, ибо ошибается она редко. Аналогично, если тест показывает высокую специфичность, то она часто выбирает верно «неправильный» чай, если она ошибается, тест более информативен.

Способность дамы диагностировать чай позволяет разбить совокупность наблюдений на две выборки:

1) чашки с «правильным» чаем, которые разбиты на две подвыборки: а) верно классифицированные чашки с «правильным» чаем (TP), б) ошибочно классифицированные как «неправильные» (FP);

2) чашки с «неправильным» чаем, которые также разбиты на две подвыборки: а) верно классифицированные чашки с «неправильным» чаем (TN), б) ошибочно классифицированные как «правильные» (FP).

Тест, имеющий высокую чувствительность при отрицательном результате, исключает «правильный» чай. Тест, имеющий высокую специфичность при положительном результате, подтверждает диагностику «неправильного» чая.

Отношение правдоподобия позволяет одновременно учитывать и чувствительность, и специфичность теста. Пусть событие A — дама диагностировала чай как «правильный», может произойти при наступлении одной из гипотез: H_0 — чай «правильный», H_1 − чай «неправильный».

Для положительного результата:

$$LR_+ = \frac{P(A/H_0)}{P(A/H_1)} = \frac{Se}{1-Sp} = \frac{7/10}{1-8/10} = \frac{7}{2} = 3{,}5.$$

Для отрицательного результата:

$$LR_- = \frac{P(\bar{A}/H_0)}{P(\bar{A}/H_1)} = \frac{1-Se}{Sp} = \frac{1-7/10}{8/10} = \frac{3}{8} = 0{,}375.$$

Таким образом, из значения LR_+ следует, что верная классификация «правильного» чая в 3,5 более вероятна, чем возможность отнесения «неправильного» чая к «правильному».

Отношение рисков верной классификации:

$$\text{«чай + молоко»: } RR = \frac{PV_+}{PV_-} = \frac{7/9}{3/11} = \frac{77}{27} \approx 2{,}852;$$

$$\text{«молоко + чай»: } RR = \frac{(1-PV_+)}{(1-PV_-)} = \frac{2/9}{8/11} = \frac{22}{72} \approx 0{,}306.$$

Отношение шансов при условии однородности данных позволяет судить о степени сопряженности качественных признаков ($OR \gg 1$ — положительная ассоциация, $OR \ll 1$ — отрицательная ассоциация, $OR = 1$ — связи нет). Для рассматриваемого примера $OR = \frac{ad}{bc} = \frac{7\cdot 8}{2\cdot 3} = 9{,}33$ значительно превышающее единицу говорит в пользу верной классификации (обычно достаточно если $OR \geq 5$).

3. *Пример* 13.18 описывает задачу бинарной классификации, которая часто хорошо представляется моделью логистической регрессии. Рассмотрим событие A, принимающее два значения (событие $A_i = 1$ — «правильный» чай, $A_i = 0$ — «правильный» чай), причем

$$A_i = \begin{cases} 1 \text{ при } y_i \geq 0, \\ 0 \text{ при } y_i < 0, \end{cases} \tag{13.83}$$

где

$$y_i = b_0 + b_1 x_i + \varepsilon_i, \tag{13.84}$$

x_i — некоторая скрытая непрерывная или бинарная переменная, например, $x_i = 1$ — прогноз «правильного чая», $x_i = 0$ — прогноз «неправильного чая».

Если $\varepsilon_i \to f(t) = \frac{e^{-t}}{(1+e^{-t})^2} = \frac{e^t}{(1+e^t)^2}$, то (13.83) — логит-модель (функция четная), если $\varepsilon_i \to f(t) = N(0,1)$, то (13.83) — пробит-модель.

Интегральная функция для логит-модели:

$$F(y) = \int_{-\infty}^{y} f(t)dt = \int_{-\infty}^{y} \frac{e^t}{(1+e^t)^2} dt = \int_{-\infty}^{y} \frac{d(1+e^t)}{(1+e^t)^2} dt =$$

$$= \int_{-\infty}^{y} \frac{d(1+e^t)}{(1+e^t)^2} dt = -\frac{1}{1+e^t}\Big|_{-\infty}^{y} = -\left(\frac{1}{1+e^y} - \lim_{u\to-\infty} \frac{1}{1+e^u}\right) =$$

$$= 1 - \frac{1}{1+e^y} = \frac{e^y}{1+e^y} — \tag{13.85}$$

логистическая функция распределения (ее называют также сигмоидой или *S*-функцией, так как она имеет *S*-образную форму).

Обычно вводят обозначение

$$\frac{e^y}{1+e^y} =: \Lambda(y), \tag{13.86}$$

где

$$\Lambda(-y) = 1 - \Lambda(y). \tag{13.87}$$

Таким образом, если рассматривать логит-модель (13.83), то вероятность того, что чай «правильный»:

$$P(A_i = 1) = P(y_i \geq 0) = 1 - P(y_i < 0) = 1 - P(b_0 + b_1 x_i + \varepsilon_i < 0) =$$
$$= 1 - P(\varepsilon_i \leq -(b_0 + b_1 x_i)) = 1 - F(-(b_0 + b_1 x_i)) = 1 - \Lambda(-y) =$$
$$= 1 - \big(1 - \Lambda(y)\big) = \Lambda(y).$$

Итак, учитывая формулы (13.82–13.86) получим, что

$$P(A_i = 1) = \Lambda(\beta_0 + \beta_1 x_i) = \frac{e^{b_0 + b_1 x_i}}{1+e^{b_0 + b_1 x_i}}. \tag{13.88}$$

Очевидно, что

$$P(A_i = 0) = 1 - P(A_i = 1) = 1 - \Lambda(b_0 + b_1 x_i),$$
$$P(A_i = 0) = \frac{1}{1+e^{b_0 + b_1 x_i}}. \tag{13.89}$$

Отношение шансов:

$$OR = \frac{P(A_i=1)}{P(A_i=0)} = e^{b_0 + b_1 x_i},$$

(в том числе $OR(x_i = 0) = e^{b_0}$, $OR(x_i = 1) = e^{b_0 + b_1}$).

Отсюда, логарифм отношения шансов:

$$ln\left(\frac{P(A_i=1)}{P(A_i=0)}\right) = ln\left(\frac{P(A_i=1)}{1-P(A_i=1)}\right) = b_0 + b_1 x_i \tag{13.90}$$

(модель (13.88) или $y_i = b_0 + b_1 x_i$ — называется моделью логистической регрессии).

Логит-модель (логит-преобразование) используется, если классы ($A_i = 1$ и $A_i = 0$) линейно разделимы, и позволяет заменить «жесткую» линейную модель вероятностной вида

$$P = \Lambda(y) = \frac{e^y}{1+e^y}, \tag{13.91}$$

где $y_i = b_0 + b_1 x_i$, а бинарные значения $A = \{0, 1\}$ естественным образом заменяются на значения из интервала $\Lambda \in (0, 1)$ и фактически представляют собой вероятности, при условии, что дама верно диагностирует ($x_i = 1$), ошибается ($x_i = 0$), где i — номер опыта ($i = \overline{1, n}$).

Используем метод максимального правдоподобия при оценке логит-модели для прогнозирования правильного чая, опираясь на способность диагностики дамы.

Функция правдоподобия в нашем случае будет иметь следующий вид:

$$L(b_0, b_1) = \prod_{i=1}^{n} P(A(y_i)) = \prod_{i=1}^{n} P(A(b_0 + b_1 x_i)) =$$
$$= \prod_{x_i=0,} P(b_0, b_1 / A = 0) \prod_{x_i=0} P(b_0, b_1 / A = 1) \times$$
$$\times \prod_{x_i=1} P(b_0, b_1 \,/\, A = 0) \prod_{x_i=1} P(b_0, b_1 / \, A = 1). \tag{13.92}$$

Учитывая формулы (13.85)–(13.86) перепишем (13.92) в виде

$$L(b_0, b_1) = \prod_{x_i=0,\ A=0} \Lambda(b_0 + b_1 x_i) \prod_{x_i=0,\ A=1} \Lambda(b_0 + b_1 x_i) \times$$
$$\times \prod_{x_i=1,\ A=0} \Lambda(b_0 + b_1 x_i) \prod_{x_i=1,\ A=1} \Lambda(b_0 + b_1 x_i) =$$
$$= \prod_{x_i=0,\ A=0} (1 - \Lambda(b_0)) \prod_{x_i=0,\ A=1} \Lambda(b_0) \times$$
$$\times \prod_{x_i=1,\ A=0} \big(1 - \Lambda(b_0 + b_1)\big) \prod_{x_i=1,\ A=1} \Lambda(b_0 + b_1) =$$
$$= (1 - \Lambda(b_0))^{\sum_{x_i=0} I(A_i=0)} (\Lambda(b_0))^{\sum_{x_i=0} I(A_i=1)} \times$$
$$\times \big(1 - \Lambda(b_0 + b_1)\big)^{\sum_{x_i=1} I\ \ (A_i=0)} \big(\Lambda(b_0 + b_1)\big)^{\sum_{x_i=1} I(A_i=1)}, \quad (13.93)$$

где, например, $\sum_{x_i=1} I(A_i = 0)$ — сумма индикаторных величин при условии, что $A_i = 0$ и $x_i = 1$ и т. д.

Пусть

$$\Lambda(b_0) = \frac{e^{b_0}}{1+e^{b_0}} =: a, \quad \Lambda(b_0 + b_1) = \frac{e^{b_0+b_1}}{1+e^{b_0+b_1}} =: b, \quad (13.94)$$

что можно представить и объяснить в виде соответствия двух таблиц.

	$A_i = 1$	$A_i = 0$
$x_i = 1$	7	2
$x_i = 0$	3	8

$\leftrightarrow$

	$A_i = 1$	$A_i = 0$
$x_i = 1$	$\Lambda(b_0 + b_1 x_i) =: b$	$1 - b$
$x_i = 0$	$\Lambda(b_0 + b_1 x_i) =: a$	$1 - a$

Тогда, опираясь на данные этих таблиц и формулы (13.91–13.92), получим

$$L(b_0, b_1) = b^7 (1 - b)^2 a^3 (1 - a)^8.$$

Следовательно, логарифмическая функция правдоподобия имеет вид

$$l(a, b) = 7 \ln(b) + 2 \ln(1 - b) + 3 \ln(a) + 8 \ln(1 - a),$$

отсюда, взяв частные производные по a и b, и приравняв их к нулю, получим

$$\begin{cases} \frac{7}{b} + \frac{-2}{1-b} = 0, \\ \frac{3}{a} + \frac{-8}{1-a} = 0, \end{cases}$$

следовательно, $\hat{a} = \frac{3}{11}, \hat{b} = \frac{7}{9}$.

Из формул (13.94) легко получить, что

$$\hat{b}_{01} = \ln\left(\frac{\hat{a}}{1-\hat{a}}\right) = ln\left(\frac{3/11}{1-3/11}\right) = \ln\left(\frac{3}{8}\right) = -0{,}981,$$
$$\hat{b}_{01} + \hat{b}_{11} = \ln\left(\frac{\hat{b}}{1-\hat{b}}\right) = ln\left(\frac{7/9}{1-7/9}\right) = \ln\left(\frac{7}{2}\right) = 1{,}253, \quad (13.95)$$

значит $\hat{b}_{11} = 2{,}234$.

Имеем

$$P = \Lambda(-0{,}981 + 2{,}234x) = \frac{e^{-0{,}981+2{,}234x}}{1+e^{-0{,}981+2{,}234x}}. \quad (13.96)$$

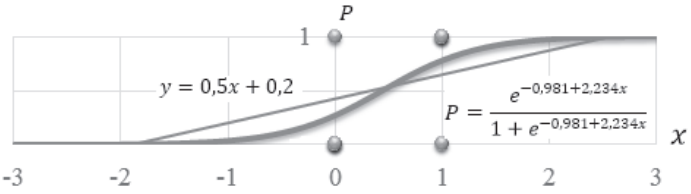


Рис. 13.9 — Логит-модель (задача линейной классификации)

Используя опцию MS Excel «добавить линию тренда», реализующую метод наименьших квадратов (МНК, см. раздел 15, 15.3), по данным таблицы 13.15 получим прямую линию $y = 0{,}5x_i + 0{,}2$ (рис. 13.9), разделяющую два класса ($A_i = 1$ и $A_i = 0$) с учетом прогнозов дамы ($x_i = 1$ и $x_i = 0$), то есть множество точек (0, 0), (0, 1), (1, 0), (1, 1) с учетом частот. Логистическая модель (13.96) позволяет перевести «жесткую» линейную модель в вероятностную интерпретацию, в нашем случае эта вероятность принадлежности к классу $A_i = 1$ (чай «правильный»).

4. Рассмотрим тест отношения правдоподобия для гипотезы о том, что «правильность» чая не связана со способностями дамы к диагностике, то есть $H_0: b_1 = 0$, соответственно, согласно данным опыта $H_1: b_1 = \frac{7}{9}$. Тогда если верна гипотеза H_0, то $a = b$ и, следовательно,

$$L(b_0, b_1 / \, b = a) = (1-a)^{2+8} a^{3+7},$$

значит логарифмическая функция правдоподобия

$$l(a, b = a) = 10\ln(1-a) + 10\ln(a),$$

$$\frac{\partial l}{\partial a} = \frac{-10}{1-a} + \frac{10}{a} = 0,$$

$$\hat{a} = 1\,/2.$$

Таким образом, $\hat{b}_{02} = \ln\left(\frac{\hat{a}}{1-\hat{a}}\right) = \ln\left(\frac{0{,}5}{1-0{,}5}\right) = 0$, $\hat{b}_{12} = 0$.

Найдем наблюдаемые значения статистик отношения правдоподобия:

$$l(\hat{b}_{01} = -0{,}981, \hat{b}_{11} = 2{,}234 \,/\, H_1) = l\left(\hat{a} = \frac{3}{11}, \hat{b} = 7/9 \,/H_1\right) =$$

$$= 7\,ln\left(\frac{7}{9}\right) + 2\,ln\left(1 - \frac{7}{9}\right) + 3\,ln\left(\frac{3}{11}\right) + 8\,ln\left(1 - \frac{3}{11}\right) + =$$

$$= -11{,}2128;$$

$$l(\hat{b}_{02} = 0, \hat{b}_{12} = 0 \,/\, H_0) = l(\hat{a} = 1/2, b = a \,/H_0) =$$

$$= 10\,ln(1 - 1/2) + 10\,ln(1/2) = -13{,}8629.$$

$$LR_{\text{н}} = 2\left[l\left(\hat{b}_{01} = -0{,}981, \hat{b}_{11} = 2{,}234/\, H_1\right) - l\left(\hat{b}_{02} = 0, \hat{b}_{12} = 0/\, H_0\right)\right],$$

$$LR_{\text{н}} = 2[-11{,}2128 - (-13{,}8629)] = 5{,}300,$$

так как

$$LR_{\text{кр}} \to \chi_1^2,$$

то при уровне значимости $\alpha = 0{,}05$ критическое значение распределения хи-квадрат Пирсона с одной степенью свободы (так как оценивается один параметр при переменной) принимает значение 3,84. Следовательно нулевая гипотеза $H_0: b_1 = 0$ отвергается — «правильность» чая связана с диагностическими возможностями дамы. Мы получили еще одно подтверждение способностей дамы к диагностике чая.

Построение логит-модели по данным примера 13.18 (табл. 13.15) в эконометрическом пакете *gretl* позволяет получить модель, совпадающую с нашей (13.96). Аналитическая платформа *Deductor* 5.3 с использованием инструмента «Логистическая регрессия» без тестового множества (то есть без обучения) выдают результат логит-модели с коэффициентами $\hat{b}_{11} = 2{,}234$ и $\hat{b}_{01} = -1{,}386$. Отличия в незначимом для нас коэффициенте $\hat{b}_{01}$ можно отнести к особенностям настройки алгоритма вычислений коэффициентов (разным численным методам, в том числе небольшим объемом имеющихся данных, так как для больших наборов — порядка

100 и более наблюдений, результаты совпадают). *ROC*-кривая с точкой отсечения 0,2, соответствующей балансу в данных (соотношению количества негативных и позитивных примеров, у нас 10:10), отражена на рисунке 13.10. Результаты моделирования ($AUC = 0{,}75$; $Sp = 72{,}73$; $Se = 77{,}78$) способностей дамы с использованием логит-модели достаточно хорошо описывают способности дамы, рассмотренные ранее (п. 2).

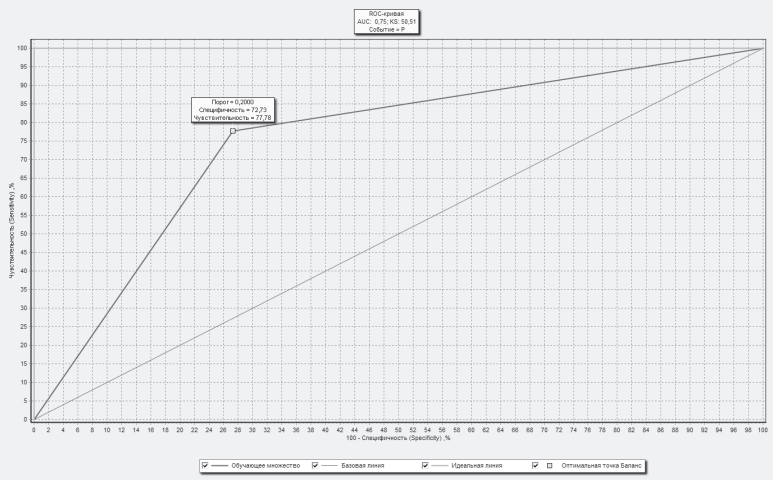


Рис. 13.10 — *ROC*-кривая

4. Рассмотрим возможные варианты вероятностей «правильного и неправильного чая» при прогнозе дамы из примера 13.18, учитывая формулы (13.86)–(13.89).

$$\hat{P}(A_i = 1/x_i = 1) = \Lambda(\beta_0 + \beta_1 \cdot 1) = \Lambda(-0{,}981 + 2{,}234) =$$

$$= \frac{e^{1{,}253}}{1 + e^{1{,}253}} = 0{,}7778;$$

$$\hat{P}(A_i = 0/x_i = 1) = 1 - \Lambda(\beta_0 + \beta_1 \cdot 1) = 1 - 0{,}7778 = 0{,}2222;$$

$$\hat{P}(A = 1/x_i = 0) = \Lambda(\beta_0) = \Lambda(-0{,}981) = \frac{e^{-0{,}981}}{1 + e^{-0{,}981}} = 0{,}2727;$$

$$\hat{P}(A_i = 0/x_i = 0) = 1 - \Lambda(\beta_0) = 1 - 0{,}2727 = 0{,}7273.$$

Итак, если дама прогнозирует «правильный чай», то с вероятностью 0,7778 она права (что совпадает со значением чувствительности, рис. 13.10) и с вероятностью 0,2222 — ошибается, отношение шансов:

$$OR(A = 1) = \frac{P(A=1)}{1-P(A=1)} = \frac{0{,}7778}{0{,}2222} = 3{,}5;$$

если дама прогнозирует «неправильный чай», то с вероятностью 0,7273 она права (что совпадает со значением специфичности, рис. 13.10) и с вероятностью 0,2727 — ошибается, отношение шансов:

$$OR(A = 0) = \frac{P(A = 0)}{1 - P(A = 0)} = \frac{0{,}7273}{0{,}2727} = 2{,}267.$$

Заметим, что $OR(A=1)\cdot OR(A=0)=3{,}5\cdot 2{,}267=9{,}33$ — равно отношению шансов, найденному выше ($OR=\frac{ad}{bc}=\frac{7\cdot 8}{2\cdot 3}=9{,}33$).

Оптимальный последовательный алгоритм А. Вальда систематически был разработан в годы Второй мировой войны (хотя первые подходы к практической реализации известны с 1929 г.) и определенное время был засекречен, опубликован в 1950-е годы. В процессе сбора данных проводится тест, опирающийся на критерий отношения правдоподобия, после каждого нового наблюдения и пересчета значения критерия принимается одно из трех решений: отвергнуть нулевую гипотезу при определенном уровне значимости в пользу альтернативной, остановив сбор данных; не отвергать нулевую гипотезу, остановив сбор данных; продолжить сбор данных, пока нет возможности принятия решения.

Пусть рассматривается простая гипотеза, альтернативная гипотеза. Имеется функция плотности вероятности $f(X,\theta)$ и n последовательных наблюдений.

После каждого наблюдения рассматривается неравенство

$$\frac{\beta}{1-\alpha}<\lambda<\frac{1-\beta}{\alpha}, \tag{13.97}$$

где $\lambda=\frac{l_1(X)}{l_0(X)}$ — коэффициент отношения правдоподобия.

Алгоритм:

1) если

$$\lambda\le\frac{\beta}{1-\alpha}, \tag{13.98}$$

то наблюдения прекращаются и принимается нулевая гипотеза;

2) если

$$\lambda\ge\frac{1-\beta}{\alpha}, \tag{13.99}$$

то наблюдения прекращаются и принимается альтернативная гипотеза;

3) если выполняется неравенство (13.97), то оснований для вывода нет, необходимы дополнительные наблюдения.

Последовательный анализ требует значительно меньше наблюдений, чем другие методы.

Рассмотрим пример А. Вальда, приводимый в книге А. К. Митропольского [81].

Пример 13.19. Проведенные испытания относительно доли бракованной продукции дали результаты, приведенные в таблице 13.16.

Таблица 13.16

Последовательный анализ проверки гипотезы

n	x_n	k	n	x_n	k
1	0	0	12	1	4
2	0	0	13	0	4
3	1	1	14	1	5
4	0	1	15	0	5
5	0	1	16	0	5
6	0	1	17	0	5
7	0	1	18	1	6
8	0	1	19	0	6
9	1	2	20	0	6
10	0	2	21	0	6
11	1	3	22	1	7

Результат каждого испытания соответствует закону Бернулли, если событие (появление брака) происходит, то $P(X = 1) = p$, если событие не происходит, то $P(X = 0) = 1 - p = q$.

Необходимо проверить нулевую гипотезу $H_0: p = p_0$. Альтернативная гипотеза $H_1: p = p_1$ $(p_1 > p_0)$.

Пусть событие произошло k раз. Если проверяемая гипотеза верна, то вероятность этого события будет равна

$$P_{n0} = p_0^k q_0^{n-k}.$$

Соответственно, если верна альтернативная гипотеза, то вероятность будет равна

$$P_{n1} = p_1^k q_1^{n-k}.$$

Коэффициент отношения правдоподобия будет равен

$$\lambda = \frac{l_1(X)}{l_0(X)} = \frac{p_1^k q_1^{n-k}}{p_0^k q_0^{n-k}} = \left(\frac{p_1}{p_0}\right)^k \left(\frac{q_1}{q_0}\right)^{n-k},$$

логарифм равен

$$lg\,\lambda = klg\,\frac{p_1}{p_0} + (n-k)lg\,\frac{q_1}{q_0}.$$

Если $p_0 = 0{,}1;\ p_1 = 0{,}3;\ \alpha = 0{,}01;\ \beta = 0{,}02,$

то

$$\frac{\beta}{1-\alpha} = \frac{2}{99}, \frac{1-\beta}{\alpha} = \frac{98}{1}.$$

Отношение правдоподобия равно

$$\lambda = \left(\frac{3}{1}\right)^k \left(\frac{7}{9}\right)^{n-k}.$$

Подставим значение λ в (13.97), получим

$$\left(\frac{3}{1}\right)^k \left(\frac{7}{9}\right)^{n-k} \le \frac{2}{99},$$

$$\left(\frac{27}{7}\right)^k \left(\frac{7}{9}\right)^n \le \frac{2}{99},$$

логарифмируя, получим неравенство

$$k \le \frac{lg\frac{2}{99}}{lg\frac{27}{7}} + n\frac{lg\frac{9}{7}}{lg\frac{27}{7}}$$

или

$$k \le -2{,}583 + 0{,}186\,n. \qquad (13.100)$$

Аналогично, опираясь на неравенство (13.98), получим

$$\left(\frac{27}{7}\right)^k \left(\frac{7}{9}\right)^n \ge \frac{98}{1}.$$

Отсюда имеем

$$k \ge \frac{lg\frac{2}{99}}{lg\frac{27}{7}} + n\frac{lg\frac{9}{7}}{lg\frac{27}{7}}$$

или

$$k \ge 2{,}875 + 0{,}186\,n. \qquad (13.101)$$

Рассмотрим величины n и k как координаты точки и проведем прямые $k1 = -2{,}583 + 0{,}186\,n$ и $k2 = 2{,}875 + 0{,}186\,n$. В результате получим на плоскости три области, на которые прямые разбивают плоскость. Каждое испытание отмечается точкой (n, k). Если точка ниже прямой $k1 = -2{,}583 + 0{,}186\,n$,

то принимается гипотеза $H_0: p = 0{,}1$ и испытания прекращаются. Если точка выше прямой $k2 = 2{,}875 + 0{,}186\,n$, то принимается альтернативная гипотеза $H_1: p = 0{,}3$. Точка между прямыми не дает основания для выбора, испытания следует продолжить (рис. 13.11).

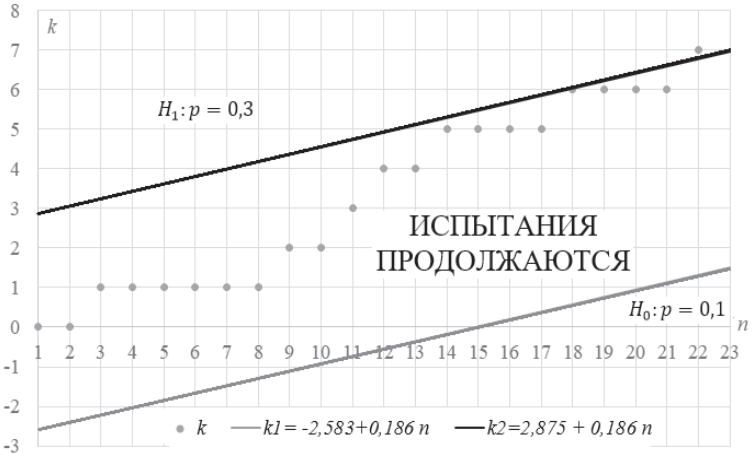


Рис. 13.11 — Последовательный анализ для проверки гипотезы о доле брака

В нашем случае последнее испытание позволяет отвергнуть нулевую гипотезу в пользу альтернативной. То есть гипотеза $p = 0{,}3$ принимается, и партия выпущенной продукции должна быть забракована, если допустимая доля брака не должна превышать 0,1.

Темы (вопросы) для самоконтроля

1. Статистическая гипотеза.
2. Статистический критерий.
3. Критическая область.
4. Алгоритм проверки статистических гипотез.
5. Вероятность ошибки первого рода, вероятность ошибки второго рода, мощность критерия, $p - value$.
6. Проверка гипотезы о среднем нормально распределенной генеральной совокупности.
7. Проверка гипотезы о числовом значении генеральной доли.
8. Проверка гипотезы о равенстве дисперсий.
9. Проверка гипотезы о равенстве средних независимых совокупностей.
10. Проверка гипотезы о равенстве средних зависимых совокупностей.
11. Проверка гипотезы о равенстве долей.
12. Проверка гипотезы о виде распределения.
13. Проверка гипотезы об однородности выборок.
14. Проверка гипотезы о независимости выборок.
15. Понятие об оптимальных критериях.
16. Принцип отношения правдоподобия и критерий отношения правдоподобия.
17. Теорема Уилкса.

Глава 14
Дисперсионный анализ

14.1. Постановка задачи и сущность дисперсионного анализа

Дисперсионный анализ (*ANOVA*[23]) как метод обработки результатов исследования разработан Р. Фишером (1918–1935 гг.) в связи с исследованиями в сельском хозяйстве для количественной оценки влияния факторов на урожайность сельскохозяйственных культур и выявления условий, при которых испытываемый сорт сельскохозяйственной культуры даёт максимальный урожай. Дальнейшее развитие дисперсионный анализ получил в работах Йейтса [122, 130].

Дисперсионный анализ позволяет ответить на вопрос о существенности влияния одного или нескольких факторов на изменчивость результативного признака, значения которого могут быть получены в результате опыта или эксперимента. При проверке статистических гипотез предполагается случайность вариации изучаемых факторов. В дисперсионном анализе один или несколько факторов изменяются заданным образом, причем эти изменения могут влиять на результаты наблюдений. Исследование такого влияния и является целью дисперсионного анализа.

В настоящее время наблюдается все более широкое использование дисперсионного анализа в экономике, социологии, биологии и др., особенно после появления программных средств, снявших проблемы громоздкости статистических вычислений.

В практической деятельности, а также в различных областях науки мы часто сталкиваемся с необходимостью оценить влияние различных факторов на те или иные признаки. Часто эти факторы имеют качественный характер (например, качественным фактором, влияющим на экономический эффект, может быть введение новой системы управления производством), и тогда дисперсионный анализ приобретает особую ценность, так как становится единственным статистическим способом исследования, позволяющим проводить количественную оценку влияния факторов.

Теорию дисперсионного анализа можно считать в достаточной мере сформировавшейся, но способы организации эксперимента и вычислительные схемы продолжают совершенствоваться.

Постановка задачи. В любой совокупности испытаний имеется несколько факторов, вызывающих изменчивость средних значений наблюдаемых случайных величин — результативных признаков. Эти факторы могут принадлежать одному или нескольким источникам изменчивости (например, расположение торговых заведений в центре и на окраине города, изменения в законодательстве, разные климатические условия, разные уровни образования и т. п.). Очевидно, что даже при самом тщательном исследовании не удается выявить все источники изменчивости, а иногда в этом нет необходимости или смысла. Но при наличии опыта эксперта и в зависимости от цели исследования всегда можно выдвинуть гипотезу о возможности влияния тех или иных факторов на результативный признак.

[23] ANOVA — ANalysis Of Variance (*англ.*) — дисперсионный анализ.

Дисперсионный анализ дает возможность установить, существенное ли влияние оказывает тот или иной из рассматриваемых факторов на изменчивость признака, а также определить количественно «удельный вес» каждого из источников в их общей изменчивости. Но дисперсионный анализ позволяет дать положительный ответ лишь о наличии существенного влияния, в противном случае вопрос остается открытым и требует дополнительных исследований (чаще всего — увеличения числа наблюдений).

В дисперсионном анализе используются следующие термины.

Фактор (A, B, C, …) — признак, который предположительно должен оказывать влияние на результат (результативный признак).

Уровень фактора — значения (A_i, i = 1, 2, …, p) (сорт, доза внесения удобрений, способ обработки почвы), которые может принимать факторный признак.

Отклик — значение измеряемого результативного признака (величина результата X).

Техника дисперсионного анализа меняется в зависимости от числа изучаемых независимых факторов. Если факторы, вызывающие изменчивость среднего значения признака, принадлежат одному источнику, то мы имеем простую группировку, или однофакторный дисперсионный анализ и далее, соответственно, двойная группировка — двухфакторный дисперсионный анализ, трехфакторный дисперсионный анализ, …, m-факторный. Факторы в многофакторном анализе принято обозначать латинскими буквами: A, B, C и т. д.

З*адача дисперсионного анализа* — количественная оценка влияния тех или иных факторов (или уровней факторов) и их взаимодействий на изменчивость средних значений результативного признака.

Сущность дисперсионного анализа состоит в разложении общей дисперсии (σ^2) результативного признака наблюдаемой совокупности, вызванной всеми источниками изменчивости, на составляющие дисперсии, порожденные независимыми факторами. Каждая из этих составляющих дает оценку дисперсии σ_A^2, σ_B^2, вызванную конкретным источником изменчивости в общей совокупности. Для проверки значимости этих составляющих оценок дисперсии их сравнивают с остаточной дисперсией в общей совокупности (по критерию Фишера).

Например, в двухфакторном анализе мы получим разложение вида:

$$\sigma_o^2 = \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma_\varepsilon^2, \qquad (14.1)$$

где σ_o^2 — общая дисперсия результативного признака X; σ_A^2 — часть дисперсии, вызванной влиянием фактора A; σ_B^2 — часть дисперсии, вызванной влиянием фактора B; σ_{AB}^2 — часть дисперсии, вызванной влиянием взаимодействия факторов A и B; σ_ε^2 — дисперсия, вызванная неучтенными случайными причинами (случайная дисперсия).

В дисперсионном анализе рассматриваются эксперименты трех видов:

а) эксперименты, в которых все факторы имеют систематические (фиксированные) уровни;

б) эксперименты, в которых все факторы имеют случайные уровни;

в) эксперименты, в которых есть факторы, имеющие случайные уровни, а также факторы, имеющие фиксированные уровни.

Эти три вида соответствуют трем моделям, которые рассматриваются в дисперсионном анализе.

Если все контролируемые факторы имеют фиксированные уровни, то модель дисперсионного анализа называется моделью 1 или линейной детерминированной моделью. Если все контролируемые факторы имеют случайные уровни, возникающие в процессе проведения эксперимента, то модель называется моделью 2 или случайной моделью.

Модель называется смешанной, если в многофакторном эксперименте одни факторы имеют фиксированные уровни, а другие — случайные.

Рассмотрим единичный фактор, который принимает p различных уровней, и предположим, что на каждом уровне сделано n наблюдений, что дает $N = np$ наблюдений. (Ограничимся рассмотрением первой модели дисперсионного анализа — все факторы имеют фиксированные уровни.) Пусть результаты представлены в виде x_{ij} $(i = 1, 2, \dots, p;\ \ j = 1, 2, \dots, n)$. Данные обычно располагают в виде таблицы (табл. 14.1).

Таблица 14.1

Значения результативного признака

Уровень фактора	Номер наблюдения							$\sum_{j=1}^{n} X_{ij}$	Среднее, $\overline{X}_i$
	1	2	3	…	j	…	n		
A_1	X_{11}	X_{12}	X_{13}	…	X_{1j}	…	X_{1n}		$\overline{X}_1$
A_2	X_{21}	X_{22}	X_{23}	…	X_{2j}	…	X_{2n}		$\overline{X}_2$
…	…	…	…	…	…	…	…		
A_i	X_{i1}	X_{i2}	X_{i3}	…	X_{ij}	…	X_{in}		$\overline{X}_i$
…	…	…	…	…	…	…	…		
A_p	X_{p1}	X_{p2}	X_{p3}	…	X_{pj}	…	X_{pn}		$\overline{X}_p$
$\sum_{i=1}^{p} X_{ij}$									$\overline{X}$

Предполагается, что для каждого уровня, состоящего из n наблюдений, имеется средняя, которая равна сумме общей средней и ее вариации, обусловленной выбранным уровнем фактора:

$$x_{ij} = \mu + A_i + \varepsilon_{ij}, \qquad (14.2)$$

где μ — общая средняя по опыту или эксперименту; A_i — эффект, обусловленный i-м уровнем фактора A; ε_{ij} — вариация результативного признака внутри отдельного уровня фактора. С помощью члена ε_{ij} принимаются в расчет все неконтролируемые факторы.

Применение дисперсионного анализа предполагает, что:

1) $M(\varepsilon_{ij}) = 0$,

2) $D(\varepsilon_{ij}) = \sigma^2 = const$,

3) $\varepsilon_{ij} \rightarrow N(0, \sigma^2)$ или $x_{ij} \rightarrow N(a, \sigma^2)$.

В однофакторном дисперсионном анализе рассматривается нулевая гипотеза H_0: $\bar{X}_1 = \bar{X}_2 = \cdots = \bar{X}_p = \mu$, средние значения по уровням фактора равны, т. е. изучаемый фактор не оказывает влияния на изменчивость признака.

Конкурирующая гипотеза, H_1: не все средние по уровням фактора A равны.

Тогда общая сумма квадратов[24]:

$$SS_o = \sum_{j=1}^{n}(X_{1j} - \bar{X}_{..})^2 + \sum_{j=1}^{n}(X_{2j} - \bar{X}_{..})^2 + \cdots + \sum_{j=1}^{n}(X_{pj} - \bar{X}_{..})^2, \quad (14.3)$$

где SS_o — общая сумма квадратов отклонений всех значений признака от общей средней, $\bar{X}_{..}$ — общая средняя для всех X_{ij}:

$$\bar{X}_{..} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n} X_{ij}}{np}.$$

SS_o, согласно правилу сложения дисперсий (11.28), может быть представлена как сумма внутригрупповой (SS_z) и межгрупповой (SS_v) дисперсий, которые затем сравниваются по F-критерию Фишера — Снедекора.

Пусть наблюдения на фиксированном уровне фактора нормально распределены относительно среднего значения $\mu + A_i$ с общей дисперсией σ^2. Тогда (точка вместо индекса обозначает усреднения соответствующих наблюдений по этому индексу):

$$X_{ij} - \bar{X}.. = (\bar{X}_{i.} - \bar{X}..) + (X_{ij} - \bar{X}_{i.}). \quad (14.4)$$

После возведения обеих частей равенства (14.4) в квадрат и суммирования по i и j, получим

$$\sum_{i,j}(X_{ij} - \bar{X}_{..})^2 = \sum_{i,j}(\bar{X}_{i.} - \bar{X}_{..})^2 + \sum_{i,j}(X_{ij} - \bar{X}_{i.})^2 + \\ + \sum_{i,j}(\bar{X}i. - \bar{X}..)(Xij - \bar{X}i.), \quad (14.5)$$

но $\sum_{i,j}(\bar{X}_{i.} - \bar{X}_{..})(X_{ij} - \bar{X}_{i.}) = \sum_i(\bar{X}_{i.} - \bar{X}_{..})\sum_j(X_{ij} - \bar{X}_{i.}) = 0$, так как по свойству средней арифметической $\sum_j(X_{ij} - \bar{X}_{i.}) = 0$.

Иначе сумму квадратов отклонений можно записать как факторную сумму квадратов, плюс остаточную сумму квадратов

$$SS_o = SS_v + SS_z, \quad (14.6)$$

где

$$SS_o = \sum_{i=1}^{p}\sum_{j=1}^{n}(X_{ij} - \bar{X}_{..})^2,$$
$$SS_v = \sum_{i=1}^{p}\sum_{j=1}^{n}(\bar{X}_{i.} - \bar{X}_{..})^2 = n\sum_{i=1}^{p}(\bar{X}_{i.} - \bar{X}_{..})^2,$$
$$SS_z = \sum_{i=1}^{p}\sum_{j=1}^{n}(X_{ij} - \bar{X}_{i.})^2.$$

Для оценки числа степеней свободы SS_z и SS_v используем ортогональное преобразование (12.28) и идеологию вывода леммы Фишера. Учитывая, что $y_p = \bar{x}\sqrt{n}$, получим

$$\sum_{j=1}^{n-1} y_{1j}^2 = \sum_{j=1}^{n} X_{1j}^2 - n(\bar{X}_1)^2 = \sum_{j=1}^{n}\left(X_{1j} - \bar{X}_1\right)^2,$$
$$\sum_{j=1}^{n-1} y_{2j}^2 = \sum_{j=1}^{n} X_{2j}^2 - n(\bar{X}_2)^2 = \sum_{j=1}^{n}\left(X_{2j} - \bar{X}_2\right)^2, \quad (14.7)$$
$$\sum_{j=1}^{n-1} y_{pj}^2 = \sum_{j=1}^{n} X_{pj}^2 - n\left(\bar{X}_p\right)^2 = \sum_{j=1}^{n}\left(X_{pj} - \bar{X}_p\right)^2.$$

Таким образом, *каждое из равенств* (14.7) *имеет* $(n-1)$ *степень свободы, равную размерности ортогональной системы векторов* y_i, *необходимой для*

[24] *SS* — Sum of Squares (*англ.*) — сумма квадратов.

описания суммы квадратов отклонений, n наблюдений от выборочной средней ($\sum_{j=1}^{n}\left(X_{ij}-\bar{X}_{i.}\right)^2$). Складывая равенства (14.7), получим, что различия внутри уровней или, иначе, остаточная сумма квадратов, равна

$$SS_z=\sum_{i=1}^{p}\sum_{j=1}^{n-1}y_{ij}^2=\sum_{i=1}^{p}\sum_{j=1}^{n}\left(X_{ij}-\bar{X}_{i.}\right)^2. \qquad (14.8)$$

Остаточная сумма квадратов отклонений SS_z вычисляется по отклонениям N наблюдений от p выборочных средних и, следовательно, имеет $N-p=np-p=$ $=p(n-1)$ степеней свободы.

Следовательно, внутригрупповую дисперсию можно оценить как

$$s_z^2=\frac{SS_z}{np-p}. \qquad (14.9)$$

Если число наблюдений на $i-$ ом уровне фактора равно n_i, то остаточная сумма квадратов будет иметь ($\sum_{i=1}^{p}n_i-p$) степеней свободы, и вместо формулы (14.9) получим формулу

$$s_z^2=\frac{SS_z}{\sum_{i=1}^{p}n_i-p}. \qquad (14.10)$$

Для вычисления факторной суммы квадратов и оценки межгрупповой дисперсии ортогональную систему векторов y_i подвергают новому ортогональному преобразованию, при котором они переходят в v_i , где v_p пропорционально общей выборочной средней $\bar{X}_{..}$ для всех X_i:

$$v_p=\bar{X}_{..}\sqrt{N}=\frac{\sum_{j=1}^{n_1}X_{1j}+\sum_{j=1}^{n_2}X_{2j}+\cdots+\sum_{j=1}^{n_p}X_{pj}}{\sqrt{N}}=$$
$$=\frac{y_1\sqrt{n_1}+y_2\sqrt{n_2}+\cdots+y_p\sqrt{n_p}}{\sqrt{N}}, \qquad (14.11)$$

сумма квадратов коэффициентов равна единице:

$$\frac{n_1+n_2+\cdots+n_p}{N}=1,$$

следовательно, такое ортогональное преобразование возможно. Преобразование ортогонально, поэтому

$$\sum_{i=1}^{p}y_i^2=\sum_{i=1}^{p}v_i^2. \qquad (14.12)$$

Пусть

$$\sum_{i=1}^{p-1}v_i^2=:SS_v, \qquad (14.13)$$

следовательно,

$$SS_v=\sum_{i=1}^{p}y_i^2-v_p^2=\sum_{i=1}^{p}y_i^2-N(\bar{X}_{..})^2. \qquad (14.14)$$

Факторная сумма квадратов отклонений SS_v вычисляется по отклонениям p средних от общей средней $\bar{X}_{..}$, поэтому SS_v имеет (p – 1) степеней свободы. Общая сумма квадратов отклонений SS_o имеет (N – 1) степеней свободы.

Сложим SS_v и SS_z, получим формулы (14.4)–(14.6), считая, что $n_i=n$:

$$SS_v+SS_z=\sum_{i=1}^{p}y_i^2-N(\bar{X}_{..})^2+\sum_{i=1}^{p}\sum_{j=1}^{n}\left(X_{ij}-\bar{X}_{i.}\right)^2=$$
$$=\sum_{i=1}^{p}X_i^2-N(\bar{X}_{..})^2+\sum_{i=1}^{p}\sum_{j=1}^{n}\left(X_{ij}-\bar{X}_{i.}\right)^2=$$
$$=n\sum_{i=1}^{p}(\bar{X}_{i.}-\bar{X}..)^2+\sum_{i=1}^{p}\sum_{j=1}^{n}\left(X_{ij}-\bar{X}_{i.}\right)^2. \qquad (14.15)$$

По результатам вычислений строится таблица дисперсионного анализа (табл. 14.2).

В случае справедливости гипотезы о том, что влияние всех уровней фактора одинаково — обе величины s_v^2 и s_z^2 (средние квадраты) будут несмещенными оценками σ^2. Значит, гипотезу можно проверить, вычислив отношение $F_н = (s_v^2/s_z^2)$ и сравнив его с $F_{кр}$ с $k_1 = (p-1)$ и $k_2 = (N-p)$ степенями свободы.

Если $F_н > F_{кр}$, то нулевая гипотеза отвергается, в противном случае принимается.

Для оценки существенности частных различий между средними, когда $F_н > F_{кр}$ (когда нулевая гипотеза отвергается) вычисляют:

а) *среднюю ошибку опыта*:

$$S_{\bar{x}} = \sqrt{\frac{s_z^2}{n}}, \tag{14.16}$$

б) *ошибку разности средних*:

$$S_d = S_{\bar{x}}\sqrt{2}, \tag{14.17}$$

в) *наименьшую существенную разность*:

$$\text{НСР}_{\alpha,к} = t_{\alpha,к}S_d. \tag{14.18}$$

Таблица 14.2

Схема однофакторного дисперсионного анализа

Источник изменчивости	Сумма квадратов (SS)	Степени свободы (k)	Средний квадрат отклонений (s^2)
Различия между уровнями (факторная)	$SS_v = n\sum_{i=1}^{p}(\bar{X}_{i.} - \bar{X}..)^2$	$p-1$	$s_v^2 = \frac{SS_v}{p-1}$
Различия внутри уровней (остаточная)	$SS_z = \sum_{i=1}^{p}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_{i.}\right)^2$	$N-p$	$s_z^2 = \frac{SS_z}{N-p}$
Сумма (общая)	$SS_o = \sum_{i=1}^{p}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_{..}\right)^2$	$N-1$	

Критическое значение t находится по таблице значений критерия Стьюдента при заданном уровне значимости α и числе степеней свободы $k = N-p$. Сравнивая разность средних значений $\bar{X}_i$. по вариантам опыта с НСР, делают вывод о существенности различий в уровне частных средних.

Важным фактом в дисперсионном анализе является приводимая ниже теорема Кочрена (Кохрана или Кокрана — *Cochran*, 1934), доказательство которой рассматривается, например, у Г. Шеффе [130].

Теорема (**Кочрена**). Пусть вектор $X = \{X_1, X_2, \dots, X_i, \dots, X_n\}$ состоит из n независимых нормально распределенных стандартизированных переменных $x_i \to N(a_i, 1)$ и пусть сумма квадратов $Q = X^TX$ разложена на s квадратичных форм $Q_j = X^TA_jX$ с рангами[25] $n_j = r(Q_j)$, то есть

[25] Ранг квадратичной формы Q будем обозначать $r(Q)$.

$$Q = X^T X = \sum_{i=1}^{n} X_i^2 = \sum_{j=1}^{s} Q_j = \sum_{j=1}^{s} X^T A_j X. \qquad (14.19)$$

Тогда любое из следующих трех условий следует из двух других:

1) сумма рангов r_j форм Q_j равна рангу формы Q:

$$r(Q) = \sum_{j=1}^{s} r(Q_j) = n_1 + n_2 + \cdots + n_s = n, \qquad (14.20)$$

2) каждая форма Q_j имеет (нецентральное или центральное, если $a_i = 0$) распределение хи-квадрат Пирсона с $n_j = r(Q_j)$ степенями свободы,

3) каждая форма Q_j независима от любой другой.

Следствие 1. Если каждую квадратичную форму представить в виде суммы квадратов линейных форм от переменных X_i, то эти формы ортогональны.

Следствие 2. Если $\sum_{i=1}^{n} X_i^2 = \sum_{j=1}^{s} Q_j$, где ранг квадратичной формы Q_j не превосходит m_j и $\sum_{j=1}^{s} m_j = n$, то $r(Q_j) = m_j$.

Рассмотрим применение теоремы Кочрена для приведенного выше примера классификации по одному признаку с равным числом наблюдений — случай однофакторного анализа с одним наблюдением в ячейке.

Выше были получены следующие формы SS:

$$SS_v = n\sum_{i=1}^{p}(\bar{X}_{i.} - \bar{X}..)^2,$$

$$SS_z = \sum_{i=1}^{p}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_{i.}\right)^2,$$

$$SS_o = \sum_{i=1}^{p}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_{..}\right)^2.$$

Из выражения (14.4) получим равенство

$$X_{ij} = \bar{X}.. + (\bar{X}_{i.} - \bar{X}..) + (X_{ij} - \bar{X}_{i.}). \qquad (14.21)$$

Сохраним скобки при возведении в квадрат (14.21) и суммировании:

$$\sum_{i,j} X_{ij}^2 = np(\bar{X}..)^2 + n\sum_{i=1}^{p}(\bar{X}_{i.} - \bar{X}..)^2 + \sum_{i=1}^{p}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_{i.}\right)^2. \qquad (14.22)$$

По следствию 2 получаем, что

$$r(np(\bar{X}..)^2) = 1, \quad r(SS_v) = p - 1, \quad r(SS_z) = p(n-1).$$

Запишем (14.22) в виде

$$\sum_{i,j}\left(\frac{X_{ij}}{\sigma}\right)^2 = \frac{np(\bar{X}..)^2}{\sigma^2} + \frac{n\sum_{i=1}^{p}(\bar{X}_{i.} - \bar{X}..)^2}{\sigma^2} + \frac{\sum_{i=1}^{p}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_{i.}\right)^2}{\sigma^2} \qquad (14.23)$$

или

$$\sum_{i,j}\left(\frac{X_{ij}}{\sigma}\right)^2 = \frac{np(\bar{X}..)^2}{\sigma^2} + \frac{SS_v}{\sigma^2} + \frac{SS_z}{\sigma^2}. \qquad (14.24)$$

По теореме Кочрена слагаемые (14.24) имеют независимые нецентральные распределения хи-квадрат с указанными степенями свободы, кроме того, согласно следствию 1, все линейные формы, входящие в (14.21), взаимно ортогональны (например, $X_{i.} - \bar{X}..$ и $X_{ij} - \bar{X}_{i.}$).

Теорема Кочрена часто используется в дисперсионном анализе, поскольку ее применение заменяет проверку ортогональности установлением равенств (14.19)–(14.20).

Замечание. В основе дисперсионного анализа, как, впрочем, и большинства методов математической статистики, лежит метод наименьших квадратов (МНК), который как метод получения оценок параметров распределений уже обсуждался в теме «выборочный метод» и тесно связан с теорией квадратичных форм в линейной алгебре. *Вычислительная сторона МНК* относится к теории

численных методов линейной алгебры (в которой среди множества книг следует порекомендовать [5, 6, 52, 106]) и сводится к решению систем линейных уравнений, поэтому мы введем некоторые понятия и поясним основные аспекты на соответствующих простых примерах [6, 106]. ■

1. *Системы линейных уравнений*. Пусть рассматривается n-мерное векторное (линейное) евклидово пространство R^n, в котором определены операции сложения векторов, умножения вектора на число, скалярное умножение векторов. Тогда произвольная матрица $A_{mn} =: A$, представляет собой *линейный оператор* (обобщение линейной функции на случай большего числа аргументов и значений) из линейного пространства $\mathbb{R}^n$ в линейное пространство $\mathbb{R}^m$, так как для любого числа λ и векторов $x, y \in \mathbb{R}^n$ выполняются свойства:

$$A(x + y) = Ax + Ay,\ A(\lambda x) = \lambda A(x).$$

Таким образом, если $Ax = b$, то вектор $b \in R^m$ — образ вектора $x \in R^n$, а сам вектор x — прообраз вектора b. Множество всех векторов $x \in R^n$— область определения оператора A. Множество значений всех образов линейного оператора — всех векторов b, для которых имеет решение система $Ax = b$, обозначается как $im\ A$ и является подмножеством в линейном пространстве $\mathbb{R}^m$: $im\ A \subseteq \mathbb{R}^m$. Образ матрицы — пространство ее столбцов.

Аналогично, образ транспонированной матрицы A^T — пространство ее строк, обозначается как $im\ A^T$. Размерности $im\ A$ и $im\ A^T$ совпадают с рангом матрицы A (числом линейно независимых строк или столбцов):

$$dim\ Im\ A = im\ A^T = rang\ A = r \le n.$$

В матричной форме действие линейного оператора на вектор $x = \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix}$ можно записать как систему линейных уравнений

$$Ax = b, \tag{14.25}$$

или

$$\begin{bmatrix} a_{11} & a_{1r} & a_{1n} \\ a_{r1} & a_{rr} & a_{rn} \\ a_{m1} & a_{mr} & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_r \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_r \\ b_m \end{bmatrix}. \tag{14.26}$$

В результате использования метода Гаусса система (14.26) может быть приведена к трапециевидному виду

$$Ux = \mathrm{C}, \tag{14.27}$$

или

$$\begin{bmatrix} u_{11} & u_{1r} & u_{1n} \\ 0 & u_{rr} & u_{rn} \\ 0 \quad 0 & 0 \quad 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_r \\ x_n \end{bmatrix} = \begin{bmatrix} c_1 \\ c_r \\ 0 \\ 0 \end{bmatrix}. \tag{14.28}$$

Множество решений однородной системы

$$AX = 0, \qquad (x \neq 0) \tag{14.29}$$

или ей эквивалентной

$$Ux = \mathrm{C}, \qquad (x \neq 0) \tag{14.30}$$

позволяет определить $(n - r)$ свободных вектор-столбцов в пространстве столбцов матриц A и U, которые линейно зависимы, в отличие от r базисных, и их образ в пространстве $\mathbb{R}^m$ — нулевой вектор. Эти столбцы образуют нуль-пространство или ядро матрицы A, которое обозначается как $ker\ A$, $ker\ A \in R^n$. Размерность ядра называется деффектом матрицы: $dim\ ker\ A = n - r$.

Аналогично определяется *левое нуль-пространство матрицы A* (или нуль-пространство матрицы A^T) — множество решений однородной системы $YA = 0$, где $Y \neq 0$,

$$[y_1, y_2, \dots, y_m][A] = 0$$

(ядро оператора A^T: $ker\ A^T \in R^m$; размерность ядра: $dim\ ker\ A^T = m - r$).

С матрицей A и матрицей U связаны четыре основные векторные пространства (табл. 14.3, 14.4), рассмотрение которых позволяет понять различные случаи решения систем линейных уравнений.

Таблица 14.3

Четыре подпространства матрицы $A = A_{mn}$

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="6">Вектор-столбцы (пространство столбцов)</td></tr>
<tr><td colspan="3">базисные
(образ оператора A,
$im\ A$, $dim\ A = r$)</td><td colspan="3">свободные
(ядро оператора A, $ker\ A$,
$dim\ ker\ A = n - r$)</td></tr>
<tr><td colspan="2"></td><td>a_1</td><td>…</td><td>a_r</td><td>a_{r+1}</td><td>…</td><td>a_n</td></tr>
<tr><td rowspan="6">Пространство строк</td><td rowspan="3">образ оператора A^T, $im\ A^T$, $(dim\ im\ A^T = r)$</td><td>a_{11}</td><td>…</td><td>a_{1r}</td><td>a_{1r+1}</td><td>…</td><td>a_{1n}</td></tr>
<tr><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td></tr>
<tr><td>a_{r1}</td><td>…</td><td>a_{rr}</td><td>a_{rr+1}</td><td>…</td><td>a_{rn}</td></tr>
<tr><td rowspan="3">левое нуль-пространство, ядро A^T, $(ker\ A^T$, $dim\ ker\ A^T = m - r)$</td><td>$a_{(r+1)1}$</td><td>…</td><td>$a_{(r+1)r}$</td><td>a_{1r}</td><td>…</td><td>a_{1n}</td></tr>
<tr><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td></tr>
<tr><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td></tr>
<tr><td colspan="2"></td><td>a_{m1}</td><td>…</td><td>a_{1r}</td><td>a_{1r}</td><td>…</td><td>a_{1n}</td></tr>
</table>

Итак, четыре основные подпространства матрицы A:

1) *пространство столбцов матрицы A* (образ матрицы A: $im\ A \in R^m$; размерность $dim\ im\ A = r$),

2) *нуль-пространство матрицы A* — множество решений однородной системы

$AX = 0$, где $x \neq 0$ (ядро оператора A: $ker\ A \in R^n$; размерность ядра: $dim\ ker\ A = n - r$),

3) *пространство строк матрицы A*, имеющее размерность r и совпадающее с пространством (ненулевых) строк матрицы U (образ оператора A^T: $im\ A^T \in R^n$; размерность образа: $dim\ im\ A^T = r$),

Таблица 14.4

Четыре подпространства матрицы U, полученной из A методом Гаусса

		Вектор-столбцы (пространство столбцов)					
		базисные (образ оператора A, $im\ A$, $dim\ A = r$)			свободные (ядро оператора A, $ker\ A$, $dim\ ker\ A = n - r$)		
		u_1	…	u_r	u_{r+1}		u_n
Пространство строк	образ оператора A^T, $im\ A^T$	u_{11}	…	u_{1r}	u_{1r+1}	…	u_{1n}
		…	…	…	…	…	
		0	…	u_{rr}	$u_{r(r+1)}$	…	u_{rn}
	левое нуль-пространство (ядро A^T, $ker\ A^T$)	0	…	0	0	…	0
		…	…	…	…	…	
		…	…	…	…	…	
		0	…	0	0	…	0

4) *левое нуль-пространство матрицы* A (или нуль-пространство матрицы A^T) — множество решений однородной системы $YA = 0$, где $Y \neq 0$ (ядро оператора A^T: $ker\ A^T \in R^m$; размерность ядра: $dim\ ker\ A^T = m - r$).

Если некоторое линейное пространство V_r является подпространством R^n ($V_r \subset R^n$), то множество всех векторов, ортогональных к V_r, образует ортогональное дополнение к V_r и обозначается $V_r{}^{\perp}$ ($\dim V_r + dim V_r{}^{\perp} = n$).

Любой вектор одного пространства ортогонален любому вектору другого: если $w \in V_r$ и $v \in V_{n-r}$, то их скалярное произведение:

$$(v, w) = v^T w = 0,$$

где

$$(v, w) = v^T w = [v_1 \dots v_n] \begin{bmatrix} w_1 \\ \cdot \\ \cdot \\ \cdot \\ w_n \end{bmatrix} = v_1 w_1 + \cdots + v_n w_n = \sum v_i w_i. \tag{14.31}$$

Для четырех, описанных выше подпространств матрицы A выполняются следующие свойства:

5) $im\ A = (ker\ A^T)^{\perp}$,

6) $im\ A^T = (ker\ A)^{\perp}$,

7) $ker\ A = (im\ A^T)^{\perp}$,

8) $ker\ A^T = (im\ A)^{\perp}$.

Свойства 1)–8) представляют собой содержание основной теоремы линейной алгебры, в частности, из свойства 5) следует, что система $AX = b$ имеет решение ($b \in im\ A$ — вектор B принадлежит пространству столбцов), когда он ортогонален каждому вектору нуль-пространства матрицы A^T.

2. *МНК с точки зрения линейной алгебры*. Рассмотрим несколько задач.

1) Есть два вектора $a(a_1, a_2, \dots, a_n)$ и $b(b_1, b_2, \dots, b_n)$. Найдем проекцию вектора b на вектор a. Положим, что это вектор $\bar{x}a$, где $\bar{x}$ — скаляр. Тогда вектор $(b - \bar{x}a)$ перпендикулярен вектору a:

$$(b - \bar{x}a) \perp a.$$

Значит, их скалярное произведение равно нулю:

$$\left(a^T, (b - \bar{x}a)\right) = 0$$

или

$$a^T b - \bar{x}a^T a = 0, \qquad \overline{x} = \frac{a^T b}{a^T a}. \tag{14.32}$$

То есть *проекцию вектора b на вектор a можно задать как* $p = \frac{a^T b}{a^T a} a$ (рис. 14.1).

2) Если система $Ax = b$ несовместна, то рекомендуется минимизировать среднюю ошибку для всех уравнений $E = Ax - b$, которая равна расстоянию от конца вектора b до точки Ax, лежащей в пространстве столбцов матрицы A.

Геометрическая интерпретация утверждает, что вектор b является проекцией на пространство столбцов матрицы A ($p = A\bar{x}$) и вектор $E = p - b = Ax - b$ ортогонален этому пространству (рис. 14.2). Пространство столбцов матрицы A — это множество линейных комбинаций столбцов матрицы A с коэффициентами — координатами ненулевого вектора-строки $y(y_1, y_2, \dots, y_m)$. Имеем

$$y^T A^T (Ax - b) = 0$$

или

$$y^T (A^T Ax - A^T b) = 0.$$

Следовательно, решение системы m линейных уравнений с n неизвестными приводит к системе «*нормальных уравнений*» Гаусса, где решение в виде среднего вектора $\bar{x}$, минимизирующего ошибку, удовлетворяет соотношению

$$A^T A\bar{x} = A^T b. \tag{14.33}$$

Если столбцы матрицы А ранга n линейно независимы, нуль-пространство матрицы А содержит только нулевой вектор, то матрица $A^T A$ обратима и система имеет единственное решение:

$$\bar{x} = (A^T A)^{-1} A^T b. \tag{14.34}$$

Проекция вектора b на пространство столбцов (рис. 14.2):

$$p = A\bar{x} = A(A^T A)^{-1} A^T b. \tag{14.35}$$

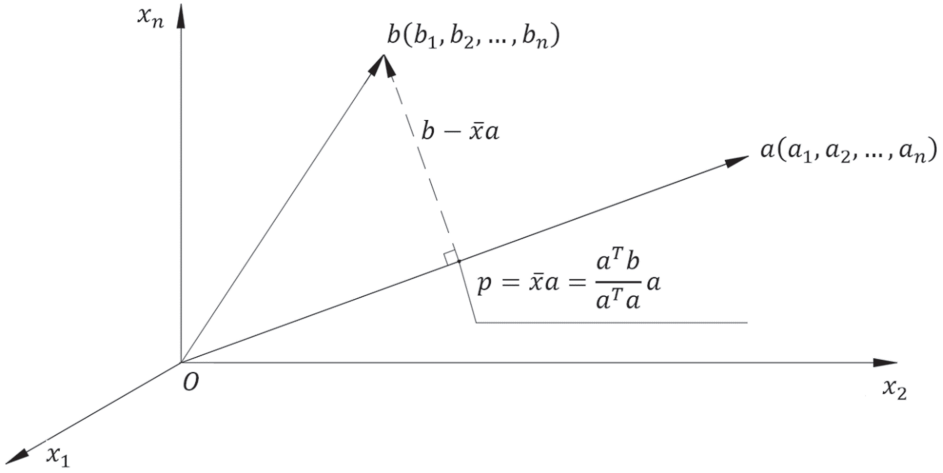


Рис. 14.1 — Проекция вектора b на вектор a [106]

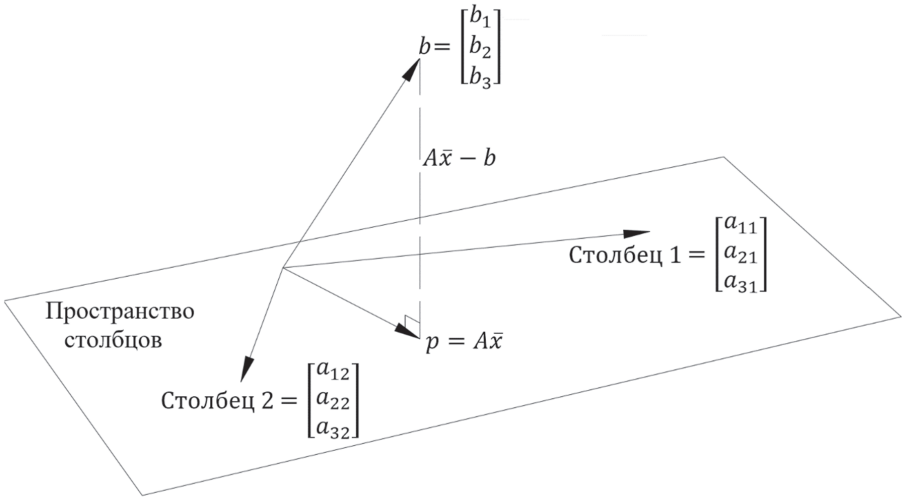


Рис. 14.2 — Проекция вектора b на пространство столбцов матрицы A_{32} [106]

Формула (14.35) представляет собой матричную запись перпендикуляра, опущенного из конца вектора b на пространство столбцов матрицы A.

Матрица

$$P = A(A^T A)^{-1} A^T \tag{14.36}$$

называется матрицей проектирования, дающая вектор $p = Pb$.

Если матрица A необратима, то вводится *понятие псевдообратной матрицы* A^+ (обратной матрицы Мура — Пенроуза). Матрица A^+ имеет размер $n \times m$, ее пространство столбцов совпадает с пространством строк матрицы A, и наоборот, пространство строк совпадает с пространством столбцов матрицы A.

$A^+ A = P$, в случае обратимости матрицы A, $A^+ = A^{-1}$.

$\bar{x} = A^+ b$ — решение несовместной системы определяется как решение с минимальной длиной проекции вектора b на пространство столбцов $p = A\bar{x}$, полученное приравниванием компоненты из нульмерного пространства нулю.

Пример 1. Решить несовместную систему:

$$\begin{cases} 2x_1 + 0x_2 + 0x_3 = 6, \\ 0x_1 + 3x_2 + 0x_3 = 15, \\ 0x_1 + 0x_2 + 0x_3 = 13. \end{cases}$$

Решение. Из первых двух уравнений находим решение $x_1 = 3, x_2 = 5$ и полагаем, что $x_3 = 0$ (как решение, имеющее наименьшую длину). Пространство столбцов матрицы A совпадает с плоскостью xy, а нуль-пространство, ортогональное пространству строк — ось z.

Проекция вектора b на пространство столбцов:

$$p = A\bar{x} = Pb = \begin{bmatrix} 6 \\ 15 \\ 0 \end{bmatrix},$$

где $b = \begin{bmatrix} 6 \\ 15 \\ 13 \end{bmatrix}$.

Имеем

$$\begin{bmatrix}2 & 0 & 0\\ 0 & 3 & 0\\ 0 & 0 & 0\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}=\begin{bmatrix}6\\ 15\\ 13\end{bmatrix},\qquad \bar{x}=\begin{bmatrix}3\\ 5\\ 0\end{bmatrix},$$

или

$$A^{+}b=\begin{bmatrix}\frac{1}{2} & 0 & 0\\ 0 & \frac{1}{3} & 0\\ 0 & 0 & 0\end{bmatrix}\begin{bmatrix}6\\ 15\\ 13\end{bmatrix}.$$

В приведенном выше равенстве первая матрица называется *диагональной*, подобные матрицы дают способ *сингулярного разложения матрицы* $A=A_{mn}$:

$$A_{mn}=Q_1\Sigma Q_2^T, \qquad (14.37)$$

где $Q_1=Q_{mm}$, $Q_2=Q_{nn}$ — ортогональные матрицы, имеющие ортонормированные столбцы, а Σ — диагональная матрица вида

$$\Sigma=\begin{pmatrix}\mu_1 & \cdots & \\ \vdots & \ddots & \vdots\\ & \cdots & \mu_r\end{pmatrix},\Sigma^{+}=\begin{pmatrix}\mu_1^{-1} & \cdots & \\ \vdots & \ddots & \vdots\\ & \cdots & \mu_r^{-1}\end{pmatrix},\ (\Sigma^{+})^{+}=\Sigma,$$

μ_i — *сингулярные числа матрицы* A, их квадраты — собственные значения симметрической матрицы A^TA, ортонормированный набор собственных векторов которой образуют столбцы матрицы Q_2. Из формулы (14.37) следует, что $Q_1^TA_{mn}Q_2=\Sigma$. Матрицы вращения Q_1 и Q_2 преобразуют пространство столбцов матрицы A_{mn} до совпадения с пространством строк и перехода в диагональную матрицу Σ. Пусть A^TA — ковариационная матрица, тогда сингулярное разложение дает перевод ее в диагональную форму (ортогональных векторов), что соответствует методу главных компонент (гл. 21 в [53]).

Сингулярное разложение приводит к формуле для псевдообратной матрицы:

$$A^{+}=Q_2\Sigma^{+}Q_1^T. \qquad (14.38)$$

3. *Вычислительная сторона МНК.* Обычно компьютерные вычисления сопровождаются ошибками, источниками которых являются исходные данные, полученные в результате измерений, вычислений, а также алгоритмы, оперирующие приближенными числами.

Стандартное представление числа с плавающей запятой (или точкой):

$$(-1)^s\times M\times B^E, \qquad (14.39)$$

где s — знак, B — основание (обычно 2 или 10), E — порядок, M — мантисса.

Например, $0{,}123\times 10^{00}+4{,}576\times 10^{-04}=0{,}123\times 10^{00}$ — точный результат $(0{,}123456)$ округляется до ближайшего числа с плавающей точкой.

Пример 2. Округляя результаты вычислений с плавающей точкой до семи значащих цифр, решить методом Гаусса следующую систему уравнений:

$$\begin{cases}1\times 10^{-7}x_1+x_2=1,\\ x_1+x_2=2.\end{cases}$$

Или в матричной форме

$$AX=B,$$

где

$$A=\begin{bmatrix}[1\times 10^{-7}] & 1\\ 1 & 1\end{bmatrix},X=\begin{bmatrix}x_1\\ x_2\end{bmatrix},B=\begin{bmatrix}1\\ 2\end{bmatrix}.$$

Решение. Заметим, что при решении системы линейных уравнений методом Гаусса выбор ведущего элемента произвольно (особенно при разном порядке точности коэффициентов) приводит к ошибкам.

I *вариант*. Выделим *разрешающий* (*главный, ведущий*) элемент матрицы A в квадратные скобки ([∗]), выполним этапы метода исключения Гаусса:

$$\begin{pmatrix}[10^{-7}] & 1 & 1\\ 1 & 1 & 2\end{pmatrix} \to \begin{pmatrix}[1] & 10^{7} & 10^{7}\\ 1 & 1 & 2\end{pmatrix} \to \begin{pmatrix}1 & 10^{7} & 10^{7}\\ 0 & 10^{7} & 10^{7}\end{pmatrix} \to \begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 1\end{pmatrix}.$$

Округление, по условию задачи, до семи значащих цифр дает решение $x_1 = = 0, x_2 = 1$, так как $1{,}0 \times 10^{7} - 2{,}0 \times 10^{0} = 1{,}0 \times 10^{7}$.

II *вариант*. Переставим строки матрицы A, тогда разрешающий элемент будет 1:

$$\begin{pmatrix}[1] & 1 & 2\\ 10^{-7} & 1 & 1\end{pmatrix} \to \begin{pmatrix}[1] & 1 & 2\\ 1 & 10^{7} & 10^{7}\end{pmatrix} \to \begin{pmatrix}[1] & 1 & 2\\ 0 & 1 & 1\end{pmatrix} \to \begin{pmatrix}1 & 0 & 1\\ 0 & 1 & 1\end{pmatrix}.$$

Получили верное решение $x_1 = 1, x_2 = 1$.

Поэтому обычно *рекомендуется выбирать в качестве разрешающего элемента наибольший по модулю элемент, преобразуемой части матрицы* (или хотя бы очередной строки или столбца). В противном случае компьютерное решение будет содержать ошибки.

Пример 3 [106]. Решить систему уравнений $AX = B$, где

$$A = \begin{bmatrix}[2] & 1 & 1\\ 4 & 1 & 0\\ -2 & 2 & 1\end{bmatrix}, X = \begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}, B = \begin{bmatrix}1\\ -2\\ 7\end{bmatrix}.$$

В результате получим новую систему вида $UX = C$:

$$U = \begin{bmatrix}2 & 1 & 1\\ 0 & [-1] & -2\\ 0 & 0 & -4\end{bmatrix}, X = \begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}, C = \begin{bmatrix}1\\ -4\\ -4\end{bmatrix}.$$

С помощью обратной подстановки легко найти решение, которое в виде вектора-строки можно записать как $X^T = (-1, 2, 1)$. *Матрица U называется верхней треугольной*.

3. *Получение среднего квадрата ошибки*. Пусть рассматривается V_n — n-мерное евклидово пространство, n нормально распределенных случайных величин X_i с одинаковой, хотя и неизвестной дисперсией. В V_n определено скалярное произведение и определена квадратичная форма

$$Q = \sum_{i=1}^{n} X_i^2.$$

С теоремой Кочрена связан факт, согласно которому для неотрицательной квадратичной формы Q существуют единственные подпространства — пространство оценок $V_r \subset V_n$ и пространство ошибок $V_{n-r} \subset V_n$ такие, что $V_n = V_r \cup V_{n-r}$.

Каждый вектор в пространстве оценок перпендикулярен каждому вектору в пространстве ошибок, то есть пространство оценок и пространство ошибок являются ортогональными дополнениями друг к другу:

$$V_r{}^{\perp} = V_{n-r}, \quad V_{n-r}{}^{\perp} = V_r, \quad V_n = V_r \cup V_{n-r}.$$

Положим, что r переменных из n линейно независимы, тогда путем соответствующих линейных преобразований квадратичную форму Q можно представить в виде

$$Q = \sum_{i=1}^{n} X_i^2 = \sum_{i=r+1}^{n} Y_i^2 . \qquad (14.40)$$

Геометрически этот факт означает выбор ядра оператора в соответствии с направлением вектора $(X_1, X_2, \dots, X_n)$, длина которого равна Q (например, в трехмерном случае ядро совпадает с диагональю параллелепипеда, построенного на векторах X_1, X_2, X_3).

То есть существует матрица линейного оператора A, преобразующего пространство V_n в V_n, причем размерность образа оператора A равна (r), а размерность ядра $(n-r)$.

Положим, что V_n — выборочное пространство, причем $M(X_i) = M(Y_i) = 0$, тогда $M(Y_i^2) = D(Y_i) = \sigma^2$, при $i > r$. То есть проекция квадратичной формы Q в V_r будет равна нулю, а в V_{n-r}

$$M(Q) = (n-r)\sigma^2.$$

Отсюда несмещенная оценка дисперсии (средний квадрат ошибки) будет равна

$$s^2 = \frac{Q}{n-r}, \qquad (14.41)$$

где n — размерность признакового пространства (V_n), r — размерность вектора оцениваемых параметров (θ_r), лежащего в V_r, s^2 — средний квадрат ошибок (обозначается как SS_e или SS_Z).

Величина s^2 при увеличении числа опытов стремится к распределению хи-квадрат Пирсона с $\nu = n - r$ степенями свободы.

$$s^2 = \frac{Q}{n-r} \to \chi^2_{n-r}. \qquad (14.42)$$

V_r — носитель квадратичной формы квадратичной формы Q (пространство оценок или, иначе, образ линейного оператора, отображающего векторное пространство V_n в V_r).

V_{n-r} — нуль − пространство квадратичной формы Q (пространство ошибок или, иначе, ядро линейного оператора, отображающего векторное пространство V_n в V_r).

Рассмотрим несколько равносильных определений.

Число степеней свободы[26] *квадратичной формы* Q равно $(n-r)$:

– числу независимых элементов информации, необходимых для образования данной суммы квадратов Q;

– размерности ядра линейного оператора — пространства V_{n-r} (или *коразмерности носителя квадратичной формы* Q-подпространства V_r);

– рангу ядра матрицы симметричной матрицы квадратичной формы от наблюдений [130].

4. Рассмотрим *геометрическую интерпретацию МНК*. Пусть вектор $y(y_1, y_2, \dots, y_n)$ принадлежит *n*-мерному пространству V_n. Независимые векторы $x_1, x_2, \dots, x_r$ порождают пространство V_r. Вектор $z \in V_r$, когда существуют коэффициенты $\beta_1, \beta_2, \dots, \beta_r$, такие, что $z = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_r x_r$. Тогда длина вектора $(y - z)$ примет наименьшее значение, если величина

$$S(y, \beta) = |y - z| = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{r} x_{ji}\beta_j\right)^2 \to min, \qquad (14.43)$$

то есть выполняются условия метода наименьших квадратов.

---

[26] Число степеней свободы механической системы есть размерность пространства ее состояний с учетом наложенных связей.

Коэффициенты b_j вектора $\hat{z} = b_1x_1 + b_2x_2 + \ldots + b_rx_r$ являются МНК-оценкой, что геометрически представлено на рисунке 14.3.

Разложение дисперсионного анализа — это фактически расщепление по теореме Пифагора:

$$O\hat{Y}^2 = OA^2 + OB^2, \qquad (14.44)$$

где вектор $O\hat{Y}$ — ортогональная проекция на вектора OA и OB, принадлежащие ортогональным подпространствам, объединение которых дает V_r.

Далее

$$(OA^2 + OB^2) + Y\hat{Y}^2 = OY^2. \qquad (14.45)$$

F-критерий для проверки гипотезы

$$\beta_1 = \cdots = \beta_r = 0, \quad (\bar{X}_j = \mu, \quad j = 1, \ldots, r),$$

против H_1: хотя бы $\beta_j \neq 0$ представляет собой сравнение сумы квадратов длин OA и OB или факторной суммы

$$OA^2 + OB^2 = SS_v = n\sum_{i=1}^{r}(X_{i.} - \bar{X}..)^2$$

против квадрата длины $Y\hat{Y}$ или остаточной суммы квадратов

$$\left(Y\hat{Y}^2 = SS_z = \sum_{i=1}^{r}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_{i.}\right)^2\right),$$

деленных на соответствующее число степеней свободы $(r-1)$ и $(n-r)$ соответственно.

5. *Геометрическая интерпретация F-критерия.*

Рисунок 14.3 иллюстрирует геометрическое представление хи-квадрат распределения Пирсона или, иначе, квадратичной формы Q с n степенями свободы в виде n-мерной сферы с центром O.

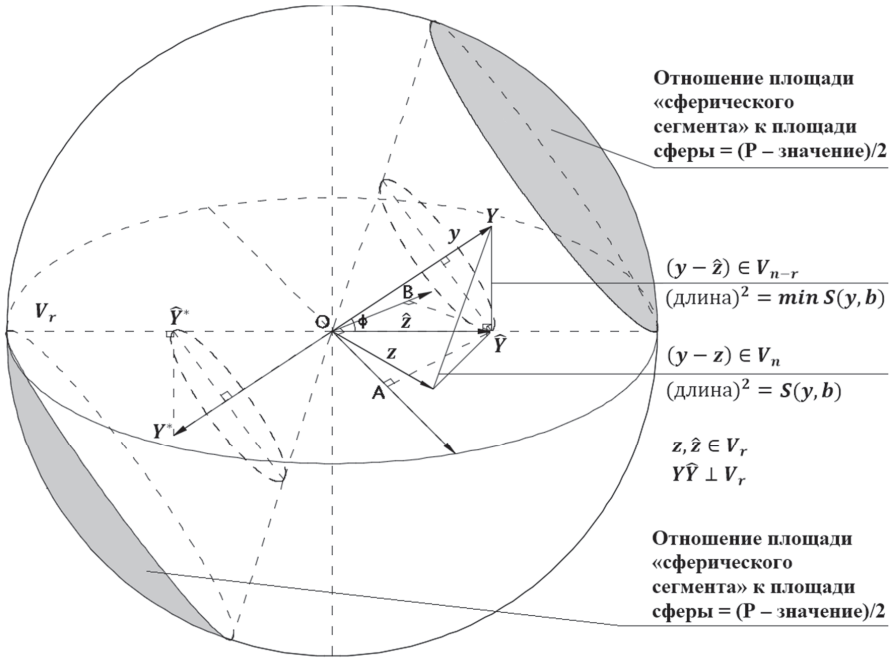


Рис. 14.3 — Геометрическая интерпретация МНК и F −критерия

Например, если радиус сферы $R = OY$ ($R = O\hat{Y}$ или любой радиус на прямой $O\hat{Y}$), то

$$\frac{\text{площадь "сферического сегмента "}}{\text{площадь сферы}} = 2\frac{2\pi R(R-h)}{4\pi R^2} = 2\frac{R-h}{2R} = 2\frac{OY-OY cos\phi}{2OY} = \\ = (1 - cos\phi). \qquad (14.46)$$

То есть

$$\frac{\text{площадь "сферического сегмента"}}{\text{площадь сферы}} = (1 - cos\phi) = P - \text{значение}. \qquad (14.47)$$

14.2. Модели однофакторного и многофакторного дисперсионного анализа

Рассмотрим несколько наиболее распространенных вариантов эксперимента, организуемого для проведения дисперсионного анализа: однофакторный, двухфакторный и трехфакторный анализ с разным числом уровней факторов и разным числом опытов на каждом уровне. Для расчетов используются преобразованные формулы сумм квадратов.

А. Однофакторный эксперимент (*один фактор* А)

Значения измеряемого признака — X_{im}.

1. *Эксперимент на двух уровнях*, $i = 1, 2$ (рис. 14.4*а*):

– без повторных опытов, $m = 1$;

– с повторными опытами, одинаковое число опытов на каждом уровне, $m = 1, 2, \ldots, n$;

– с повторными опытами, разное число опытов на каждом уровне $m = 1, 2, \ldots, n_i$.

2. *Эксперимент на нескольких уровнях*, $i = 1, 2, \ldots, a$ (рис. 14.1*б*):

– без повторных опытов, $m = 1$;

– с повторными опытами, одинаковое число опытов на каждом уровне $m = 1, 2, \ldots, n$;

– с повторными опытами, разное число опытов на каждом уровне $m = 1, 2, \ldots, n_i$.

а) б)

а) два уровня A_i, $i = 1, 2$; б) несколько уровней A_i, $i = 1, 2, \ldots, a$

Рис. 14.4 — Точки эксперимента в однофакторном анализе

Таблица 14.5 представляет исходные данные однофакторного эксперимента на двух уровнях с одинаковым числом повторных опытов. Число групп (H) равно числу уровней: A_i, $i = 1, 2$.

Расчет непосредственно по вышеприведенным формулам удобен только в случае малого числа уровней и опытов. В противном случае используются преобразованные формулы сумм квадратов. В основе вычислительных формул лежат

преобразования, рассмотренные ранее в теме 10 (вариационные ряды), которые проиллюстрируем на общей сумме квадратов однофакторного эксперимента.

Таблица 14.5

Данные для однофакторного анализа, равное число опытов

Уровень (групп) фактора	Результаты опытов: $X_{im},\ m = 1, 2, \dots, n$				
	X_{i1}	...	X_{im}	...	X_{in}
A_1	X_{11}	...	X_{1m}	...	X_{1n}
A_2	X_{21}	...	X_{2m}	...	X_{2n}

$$SS = \sum_{i=1}^{a}\sum_{m=1}^{n}(X_{im} - \bar{X}_{..})^2 = \sum_{i=1}^{a}\sum_{m=1}^{n}(X_{im}^2 - 2X_{im}\bar{X}_{..} + (\bar{X}_{..})^2) =$$
$$= \sum_{i=1}^{a}\sum_{m=1}^{n}(X_{im})^2 - \frac{\left(\sum_{i=1}^{a}\sum_{m=1}^{n}X_{im}\right)^2}{N} = C_0 - \frac{C^2}{N}. \qquad (14.48)$$

В результате аналогичных преобразований получим нижеследующие формулы.

Однофакторный эксперимент, разное число параллельных опытов; общее число опытов $N = a\sum_{i=1}^{a} n_i$.

$$SS_A = \sum_{i=1}^{a} n_i\,(\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_{i=1}^{a} n_i\,(\bar{X}_{i.})^2 - \frac{1}{N}\left(\sum_{i=1}^{a}\sum_{m=1}^{n}X_{im}\right)^2 =$$
$$= \frac{a}{N}\sum_{i=1}^{a} C_i^2 - \frac{C^2}{N}; \qquad (14.49)$$
$$SS_z = \sum_{i=1}^{a}\sum_{m=1}^{n}(X_{im} - \bar{X}_{i.})^2 = \sum_{i=1}^{a}\sum_{m=1}^{n}(X_{im})^2 -$$
$$-\sum_{i=1}^{a} n_i(\bar{X}_{i.})^2 = C_0 - \frac{1}{n}\sum_{i=1}^{a} C_i^2; \qquad (14.50)$$
$$C = \sum_{i=1}^{a}\sum_{m=1}^{n} X_{im} = \sum_{i=1}^{a} C_i;\ \ \mathrm{C}_i = \sum_{m=1}^{n} X_{im};$$
$$\mathrm{C}_0 = \sum_{i=1}^{a}\sum_{m=1}^{n} X_{im}^2. \qquad (14.51)$$

Оценки дисперсий и определение числа степеней свободы.

$s^2 = \frac{SS}{k}$ — *оценка общей дисперсии;* $k = N - 1$ — *число степеней свободы при определении общей дисперсии;*

$s_A^2 = \frac{SS_A}{k_A}$ — *оценка дисперсии по уровням фактора A*;

$k_A = \alpha - 1$ — *число степеней свободы фактора A;*

$s_z^2 = \frac{SS_z}{k_z}$ — *остаточная оценка дисперсии (дисперсия ошибки);*

$k_Z = N - a$ — *число степеней свободы при определении ошибки.*

$$k = k_A + k_z = N - 1 = (a - 1) + (N - a). \qquad (14.52)$$

Проверка нулевой гипотезы.

Расчетное или фактически наблюдаемое значение *F*-критерия:

$$F_{\text{н.}} = \frac{s_A^2}{s_Z^2}. \qquad (14.53)$$

Значение $F_{\text{кр.}}$ определяется по приложению 4 при уровне значимости α и числе степеней свободы $k_1 = k_A$ и $k_2 = k_Z$. Если $F_{\text{н.}} \le F_{\text{кр.}}$ при α, k_1, k_2, то нулевая гипотеза о равенстве средних значений по уровням фактора *A* принимается. В противном случае — отклоняется.

А.В. Двухфакторный эксперимент (*факторы* А *и* В)

Значения измеряемого признака — X_{ijm}.

1. *Эксперимент на двух уровнях,* $i = 1,2;\ j = 1,2$ (рис. 14.5):

– без повторных опытов, $m = 1$;

– с повторными опытами, одинаковое число опытов на каждом уровне $m = 1,2,\dots,n$;

– с повторными опытами, разное число опытов на каждом ij −уровне, $m_{ij} = 1,2,\dots,n_{ij}$.

2. *Эксперимент на нескольких уровнях,* $i = 1,2,\dots,a;\ j = 1,2,\dots,b$:

– без повторных опытов, $m = 1$;

– с повторными опытами, одинаковое число опытов на каждом ij −уровне, $m_{ij} = 1,2,\dots,n$;

– с повторными опытами, разное число опытов на каждом ij −уровне, $m_{ij} = 1,2,\dots,n_{ij}$.

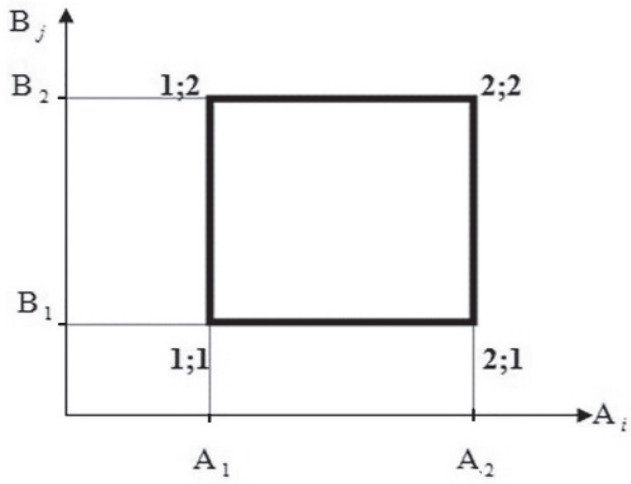


Рис. 14.5 — Точки эксперимента двухуровневого двухфакторного анализа

Приведем различные формы представления исходных данных в зависимости от вида эксперимента (табл. 14.6 и 14.7).

Число групп (H) равно числу перестановок уровней:

$ij = 1,2,\dots,H;\ H = ab$.

Двухфакторный эксперимент, равное число параллельных опытов; общее число опытов $N = abn$. Пользуясь свойствами средней арифметической, аналогично случаю однофакторного дисперсионного анализа получим следующие формулы.

$$SS = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{m=1}^{n}\left(X_{ijm} - \bar{X}_{...}\right)^2 = \sum_{i,j,m} X_{ijm}^2 - \frac{1}{N}\sum_{i,j,m}\left(X_{ijm}\right)^2. \quad (14.54)$$

Таблица 14.6

Данные для двухфакторного анализа на двух уровнях, разное число опытов

№ строки (группы)	Сочетания уровней факторов A и B	Результаты опытов: X_{ijm}; $m = 1, 2, \ldots, n_{ij}$					
		X_{ij1}	…	X_{ijm}	…	$X_{ij(n-1)}$	X_{ijn}
1	1; 1	X_{111}	…	X_{11m}	…	$X_{11(n-1)}$	X_{11n}
2	1; 2	X_{121}	…	X_{12m}	…	$X_{12(n-1)}$	X_{12n}
3	2; 1	X_{211}	…	X_{21m}	…	$X_{21(n-1)}$	X_{21n}
4	2; 2	X_{221}	…	X_{22m}	…	$X_{22(n-1)}$	X_{22n}

Таблица 14.7

Данные для двухфакторного анализа на нескольких уровнях, равное число опытов

№ строки	Сочетания уровней A и B	Наблюдаемые значения признака в группах, X_{ijm}				
		1-й опыт	…	m-опыт	…	n-опыт
1	1; 1	X_{111}	…	X_{11m}	…	X_{11n}
2	1; 2	X_{121}	…	X_{12m}	…	X_{12n}
…	…	…	…	…	…	…
ij	i; j	X_{ij1}	…	X_{ijm}	…	X_{ijn}
…	…	…	…	…	…	…
H	a; b	X_{ab1}	…	X_{abm}	…	X_{abn}

$$SS_A = bn\sum_{i=1}^{a}(\bar{X}_{i..} - \bar{X}_{...})^2 =$$

$$= \frac{a}{N}\sum_{i=1}^{a}\left(\sum_{j=1}^{b}\sum_{m=1}^{n}X_{ijm}\right)^2 - \frac{1}{N}\left(\sum_{ijm}X_{ijm}\right)^2 = \frac{a}{N}C_i^2 - \frac{C^2}{N}; \quad (14.55)$$

$$SS_B = an\sum_{j=1}^{b}\left(\bar{X}_{.j.} - \bar{X}_{...}\right)^2 =$$

$$= \frac{b}{N}\sum_{j=1}^{b}\left(\sum_{i=1}^{a}\sum_{m=1}^{n}X_{ijm}\right)^2 - \frac{1}{N}\left(\sum_{ijm}X_{ijm}\right)^2 = \frac{b}{N}C_j^2 - \frac{C^2}{N}; \quad (14.56)$$

$$SS_{AB} = n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{X}_{...} + \bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.}\right)^2 = \frac{ab}{N}\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\sum_{m=1}^{n}X_{ijm}\right)^2 -$$

$$-SS_A - SS_B - \frac{C^2}{N} = \frac{ab}{N}\sum_{i=1}^{a}\sum_{j=1}^{b}C_{ij}^2 - SS_A - SS_B - \frac{C^2}{N}; \quad (14.57)$$

$$SS_z = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{m=1}^{n}\left(X_{ijm} - \bar{X}_{ij.}\right)^2 = C - \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}C_{ij}^2, \quad (14.58)$$

где

$$\sum_{i,j,m} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{m=1}^{n}; \ \mathrm{C} = \sum_{i,j,m}X_{ijm}; \ \mathrm{C}_0 = \sum_{i,j,m}X_{ijm}^2;$$

$$\mathrm{C}_i = \sum_{j=1}^{b}\sum_{m=1}^{n}X_{ijm};$$

$$\mathrm{C}_j = \sum_{i=1}^{a}\sum_{m=1}^{n}X_{ijm}; \ \mathrm{C}_{ij} = \sum_{m=1}^{n}X_{ijm}. \quad (14.59)$$

Оценки дисперсий и определение числа степеней свободы:

$$s^2 = \frac{SS}{k}, \tag{14.60}$$

где s^2 — оценка общей дисперсии; $k = N - 1$ — число степеней свободы при определении общей дисперсии;

$$s_A^2 = \frac{SS_A}{k_A}, s_B^2 = \frac{SS_B}{k_B}, \tag{14.61}$$

где s_A^2 — оценка дисперсии по уровням фактора *A;* $k_A = a - 1$ число степеней свободы фактора *A*, s_B^2 — оценка дисперсии по уровням фактора *B;* $k_B = b - 1$ — число степеней свободы фактора *B*;

$$s_{AB}^2 = \frac{SS_{AB}}{k_{AB}}, \tag{14.62}$$

где s_{AB}^2 — оценка дисперсии по уровням факторов *A* и *B*; $k_{AB} = (a - 1)(b - 1)$ — число степеней свободы взаимодействия факторов *A* и *B;*

$$s_Z^2 = \frac{SS_Z}{k_Z}, \tag{14.63}$$

где s_Z^2 — оценка остаточной дисперсии (дисперсия ошибки); $k_z = N - ab = ab(n - 1)$ — число степеней свободы при определении ошибки.

Общее число степеней свободы:

$$k = k_A + k_B + k_{AB} + k_Z\,;\; k = N - 1 = abn - 1. \tag{14.64}$$

Проверка H_0 −гипотезы.

Определение расчетного или наблюдаемого значения *F*-критерия:

$$F_{нA} = \frac{s_A^2}{s_Z^2}; \qquad F_{нB} = \frac{s_B^2}{s_Z^2}; \quad F_{нAB} = \frac{s_{AB}^2}{s_Z^2}. \tag{14.65}$$

Критическое значение $F_{кр}$ определяется при уровне значимости α и числе степеней свободы:

− для фактора A при $k_1 = k_A$ и $k_2 = k_z$;

− для фактора B при $k_1 = k_B$ и $k_2 = k_z$;

− для взаимодействия факторов AB при $k_1 = k_{AB}$ и $k_2 = k_z$.

Если по вариантам сравнения $F_н \le F_{кр}$ при α, k_1, k_2, то гипотеза H_0 принимается. В противном случае нулевая гипотеза отклоняется.

A.B.C. Трехфакторный эксперимент (*факторы* A, B и C)

Значения измеряемого признака — X_{ijkm}.

1. *Эксперимент на двух уровнях*, $i = 1, 2;\; j = 1, 2;\; k = 1, 2$ (рис. 14.6, табл. 14.8):

– без повторных опытов, $m = 1$;

– с повторными опытами, одинаковое число опытов на каждом ijk −уровне, $m_{ijk} = 1, 2, \ldots, n$.

– с повторными опытами, разное число опытов на каждом ijk-уровне, $m_{ijk} = 1, 2, \ldots, n_{ijk}$;

2. *Эксперимент на нескольких уровнях*, $i = 1, 2, \ldots, a;\; j = 1, 2, \ldots, b$; k =1, 2, …, c:

– без повторных опытов, $m = 1$;

– с повторными опытами, одинаковое число опытов на каждом ijk − уровне, $m_{ijk} = 1, 2, \ldots, n$;

– с повторными опытами, разное число опытов на каждом ijk-уровне, $m_{ijk} = 1, 2, \ldots, n_{ijk}$.

Таблица 14.8

Данные для трехфакторного анализа на двух уровнях, разное число опытов

№ строки (группы)	Сочетания уровней A B C	Результаты опытов: X_{ijkn}; $m = 1, 2, \ldots, n_{ijk}$					
		X_{ijk1}	…	X_{ijkm}	…	$X_{ijk(n-1)}$	X_{ijkn}
1	1; 1; 1	X_{1111}	…	X_{111m}	…	$X_{111(n-1)}$	X_{111n}
2	1; 2; 1	X_{1211}	…	X_{121m}	…	$X_{121(n-1)}$	X_{121n}
3	2; 1; 1	X_{2111}	…	X_{211m}	…	$X_{211(n-1)}$	X_{211n}
4	2; 2; 1	X_{2211}	…	X_{221m}	…	$X_{221(n-1)}$	X_{221n}
5	1; 1; 2	X_{1121}	…	X_{112m}	…	$X_{112(n-1)}$	X_{112n}
6	1; 2; 2	X_{1221}	…	X_{122m}	…	$X_{122(n-1)}$	X_{122n}
7	2; 1; 2	X_{2121}	…	X_{212m}	…	$X_{212(n-1)}$	X_{212n}
8	2; 2; 2	X_{2221}	…	X_{222m}	…	$X_{222(n-1)}$	X_{222n}

Число групп (H) равно числу перестановок уровней: $ijk = 1, 2, \ldots, H;\ H = 8$.

Трехфакторный эксперимент, равное число параллельных опытов, общее число опытов $N = abcn$.

$$SS = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n}\left(X_{ijkm} - \bar{X}_{....}\right)^2 = C_0 - \frac{C^2}{N}; \qquad (14.66)$$

$$SS_A = bcn\sum_{i=1}^{a}(\bar{X}_{i...} - \bar{X}_{....})^2 = \frac{a}{N}\sum_{i=1}^{a} C_i^2 - \frac{C^2}{N}; \qquad (14.67)$$

$$SS_B = acn\sum_{i=1}^{b}\left(\bar{X}_{.j..} - \bar{X}_{....}\right)^2 = \frac{b}{N}\sum_{j=1}^{b} C_j^2 - \frac{C^2}{N}; \qquad (14.68)$$

$$SS_C = abn\sum_{i=1}^{b}(\bar{X}_{..k.} - \bar{X}_{....})^2 = \frac{c}{N}\sum_{j=1}^{c} C_k^2 - \frac{C^2}{N}; \qquad (14.69)$$

$$SS_{AB} = cn\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{X}_{....} + \bar{X}_{ij..} - \bar{X}_{i...} - \bar{X}_{.j..}\right)^2 =$$

$$= \frac{ab}{N}\sum_{i=1}^{a}\sum_{j=1}^{b} C_{ij}^2 - SS_A - SS_B - \frac{C^2}{N}; \qquad (14.70)$$

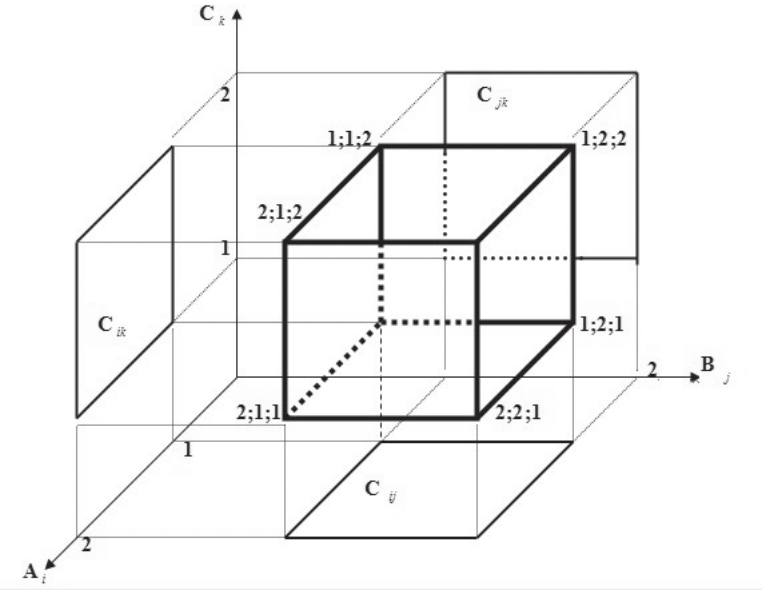


Рис. 14.6 — Точки эксперимента двухуровневого трехфакторного анализа

$$SS_{AC} = bn\sum_{i=1}^{a}\sum_{k=1}^{c}(\bar{X}_{....} + \bar{X}_{i.k.} - \bar{X}_{i...} - \bar{X}_{..k.})^2 =$$
$$= \frac{ac}{N}\sum_{i=1}^{a}\sum_{k=1}^{c} C_{ik}^2 - SS_A - SS_C - \frac{C^2}{N}; \quad (14.71)$$
$$SS_{BC} = an\sum_{j=1}^{b}\sum_{k=1}^{c}\left(\bar{X}_{....} + \bar{X}_{.jk.} - \bar{X}_{.j..} - \bar{X}_{..k.}\right)^2 =$$
$$= \frac{bc}{N}\sum_{j=1}^{b}\sum_{k=1}^{c} C_{jk}^2 - SS_B - SS_C - \frac{C^2}{N}; \quad (14.72)$$
$$SS_{ABC} = n\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(\bar{X}_{ijk.} + \bar{X}_{i...} - \bar{X}_{.j..} - \bar{X}_{..k.} - \bar{X}_{ij..} - \bar{X}_{i.k.} - \bar{X}_{.jk.} - \bar{X}_{....}\right)^2 =$$
$$= \frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} C_{ijk}^2 - SS_A - SS_B - SS_C - SS_{AB} - SS_{AC} - SS_{BC} - \frac{C^2}{N}; \quad (14.73)$$
$$SS_z = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n}\left(X_{ijkm} - \bar{X}_{ijk.}\right)^2 = C_0 - \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} C_{ijk}^2, \quad (14.74)$$

где

$$C = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n} X_{ijkm},$$
$$C_0 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n} X_{ijkm}^2, \quad C_i = \sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n} X_{ijkm},$$
$$C_j = \sum_{i=1}^{a}\sum_{k=1}^{c}\sum_{m=1}^{n} X_{ijkm}, \quad C_k = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{m=1}^{n} X_{ijkm},$$
$$C_{ij} = \sum_{k=1}^{c}\sum_{m=1}^{n} X_{ijkm},$$
$$C_{ik} = \sum_{j=1}^{b}\sum_{m=1}^{n} X_{ijkm}, \quad C_{jk} = \sum_{i=1}^{a}\sum_{m=1}^{n} X_{ijkm},$$
$$C_{ijk} = \sum_{m=1}^{n} X_{ijkm}. \quad (14.75)$$

Оценки дисперсий и определение числа степеней свободы.

$s^2 = \frac{SS}{k}$ *— оценка общей дисперсии* $k = N - 1$ *— число степеней свободы при определении общей дисперсии;*

$s_A^2 = \frac{SS_A}{k_A}$; $s_B^2 = \frac{SS_B}{k_B}$; $s_C^2 = \frac{SS_C}{k_C}$ *— оценки дисперсий по уровням соответственно факторов A, B и C;* $k_A = a - 1$; $k_B = b - 1$; $k_C = c - 1$ *— число степеней свободы факторов A, B и C соответственно;*

$s_{AB}^2 = \frac{SS_{AB}}{k_{AB}}$; $s_{AC}^2 = \frac{SS_{AC}}{k_{AC}}$; $s_{BC}^2 = \frac{SS_{BC}}{k_{BC}}$; $s_{ABC}^2 = \frac{SS_{ABC}}{k_{ABC}}$ *— оценки дисперсий по взаимосвязям уровней факторов;*

$k_{AB} = (a-1)(b-1)$; $k_{AC} = (a-1)(c-1)$; $k_{BC} = (b-1)(c-1)$;

$k_{ABC} = (a-1)(b-1)(c-1)$ *— числа степеней свободы взаимодействий факторов;*

$s_Z^2 = \frac{SS_Z}{k_Z}$ *— остаточная оценка дисперсии (дисперсия ошибки);*

$k_z = N - abc = abc(n-1)$ *— число степеней свободы при определении ошибки.*

$$k = k_A + k_B + k_C + k_{AB} + k_{AC} + k_{BC} + k_{ABC} + k_Z.$$

Проверка H_0 *−гипотезы.*

Определение расчетных значений критерия:

$$F_{нA} = \frac{s_A^2}{s_Z^2}; F_{нB} = \frac{s_B^2}{s_Z^2}; F_{нC} = \frac{s_C^2}{s_Z^2}; F_{нAB} = \frac{s_{AB}^2}{s_Z^2}; F_{нAC} = \frac{s_{AC}^2}{s_Z^2};$$
$$F_{нBC} = \frac{s_{BC}^2}{s_Z^2}; F_{нABC} = \frac{s_{ABC}^2}{s_Z^2}. \quad (14.76)$$

Критическое значение $F_{кр}$ определяется по приложению 4 при уровне значимости α и числе степеней свободы k_1 и $k_2 = k_Z$.

Если $F_н \leq F_{кр}$ при α, k_1, k_2, то гипотеза H_0 принимается, а различия между средними по всем вариантам опыта статистически не значимы. В противном случае нулевая гипотеза отклоняется, т. е. имеется хотя бы одна существенная разность между средними (устанавливается существенное влияние факторов и их взаимодействий на изменчивость признака) и исследуется значимость средних признака на отдельных уровнях.

Результаты вычислений принято представлять в виде таблицы дисперсионного анализа, например схема двухфакторного дисперсионного анализа (табл. 14.9).

Алгоритм расчетов

1. Построение вспомогательной таблицы.
2. Вычисление коэффициентов (вспомогательных сумм), например для трехфакторного эксперимента: $C, C_0, C_i, C_j, C_k, C_{ij}, C_{jk}, C_{ik}, C_{ijk}$.
3. Вычисление сумм квадратов.
4. Вычисление оценок дисперсий.
5. Проверка гипотезы H_0.
6. Если H_0 не отклоняется, то необходима проверка значимости уровней факторов.

Таблица 14.9

Схема двухфакторного дисперсионного анализа

Источник изменчивости	Сумма квадратов эффектов,	k	Оценка дисперсии	$F_н$	η·100, %
Общая	$SS = \sum_{i,j,m} \left(X_{ijm} - \bar{X}_{...}\right)^2$	$k = N - 1$	$s^2 = \frac{SS}{k}$		
Фактор А	$SS_A = bn \sum_{i=1}^{a} (\bar{X}_{i..} - \bar{X}_{...})^2$	$k_A = a - 1$	$s_A^2 = \frac{SS_A}{k_A}$	$F_A = \frac{s_A^2}{s_Z^2}$	$\eta_A = \frac{SS_A}{SS}$
Фактор В	$SS_B = an \sum_{j=1}^{b} \left(\bar{X}_{.j.} - \bar{X}_{...}\right)^2$	$k_B = b - 1$	$s_B^2 = \frac{SS_B}{k_B}$	$F_B = \frac{s_B^2}{s_Z^2}$	$\eta_B = \frac{SS_B}{SS}$
Взаимодействие факторов АВ	$SS_{AB} = n \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{X}_{...} + \bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.})^2$	$k_{AB} = (a-1) \times (b-1)$	$s_{AB}^2 = \frac{SS_{AB}}{k_{AB}}$	$F_{AB} = \frac{s_{AB}^2}{s_Z^2}$	$\eta_{AB} = \frac{SS_{AB}}{SS}$
Остаточная (внутри групп)	$SS_Z = \sum_{i,j,m} \left(X_{ijm} - \bar{X}_{ij.}\right)^2$	$k_Z = ab \times (n-1)$	$s_Z^2 = \frac{SS_Z}{k_Z}$		$\eta_Z = \frac{SS_Z}{SS}$

14.3. Примеры применения дисперсионного анализа

Пример 14.1. *Однофакторный дисперсионный анализ, равное число наблюдений*. Изучались четыре сорта озимого ячменя. Каждый сорт высевался на 5 участках одинаковой площади, при случайном размещении сортов и повторений. Получены данные об урожайности озимого ячменя в расчете на 1 га посевной площади (табл. 14.10).

Таблица 14.10

Урожайность озимого ячменя, ц с 1 га

Сорт	Повторения					Сумма	Средняя урожайность, ц/га
	1	2	3	4	5	C_i	
Павел	57,4	55,2	56,6	53,9	56,9	280,0	56,0
Сармат	55,3	57,1	50,2	52,1	52,7	267,4	53,5
Секрет	51,6	55,7	53,2	56,7	54,6	271,8	54,4
Хуторок	68,7	57,8	59,3	64,7	58,3	308,8	61,8
Сумма	233,0	225,8	219,3	227,4	222,5	1128,0	56,4

Проверяемая гипотеза H_0: отсутствие влияния фактора A — сортов на урожайность озимого ячменя.

Решение. 1) Построение вспомогательной таблицы.

Построим вспомогательную таблицу (табл. 14.11) для промежуточных вычислений сумм квадратов.

Таблица 14.11

Вспомогательные вычисления при однофакторном дисперсионном анализе

i	C_i^2	X_{im}^2				
		1	2	3	4	5
1	78400,00	3294,76	3047,04	3203,56	2905,21	3237,61
2	71502,76	3058,09	3260,41	2520,04	2714,41	2777,29
3	73875,24	2662,56	3102,49	2830,24	3214,89	2981,16
4	95357,44	4719,69	3340,84	3516,49	4186,09	3398,89
Сумма	319135,44	13735,1	12750,78	12070,33	13020,6	12394,95

2) Вычисление вспомогательных сумм.

$$C^2 = \left(\sum_{i=1}^{a}\sum_{m=1}^{n} X_{ij}\right)^2 = \left(\sum_{i=1}^{a} C_i\right)^2 = 1128^2 = 1272384;$$

$$C_0 = \sum_{i=1}^{a}\sum_{m=1}^{n} X_{ij}^2 =$$

$$= 13735{,}1 + 1270{,}78 + 120750{,}33 + 13020{,}6 + 12394{,}95 =$$

$$= 63971{,}76;$$

$$\sum_{i=1}^{a} C_i^2 = 280^2 + 267{,}4^2 + 271{,}8^2 + 308{,}8^2 = 319135{,}44.$$

3) Вычисление сумм квадратов.

Общая сумма квадратов:

$$SS = \sum_{i=1}^{a}\sum_{m=1}^{n}\left(X_{ij} - \bar{X}\right)^2 = \sum_{i=1}^{a}\sum_{m=1}^{n} X_{ij}^2 - \frac{1}{N}\left(\sum_{i=1}^{a}\sum_{m=1}^{n} X_{ij}\right)^2 =$$

$$= C_0 - \frac{C^2}{N} = 63971{,}76 - \frac{1272384}{20} = 352{,}56.$$

Сумма квадратов между группами:

$$SS_A = \sum_{i=1}^{a} n_i(\bar{X}_i - \bar{X})^2 = n\sum_{i=1}^{a}(\bar{X}_i - \bar{X})^2 = \sum_{i=1}^{a} n_i\bar{X}_i^2 - \frac{1}{N}\left(\sum_{i=1}^{a}\sum_{m=1}^{n} X_{ij}\right)^2 =$$

$$= \frac{a}{N}\sum_{i=1}^{a} C_i^2 - \frac{C^2}{N} = \frac{4}{20}\cdot 319135{,}44 - \frac{1272384}{20} = 207{,}888.$$

Сумма квадратов внутри групп:

$$SS_Z = \sum_{i=1}^{a}\sum_{m=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2 = \sum_{i=1}^{a}\sum_{m=1}^{n} X_{ij}^2 - \sum_{i=1}^{a} n_i\bar{X}_i^2 = C_0 - \frac{1}{n}\sum_{i=1}^{a} C_i^2 =$$

$$= 63971{,}76 - \frac{1}{5}319135{,}44 = 144{,}672\ .$$

Проверка: $SS_Z = SS - SS_a = 352{,}56 - 207{,}888 = 144{,}672.$

4) Вычисление оценок дисперсий.

$$s_A^2 = \frac{SS_A}{k_A} = \frac{207{,}888}{3} = 69{,}296;$$

$$s_Z^2 = \frac{SS_Z}{k_Z} = \frac{144{,}672}{16} = 9{,}042;$$

$$k_A = a - 1 = 4 - 1 = 3; \quad k_Z = N - a = 20 - 4 = 16.$$

5) Проверка гипотезы:

$$F_{A\text{ н.}} = \frac{s_A^2}{s_Z^2} = \frac{69{,}296}{9{,}042} = 7{,}66;\ F_{\text{кр.}} = 3{,}24,$$

при $\alpha = 0{,}05; k_1 = 3; k_2 = 16.$

Результаты расчетов обычно заносятся в таблицу (табл. 14.12).

Таблица 14.12

Дисперсионный анализ различий в урожайности

Источник изменчивости	Сумма квадратов отклонений	Степени свободы	Средний квадрат отклонений	*F*	
				наблюдаемое	критическое
Различия между уровнями (сортов)	207,888	3	69,296	7,66	3,24
Различия внутри уровней (остаточная)	144,672	16	9,042		
Общая	352,56	19			

Наблюдаемое значение критерия сравнивается с критическим. Так как $F_{\text{н.}} > F_{\text{кр.}}$, то нулевая гипотеза об отсутствии влияния фактора сорта на урожайность озимого ячменя отвергается. Значит, имеется хотя бы одна статистически значимая разность в средней урожайности между сортами. Оценку частных различий между средними урожайностями проведем с помощью расчета наименьшей существенной разности.

Средняя ошибка опыта:

$$s_{\bar{X}} = \sqrt{\frac{s_Z^2}{n}} = \sqrt{\frac{9{,}042}{5}} = 1{,}345.$$

Ошибка разности средних значений:

$$s_d = s_{\bar{X}}\sqrt{2} = 1{,}345 \cdot 1{,}414 = 1{,}902.$$

Наименьшая существенная разность:

$$\text{НСР} = t_{0,05;16}s_d = 2{,}12 \cdot 1{,}902 = 4{,}032 \approx 4{,}0.$$

Сравнивая средние урожайности по сортам между собой, видно, что сорт озимого ячменя «Хуторок» существенно превосходит все другие по уровню средней урожайности. Различия в средней урожайности сортов «Сармат», «Павел» и «Секрет» не значимы.

Пример 14.2. *Однофакторный дисперсионный анализ, неравное число наблюдений по уровням.* Срок службы электрических ламп [81].

Для изготовления каждой партии ламп была взята проволока разных заводов изготовителей. Все же прочие условия производства были одинаковы для каждой партии. Требуется выяснить, отличаются ли партии ламп между собой по сроку службы. Если ответ будет положительным, то можно думать, что качество проволоки варьирует реально, и, следовательно, для достижения стандартизации производства электрических ламп необходимо достигнуть большей однородности проволоки во всех партиях и пересмотреть контракты с поставщиками. Данные наблюдений представлены в таблице 14.13.

Таблица 14.13

Срок службы электрических ламп

Партия ламп (группа)	Результаты наблюдений: X_{im} — срок службы, тыс. час, $m = 1,2,\dots,n_i$							
	1	2	3	4	5	6	7	8
№ 1	1,60	1,61	1,65	1,68	1,70	1,72	1,80	–
№ 2	1,58	1,64	1,64	1,70	1,75	–	–	–
№ 3	1,46	1,55	1,60	1,62	1,64	1,66	1,74	1,82
№ 4	1,51	1,52	1,53	1,60	1,67	1,68	–	–

Проверяемая гипотеза H_0: «партии ламп не отличаются между собой по сроку службы».

Решение. 1) Построим вспомогательную таблицу (табл. 14.14) для вычисления вспомогательных сумм, округляя результаты до трех цифр.

Пояснения к вычислениям в таблице 14.14:

$$C_i = \sum_{m=1}^{n_i} X_{im};\ \ \bar{X} = \frac{\sum_{i=1}^{a} C_i}{N};\ \ \bar{X} = \frac{42{,}67}{26} = 1{,}641.$$

$$\bar{X}_i = \frac{C_i}{n_i};\ \ i = 1;\ \ \bar{X}_{i=1} = \frac{11{,}76}{7} = 1{,}68;\ \ \bar{X}_2 = 1{,}662 \text{ и т. д.}$$

$$x_{im} = X_{im} - \bar{X}\text{:}\quad i = 1;\ \ x_{i=1\,m=1} = x_{11} = 1{,}6 - 1{,}641 = -0{,}041;$$

$$x_{12} = 1{,}61 - 1{,}641 = -0{,}031 \text{ и т. д.}$$

$$\bar{x}_i = \frac{\sum_{m_i}^{n_i} x_{im}}{n_i};\quad i = 1;\quad \bar{x}_1 = \frac{0{,}273}{7} = 0{,}039;\quad \bar{x}_2 = 0{,}021 \text{ и т. д.}$$

Таблица 14.14

Вычисления средних значений и сумм квадратов

Партия ламп (группа)	n_i	C_i	$\bar{X}_i$	$x_{im} = X_{im} - \bar{X}$	$\sum_{m=1}^{n_i} x_{im}$	$\bar{x}_i$	$\sum_{m=1}^{n_i} x_{im}^2$	$n_i \bar{x}_i^2$
№ 1	7	11,76	1,680	–0,041; –0,031; 0,009; 0,039; 0,059; 0,079; 0,159	0,273	0,039	0,0392	0,0106
№ 2	5	8,31	1,662	–0,061; –0,001; –0,001; 0,059; 0,109	0,105	0,021	0,0191	0,0022
№ 3	8	13,09	1,636	–0,181; –0,091; –0,041; –0,021; –0,001; 0,019; 0,099; 0,179	–0,038	–0,005	0,0854	0,0002
№ 4	6	9,51	1,585	–0,131; –0,121; –0,111; –0,041; 0,029; 0,039	–0,336	–0,056	0,0482	0,0188
∑	26	42,67	1,641		–	–	0,1919	0,0318

2) Вычисления сумм квадратов.

Учитывая, что $\bar{x} = 0$, найдем суммы квадратов:

$$SS = \sum_{i=1}^{a}\sum_{m=1}^{n_i}(X_{im} - \bar{X})^2 = \sum_{i=1}^{a}\sum_{m=1}^{n_i}(x_{im})^2 - n\bar{x}^2 =$$

$$= 0{,}0392 + 0{,}0191 + 0{,}0854 + 0{,}0482 = 0{,}1919,$$

$$SS_A = \sum_{i=1}^{a} n_i\,(X_{im} - \bar{X})^2 = \sum_{i=1}^{a} n_i\,\bar{x}_i^2 - n\bar{x}^2 = 0{,}0318;$$

$$SS_Z = \sum_{i=1}^{a}\sum_{m=1}^{n_i}(X_{im} - X)^2 = SS - SS_A = 0{,}1919 - 0{,}0318 = 0{,}1601.$$

3) Оценки дисперсий.

$$s^2 = 0{,}1919/(26-1) = 0{,}0077;\ s_A^2 = 0{,}0318/(4-1) = 0{,}0106;$$

$$s_Z^2 = 0{,}1601/(26-4) = 0{,}0073.$$

4) Проверка гипотезы.

$$F_{\text{н.}} = 0{,}0106/0{,}0073 = 1{,}45;$$

при $k_1 = 3, k_2 = 22$ и $\alpha = 0{,}05$:

$$F_{\text{кр.}} = 3{,}05.$$

Так как $F_{\text{н.}} < F_{\text{кр.}}$, то нулевая гипотеза, что партии ламп не отличаются между собой по сроку службы, — принимается.

Пример 14.3. *Двухфакторный дисперсионный анализ.* В двухфакторном опыте, поставленном методом рандомизированных повторений, изучалось влияние систем удобрений и способов обработки почвы на урожайность озимой пшеницы. Выделено три уровня применения удобрений, три способа обработки почвы в четырех специально организованных повторениях.

Решение. Результаты опыта представлены в таблице 14.15.

Таблица 14.15

Урожайность озимой пшеницы при различных сочетаниях систем удобрений (А) и способов обработки почвы (В), ц/га

Фактор *A* (удобрения)	Фактор *B* (способ обработки)	Повторения, x_{ijm}				Сумма (V_{ij})	Средняя, $\overline{x}_{ij}$
		I	II	III	IV		
Без удобрений	Отвальная	38,4	38,1	37,5	38,7	152,7	38,2
	Мелкая	37,9	37,2	36,8	38,4	150,3	37,6
	Комбинированная	36,5	36,9	37,3	38,1	148,8	37,2
Минеральные удобрения	Отвальная	51,4	50,1	54,3	55,8	211,6	52,9
	Мелкая	49,6	51,4	50,8	53,4	205,2	51,3
	Комбинированная	52,0	49,9	55,9	57,4	215,2	53,8
Органические и минеральные удобрения	Отвальная	61,8	54,6	60,7	66,5	243,6	60,9
	Мелкая	56,8	55,7	62,1	64,3	238,9	59,7
	Комбинированная	63,4	57,8	66,2	70,8	258,2	64,6
Сумма (x_{ij})		447,8	431,7	461,6	483,4	1824,5	
Средняя ($\bar{x}_m$)		49,8	46,8	53,5	53,7		50,7

1) Определяются частные средние, средние значения по градациям факторов *A* и *B*, по повторениям, а также по опыту в целом, используя формулы:

$$\overline{x}_{ij} = \frac{\sum_{m=1}^{n} x_{ijm}}{n}; \overline{x}_i = \frac{\sum_{j=1}^{b}\sum_{m=1}^{n} x_{ijm}}{bn}; \overline{x}_j = \frac{\sum_{i=1}^{a}\sum_{m=1}^{n} x_{ijm}}{an};$$

$$V_{ij} = \sum_{m=1}^{n} x_{ijm};\ \overline{x}_{ij} = \frac{V_{ij}}{m};\ \overline{x}_m = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} x_{ijm}}{ab};$$

$$\overline{x} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{m=1}^{n} x_{ijm}}{abn};\ i = 1,2,\dots,a;\ j = 1,2,\dots,b;$$

$$m = 1,2,\dots,n;\ a = 3;\ b = 3, n = 4;$$

$$N = abn = 3 \cdot 3 \cdot 4 = 36.$$

Средние по градациям фактора *A*:

$$\overline{x}_1 = \frac{152{,}7+150{,}3+148{,}8}{12} = \frac{38{,}2+37{,}6+37{,}2}{3} = 37{,}7;$$

$$\overline{x}_2 = 52{,}7;\ \overline{x}_3 = 61{,}7.$$

Средние по градациям фактора *B*:

$$\overline{x}_1 = \frac{152{,}7+211{,}6+243{,}6}{12} = 50{,}6;\ \overline{x}_2 = 49{,}5;\ \overline{x}_3 = 51{,}9.$$

2) Находятся суммы квадратов отклонений (табл. 14.16):

$$C = \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{m=1}^{n} x_{ijm}\right)^2}{N} = \frac{1824.5^2}{36} = 92466{,}67.$$

Общая сумма квадратов отклонений:

$$SS = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{m=1}^{n} x_{ijm}^2 - C = 96360{,}17 - 92466{,}67 = 3893{,}5.$$

Сумма квадратов отклонений повторений:

$$SS_p = \sum_{m=1}^{n}\left(\sum_{i=1}^{a}\sum_{j=1}^{b} x_{ijm}\right)^2 : (ab) - C =$$

$$= (447{,}8^2 + 431{,}7^2 + 461{,}6^2 + 483{,}4^2) : (3 \cdot 3) - 92466{,}67 =$$

$$= 92626{,}65 - 92466{,}67 = 159{,}98.$$

Таблица 14.16

Вспомогательная таблица для сумм квадратов отклонений

Фактор *A* (удобрения)	Фактор *B* (способ обработки)	Повторения, x_{ijm}^2				Сумма x_{ij}^2	Сумма V_{ij}^2
		I	II	III	IV		
1	1	1474,56	1451,61	1406,25	1497,69	5830,11	23317,29
	2	1436,41	1383,84	1354,24	1474,56	5649,05	22590,09
	3	1332,25	1361,61	1391,29	1451,61	5536,76	22141,44
2	1	2641,96	2510,01	2948,49	3113,64	11214,1	44774,56
	2	2460,16	2641,96	2580,64	2851,56	10534,32	42107,04
	3	2704,0	2490,01	3124,81	3294,76	11613,58	46311,04
3	1	3819,24	2981,16	3684,49	4422,25	14907,14	59340,96
	2	3226,24	3102,49	3856,41	4134,49	14319,63	57073,21
	3	4019,56	3340,84	4382,44	5012,64	16755,48	66667,24
Сумма		23114,38	21263,53	24729,06	27253,2	96360,17	384322,87

Факторная сумма квадратов отклонений:

$$SS_V = \sum_{i=1}^{a}\sum_{j=1}^{b} V_{ij}^2 : n - C = 384322{,}87 : 4 - 92466{,}67 = 3614{,}05.$$

Остаточная сумма квадратов отклонений:

$$SS_Z = SS - SS_V - SS_P = 3893{,}5 - 3614{,}05 - 159{,}98 = 119{,}47.$$

3) Определяется сумма квадратов отклонений эффектов факторов *A* и *B* (сумма квадратов между группами) и взаимодействия *AB*, предварительно составляется вспомогательная таблица 14.17.

Таблица 14.17

Определение главных эффектов факторов и взаимодействий

Удобрения, *A*	Способ обработки, *B*			Сумма, *A*
	1	2	3	
1	152,7	150,3	148,8	451,8
2	211,6	205,2	215,2	632,0
3	243,6	238,9	258,2	740,7
Сумма, *B*	607,9	594,4	622,2	1824,5

$$SS_A = \frac{\sum_{i=1}^{a}\left(\sum_{j=1}^{b}\overline{x}_{ij}\right)^2}{bn} - C = \sum_{i=1}^{a} A_i^2 : bn - C =$$

$$= (451{,}8^2 + 632^2 + 740{,}7^2) : (3 \cdot 4) - 92466{,}67 = 3548{,}64;$$

$$SS_B = \frac{\sum_{j=1}^{b}\left(\sum_{i=1}^{a}\overline{x}_{ij}\right)^2}{an} - C = \sum_{j=1}^{b} B_j^2 : an - C =$$

$$= (607{,}9^2 + 594{,}4^2 + 622{,}2^2) : (3 \cdot 4) - 92466{,}67 = 32{,}21;$$

$$SS_{AB} = SS_V - SS_A - SS_B = 3614{,}05 - 3548{,}64 - 32{,}21 = 33{,}2.$$

Проверка:

$$SS_Z = SS - SS_A - SS_B - SS_{AB} - SS_P =$$
$$= 3893{,}5 - 3548{,}64 - 32{,}21 - 33{,}2 - 159{,}98 = 119{,}47.$$

4) Находится число степеней свободы:

– общее $k_o = N - 1 = 36 - 1 = 35$;

– повторений $k_p = n - 1 = 4 - 1 = 3$;

– фактора A $k_A = a - 1 = 3 - 1 = 2$;

– фактора B $k_B = b - 1 = 3 - 1 = 2$;

– взаимодействия AB $k_{AB} = (a-1)(b-1) = (3-1)(3-1) = 4$;

– остаточное $k_z = k_o - k_A - k_B - k_{AB} - k_P = 35 - 2 - 2 - 4 - 3 = 24$.

5) Оценки дисперсий:

$$s^2 = \frac{SS}{k_o} = \frac{3893{,}5}{35} = 111{,}29;$$

$$s_A^2 = \frac{SS_A}{k_A} = \frac{3548{,}64}{2} = 1774{,}32;\ s_B^2 = \frac{SS_B}{k_B} = \frac{32.21}{2} = 16{,}11;$$

$$s_{AB}^2 = \frac{SS_{AB}}{k_{AB}} = \frac{33{,}2}{4} = 8{,}05;;\ s_Z^2 = \frac{SS_Z}{k_Z} = \frac{119.47}{24} \approx 4{,}97.$$

6) Проверка гипотез:

$$F_{A\text{н.}} = \frac{s_A^2}{s_Z^2} = \frac{1774{,}32}{4{,}97} = 347{,}01;$$

при $\alpha = 0{,}05$, $k_1 = 2$, $k_2 = 24$;

табличное значение $F_{\text{кр}} = 3{,}40$.

Так как $F_{A\text{н.}} = 357{,}01 > F_{\text{кр.}} = 3{,}40$, то нулевая гипотеза о равенстве средних по вариантам фактора *A* отклоняется.

$F_{\text{Вн.}} = \frac{s_B^2}{s_Z^2} = \frac{16{,}11}{4{,}97} = 3{,}24$; при $\alpha = 0{,}05$, $k_1 = 2$, $k_2 = 24$ табличное значение $F_{\text{кр.}} = 3{,}40$. Так как $F_{\text{Вн.}} = 3{,}24 < F_{\text{кр.}} = 3{,}40$, то нулевая гипотеза о равенстве средних по вариантам фактора *B* принимается.

$F_{AB\text{н.}} = \frac{s_{AB}^2}{s_Z^2} = \frac{8{,}05}{4{,}97} = 1{,}62$; при $= 0{,}05$, $k_1 = 4$, $k_2 = 24$ табличное значение $F_{\text{кр.}} = 2{,}78$. Так как $F_{AB\text{н.}} = 1{,}62 < F_{\text{кр.}} = 2{,}78$, то нулевая гипотеза о равенстве средних по уровням взаимодействия факторов принимается.

Таким образом, дисперсионный анализ выявил существенное влияние на результативный признак только фактора *A*.

Результаты расчетов можно представить в таблице дисперсионного анализа 14.18.

Из таблицы дисперсионного анализа (табл. 14.18) по величине η видно также, что доля влияния фактора *A* составляет 91,1%, доля фактора *B* — 0,8%, взаимодействия факторов — 0,8%, не учтенных в опыте факторов — 3,2%.

После того как установлено существенное влияние факторов на изменчивость результативного признака, проводится оценка значимости различий между средними на отдельных уровнях изучаемых факторов и их взаимодействий.

7) Оценка значимости частных различий между средними:

$$s_{\overline{x}} = \sqrt{\frac{s_Z^2}{n}} = \sqrt{\frac{4{,}97}{4}} = 1{,}115;\ s_d = \sqrt{2}s_{\overline{x}} = 1{,}414 \cdot 1{,}115 = 1{,}58;$$

при $\alpha = 0{,}05$ и $k_z = 24$ по таблице значений *t*-критерия Стьюдента

$$t_{\text{кр.}} = 2{,}06\ ;\ HCP = t_{\text{кр.}} \cdot s_d = 2{,}05 \cdot 1{,}58 = 3{,}25.$$

Таблица 14.18

Результаты двухфакторного дисперсионного анализа

Источник изменчивости	Сумма квадратов отклонений	Степени свободы	Средний квадрат отклонений	$F_{\text{набл}}$	η·100, %
Общая	3893,5	35	111,29		
Повторений	159,98	3			4,1
Фактор *A*	3548,64	2	1774,32	357,01	91,1
Фактор *B*	32,21	2	16,11	3,24	0,8
Взаимодействие факторов *AB*	33,2	4	8,05	1,62	0,8
Остаточная	119,47	24	4,97		3,2

Если разность между средними урожайностями озимой пшеницы, приведенными в таблице 4.13, больше НСР, то она является значимой или существенной, если же меньше НСР, то не значимой.

8) Оценка значимости главных эффектов факторов *A* и *B* и их взаимодействий:

для фактора *A*

$$\text{НСР}_{\alpha,k_Z} = t_{\alpha,k_Z}\sqrt{\frac{2s_Z^2}{nb}};\ \text{НСР}_{05,24} = 2{,}06\sqrt{\frac{2\cdot 4{,}97}{4\cdot 3}} = 1{,}87;$$

для фактора *B*

$$\text{НСР}_{\alpha,k_Z} = t_{\alpha,k_Z}\sqrt{\frac{2s_Z^2}{na}};\ \text{НСР}_{05,24} = 2{,}06\sqrt{\frac{2\cdot 4{,}97}{4\cdot 3}} = 1{,}87.$$

По фактору *A* (системы удобрений) разности между средними урожайностями больше НСР, то есть применение минеральных удобрений, органических и минеральных удобрений, по сравнению с вариантом без применения удобрений, дает существенную прибавку средней урожайности

$$(52{,}7 - 37{,}7 = 15{,}0;\ 61{,}7 — 37{,}7 = 24{,}0;\ 61{,}7 — 52{,}7 = 9{,}0).$$

По фактору *B* и взаимодействию *AB* все разности между средними меньше НСР, и они являются не значимыми $(51{,}9 - 49{,}5 = 1{,}4)$.

Если опыт поставлен способом случайного размещения повторений, то дисперсионный анализ проводится по схеме, приведенной в таблице 14.9.

Пример 14.4. Трехфакторный эксперимент на двух уровнях с равным числом повторных опытов, $n = 3$ (табл. 14.19).

Менеджер фирмы, занимающийся торговлей промышленными товарами, решил провести анализ причин, влияющих на выручку сети магазинов, принадлежащих фирме. В первую очередь он решил выяснить, зависит ли средняя по магазинам еженедельная выручка (результативный признак X_{ijkm}) от уровня образования персонала (фактор *A*), от поставщиков (фактор *B*), от дня недели (фактор *C*). Каждый из факторов варьировался первоначально только на двух уровнях: для *A* — среднее (уровень 1) и высшее профессиональное образование (уровень 2), для фактора *B* — фирма поставщик № 1 (уровень 1) и фирма поставщик № 2 (уровень 2), для *C* — начало недели (уровень 1) и конец недели (уровень 2). Анализ проводился по данным трех недель наблюдений: $n = 3$ (табл. 14.19).

Таблица 14.19

Данные о выручке сети промтоварных магазинов
(часть промежуточных вычислений)

№	*A B C*	X_{ijkm} у. е.			C_{ijk}	C_{ijk}^2	C_{ij}	C_i	X_{ijkm}^2			$\sum_{m=1}^{n} X_{ijkm}^2$
1	1 1 1	1	3	5	9	81	18	31	1	9	25	35
2	1 1 2	2	4	3	9	81			4	16	9	29
3	1 2 1	2	2	1	5	25	13		4	4	1	9
4	1 2 2	1	5	2	8	64			1	25	4	30
5	2 1 1	5	3	4	12	144	19	41	25	9	16	50
6	2 1 2	4	2	1	7	49			16	4	1	21
7	2 2 1	3	1	4	8	64	22		9	1	16	26
8	2 2 2	5	4	5	14	196			25	16	25	66
∑		23	24	25	*C*=72	704	72	72	85	84	97	$C_0 = 266$

Проверяемая гипотеза H_0: ни уровень образования работников торговли, ни поставщики, ни начало и конец недели не влияют на выручку магазинов.

Решение. 1) Для облегчения вычислений изобразим пространство трехфакторного эксперимента (рис. 14.7).

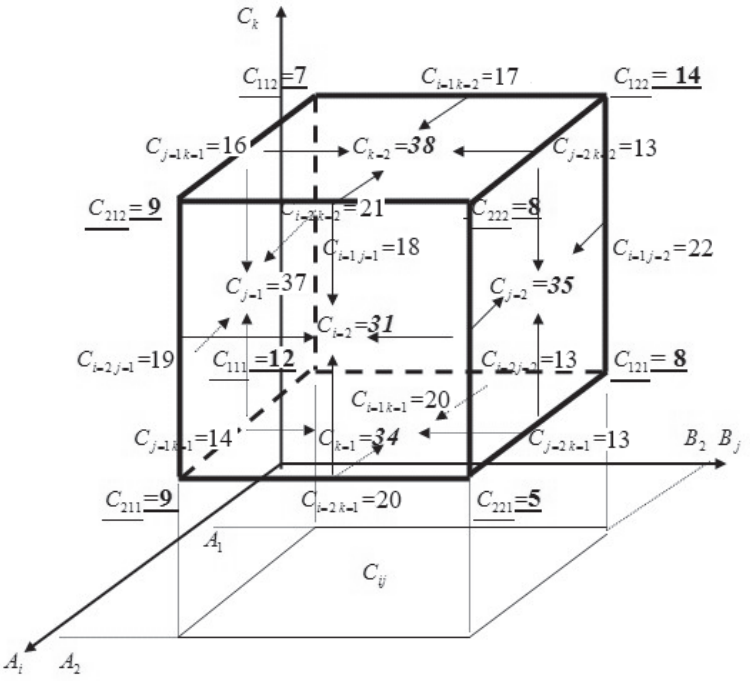


Рис. 14.7 — Вычисления в пространстве
трехфакторного дисперсионного анализа

2) Вычисления:

$$C = 72;\ \ \frac{C^2}{N} = \frac{5184}{24} = 216;\ \ C_0 = 266.$$

Вычисления C_{ij}:

$$C_{i=1\ j=1} = 9 + 9 = 18; \quad C_{i=1\ j=2} = 5 + 8 = 13;$$
$$C_{i=2\ j=1} = 12 + 7 = 19; \quad C_{i=2\ j=2} = 8 + 14 = 22.$$

Вычисления C_{ik}:

$$C_{i=1\ k=1} = 9 + 5 = 14; \ C_{i=1\ k=2} = 9 + 8 = 17;$$
$$C_{i=2\ k=1} = 12 + 8 = 20; \quad C_{i=2\ k=2} = 7 + 14 = 21.$$

Вычисления C_{jk}:

$$C_{j=1\ k=1} = 9 + 12 = 21; \quad C_{j=1\ k=2} = 9 + 7 = 16;$$
$$C_{j=2\ k=1} = 5 + 8 = 13; \quad C_{j=2\ k=2} = 8 + 14 = 22.$$

Вычисления C_i:

$$C_{i=1} = 18 + 13 = 31; \ C_{i=2} = 19 + 22 = 41.$$

Вычисления C_j:

$$C_{j=1} = 18 + 19 = 37; \ C_{j=2} = 13 + 22 = 35.$$

Вычисления C_k:

$$C_{k=1} = 20 + 14 = 34; \ C_{k=2} = 21 + 17 = 38.$$

3) Вычисления сумм квадратов.

$$SS = C_0 - \frac{C^2}{N} = 266 - \frac{5184}{24} = 266 - 216 = 50;$$

$$SS_A = \frac{a}{N}\sum_{i=1}^{a} C_i^2 - \frac{C^2}{N} = \frac{2}{24}(31^2 + 41^2) - 216 = 220{,}17 - 216 = 4{,}17;$$

$$SS_B = \frac{b}{N}\sum_{j=1}^{b} C_j^2 - \frac{C^2}{N} = \frac{2}{24}(37^2 + 35^2) - 216 = 216{,}17 - 216 = 0{,}17;$$

$$SS_C = \frac{c}{N}\sum_{k=1}^{c} C_k^2 - \frac{C^2}{N} = \frac{2}{24}(34^2 + 38^2) - 216 = 216{,}67 - 216 = 0{,}67;$$

$$SS_{AB} = \frac{ab}{N}\sum_{i=1}^{a}\sum_{j=1}^{b} C_{ij}^2 - SS_A - SS_B - \frac{C^2}{N} =$$

$$= \frac{4}{24}(18^2 + 13^2 + 19^2 + 22^2) - 4{,}17 - 0{,}17 - 216 =$$

$$= 223 - 220{,}33 = 2{,}66;$$

$$SS_{AC} = \frac{ac}{N}\sum_{i=1}^{a}\sum_{k=1}^{c} C_{ik}^2 - SS_A - SS_C - \frac{C^2}{N} =$$

$$= \frac{4}{24}(20^2 + 14^2 + 21^2 + 17^2) - 4{,}17 - 0{,}67 - 216 = 221 - 220{,}83 = 0{,}16;$$

$$SS_{BC} = \frac{bc}{N}\sum_{j=1}^{b}\sum_{k=1}^{c} C_{jk}^2 - SS_B - SS_C - \frac{C^2}{N} =$$

$$= \frac{4}{24}(21^2 + 13^2 + 16^2 + 22^2) - 0{,}17 - 0{,}67 - 16 =$$

$$= 225 - 216{,}84 = 8{,}16;$$

$$SS_{ABC} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} C_{ijk}^2 - SS_A - SS_B - SS_{AB} - SS_{AC} - SS_{BC} -$$
$$- \frac{C^2}{N} = \frac{1}{3}704 - 15{,}99 - 216 = 234{,}67 - 231{,}99 = 2{,}68;$$

$$SS_Z = C_0 - \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} C_{ijk}^2 = 266 - 234{,}67 = 31{,}33.$$

Проверка:

$$SS_Z = SS - SS_A - SS_B - SS_C - SS_{AB} - SS_{AC} - SS_{BC} - SS_{ABC} =$$
$$= 50 - 18{,}67 = 31{,}33.$$

4) Оценка дисперсий.

$$S^2 = \frac{SS}{k} = \frac{50}{23} = 2{,}17; \quad k = N - 1 = 24 - 1 = 23;$$

$$S_A^2 = \frac{SS_A}{k_A} = 4{,}17; \quad k_A = a - 1 = 2 - 1 = 1;$$

$$S_B^2 = \frac{SS_B}{k_B} = 0{,}17; \quad k_B = b - 1 = 2 - 1 = 1;$$

$$S_C^2 = \frac{SS_C}{k_C} = 0{,}67; \quad k_C = c - 1 = 2 - 1 = 1;$$

$$S_{AB}^2 = \frac{SS_{AB}}{k_{AB}} = 2{,}66; \quad k_{AB} = (a-1)(b-1) = 1;$$

$$S_{AC}^2 = \frac{SS_{AC}}{k_{AC}} = 0{,}16; \quad k_{AC} = (a-1)(c-1) = 1;$$

$$S_{BC}^2 = \frac{SS_{BC}}{k_{BC}} = 8{,}16; \quad k_{BC} = (b-1)(c-1) = 1;$$

$$S_{ABC}^2 = \frac{SS_{ABC}}{k_{ABC}} = 2{,}68; \quad k_{ABC} = (a-1)(b-1)(c-1) = 1.$$

$$S_Z^2 = \frac{SS_Z}{k_Z} = \frac{31{,}33}{16} = 1{,}96; \quad k_Z = N - abc = abc(n-1) = 24 - 8 = 16.$$

5) Проверка H_0-гипотезы.

Определение расчетных значений критерия:

$$F_{a\text{ н.}} = \frac{S_a^2}{S_Z^2} = \frac{4{,}17}{1{,}96} = 2{,}13; \; F_{b\text{ н.}} = \frac{S_b^2}{S_Z^2} = \frac{0{,}17}{1{,}96} = 0{,}09;$$

$$F_{c\text{ н.}} = \frac{S_c^2}{S_Z^2} = \frac{0{,}67}{1{,}96} = 0{,}34; \; F_{ab\text{ н.}} = \frac{S_{ab}^2}{S_Z^2} = \frac{2{,}66}{1{,}96} = 1{,}36;$$

$$F_{ac\text{ н.}} = \frac{S_{ac}^2}{S_Z^2} = \frac{0{,}16}{1{,}96} = 0{,}08; \; F_{bc\text{ н.}} = \frac{S_b^2}{S_Z^2} = \frac{0{,}16}{1{,}96} = 4{,}17;$$

$$F_{abc\text{ н.}} = \frac{S_{abc}^2}{S_Z^2} = \frac{2{,}86}{1{,}96} = 1{,}37.$$

Результаты расчетов заносятся таблицу дисперсионного анализа (табл. 14.20).

Таблица 14.20

Результаты трехфакторного дисперсионного анализа

Источник изменчивости	Сумма квадратов отклонений	Степени свободы	Оценка дисперсии	$F_{\text{н.}}$	$\eta \cdot 100{,}\%$
Общая	$SS = 50$	23	$S^2 = 2{,}17$		100
Фактор *A*	$SS_a = 4{,}17$	1	$S_a^2 = 4{,}17$	2,13	8,34
Фактор *B*	$SS_b = 0{,}17$	1	$S_b^2 = 0{,}17$	0,09	0,34
Фактор *C*	$SS_c = 0{,}67$	1	$S_c^2 = 0{,}67$	0,34	1,34
Взаимодействие *AB*	$SS_{ab} = 2{,}66$	1	2,66	1,36	5,32
Взаимодействие *AC*	$SS_{ac} = 0{,}16$	1	0,16	0,08	0,32
Взаимодействие *BC*	$SS_{bc} = 8{,}16$	1	8,16	4,17	16,32
Взаимодействие *ABC*	$SS_{abc} = 2{,}68$	1	2,68	1,37	5,36
Факторная	18,67	7	2,67	1,36	37,34
Остаточная	$SS_Z = 31{,}33$	$k_Z = 16$	$S_Z^2 = 1{,}96$		62,66

Критическое значение $F_{кр.}$ определено при уровне значимости $\alpha = 0{,}05$, числе степеней свободы $k_1 = 1$ и $k_2 = 16$; $F_{кр.} = 4{,}49$.

Так как $F_{н.} < F_{кр.}$ при α, k_1, k_2, то нулевая гипотеза принимается по всем источникам изменчивости. Нет необходимости проводить исследование значимости средних значений признака на отдельных уровнях. Таким образом, ни один из факторов: уровень образования торговых работников (A), поставщики товара (B), день недели (C) — не влияли на выручку исследуемых промтоварных магазинов.

Пример 14.5. Трехфакторный дисперсионный анализ с разным числом уровней факторов и равным числом повторений опытов на каждом уровне (исходные данные и расчеты в табл. 14.21).

На урожайность озимой пшеницы изучалось влияние трех факторов:

A — плодородие почвы, по которому было выделено два уровня: A_1 — низкое плодородие с содержанием гумуса до 3,2%; A_2 — высокое плодородие с содержанием гумуса свыше 3,2%;

B — система удобрений в севообороте, по которому изучалось четыре уровня: B_1 — без удобрений; B_2 — минимальная норма удобрений (95 кг д. в./га); B_3 — средняя норма (190 кг д. в./га); B_4 — высокая норма (380 кг. д. в./га);

C — система защиты растений от сорняков, вредителей и болезней на трех уровнях: C_1 — без применения средств защиты; C_2 — защита растений от сорняков; C_3 — защита растений от вредителей, болезней и сорняков.

Опыт поставлен на трех полях равной площади, со случайным размещением делянок по вариантам опыта. Выдвигается нулевая гипотеза — на урожайность озимой пшеницы не оказывает значимого влияния ни один из вышеперечисленных факторов.

Решение. Модель трехфакторного опыта имеет вид

$$x_{ijkm} = \overline{x} + A_i + B_j + C_k + AB_{ij} + AC_{ik} + BC_{jk} + ABC_{ijk} + P_m + \varepsilon_{ijkm};$$

где $\overline{x}$ — средняя урожайность по опыту;

A_i — эффект фактора А, плодородия почв;

B_j — эффект фактора В, систем удобрений;

C_k — эффект фактора С, систем защиты растений;

$AB_{ij}, AC_{ik}, BC_{jk}, ABC_{ijk}$ — эффекты взаимодействия факторов;

P_m — эффект повторений;

ε_{ijkm} — остаток, характеризует влияние прочих факторов.

$$i = 1, 2, \dots, a, a = 2;\ j = 1, 2, \dots, b, b = 4;\ k = 1, 2, \dots, c, c = 3;$$
$$m = 1, 2, \dots, n, n = 3.$$

Всего наблюдений $N = abcn = 2 \cdot 4 \cdot 3 \cdot 3 = 72$.

1) Для характеристики влияния факторов и их взаимодействий сначала определяются средние по всем градациям факторов и их взаимодействий:

Таблица 14.21

Урожайность озимой пшеницы по вариантам опыта, ц с 1 га

№ п/п	A	*B*	*C*	Повторения, *p*			Сумма, $\sum_{m=1}^{n} X_{ijkm}$	Средняя, $\overline{x}_{ijk}$	Сумма, $\sum_{k=1}^{c}\sum_{m=1}^{n} x_{ij}$	Средняя, $\overline{x}_{ij}$
				1	2	3				
1	1	1	1	37,6	35,5	33,7	106,8	35,60	324,4	36,04
2	1	1	2	35,5	36,6	33,5	105,6	35,20		
3	1	1	3	38,2	36,2	37,6	112,0	37,33		
4	1	2	1	49,9	50,9	52,7	153,5	51,17	517,8	57,53
5	1	2	2	56,6	55,8	52,2	164,6	54,87		
6	1	2	3	63,2	68,2	68,3	199,7	66,57		
7	1	3	1	66,7	63,2	64,4	194,3	64,77	614,1	68,23
8	1	3	2	65,0	67,0	66,8	198,8	66,27		
9	1	3	3	75,8	71,2	74,0	221,0	73,67		
10	1	4	1	65,3	66,9	69,5	201,7	67,23	639,1	71,01
11	1	4	2	69,5	67,8	71,4	208,7	69,57		
12	1	4	3	77,4	76,6	74,7	228,7	76,23		
13	2	1	1	42,1	38,3	42,0	122,4	40,80	380,7	42,3
14	2	1	2	43,3	40,5	39,4	123,2	41,07		
15	2	1	3	47,4	44,5	43,2	135,1	45,03		
16	2	2	1	64,8	63,2	65,7	193,7	64,57	623,3	69,26
17	2	2	2	70,7	67,2	64,0	201,9	67,30		
18	2	2	3	74,3	76,8	76,6	227,7	75,90		
19	2	3	1	71,2	65,3	69,4	205,9	68,63	634,2	70,47
20	2	3	2	69,8	65,4	68,4	203,6	67,87		
21	2	3	3	76,9	73,2	74,6	224,7	74,90		
22	2	4	1	65,3	61,4	64,3	191,0	63,67	615,2	68,36
23	2	4	2	67,1	66,2	65,9	199,2	66,40		
24	2	4	3	76,9	75,4	72,7	225,0	75,00		
$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} X_{ijkm}$				1470,5	1433,3	1445,0	4348,8	60,40	4348,8	
Средняя, $\overline{X}_m$				61,27	59,72	60,21				

a) средние значения по всем строкам в таблице 14.21:

$$\overline{x}_{ijk} = \frac{\sum_{m=1}^{n} X_{ijkm}}{n},\ \overline{x}_{111} = \frac{37{,}6+35{,}5+33{,}7}{3} = \frac{106{,}8}{3} = 35{,}6 \text{ и т. д.;}$$

b) средние значения по уровням фактора *A*:

$$\overline{x}_i = \frac{\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n} X_{ijkm}}{bcn},\ \overline{x}_1 = \frac{106{,}8+105{,}6+\cdots+208{,}7+228{,}7}{4\cdot 3\cdot 3} = \frac{2095{,}4}{36} = 58{,}20;$$

$$\overline{x}_2 = \frac{122{,}4+123{,}2+\cdots+199{,}2+225{,}0}{4\cdot 3\cdot 3} = \frac{2253{,}4}{36} = 62{,}594;$$

c) средние значения по уровням фактора *B*:

$$\overline{x}_j = \frac{\sum_{i=1}^{a}\sum_{k=1}^{c}\sum_{m=1}^{n} X_{ijkm}}{acn},$$

$$\overline{x}_1 = \frac{324{,}4 + 380{,}7}{2\cdot 3\cdot 3} = \frac{705{,}1}{18} = 39{,}17;\ \overline{x}_2 = \frac{517{,}8 + 623{,}3}{2\cdot 3\cdot 3} = \frac{1141{,}1}{18} = 63{,}39;$$

$$\overline{x}_3 = \frac{614{,}1+634{,}2}{2\cdot 3\cdot 3} = \frac{1248{,}3}{18} = 69{,}35;\ \overline{x}_4 = \frac{639{,}1+615{,}2}{2\cdot 3\cdot 3} = \frac{1254{,}3}{18} = 69{,}68;$$

d) средние значения по уровням фактора *C*:

$$\overline{x}_k = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{m=1}^{n} x_{ijkm}}{abn};$$

$$\overline{x}_1 = \frac{106{,}8+153{,}5+194{,}3+201{,}7+122{,}4+193{,}7+205{,}9+191{,}0}{2\cdot 4\cdot 3} = \frac{1369{,}3}{24} = 57{,}05;$$

$$\overline{x}_2 = \frac{105{,}6+164{,}6+198{,}8+208{,}7+123{,}2+201{,}9+203{,}6+199{,}2}{2\cdot 4\cdot 3} = \frac{1405{,}6}{24} = 58{,}57;$$

$$\overline{x}_3 = \frac{112+199{,}7+221+228{,}7+135{,}1+227{,}7+224{,}7+225}{2\cdot 4\cdot 3} = \frac{1573{,}9}{24} = 65{,}58;$$

e) средние по взаимодействиям факторов *A* и *B* представлены в последнем столбце исходной таблицы:

$$\overline{x}_{ij} = \frac{\sum_{k=1}^{c}\sum_{m=1}^{n} x_{ijkm}}{cn} = \frac{c_{ij}}{cn};\ \overline{x}_{11} = 36{,}04;\ \overline{x}_{12} = 57{,}53 \text{ и т. д.};$$

f) средние по взаимодействиям факторов *A* и *C*:

$$\overline{x}_{ik} = \frac{\sum_{j=1}^{b}\sum_{m=1}^{n} x_{ijkm}}{bn} = \frac{c_{ik}}{bn};$$

$$\overline{x}_{11} = \frac{106{,}8+153{,}5+194{,}3+201{,}7}{4\cdot 3} = \frac{656{,}3}{12} = 54{,}69;$$

$$\overline{x}_{12} = \frac{105{,}6+164{,}6+198{,}8+208{,}7}{4\cdot 3} = \frac{677{,}7}{12} = 56{,}48;$$

$$\overline{x}_{13} = \frac{112+199{,}7+221+228{,}7}{4\cdot 3} = \frac{761{,}4}{12} = 63{,}45;$$

$$\overline{x}_{21} = \frac{122{,}4+193{,}7+205{,}9+191}{4\cdot 3} = \frac{712{,}4}{12} = 59{,}37;$$

$$\overline{x}_{22} = \frac{123{,}2+201{,}9+203{,}6+199{,}2}{4\cdot 3} = \frac{727{,}9}{12} = 60{,}66;$$

$$\overline{x}_{23} = \frac{135{,}1+227{,}7+224{,}7+225}{4\cdot 3} = \frac{812{,}5}{12} = 67{,}71;$$

g) средние по взаимодействиям факторов *B* и *C*:

$$\overline{x}_{jk} = \frac{\sum_{i=1}^{a}\sum_{m=1}^{n} x_{ijkm}}{an} = \frac{c_{jk}}{an};$$

$$\overline{x}_{11} = \frac{106{,}8+122{,}4}{2\cdot 3} = \frac{229{,}2}{6} = 38{,}2;$$

$$\overline{x}_{12} = \frac{105{,}6+123{,}2}{2\cdot 3} = \frac{228{,}8}{6} = 38{,}13;$$

$$\overline{x}_{13} = \frac{112+135{,}1}{2\cdot 3} = \frac{247{,}1}{6} = 41{,}18;$$

$$\overline{x}_{21} = \frac{153{,}5+193{,}7}{2\cdot 3} = \frac{347{,}2}{6} = 57{,}87;$$

$$\overline{x}_{22} = \frac{164{,}6+201{,}9}{2\cdot 3} = \frac{366{,}5}{6} = 61{,}08;$$

$$\overline{x}_{23} = \frac{199{,}7+227{,}7}{2\cdot 3} = \frac{427{,}4}{6} = 71{,}23;$$

$$\overline{x}_{31} = \frac{194{,}3+205{,}9}{2\cdot 3} = \frac{400{,}2}{6} = 66{,}7;$$

$$\overline{x}_{32} = \frac{198{,}8+203{,}6}{2\cdot 3} = \frac{402{,}4}{6} = 67{,}07;$$

$$\overline{x}_{33} = \frac{221+224{,}7}{2\cdot 3} = \frac{445{,}7}{6} = 74{,}28;$$

$$\overline{x}_{41} = \frac{201{,}7+191}{2\cdot 3} = \frac{392{,}7}{6} = 65{,}45;$$

$$\overline{x}_{42} = \frac{208{,}7+199{,}2}{2\cdot 3} = \frac{407{,}9}{6} = 67{,}98;$$

$$\overline{x}_{43} = \frac{228{,}7+225}{2\cdot 3} = \frac{453{,}7}{6} = 75{,}62.$$

2) Число степеней свободы составит:

общее $k_o = N - 1 = 72 - 1 = 71$;

фактора A $k_A = a - 1 = 2 - 1 = 1$;

фактора B $k_B = b - 1 = 4 - 1 = 3$;

фактора C $k_C = c - 1 = 3 - 1 = 2$;

взаимодействия факторов AB $k_{AB} = (a-1)(b-1) = 1 \cdot 3 = 3$;

взаимодействия факторов AC $k_{AC} = (a-1)(c-1) = 1 \cdot 2 = 2$;

взаимодействия факторов BC $k_{BC} = (b-1)(c-1) = 3 \cdot 2 = 6$;

взаимодействия факторов ABC $k_{ABC} = (a-1)(b-1)(c-1) = 1 \cdot 3 \cdot 2 = 6$;

повторений $k_P = m - 1 = 3 - 1 = 2$;

остаточное $k_Z = k_o - k_A - k_B - k_C - k_{AB} - k_{AC} - k_{BC} - k_{ABC} - k_P =$
$= 71 - 1 - 3 - 2 - 3 - 2 - 6 - 6 - 2 = 46.$

3) Расчет сумм квадратов отклонений:

a) общая:

$$SS_o = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n}(X_{ijkm} - \overleftarrow{X}_o)^2 =$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n}X_{ijkm}^2 - \frac{1}{N}\left(\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n}X_{ijkm}\right)^2 =$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n}X_{ijkm}^2 - \frac{C^2}{N};$$

$$C = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n}X_{ijkm} = 4348{,}8;$$

$$SS_o = 37{,}6^2 + 35{,}5^2 + 33{,}7^2 + 35{,}5^2 + \cdots +$$

$$+ 76{,}9^2 + 75{,}4^2 + 72{,}7^2 - \frac{4348{,}8^2}{72} = 276182{,}3 - 262667{,}52 =$$

$$= 13514{,}78;$$

b) факторная, рассматривая опыт как однофакторный, тогда

$$SS_o = SS_V + SS_P + SS_Z;$$

$$SS_V = n\left(\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(\overline{X}_{ijk} - \overline{X}_o)^2\right) =$$

$$= 3((35{,}60 - 60{,}4)^2 + (35{,}20 - 60{,}4)^2 + \cdots +$$

$$+ (75{,}00 - 60{,}4)^2 = 3 \cdot 4439{,}871 = 13319{,}613;$$

c) повторений:

$$SS_P = abc\sum_{m=1}^{n}(\overline{X}_m - \overline{X}_o)^2 = 2 \cdot 4 \cdot 3((61{,}27 - 60{,}4)^2 +$$

$$+ (59{,}72 - 60{,}4)^2 + (60{,}21 - 60{,}4)^2) = 30{,}1296;$$

d) остаточная:

$$SS_Z = SS_o - SS_V - SS_P = 13514{,}78 - 13319{,}613 - 30{,}1296 =$$

$$= 165{,}0374.$$

Проверку нулевой гипотезы о незначимости влияния изучаемых факторов проведем с помощью *F*-критерия Фишера (табл. 14.22).

Таблица 14.22

Дисперсионный анализ различий в урожайности

Источник изменчивости	Сумма квадратов отклонений	Степени свободы	Средний квадрат отклонений	*F*	
				наблюдаемое	критическое
Факторный	13319,613	23	579,114	161,40	1,77
Повторений	30,1296	2	15,065		
Остаточный	165,0374	46	3,588		
Общее	13514,78	71			

При уровне значимости α=0,05 и числе степеней свободы $k_1 = 23$ и $k_2 = 46, F_{\text{кр.}} = 1{,}77$. Так как наблюдаемое значение критерия больше критического, то нулевая гипотеза отвергается, изучаемые факторы в целом оказывают значимое влияние на урожайность озимой пшеницы.

Если наблюдаемое значение критерия меньше критического, то гипотеза об отсутствии влияния факторов на урожайность принимается, дальнейшая обработка данных не проводится. В этом случае отсутствуют статистически значимые различия между средними по изучаемым факторам и их взаимодействиям. В примере влияние факторов оказалось значимым, в этом случае необходимо провести дисперсионный анализ влияния отдельных факторов и их взаимодействий на урожайность озимой пшеницы.

Расчет сумм квадратов отклонений по факторам и их взаимодействиям:

фактора *A*:

$$SS_A = bcn\sum_{i=1}^{a}\left(\overline{X}_i - \overline{X}_o\right)^2 =$$

$$= 4 \cdot 3 \cdot 3((58{,}20 - 60{,}4)^2 + (62{,}59 - 60{,}4)^2) = 36(4{,}84 + 4{,}7961) =$$

$$= 346{,}8996;$$

фактора *B*:

$$SS_B = \frac{b}{N}\sum_{j=1}^{b}\left(\sum_{i=1}^{a}\sum_{k=1}^{c}\sum_{m=1}^{n} X_{ijkm}\right)^2 - \frac{C^2}{N} = \frac{b}{N}\sum_{j=1}^{b} C_j^2 - \frac{C^2}{N};$$

$$C_1 = 324{,}4 + 380{,}7 = 705{,}1;$$

$$C_2 = 517{,}8 + 623{,}3 = 1141{,}1;$$

$$C_3 = 614{,}1 + 634{,}2 = 1248{,}3;$$

$$C_4 = 639{,}1 + 615{,}2 = 1254{,}3;$$

$$SS_B = \frac{4}{72}(705{,}1^2 + 1141{,}1^2 + 1248{,}3^2 + 1254{,}3^2) - \frac{4348{,}8^2}{72} =$$

$$= 273933{,}2502 - 262667{,}52 = 11265{,}6244$$

или

$$SS_B = acm\ \sum_{j=1}^{b}\left(\overline{X}_j - \overline{X}_o\right)^2,$$

$$SS_B = 2 \cdot 3 \cdot 3((39{,}17 - 60{,}4)^2 + (63{,}39 - 60{,}4)^2 + (69{,}35 - 60{,}4)^2 +$$

$$+(69{,}68 - 60{,}4)^2) = 11265{,}7302$$

(небольшие расхождения в результатах расчетов объясняются округлением при определении средних значений);

фактора *C*:

$$SS_C = abm\sum_{k=1}^{c}(\overline{X}_k - \overline{X}_o)^2,$$

$$SS_C = 2 \cdot 4 \cdot 3((57{,}05 - 60{,}4)^2 + (58{,}57 - 60{,}4)^2 + (65{,}58 - 60{,}4)^2) =$$

$$= 993{,}6912;$$

взаимодействия факторов *AB*:

$$SS_{AB} = \frac{ab}{N}\sum_{i=1}^{a}\sum_{j=1}^{b} C_{ij}^2 - SS_A - SS_B - \frac{C^2}{N};\ C_{ij}^2 = \sum_{m=1}^{n} X_{ijm}^2;$$

$$C_{ij}^2 = 324{,}4^2 + 517{,}8^2 + 614{,}1^2 + 639{,}1^2 + 380{,}7^2 + 623{,}3^2 + +634{,}2^2 +$$

$$+615{,}2^2 = 2473035{,}88;$$

$$SS_{AB} = \frac{2 \cdot 4}{72} \cdot 2473035{,}88 - 346{,}8996 - 11265{,}6244 - 262667{,}52 =$$

$$= 501{,}7204;$$

взаимодействия факторов AC:

$$SS_{AC} = \frac{ac}{N}\sum_{i-1}^{a}\sum_{k=1}^{c} C_{ik}^2 - SS_A - SS_C - \frac{C^2}{N} =$$

$$= \frac{2\cdot 3}{72}(656{,}3^2 + 677{,}7^2 + 761{,}4^2 + 713{,}0^2 + 727{,}9^2 + 812{,}5^2) -$$

$$-346{,}8996 - 993{,}6912 - 262667{,}52 = 0{,}2725;$$

взаимодействия факторов BC:

$$SS_{BC} = \frac{bc}{N}\sum_{j-1}^{b}\sum_{k=1}^{c} C_{jk}^2 - SS_B - SS_C - \frac{C^2}{N};\ C_{jk}^2 =$$

$$= (\sum_{i=1}^{a}\sum_{m=1}^{n} X_{ijkm})^2;$$

$$SS_{BC} = \frac{4\cdot 3}{72}(229{,}2^2 + 228{,}8^2 + 247{,}1^2 + \cdots + 407{,}8^2 + 453{,}7^2) -$$

$$-11265{,}6244 - 993{,}6912 - 262667{,}52 = 182{,}1228;$$

взаимодействия факторов ABC:

$$SS_{ABC} = SS_V - SS_A - SS_B - SS_C - SS_{AB} - SS_{AC} - SS_{BC},$$

$$SS_{ABC} = 13319{,}613 - 346{,}8996 - 11265{,}6244 - 993{,}6912 -$$

$$-501{,}6146 - 0{,}2725 - 182{,}2286 = 29{,}2821.$$

Определяются средние квадраты отклонений как отношение сумм квадратов отклонений на соответствующее число степеней свободы:

$$s_A^2 = \frac{SS_A}{k_A} = \frac{346{,}8996}{1} = 346{,}8996;$$

$$s_B^2 = \frac{SS_B}{k_B} = \frac{11265{,}6244}{3} = 3755{,}2081;$$

$$s_C^2 = \frac{SS_C}{k_C} = \frac{993{,}6912}{2} = 496{,}8456;\ s_{AB}^2 = \frac{SS_{AB}}{k_{AB}} = \frac{501{,}7204}{3} = 167{,}2401;$$

$$s_{AC}^2 = \frac{SS_{AC}}{k_{AC}} = \frac{0{,}2725}{2} = 0{,}1362;\ s_{BC}^2 = \frac{SS_{BC}}{k_{BC}} = \frac{182{,}1228}{6} = 30{,}3538;$$

$$s_{ABC}^2 = \frac{SS_{ABC}}{k_{ABC}} = \frac{29{,}2821}{6} = 4{,}8804;\ s_P^2 = \frac{SS_P}{k_P} = \frac{30{,}1296}{2} = 15{,}0648;$$

$$s_Z^2 = \frac{SS_Z}{k_Z} = \frac{165{,}0374}{46} = 3{,}5878.$$

4) Находятся наблюдаемые значения F-критерия Фишера, разделив средние квадраты отклонений по факторам и взаимодействиям на остаточный средний квадрат:

$$F_{A\text{н.}} = \frac{s_A^2}{s_Z^2} = \frac{346{,}8996}{3{,}5878} = 96{,}69;$$

$$F_{B\text{н.}} = \frac{s_B^2}{s_Z^2} = \frac{3755{,}2081}{3{,}5878} = 1046{,}66 \text{ и т. д.}$$

По таблице распределения F-критерия Фишера находятся критические значения критерия при заданном уровне значимости α и числе степеней свободы k_1 для факторного и k_2 остаточного среднего квадрата:

$$\alpha = 0{,}05;\ k_1 = 1, k_2 = 4,\ F_{\text{кр}} = 4{,}05;\ k_1 = 2, k_2 = 46, F_{\text{кр.}} = 3{,}20;$$

$$k_1 = 3, k_2 = 46, F_{\text{кр.}} = 2{,}811;\ k_1 = 6, k_2 = 46, F_{\text{кр.}} = 2{,}30.$$

5) По результатам расчетов составляется таблица дисперсионного анализа (табл. 14.23).

Сравнение наблюдаемого значения F-критерия с критическим показывает, что значимое влияние на урожайность озимой пшеницы оказали все три фактора, т. е. плодородие почвы, система удобрений и система защиты растений, а также

взаимодействие факторов: плодородие почвы и система удобрений; система удобрений и система защиты растений. Приняв изменчивость урожайности озимой пшеницы за 100,0%, видно, что система удобрений объясняет 83,4% изменчивости урожайности, система защиты растений — 7,4%, плодородие почв — 2,6%, взаимодействие системы удобрений и системы защиты растений — 3,7%. Влияние других источников изменчивости было незначительным.

6) Оценку значимости частных различий между средними проведем с помощью расчета наименьшей существенной разности (НСР). Если разность между сравниваемыми средними больше НСР, то она является статистически значимой, если меньше, то не значимой.

Таблица 14.23

Дисперсионный анализ влияния факторов на урожайность озимой пшеницы

Источник изменчивости	Сумма квадратов отклонений	Степени свободы	Средний квадрат отклонений	Наблюдаемое значение *F*-критерия	Критическое значение *F*-критерия	η, %
A	346,8996	1	346,8996	96,69	4,05	2,57
B	11265,6244	3	3755,2081	1046,66	2,81	83,36
C	993,6912	2	496,8456	138,48	3,20	7,35
AB	501,7204	3	167,2401	46,61	2,81	3,71
AC	0,2725	2	0,1362	0,04	3,20	0,00
BC	182,1228	6	30,3538	8,46	2,30	1,35
ABC	29,2821	6	4,8804	1,36	2,30	0,22
P	30,1296	2	15,0648	4,199	3,20	0,22
Z	165,0374	46	3,5878			1,27
Общее	13514,78	71				100,00

При $\alpha = 0{,}05,\ k = k_Z = 46,\ t_{00,5:46} = 2{,}01$ — по таблице t-критерия Стьюдента

$$\text{НСР} = t_{\alpha,k_Z}\sqrt{\frac{2s_Z^2}{n}} = 2{,}01\sqrt{\frac{2\cdot 3{,}5875}{3}} = 3{,}11.$$

Например, из таблицы 14.20: $35{,}6-35{,}2 = 0{,}4 < \text{НСР}$ — разность между средними не значима; $54{,}87-51{,}17 = 3{,}7 > \text{НСР}$ — разность между средними значима и т. д.

В многофакторных опытах проводится оценка значимости разности средних по факторам и взаимодействиям:

для фактора *A*: $\text{НСР} = t_{\alpha,k_Z}\sqrt{\frac{2s_Z^2}{bcn}} = 2{,}01\sqrt{\frac{2\cdot 3{,}5875}{4\cdot 3\cdot 3}} = 0{,}90,$

разность между средними $62{,}59 - 58{,}20 = 4{,}39 > \text{НСР}$ значима;

для фактора *B*: $\text{НСР} = t_{\alpha,k_Z}\sqrt{\frac{2s_Z^2}{acn}} = 2{,}01\sqrt{\frac{2\cdot 3{,}5875}{2\cdot 3\cdot 3}} = 1{,}27,$

средние по уровням фактора *B* составили 39,17; 63,39; 69,35; 69,68, сравнение средних показывает, что урожайность озимой пшеницы по всем вариантам применения удобрений превосходит урожайность без применения удобрений, урожайность при средней и высокой норме удобрений значимо выше, чем при минимальной, различие в урожайности между средней и высокой нормой удобрений не значима;

для фактора *C*: $\text{НСР} = t_{\alpha,k_Z}\sqrt{\frac{2s_Z^2}{abn}} = 2{,}01\sqrt{\frac{2\cdot 3{,}5875}{2\cdot 4\cdot 3}} = 1{,}10,$

$\overline{x}_2 - \overline{x}_1 = 58{,}57 - 57{,}05 = 1{,}52 > \text{НСР}$ значима,

$\overline{x}_3 - \overline{x}_1 = 65{,}50 - 57{,}05 = 8{,}53 > \text{НСР}$ значима,

$\overline{x}_3 - \overline{x}_2 = 65{,}58 - 58{,}57 = 7{,}01 > \text{НСР}$ значима;

для взаимодействия факторов *AB*:

$$\text{НСР} = t_{\alpha,k_Z}\sqrt{\frac{2s_Z^2}{cn}} = 2{,}01\sqrt{\frac{2\cdot 3{,}5875}{3\cdot 3}} = 1{,}79,$$

например $\overline{x}_{21} - \overline{x}_{11} = 57{,}53 - 36{,}04 = 21{,}49 > \text{НСР}$ значима и т. д.;

для взаимодействия факторов *BC*:

$$\text{НСР} = t_{\alpha,k_Z}\sqrt{\frac{2s_Z^2}{an}} = 2{,}01\sqrt{\frac{2\cdot 3{,}5875}{2\cdot 3}} = 2{,}20,$$

например, $\overline{x}_{13} - \overline{x}_{11} = 41{,}18 - 38{,}2 = 2{,}98 > \text{НСР}$ значима и т. д.

В случае, когда число уровней по факторам и взаимодействиям больше двух или трех, то для оценки значимости различий между средними рационально использовать множественный ранговый критерий Дункана, учитывающий порядковый номер средней в их ранжированном ряду по убыванию значений средних. Данный критерий применяется после проведения дисперсионного анализа, если нулевая гипотеза о равенстве всех средних отвергается, а число сравниваемых средних больше двух. Применение критерия Дункана проводится в следующей последовательности:

а) средние располагаются в порядке убывания

$$\overline{X}_1 > \overline{X}_2 > \overline{X}_3 > \cdots > \overline{X}_k;$$

б) определяется средняя ошибка среднего по источнику изменчивости

$$s_{\overline{x}} = \sqrt{\frac{s_Z^2}{m}},$$

где m — число значений, использованных при расчете средних;

в) по таблице значений критерия Дункана при уровне значимости α, числе степеней свободы $m_1 = r + 2$, r — число средних значений между сравниваемыми средними, m_2 — число значений признака, использованных при расчете средних, устанавливаются значения рангов;

г) умножается средняя ошибка средней на выбранные значения рангов, в результате получаем совокупность наименьших значимых рангов (НЗР);

д) разности между средними сравниваются с соответствующими НЗР.

Если разность между средними значениями больше НЗР, то она является значимой, если меньше — то не значимой.

Проведем сравнение средних по уровням фактора *B*. Упорядочим средние по убыванию значений фактора *B*, которые запишем в таблицу 14.24.

По таблице при α=0,05, $m_1 = 2{,}3{,}4$, $m_2 = 18$ находятся значения рангов. Умножив значения рангов на среднюю ошибку средней, находим НЗР,

$$s_{\overline{x}} = \sqrt{\frac{s_Z^2}{m}} = \sqrt{\frac{3{,}5878}{18}} = 0{,}4464.$$

Таблица 14.24

Анализ различий между средними по критерию Дункана

Уровни фактора	B_4	B_3	B_2	B_1
Порядковый номер	0	1	2	3
Среднее значение	69,68	69,35	63,39	39,17
Значение ранга		2,97	3,12	3,21
НЗР		1,32	1,39	1,43

Разности между средними сравним с НЗР:

$\overline{X}_{B_4} - \overline{X}_{B_3} = 69{,}68 - 69{,}35 = 0{,}33 < 1{,}32$ разность не значима;

$\overline{X}_{B_4} - \overline{X}_{B_2} = 69{,}68 - 63{,}39 = 6{,}29 > 1{,}39$ разность значима;

$\overline{X}_{B_4} - \overline{X}_{B_1} = 69{,}68 - 39{,}17 = 30{,}51 > 1{,}43$ разность значима;

$\overline{X}_{B_3} - \overline{X}_{B_2} = 69{,}35 - 63{,}39 = 5{,}96 > 1{,}32$ разность значима;

$\overline{X}_{B_3} - \overline{X}_{B_1} = 69{,}68 - 63{,}39 = 6{,}29 > 1{,}39$ разность значима;

$\overline{X}_{B_2} - \overline{X}_{B_1} = 63{,}39 - 39{,}17 = 24{,}22 > 1{,}32$ разность значима.

Применение удобрений дает значимую прибавку урожайности озимой пшеницы по сравнению с вариантом без удобрений. Наибольший прирост урожайности получен при использовании средней и высокой норм удобрений по сравнению с минимальной нормой.

Темы (вопросы) для самоконтроля

1. Сущность дисперсионного анализа.
2. Условия применения дисперсионного анализа.
3. Однофакторный дисперсионный анализ.
4. Схема однофакторного дисперсионного анализа.
5. Теорема Кочрена.
6. Геометрическая интерпретация *F*-критерия.
7. Модели многофакторного дисперсионного анализа.

Глава 15
Корреляционно-регрессионный анализ

15.1. Виды и формы связей между признаками

В экономике большое значение имеет исследование зависимостей и взаимосвязей между явлениями и процессами. Оно дает возможность глубже понять сложный механизм причинно-следственных отношений между явлениями. Для исследования интенсивности, вида и формы связей между явлениями и процессами применяется корреляционно-регрессионный анализ. Он находит широкое применение при прогнозировании, при решении задач народнохозяйственного и внутрихозяйственного планирования, при выявлении факторов, воздействуя на которые можно вмешиваться в технологический, технический или экономический процесс с целью получения нужных результатов.

Под причинной связью понимают такое взаимодействие явлений и процессов реальной действительности, когда изменение одного из них — это следствие изменения другого. Часто какое-то явление может выступать как результат одной или нескольких причин и в то же время само служит причиной наступления других явлений и процессов. На рисунках 15.1–15.3 изображен ряд примеров, иллюстрирующих основные типы причинно-следственных связей между явлениями.

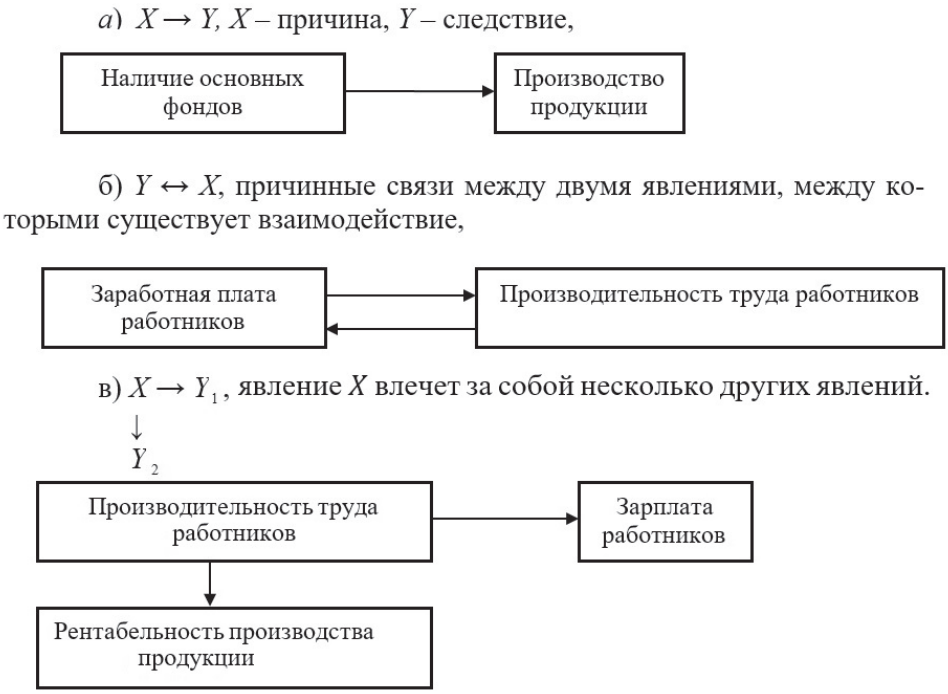


Рис. 15.1 — Основные типы причинно-следственных связей между явлениями

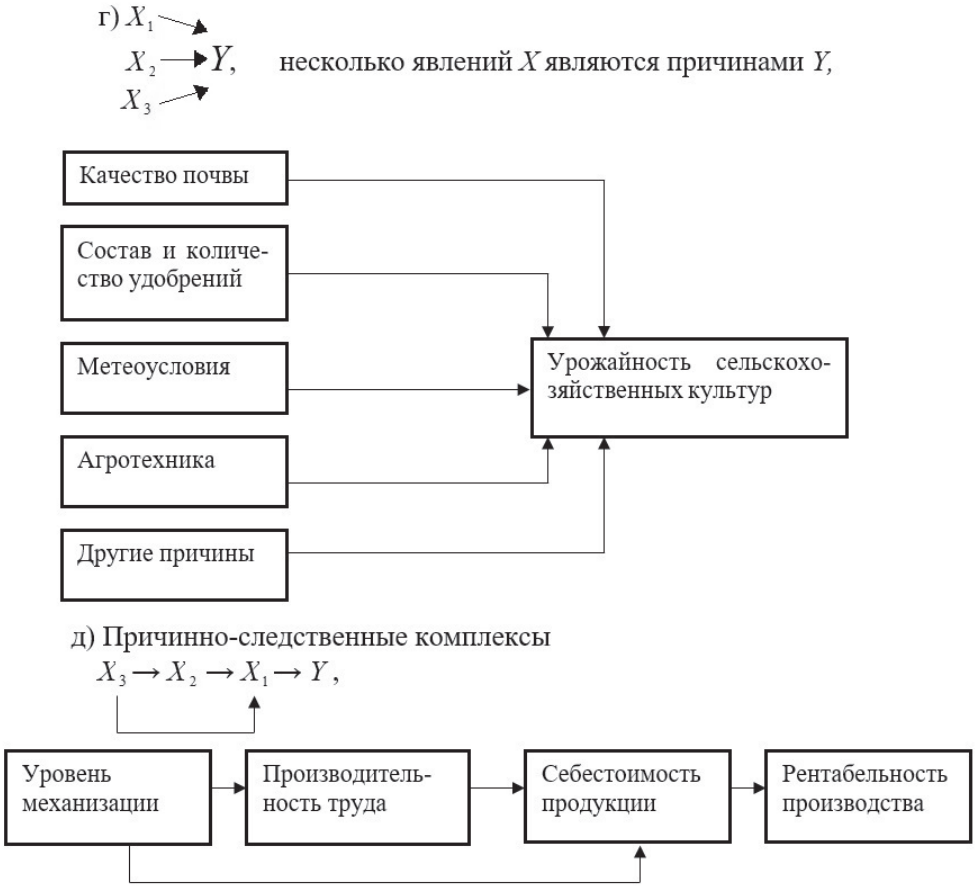


Рис. 15.2 — Основные типы причинно-следственных связей между явлениями

Различают два вида зависимостей между изучаемыми явлениями: функциональную и стохастическую.

Функциональной называется связь между переменными, когда каждому значению одной переменной величины соответствует вполне определенное значение другой переменной и, наоборот, т. е.

$$y = f(x) \text{ и } x = \varphi(y).$$

Стохастической называют зависимость, когда каждому значению переменной X может соответствовать одно или несколько различных значений переменной Y, причем до опыта нельзя предсказать возможное соответствие между ними. В случае стохастической связи изменение случайной величины Y, вследствие изменения случайной величины X, можно разбить на две компоненты: а) функциональную, связанную с зависимостью Y от X, б) случайную, связанную со случайным характером самих случайных величин X и Y. Соотношение между функциональной и случайной компонентой определяет силу связи. Отсутствие первой компоненты указывает на независимость случайных величин X и Y, отсутствие второй компоненты показывает, что между X и Y существует функциональная связь.

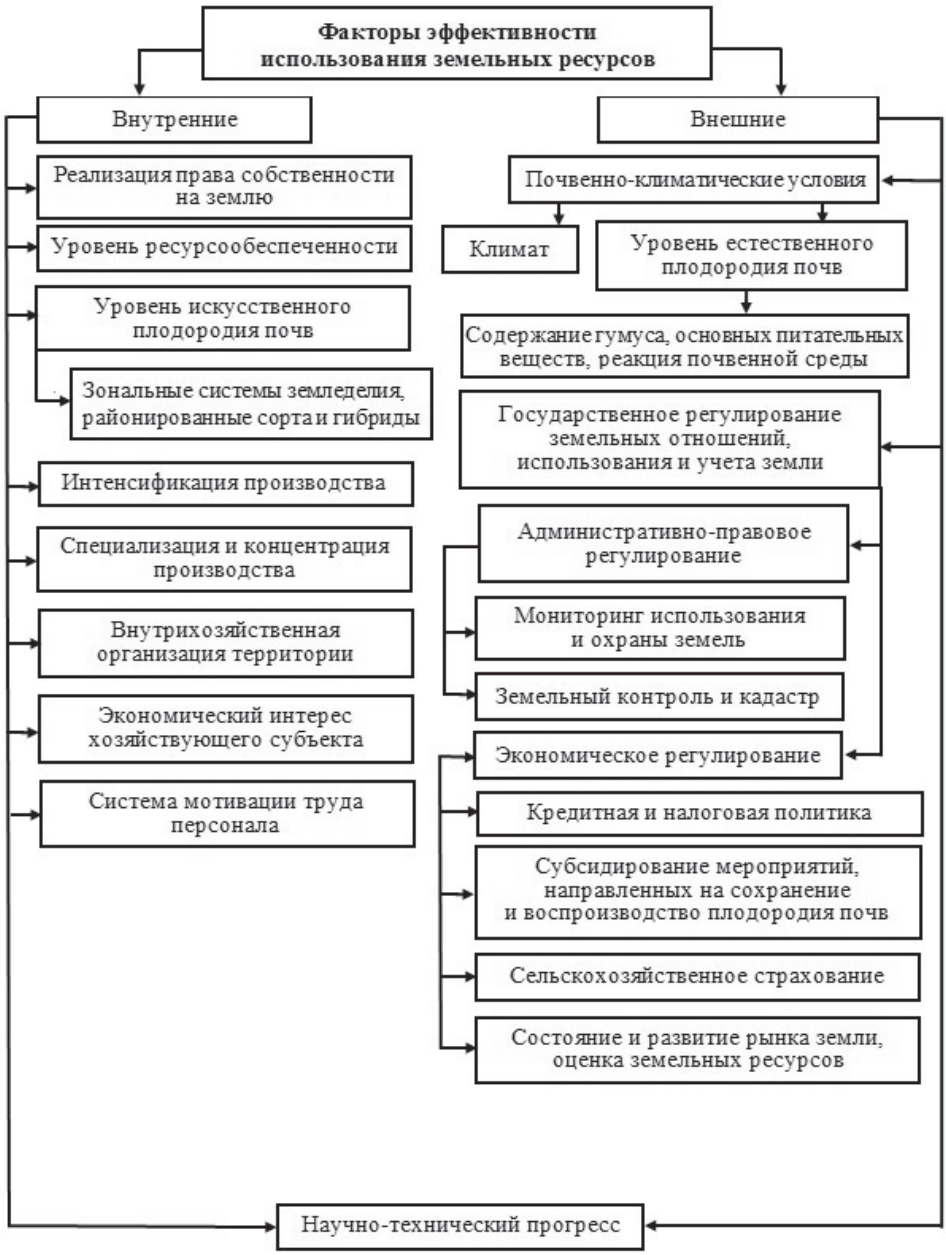


Рис. 15.3 — Основные типы причинно-следственных связей между явлениями, причинно-следственные комплексы

Статистическая или стохастическая (вероятностная) связь отражает закономерности только в массовых явлениях и процессах. Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения другой величины.

Статистическую зависимость называют корреляционной, точнее, выборочной корреляционной, если при изменении значений одной величины меняется среднее значение другой. Корреляционная зависимость проявляется тогда, когда в стохастической зависимости есть линейная функциональная компонента.

Традиционно в разных областях науки переменным X и Y дают следующие наименования, X: предикторная переменная, входная переменная, регрессор, независимая переменная, факторная переменная; Y: переменная отклика, выходная переменная, зависимая переменная, результативная переменная. В экономических исследованиях чаще всего используют последние названия переменных (X — факторная переменная, Y — результативная переменная), обозначающих изучаемые экономические признаки — факторные и результативные.

Факторными (объясняющими, независимыми) признаками называются признаки или переменные, оказывающие влияние на другие признаки. Они могут быть случайными и неслучайными.

Результативными (объясняемыми, зависимыми) называются признаки, формирующиеся под влиянием факторных признаков.

При сравнении функциональных и корреляционных зависимостей следует иметь ввиду, что при функциональной зависимости, зная X, можно точно вычислить величину Y. При корреляционной зависимости устанавливается лишь тенденция изменения Y при изменении X.

Если не известно, какой их признаков зависимый, а какой — независимый, или же это безразлично, то X и Y равноправны в этом смысле. В такой ситуации говорят о взаимосвязи корреляционного (ковариационного) типа в широком смысле. Если переменные не равноправны, т. е. четко ясно, какая из них причина, какая — следствие, то говорят о регрессионной зависимости.

Например:

− при изучении потребления электроэнергии (Y) в зависимости от объема производства (X) речь идет об односторонней связи;

− рост доходов населения ведет к увеличению потребления;

− снижение процентной ставки увеличивает инвестиции;

− увеличение валютного курса сокращает чистый экспорт.

По направлению изменения своих значений различают связи прямые и обратные. Если результативный и факторный признаки изменяются в одном направлении, то связь называется прямой, если же они изменяются в противоположных направлениях, то обратной. Например, зависимость между уровнем качества и ценой единицы продукции, уровнем доходов и величиной сбережений населения, урожайностью культуры и дозами внесенных удобрений на 1 га посева будет прямой. Связь же между себестоимостью продукции и производительностью труда, сроком эксплуатации машин и уровнем их использования, урожайностью культуры и себестоимостью продукции будет обратной.

По аналитическому выражению связи подразделяются на линейные и нелинейные. Линейная связь между переменными выражается уравнением прямой на плоскости или в пространстве, или в гиперпространстве. Нелинейные связи выражаются уравнениями кривых различного вида: парабола, гипербола, показательная, степенная, логистическая и другие.

По количеству одновременно включаемых в исследование факторов различают связи однофакторные $Y = f(X)$ и многофакторные $Y = f(X_1, X_2, \dots, X_p)$. Например:

$$y = a + bx, y = a + \frac{b}{x}, y = a \cdot x^b.$$

В экономических исследованиях результативный признак Y формируется, как правило, под влиянием нескольких факторных признаков $X_1, X_2, \dots, X_p$.

При построении уравнения множественной регрессии обычно используются следующие функции:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p,$$

$$y = b_0x_1^{b_1}x_2^{b_2} \dots x_p^{b_p},$$

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_1^2 + b_4x_2^2 + b_5x_1x_2.$$

С помощью корреляционного анализа оценивается направление и теснота связи между изучаемыми переменными. Его задачами являются:

1) выбор наиболее приемлемого показателя тесноты связи между переменными (коэффициент корреляции, корреляционное отношение, ранговый коэффициент корреляции, коэффициент конкордации, коэффициент взаимной сопряженности и т. п.);

2) точечная и интервальная оценка показателя тесноты связи по выборочным данным;

3) статистическая проверка значимости показателя тесноты связи;

4) формулирование вывода о наличии или отсутствии связи между переменными.

На практике встречается несколько типов корреляционной зависимости.

1) Факторный и результативный признаки причинно связаны. Например, урожайность и удобрения, производительность труда и его техническая вооруженность. В этих примерах однозначно понятно, какой признак факторный, а какой — результативный.

2) Оба признака являются следствием общей причины. Например, зависимость суммы товарооборота от числа торговых заведений по совокупности населенных пунктов может быть прямой и довольно существенной. Но оба эти признака зависят от одной причины — числа жителей в населенном пункте. Чем больше численность населения, тем больше требуется торговых организаций, соответственно, тем больше товарооборот.

3) Зависимость между признаками, каждый из которых является и причиной, и следствием, т. е. когда признаки являются взаимосвязанными. Например, зависимость между производительностью и оплатой труда. Ясно, что источником роста оплаты труда является рост производительности труда работников, в то же время материальная заинтересованность является фактором роста производительности труда.

4) Зависимость между признаками, при которых следствие определяется не одним фактором, а целой совокупностью факторов. Например, зависимость себестоимости продукции от уровня использования трудовых ресурсов, основных фондов, концентрации и специализации производства и др.

С корреляционным анализом тесно связан регрессионный анализ. Их объединяют методы обработки данных, отличают цели и формы установления связи. В корреляционном анализе оценивается сила стохастической связи, в регрессионном — форма связи.

Регрессионный анализ заключается в выборе и обосновании математического уравнения (совокупности уравнений), выражающего аналитически зависимость между признаками. К основным задачам регрессионного анализа относят:

1) определение аналитического вида функции, описывающей связь между результативным и факторными признаками;

2) нахождение параметров уравнения связи;

3) определение теоретических значений результативного признака по каждой единице совокупности при фактических значениях факторных признаков;

4) нахождение отклонений фактически наблюдаемых значений результативного признака от теоретических значений;

5) оценка значимости параметров и всего уравнения регрессии.

Имеется различие при исследовании корреляционных связей между экономическими явлениями и между явлениями в естественных науках и технике. В последних можно планировать эксперимент, когда добиваются элиминирования побочных факторов и поддержания условий эксперимента на неизменном уровне. В экономическом эксперименте эти действия практически невозможны — одно и то же следствие может быть порождено слишком многими причинами.

В экономике массовое исследование носит апостериорный характер, в естественных — априорный.

15.2. Корреляционный анализ

Изучение реальных процессов обычно предполагает наблюдение над целым рядом случайных величин (например, количество внесенных удобрений и урожайность, объем производства и численность работников и т. д.). Возникает задача количественного изучения взаимосвязи между случайными величинами. На первом уровне эта задача (оценка существенности влияния одного фактора на другой) может быть решена средствами дисперсионного анализа. Для численной оценки тесноты связи используется корреляционный анализ. В общем виде задача выявления и оценки силы *стохастической связи* не решена до сих пор. Важным частным случаем стохастической зависимости является *корреляционная зависимость* — это функциональная зависимость, которая существует между значениями одной переменной и групповыми средними другой. Корреляционная связь чаще всего оценивается выборочным коэффициентом корреляции r, который характеризует степень линейной функциональной зависимости между случайными величинами X и Y.

Пусть имеется n пар выборочных значений факторного и результативного признаков (x_i, y_i), которые можно представить графически в виде корреляционного поля (рис. 15.4) и (или) корреляционной таблицы.

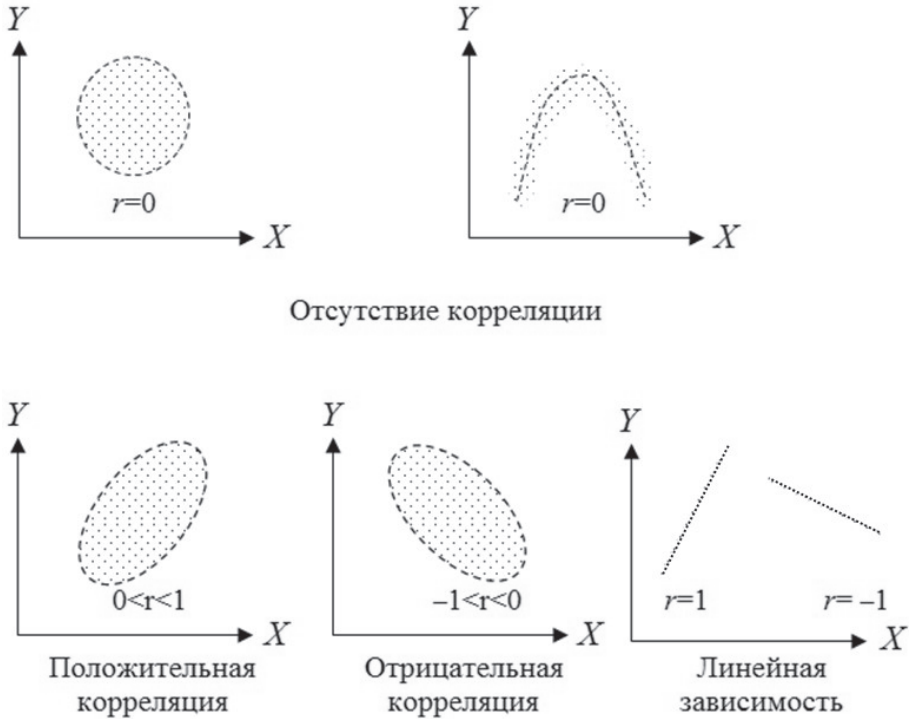


Рис. 15.4 — Оценка зависимости между выборочными значениями случайных величин X и Y с помощью выборочного коэффициента корреляции r

В случае парной зависимости между количественными переменными вычисляется коэффициент корреляции Пирсона:

$$r_{xy} = \frac{\Sigma(x-\overline{X})(y-\overline{Y})}{\sqrt{\Sigma(x-\overline{X})^2 \Sigma(y-\overline{Y})^2}} = \frac{\overline{XY}-\overline{X}\,\overline{Y}}{\sigma_x \sigma_y}, \qquad (15.1)$$

где $\overline{XY} = \frac{\Sigma xy}{n}$, $\overline{X} = \frac{\Sigma x}{n}$, $\overline{Y} = \frac{\Sigma y}{n}$, $\sigma_x = \sqrt{\overline{X^2} - \left(\overline{X}\right)^2}$, $\sigma_y = \sqrt{\overline{Y^2} - \left(\overline{Y}\right)^2}$,

$$\overline{X^2} = \frac{\Sigma x^2}{n}, \qquad \overline{Y^2} = \frac{\Sigma y^2}{n}.$$

Квадрат коэффициента корреляции (r^2) называется коэффициентом детерминации. Он показывает, какая часть общей дисперсии результативного признака Y объясняется влиянием изучаемого фактора X. Тогда ($1 - r^2$) — толерантность, которая показывает долю влияния других, неучтенных факторов.

Рассмотрим идеологию построения коэффициента Пирсона r_{xy}. Положение объекта относительно других в выборке (для парной зависимости) зависит от $\overline{X}$ и $\overline{Y}$ проявляется в величине и знаках $\left(x - \overline{X}\right)$ и $\left(y - \overline{Y}\right)$.

Если объект имеет высокий уровень по обеим переменным, то произведение $\left(x - \overline{X}\right)\left(y - \overline{Y}\right)$ будет большим и положительным. Аналогично, если он относительно низок как по x, так и по y, то $\left(x - \overline{X}\right)\left(y - \overline{Y}\right)$ для него также будет большим и положительным (поскольку произведение двух отрицательных чисел

положительно). Если x и y в основном связаны прямо (большие значения с большими, а малые — с малыми), то большинство произведений $(x-\overline{X})(y-\overline{Y})$ будет положительно, следовательно, $\sum(x-\overline{X})(y-\overline{Y})$ будет большой и положительной.

Если X и Y имеют обратную связь — большое X встречается с малым Y и наоборот, то большинство произведений $(x-\overline{X})(y-\overline{Y})$ будет отрицательно, следовательно, $\sum(x-\overline{X})(y-\overline{Y})$ будет отрицательной (когда x и y связаны обратной зависимостью).

Если между переменными нет систематической связи (большие X сочетаются с малыми Y так же часто, как и с большими и то же самое справедливо для малых X), то $\sum(x-\overline{X})(y-\overline{Y})$ будет близка к нулю.

Обычно $\sum(x-\overline{X})(y-\overline{Y})$ усредняют (для устранения зависимости от числа пар наблюдений) и получают ковариацию

$$cov(x,y)=\frac{\sum(x-\overline{X})(y-\overline{Y})}{n}— \qquad (15.2)$$

величину, которая велика и положительна, когда X и Y сильно связаны прямой связью; близка к нулю в случае отсутствия связи; велика и отрицательна, когда переменные сильно связаны обратной связью.

Для устранения влияния на ковариацию разброса случайных величин ее нормируют, делят на σ_x и σ_y. В результате получается коэффициент корреляции Пирсона (15.1).

Коэффициент корреляции имеет следующие свойства (рис. 15.4):

1) принимает значения на отрезке $[-1;1]$, то есть $-1 \le r \le 1$;

2) если $r = \pm 1$, то переменные X и Y связаны функциональной линейной зависимостью;

3) если $r = 0$, то переменные X и Y линейно не коррелированы, что не означает независимости вообще (на втором рисунке сверху рис. 15.4 случайные величины X и Y линейно не коррелированы, однако зависимость между ними можно описать параболой);

4) если X и Y образуют систему нормально распределенных случайных величин, то из их некоррелированности следует их независимость;

5) коэффициенты корреляции Y на X и X на Y совпадают;

6) знак коэффициента корреляции показывает направление связи, прямая (+), обратная (–).

Коэффициент корреляции определяется по выборочным наблюдениям, поэтому возникает задача оценки его статистической значимости. Рассматривается нулевая гипотеза — коэффициент корреляции равен нулю в генеральной совокупности, т. е. не является статистически значимым ($H_0: r = 0$) и альтернативная — коэффициент корреляции существенно отличен от нуля ($H_1: r \neq 0$) в генеральной совокупности. Проверка гипотезы осуществляется по t-критерию Стьюдента. Критическое значение $t_{кр}$ находится по таблице приложения 2 для двусторонней критической области при заданном уровне значимости α и числе степеней свободы $k = n — 2$. Наблюдаемое значения критерия определяется по формуле

$$t_{\text{набл}} = |r|\sqrt{\frac{n-2}{1-r^2}}. \quad (15.3)$$

Если $t_{\text{набл}} > t_{\text{кр}}$ — то нулевая гипотеза отвергается, коэффициент корреляции считается статистически значимым в генеральной совокупности. Если же $t_{\text{набл}} < t_{\text{кр}}$, то нулевая гипотеза принимается, коэффициент корреляции может быть равен нулю в генеральной совокупности.

Непосредственное вычисление и оценка коэффициента корреляции еще не говорит о наличии причинной зависимости между случайными величинами. Изучение корреляционной связи должно сопровождаться пониманием внутренних особенностей изучаемых процессов и возможных причинно-следственных связей.

Так М. Езекиэл и К. А. Фокс пишут по этому поводу [43]: «Исследователи думают, что они правильно поступают, когда изучают зависимость данной переменной от нескольких факторов, пренебрегая теми из них, корреляция с которыми не проявляется, и отбирая для анализа множественной корреляции лишь те факторы, простая корреляция с которыми высока. Эти действия могут привести к пренебрежению такими факторами…, действие которых может обнаружиться лишь после устранения влияния других, находящихся во взаимодействии факторов».

Различают корреляцию, обусловленную:

– причинной зависимостью *Y* от *X*;

– зависимостью *X* и *Y* от третьей переменной;

– неоднородностью выборки (например, при изучении зависимости размеров кровяных шариков от содержания гемоглобина в крови у новорождённых, женщин и мужчин коэффициент корреляции порядка нескольких сотых. Однако при объединении статистического материала коэффициент корреляции равен 0,75);

– формально (числовыми данными).

Для определения вида корреляции можно использовать схему, приведенную на рисунке 15.5.

Коэффициент корреляции характеризует тесноту связи между результативным признаком и одним фактором, но на результативный признак y могут оказывать влияние несколько факторов, например x_1 и x_2.

Для исключения влияния, например x_2, можно перейти к новым переменным:

$$x_1^* = x_1 - \lambda x_2 \ \text{ и } \ y_1^* = y_1 - \lambda x_2,$$

которые некоррелированны с x_2.

После соответствующих преобразований получим, что оставшаяся корреляция между y и x_1 будет равна числу

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right)\left(1 - r_{x_1 x_2}^2\right)}},$$

которое называется *частным коэффициентом корреляции*.

Аналогично, для случая двух факторных признаков, можно получить формулы:

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{\sqrt{(1 - r_{yx_1}^2)(1 - r_{x_1 x_2}^2)}}, \qquad (15.4)$$

$$r_{x_1 x_2 \cdot y} = \frac{r_{x_1 x_2} - r_{yx_1} r_{yx_2}}{\sqrt{(1 - r_{yx_1}^2)(1 - r_{yx_2}^2)}}.$$

Частные коэффициенты корреляции позволяют измерить тесноту связи между факторным и результативным признаком (X_j и Y) при исключении влияния всех остальных переменных (значения которых известны), что тоже не совсем соответствует истине, ибо и в этом случае на результативный признак оказывают влияние другие факторы, кроме изучаемых.

Рис. 15.5 — Схема определения вида корреляции

Как отмечено выше, коэффициент корреляции характеризует степень линейной зависимости между переменными. В случае нелинейной зависимости для оценки тесноты связи используется корреляционное отношение. *Корреляционным отношением Y* на *X* называется отношение межгруппового среднего квадратического отклонения δ_y переменной Y к ее общему среднему квадратическому отклонению σ_y:

$$\eta_{yx} = \frac{\delta_y}{\sigma_y}, \qquad (15.5)$$

где межгрупповая дисперсия определяется по формуле

$$\delta_y = \frac{\sum_i (\bar{y}_i - \bar{y})^2}{n}. \qquad (15.6)$$

Аналогично определяется корреляционное отношение X на Y. Основные свойства корреляционных отношений:

1) $0 \le \eta_{yx} \le 1,\ 0 \le \eta_{xy} \le 1$;

2) если $\eta = 0$, то корреляционная связь между переменными отсутствует;

3) если $\eta = 1$, то переменные связаны функционально;

4) для линейной зависимости между переменными X и Y необходимо и достаточно, чтобы выполнялось равенство $|r_{yx}| = \eta_{yx}$;

5) $\eta_{xy} \neq \eta_{yx}$;

6) $|r| \le \eta$.

Использование коэффициента парной корреляции r неявно предполагает нормальное распределение генеральной совокупности, из которой производится выборка. В случае парной зависимости между качественными переменными, применяются методы корреляционного анализа номинальных (значения которых можно только сравнить — равны или нет), порядковых (значения которых можно упорядочить) или измеренных в разных шкалах признаков, не предполагающие известным вид распределения генеральной совокупности (методы непараметрической статистики).

Например, в случае двух признаков, которые можно упорядочить — *коэффициент ранговой корреляции* Спирмена — r_s:

$$r_s = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}, \qquad (15.7)$$

где d_i^2 — квадраты разности рангов ($d_i = x_i - y_i$), n — число наблюдений, т. е. число пар рангов.

Ранг — это порядковый номер значений признака, расположенных в порядке возрастания или убывания их величин.

Пример 15.1. Преподавателю и студенту было предложено расположить 10 профессий в порядке их общественной значимости. Ответы перечислены ниже. Какова корреляция рангов между двумя рядами оценок? Одинаково ли мнение преподавателя и студента по этому вопросу?

Таблица 15.1

Ранги значимости профессий

Профессии	Оценка преподавателя, x_i	Оценка студента, y_i
Профессор	3	2
Врач	1	1
Учитель школы	4	7
Директор магазина	2	4
Бухгалтер	8	5
Банкир	6	3
Водитель	9	9
Журналист	5	8
Диджей	10	10
Программист	7	6

Решение. Определим разности рангов, их квадраты и суммы.

$d_i = x_i - y_i$	1	0	–3	–2	3	3	0	–3	0	1	∑=0
d_i^2	1	0	9	4	9	9	0	9	0	1	∑=42

Имеем

$$r_s = 1 - \frac{6\sum_{i=1}^{10} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 42}{10(100-1)} = 0{,}7454.$$

Проверим, существенна ли положительная корреляционная связь между мнениями преподавателя и студента, для этого (при $n \geq 10$) используем t-статистику Стьюдента с $k = (n-2)$ степенями свободы:

$$t_{\text{н}} = |r_s| \cdot \sqrt{\frac{n-2}{1-r_s^2}}. \qquad (15.8)$$

Нулевая гипотеза — коэффициент корреляции не является статистически значимым ($\text{H}_0: r_s = 0$). Альтернативная гипотеза — существует положительная корреляционная зависимость ($\text{H}_1: r_s > 0$). При уровне значимости $\alpha = 0{,}05$ для односторонней (правосторонней) критической области (прил. 3):

$$t_{\text{кр.}} = t(\alpha = 0{,}05; k = 8) = 1{,}86,$$

$$t_{\text{н.}} = |0{,}7454| \cdot \sqrt{\frac{10-2}{1-0{,}7454^2}} = 3{,}16.$$

Так как $t_{\text{н.}} > t_{\text{кр.}}$, то связь между мнениями преподавателя и студента является статистически значимой, при 5%-ном уровне значимости. Значит, мнения преподавателя и студента об общественной значимости различных профессий в основном совпадают.

15.3. Однофакторный регрессионный анализ

Регрессионный анализ — один из основных методов современной математической статистики, позволяющий аналитически представить связь между двумя или несколькими переменными. Если корреляционный анализ позволяет установить существует или не существует факт зависимости между парами наблюдений, то регрессионный анализ дает целый арсенал методов построения соответствующих зависимостей.

В настоящее время серьезные исследовательские работы с применением методов регрессионного анализа выполняются исключительно с помощью компьютера. Изучение регрессионного анализа можно провести на первых стадиях и без применения компьютера, главное — отработать навыки «правильного» принятия решений после каждого шага вычислений и уметь сделать «правильные» выводы.

Объект исследования в регрессионном анализе — экономические, социальные, политические, экологические, технические и другие системы, явления и процессы.

Предмет исследования — математические модели взаимодействия изучаемых явлений и процессов.

Цель исследования — установление по результатам статистических наблюдений (пассивных или активных) адекватной аналитической зависимости (уравнения регрессии) между результативными признаками и факторами, которые характеризуют изучаемые системы. Это соответствует одной из наиболее общих задач статистики — оценивания степени и формы связи между величинами.

Термин «регрессия» впервые появился более 100 лет назад в работе английского физиолога и антрополога Ф. Гальтона, исследовавшего 205 пар родителей и 930 их детей и пришедшего к выводу о том, что имеет место «регресс» — чем выше родители, тем ниже дети, поэтому проведенный анализ назвал регрессионным. В частности, Гальтон показал, что если Y — рост взрослого потомка, X — рост родителей (взвешенное среднее роста отца и матери), то уравнение регрессии имеет вид $\hat{Y} = \bar{Y} + \frac{2}{3}(X - \bar{X})$. Хотя анализ, который он проводил, скорее, можно назвать корреляционным, но термин исторически прижился. Кстати, термин «корреляция» также придумал он, обозначение r происходит от слова регрессия (*regression*) [38]. Рассмотрим первоначально простую задачу. Пусть изучается связь между двумя величинами X и Y, в результате наблюдения определены их попарные совместные значения (табл. 15.2).

Таблица 15.2

Результаты статистического наблюдения

X	x_1	x_2	...	x_i	...	x_n
Y	y_1	y_2	...	y_i	...	y_n

Данные таблицы 15.2 можно изобразить на графике — диаграмме рассеяния, которая имеет название — корреляционное поле — и является вспомогательным средством визуализации наблюдений для первоначального заключения о характере и форме зависимости. На рисунке 15.4 представлены точки, соответствующие парам значений наблюдений, по которым можно предположить наличие или отсутствие стохастической зависимости между переменными X и Y. Задача исследователя — найти аналитическую функцию, которая наилучшим образом описывает экспериментальные данные.

Переменные X и Y могут быть: обе случайными; одна случайная, другая — нет; обычно X — не случайная, фиксированная и управляемая, Y — случайная.

Если X и Y — случайные величины, то имеет место некоторое совместное распределение $f(x, y)$, причем степень зависимости между X и Y, как известно, может быть охарактеризована коэффициентом корреляции r_{xy} (в случае линейной зависимости), корреляционным отношением (в случае нелинейной зависимости) или двумя функциями регрессии Y на X и X на Y, т. е. зависимостями:

$$y = \mathrm{M}[Y/X], \quad x = \mathrm{M}[X/Y],$$

где M — соответствующие математические ожидания ($M(X) = a$, $M(Y) = b$).

Например, если X и Y — система двух нормально распределенных случайных величин с функцией плотности распределения вероятностей:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} \exp\left(-\frac{1}{2(1-r^2)}\right)\left[\frac{(x-a)^2}{\sigma_x^2} - \frac{2r(x-a)(y-b)}{\sigma_x\sigma_y} + \frac{(y-b)^2}{\sigma_y^2}\right], \quad (15.9)$$

то уравнение регрессии Y на X можно записать:

$$\mathrm{M}\,[Y/X] - M(Y) = r\frac{\sigma_y}{\sigma_x}\left(x - M(X)\right), \tag{15.10}$$

а регрессии X на Y:

$$\mathrm{M}\,[X/Y] - M(X) = r\frac{\sigma_x}{\sigma_y}\left(y - M(Y)\right). \tag{15.11}$$

Либо, учитывая, что оценкой математического ожидания является среднее арифметическое значение, соответственно получим:

$$y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}\,(x - \bar{X}), \tag{15.12}$$

$$x - \bar{X} = r\frac{\sigma_x}{\sigma_y}\,(y - \bar{Y}). \tag{15.13}$$

Функция нормально распределенной двумерной случайной величины $f(x, y)$ имеет колоколобразную форму (рис. 15.6), сечения которой, перпендикулярные оси аппликат, представляют собой эллипсы с *центром рассеивания* $(\bar{X}, \bar{Y})$, а сечения, перпендикулярные осям ординат и абсцисс, имеют соответствующие одномерные нормальные законы распределения переменных X и Y:

$$f_1(x) = \frac{1}{\sqrt{2\pi}\sigma_x}\exp(-\frac{(x-a)^2}{2\sigma_x^2}),\;\; f_2(y) = \frac{1}{\sqrt{2\pi}\sigma_y}\exp\,(-\frac{(y-b)^2}{2\sigma_y^2}).$$

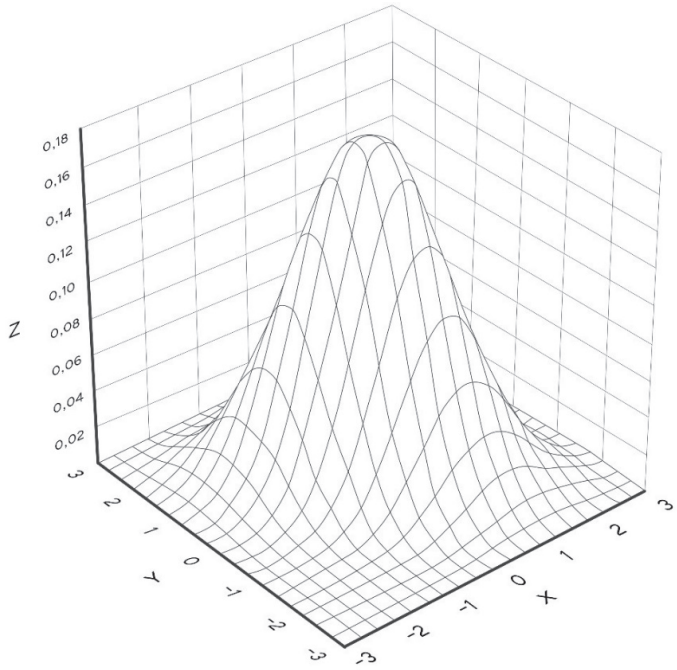


Рис. 15.6 — График плотности вероятностей
двумерного нормального распределения

Условные математические ожидания (15.10)–(15.11) представляют собой линии регрессии Y на X и X на Y, полученные по выборочным данным, представленные в виде (15.12) и (15.13) соответственно и пересекающихся в точке $(\bar{X},\ \bar{Y})$, где $\bar{X} \to a$, $\bar{Y} \to b$. Таким образом, коэффициент корреляции r представляет собой один из параметров двумерного нормального распределения. Он определяет

ориентацию эллипса рассеяния с центром (a, b), который строится на основе предположения двумерного нормального распределения двух переменных и представляет собой предсказанный интервал для нового одиночного наблюдения при заданном числе наблюдений n (рис. 15.7).

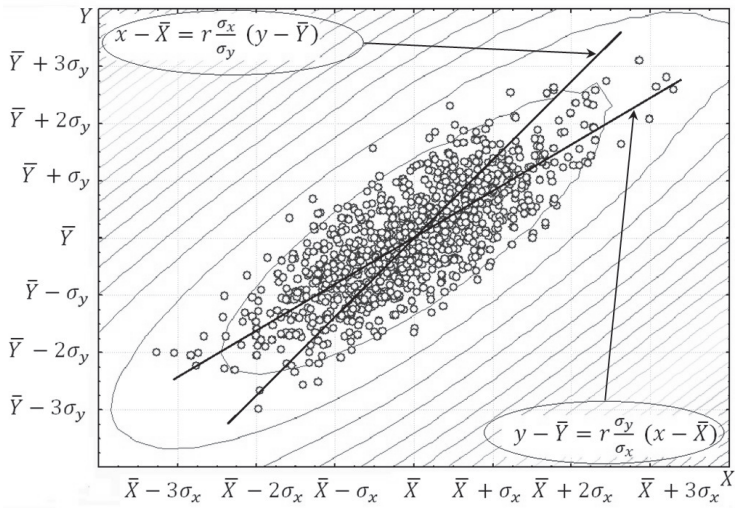


Рис. 15.7 — Корреляционное поле, эллипсы рассеяния с центром $(\bar{X},\ \bar{Y})$ и прямые регрессии $(r > 0)$ Y на X и X на Y (эллипсы рассеяния получены в виде «карты линий уровня» двумерного нормального закона в пакете Statistica 6.1)

Теоретические положения можно проиллюстрировать имитационной моделью, выполненной в одной из программ. Рассмотрим возможный способ генерации случайных чисел, подчиненных нормальному закону и имеющих определенную корреляционную зависимость. С использованием стандартной надстройки *MS Excel — Анализ данных — Генерация случайных чисел*[27] легко сгенерировать два вектора X и Y одного размера, например, $n = 1000$ и с одним распределением, случайные элементы которого попарно коррелированы с коэффициентом r.

Итак, пусть для простоты в генерируемых распределениях математическое ожидание $a = 0$, а дисперсия равна 1. Тогда если $rnom\ (n, a, \sigma)$ — генерация n чисел, подчиненных нормальному закону распределения с математическим ожиданием a и средним квадратическим отклонением σ, то алгоритм генерации коррелированных векторов X и Y выглядит так:

$n := 1000, a := 0, \sigma := 1; r := 0{,}8;$

$X := rnorm(n, a, \sigma);$

$Y := r \cdot X + \sqrt{1 - r^2} \cdot norm(n, a, \sigma).$

Реализация алгоритма позволила получить соответствующие векторы X, Y и значение выборочного коэффициента корреляции $r \approx 0{,}784$, а также проекции двумерного нормального закона распределения на плоскости XOZ, YOZ (в зави-

[27] О генерации случайных чисел см. раздел 10.3.

симости от контекста, Z — значение функции плотности распределения вероятностей от одной или двух переменных), а также уравнения регрессии Y на X и X на Y (рис. 15.8–15.11).

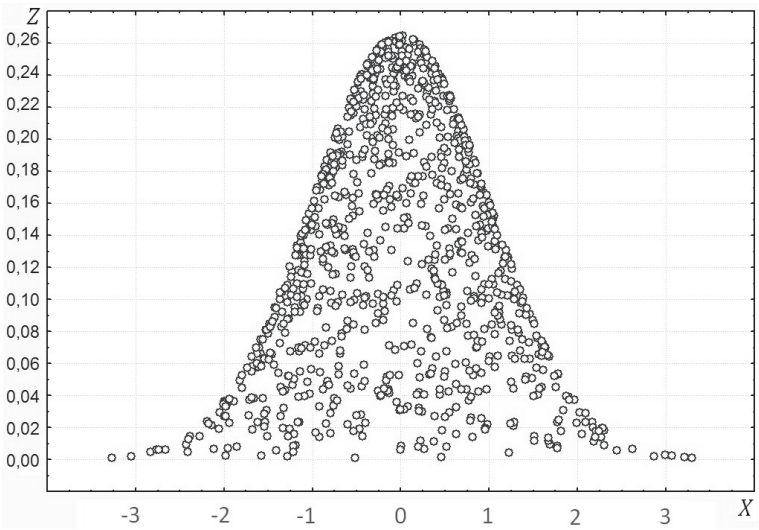


Рис. 15.8 — Проекция точек двумерного нормального распределения $(a = b = 0, \sigma_x = \sigma_y = 1, r = 0{,}8)$ на плоскость XOZ

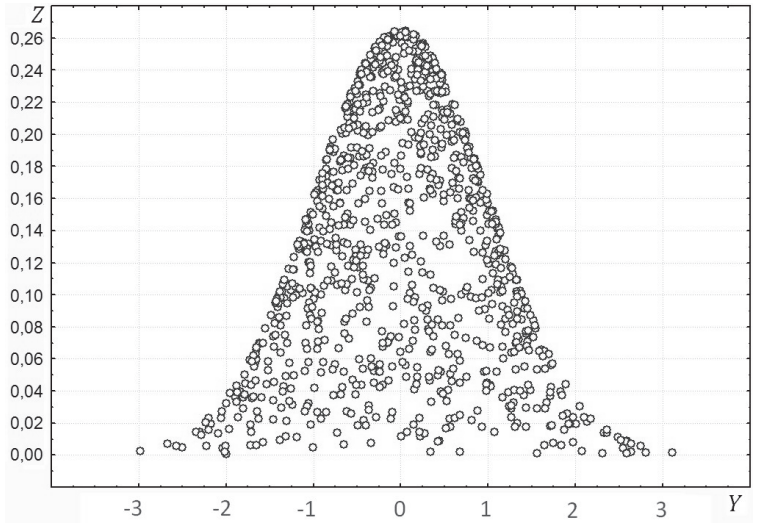


Рис. 15.9 — Проекция точек двумерного нормального распределения $(a = b = 0, \sigma_x = \sigma_y = 1, r = 0{,}8)$ на плоскость YOZ

Если переменная X — не случайная, то каждому ее значению x соответствует некоторое распределение $f(y)$ случайной величины Y. Рассмотрим этот случай для одной факторной переменной. В этом случае говорят о парной регрессии. Итак, пусть имеется n пар наблюдений (x_i, y_i).

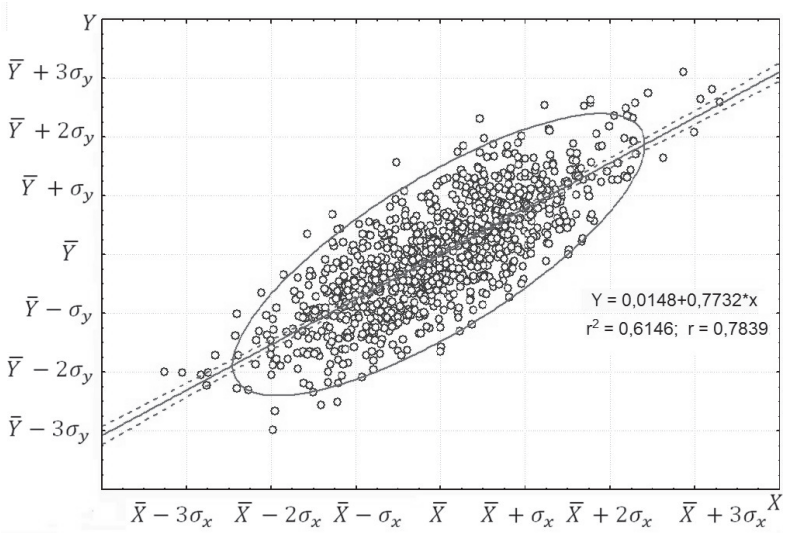


Рис. 15.10 — Регрессия Y на X

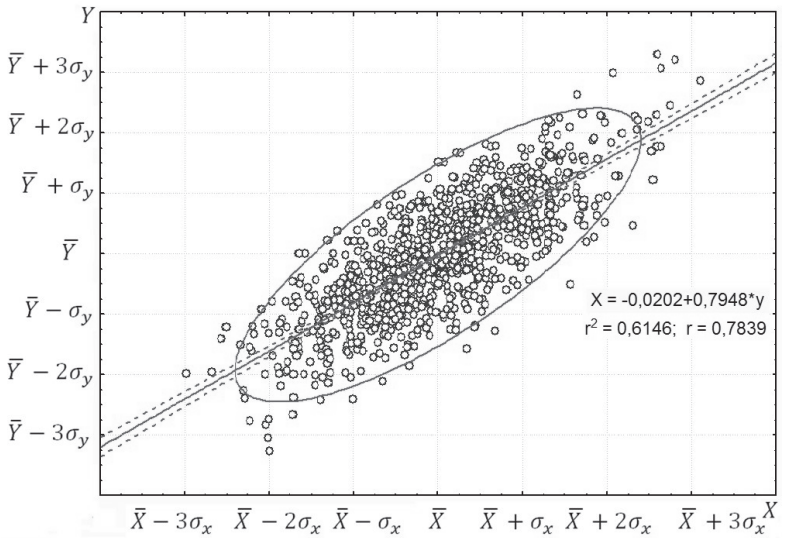


Рис. 15.11 — Регрессия X на Y

Теоретические соображения, подкрепленные визуализацией наблюдений в системе координат XOY, могут позволить сделать предположение о функциональной форме зависимости между переменными. В простейшем случае это уравнение прямой $y = \beta_0 + \beta_1 x$ (*теоретическое уравнение регрессии*). Если попытаться искать коэффициенты β_0 и β_1, подставляя наблюдения в предполагаемое уравнение линейной зависимости, то получим n уравнений $y_i = b_0 + b_1 x_i$ (*эмпирическое уравнение регрессии*) относительно переменных b_0 и b_1, причем эти уравнения окажутся несовместными. Несовместность уравнений может объясняться либо несовершенством теории, предписывающей линейность зависимости, либо погрешностями наблюдений, либо тем и другим вместе.

Можно предположить, что количественные поправки к наблюдениям и теории невелики, и попытаться если не точно, то приближенно выразить наблюдения линейной зависимостью.

В этом случае естественно искать такие значения b_0 и b_1, что абсолютные значения ошибок $\varepsilon_i = y_i - \hat{y}_i = y_i - (\, b_0 \, + \, b_1 x_i)$ были бы в каком-то смысле «малыми в совокупности». В общем случае речь идет о некоторой функции ошибок (потерь), которая должна быть минимизирована. Чаще всего, в силу удобства в вычислительном отношении и применимости к весьма широкому классу задач, рассматривается функция ошибок (потерь):

$$S(\varepsilon_i) = \sum_{i=1}^{n} \varepsilon_i^2 \to min. \tag{15.14}$$

Рассмотрим функцию от двух переменных b_0 и b_1:

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \to min. \tag{15.15}$$

Для нахождения точек (b_0, b_1), удовлетворяющих (15.15), найдем частные производные по b_0 и b_1 и приравняем их к нулю:

$$\begin{cases} \frac{\partial S(b_0,b_1)}{\partial b_0} = 0, \\ \frac{\partial S(b_0,b_1)}{\partial b_1} = 0. \end{cases} \tag{15.16}$$

Имеем

$$\begin{cases} \frac{\partial S(b_0,b_1)}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0, \\ \frac{\partial S(b_0,b_1)}{\partial b_1} = -2 \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0. \end{cases} \tag{15.17}$$

В результате получим систему уравнений, называемую системой нормальных уравнений Гаусса*:*

$$\begin{cases} b_0 n + b_1 \sum x_i = \sum y_i \ , \\ b_0 \sum x_i + b_1 \sum x_i^2 = \sum x_i \, y_i. \end{cases} \tag{15.18}$$

Разделив уравнения последней системы на число наблюдений (n) и переходя к средним арифметическим значениям, можно записать систему (15.18) в виде

$$\begin{cases} b_0 + b_1 \overline{X} = \overline{Y}, \\ b_0 \overline{X} + b_1 \overline{X^2} = \overline{XY}. \end{cases} \tag{15.19}$$

Решив систему (15.19), получим

$$\begin{cases} b_0 = \overline{Y} - b_1 \overline{X}, \\ b_1 = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - (\overline{X})^2}. \end{cases} \tag{15.20}$$

Формулу для коэффициента b_1 можно переписать в виде

$$b_1 = \frac{\sum (x_i - \overline{X})(y_i - \overline{Y})}{\sum (x_i - \overline{X})^2}, \tag{15.21}$$

тогда искомая прямая может быть найдена в виде

$$y - \overline{Y} = b_1 \big(x - \overline{X}\big),$$

что очевидно соответствует виду уравнения (15.12). Таким образом, легко получить прямую функциональную зависимость между коэффициентом регрессии (b_1) и коэффициентом корреляции (r):

$$b_1 = r \frac{\sigma_x}{\sigma_y} \, . \tag{15.22}$$

Коэффициент регрессии (b_1) показывает, на сколько единиц в среднем изменится результативный признак Y при увеличении факторного признака X на единицу. При интерпретации уравнения регрессии используется коэффициент эластичности, который определяется по формуле

$$Э = b_1 \frac{\overline{X}}{\overline{Y}}. \quad (15.23)$$

Коэффициент эластичности показывает, на сколько процентов в среднем изменится результативный признак Y при увеличении факторного признака X на 1%.

Оценка качества полученной зависимости требует дополнительных предположений относительно изучаемого явления и свойств, полученных ошибок.

Описанный способ получения параметров линейного уравнения регрессии из условия минимизации суммы квадратов ошибок (разностей эмпирических и теоретических значений результативной переменной) называется методом наименьших квадратов (МНК).

Метод наименьших квадратов — основной (но не единственный) метод регрессионного анализа, кроме него чаще всего используется метод максимального правдоподобия (см. 12.2), предполагающий известным закон распределения ошибок, в частности, в случае нормального закона получается метод наименьших квадратов.

Оценки метода наименьших квадратов. Формулы (15.20)–(15.21) позволяют утверждать, что параметры b_0 и b_1 являются линейными функциями от наблюдений $y_1, y_2, \dots, y_n$. Пусть наблюдаемые ошибки ε_i случайны и $M(\varepsilon_i) = 0$, $D(\varepsilon_i) = \sigma^2$, тогда можно получить следующие свойства МНК-оценок (которые рекомендуется доказать самостоятельно):

1) $M(b_0) = \beta_0,\ M(b_1) = \beta_1,$

2) $D(b_0) = \frac{\sigma^2}{n},\ D(b_1) = \frac{\sigma^2}{\sum(x-\overline{X})^2},$

3) $cov(b_0, b_1) = 0.$

Нормальное распределение ошибок. Учитывая, что параметры b_0 и b_1 являются линейными функциями от наблюдений $y_1, y_2, \dots, y_n$, положим, что $\varepsilon_i \to N(0, \sigma^2)$, тогда можно показать, что

$$b_0 \to N(\beta_0, \frac{\sigma^2}{n}),\ b_1 \to N\left(\beta_1, \frac{\sigma^2}{\sum(x-\overline{X})^2}\right). \quad (15.24)$$

Рассмотрим остаточную сумму квадратов $\sum(y_i - \hat{y}_i)^2$. Опираясь на идеологию доказательства леммы Фишера и теоремы Пирсона, можно показать, что величины b_0, b_1 и $\sum(y_i - \hat{y}_i)^2$ независимы, кроме того:

$$\frac{\sum(y_i - \hat{y}_i)^2}{\sigma^2} \to \chi^2_{n-2}, \quad (15.25)$$

$$M(\sigma^2) = \frac{\sum(y_i-\hat{y}_i)^2}{n-2}. \quad (15.26)$$

Из предыдущего следует, что величина

$$\frac{b_1-\beta_1}{\sqrt{\frac{\sigma^2}{\sum(x-\overline{X})^2}}} \frac{1}{\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{\sigma^2(n-2)}}} = \frac{b_1-\beta_1}{\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{(n-2)\sum(x-\overline{X})^2}}} \to t_{n-2} \quad (15.27)$$

асимптотически стремится к t-распределению Стьюдента с $k = n - 2$ степенями свободы.

Действительно, так как

$$\frac{b_1-\beta_1}{\sqrt{\frac{\sigma^2}{\sum(x-\overline{X})^2}}} \to N(0,1) \tag{15.28}$$

и верна формула (15.25), то по определению *t*-распределения Стьюдента верна формула (15.27).

Аналогично,

$$\frac{b_0-\beta_0}{\sqrt{\frac{\sigma^2}{n}}}\frac{1}{\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{\sigma^2(n-2)}}} = \frac{b_0-\beta_0}{\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n(n-2)}}} \to t_{n-2}. \tag{15.29}$$

Формулы (15.27)–(15.29) можно использовать для проверки гипотез о значимости коэффициентов (например, $H_0: b_1 = 0$ — коэффициент регрессии статистически не значим), а также при построении доверительных интервалов для коэффициентов регрессии.

Полагая, что верна гипотеза $H_0: b_1 = 0$ (или, в силу линейной связи, $H_0: r = 0$), преобразуем формулу (15.27), учитывая формулы (15.33)–(15.35):

$$\frac{b_1}{\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{(n-2)\sum(x-\overline{X})^2}}} = r\frac{\sigma_y}{\sigma_x}\sqrt{\frac{(n-2)\sum(x-\overline{X})^2}{\sum(y_i-\hat{y}_i)^2}} = r\sqrt{\frac{\sum(y_i-\bar{y})^2}{n}\frac{1}{\frac{\sum(x-\overline{X})^2}{n}}\frac{(n-2)\sum(x-\overline{X})^2}{\sum(y_i-\hat{y}_i)^2}} =$$

$$= r\sqrt{\frac{\sum(y_i-\bar{y})^2}{\sum(y_i-\hat{y}_i)^2}}\sqrt{n-2} = r\sqrt{\frac{n-2}{\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}}} = r\sqrt{\frac{n-2}{\frac{\sum(y_i-\bar{y})^2-\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}}} =$$

$$= r\sqrt{\frac{n-2}{1-\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}}} = r\sqrt{\frac{n-2}{1-r^2}}.$$

Таким образом, имеем

$$\frac{b_1}{\sqrt{\frac{\sigma^2}{\sum(x-\overline{X})^2}}} = r\sqrt{\frac{n-2}{1-r^2}} \to t_{n-2}. \tag{15.30}$$

Следовательно, для оценки значимости коэффициента корреляции ($H_0: r = 0$) следует рассматривать статистику (15.30) исходя из *t*-распределения Стьюдента с $(n-2)$ степенями свободы.

Замечание [74]. 1. Первое изложение элементов метода наименьших квадратов дано А. М. Лежандром в 1806 г. в связи с вопросами вычисления космических орбит. К. Ф. Гаусс дал вероятностное обоснование МНК (1809 г.), разработал вычислительную сторону вопроса (1810 г., 1821 г.).

2. МНК чувствителен к ошибкам округления, поэтому при больших значениях наблюдений X переходят к новой переменной $\dot{X} = X - \overline{X}$ (а иногда и $\dot{Y} = Y - \overline{Y}$). При этом лучше иметь нечетное число измерений, так как в случае равных интервалов между последовательными значениями X, сумма нечетных степеней $\dot{X}$ становится равной 0. ■

Дисперсионный анализ. После построения уравнения регрессии возникает вопрос о качестве решения. Пусть при исследовании n пар наблюдений (x_i, y_i) получено уравнение регрессии *Y* на *X*.

Рассмотрим тождество

$$(y_i - \hat{y}_i) = (y_i - \bar{y}) - (\hat{y}_i - \bar{y}). \quad (15.31)$$

Геометрически это тождество можно проиллюстрировать рисунком 15.12.

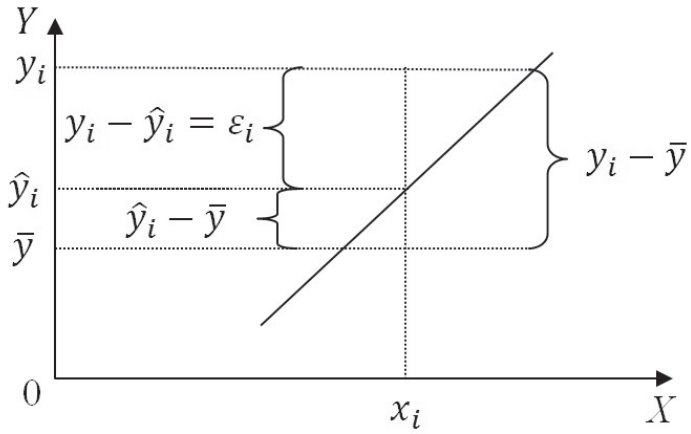


Рис. 15.12 — Геометрическая интерпретация отклонений

Видно, что остаток $\varepsilon_i = (y_i - \hat{y}_i)$ представляет собой разность наблюдаемого значения от общего среднего $(y_i - \bar{y})$ и отклонений предсказанного значения от того же общего среднего.

Если переписать (15.31) в виде

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i), \quad (15.32)$$

возвести обе части в квадрат и просуммировать по *i*, то получим

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2. \quad (15.33)$$

$2\sum(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 2\sum(\hat{y}_i - \bar{y})\sum(y_i - \hat{y}_i) = 0,$

так как в силу свойства средней арифметической $\sum(\hat{y}_i - \bar{y}) = 0.$

Уравнение (15.33) представляется в виде сумм квадратов отклонений:

$$SS_o = SS_{\text{регр}} + SS_z. \quad (15.34)$$

$Sum\ of\ Squares\ (SS)$ — сумма квадратов, где:

$SS_o = \sum(y_i - \bar{y})^2$ — общая сумма квадратов отклонений наблюдаемых значений результативного признака от общего среднего значения, или сумма квадратов отклонений относительно среднего наблюдений;

$SS_{\text{регр}} = \sum(\hat{y}_i - \bar{y})^2$ — сумма квадратов отклонений предсказанных значений результативного признака, найденных по уравнению регрессии, от общего среднего, или сумма квадратов, обусловленная регрессией;

$SS_z = \sum(y_i - \hat{y}_i)^2$ — сумма квадратов отклонений наблюдаемых значений от предсказанных по уравнению регрессии или сумма квадратов, относительно регрессии (остаточная сумма квадратов).

Из формулы (15.33) получим коэффициент (индекс) детерминации:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}. \quad (15.35)$$

В случае парной линейной зависимости $R^2 = r^2$, которая показывает долю вариации переменной y, обусловленную вариацией переменной x (уравнением регрессии).

Адекватность линии регрессии зависит от того, какая часть суммы квадратов относительно среднего обусловлена суммой квадратов относительной регрессии, а какая — суммой квадратов обусловленной регрессии. Суммы квадратов связаны с некоторым числом — числом их степеней свободы. Это число показывает, сколько независимых элементов информации (из n чисел $y_1, y_2, \ldots, y_n$) необходимо для образования данной суммы квадратов.

Например, для $\sum(y_i - \bar{y})^2$, $k = (n-1)$. Действительно, из n разностей $(y_1 - \bar{y}), (y_2 - \bar{y}), \ldots, (y_n - \bar{y})$ только $(n-1)$ независимы (или, иначе, для образования рассматриваемой суммы из $y_1, y_2, \ldots, y_n$ достаточно $(n-1)$ значение, так как оставшееся можно определить, зная $\bar{y}$).

Аналогично для $\sum(\hat{y}_i - \bar{y})^2$, $k = 1$; для $\sum(y_i - \hat{y}_i)^2$, $k = n-2$. Таким образом, каждой сумме квадратов соответствует собственное число степеней свободы. Аналогично основному уравнению дисперсионного анализа можно записать

$$k_o = k_{\text{регр}} + k_{\text{ост}}, \quad (15.36)$$

где $k_{\text{о}} = n - 1$, $k_{\text{регр}} = p - 1$, $k_{\text{ост}} = n - p$, p — число оцениваемых параметров при переменных, в случае парной линейной регрессии $p = 1$.

Для построения таблицы дисперсионного анализа необходимо определить средние квадраты (или MS $(Mean\ Square)$), для этого каждая сумма SS делится на соответствующие число степеней свободы k:

$$MS_R = s^2_{\text{регр}} = \frac{SS_{\text{регр}}}{k_{\text{регр}}} =; \;\; s^2_{\text{ост}} = \frac{SS_{\text{ост}}}{k_{\text{ост}}}. \quad (15.37)$$

Известно, что если в парном уравнении регрессии коэффициент $b_1 = 0$, то величина

$$F = \frac{MS_R}{s^2_{\text{ост}}}$$

имеет распределение Фишера с $(k_1 = 1, k_2 = n-2)$ степенями свободы.

Действительно, при выполнении гипотезы H_0: $b_1 = 0$ в силу формул (15.25) и (15.35)

$$\frac{\sum(\hat{y}_i - \bar{y})^2}{\sigma^2} \to \chi^2_1, \;\; \frac{\sum(y_i - \hat{y}_i)^2}{\sigma^2} \to \chi^2_{n-2}.$$

Отсюда, по определению *F*-распределения Фишера — Снедекора, имеем

$$F = \frac{\frac{\sum(\hat{y}_i - \bar{y})^2}{\sigma^2}}{\frac{\sum(y_i - \hat{y}_i)^2}{\sigma^2(n-2)}} = \frac{\chi^2_1/1}{\chi^2_{n-2}/(n-2)} \to F(1, n-2). \quad (15.38)$$

Этот факт используется для проверки гипотезы H_0: $b_1 = 0$ с уровнем значимости α, против альтернативы H_1: $b_1 \neq 0$.

Обобщим все в таблице дисперсионного анализа (табл. 15.3).

Таблица 15.3

Схема дисперсионного анализа

Источник вариации	Число степеней свободы, k	Сумма квадратов, SS	Средний квадрат, s^2	$F_н$	$F_{кр}$
Обусловленный регрессией	1	$\sum(\hat{y}_i - \bar{y})^2$	$s^2_{регр}$	$\frac{s^2_{регр}}{s^2_{ост}}$	$F_\alpha(1, n-2)$
Относительно регрессии (остаток)	$n-2$	$\sum(y_i - \hat{y}_i)^2$	$s^2_{ост} = \frac{SS_{ост}}{n-2}$		
Общий, скорректированный на среднее Y	$n-1$	$\sum(y_i - \bar{y})^2$			

Наблюдаемое значение $F-$критерия находится по формуле

$$F_н = \frac{s^2_{регр}}{s^2_{ост}} = \frac{\sum(\hat{y}_i - \bar{y}_i)^2}{\sum(y_i - \hat{y}_i)^2}(n-2). \qquad (15.39)$$

Если $F_{н.} > F_{кр.}$, при заданном уровне значимости α и соответствующих числах степеней свободы, то гипотеза H_0: $b_1 = 0$ отбрасывается с риском ошибиться не более чем в 100% случаев и уравнение регрессии считается статистически значимым. Если $F_{н.} < F_{кр.}$, то нулевая гипотеза принимается и уравнение регрессии считается статистически не значимым.

Анализ графического изображения наблюдений и знание природы изучаемого процесса может привести к нелинейной относительно переменной X функции регрессии. В этом случае для оценки параметров регрессии можно использовать системы уравнений, представленные в таблице 15.4. Многие парные нелинейные относительно переменной X уравнения регрессии можно свести к парным линейным путем соответствующих преобразований переменных и параметров исходных уравнений регрессии (табл. 15.5).

Таблица 15.4

Системы нормальных уравнений для определения параметров основных функций

№ п/п	Функция	Система нормальных уравнений
1	$y = a + bx$	$an + b\sum x = \sum y$ $a\sum x + b\sum x^2 = \sum(xy)$
2	$lgy = a + bx$ или $y = 10^{a+bx}$	$an + b\sum x = \sum lgy$ $a\sum x + b\sum x^2 = \sum(xlgy)$
3	$y = a + b\,lgx$	$an + b\sum lgx = \sum y$ $a\sum lgx + b\sum(lgx)^2 = \sum(ylgx)$
4	$lgy = a + b\,lgx$ или $y = 10^{a+blgx} = 10^a x^b$	$an + b\sum lgx = \sum lgy$ $a\sum lgx + b\sum(lgx)^2 = \sum(lgxlgy)$
5	$y = ab^x$ или $lgy = lga + x\,lgb$	$nlga + lgb\sum x = \sum lgy$ $lga\sum x + lg\,b\sum x^2 = \sum(lgxlgy)$

Продолжение табл. 15.4

№ п/п	Функция	Система нормальных уравнений
6	$y = a + bx + cx^2$	$an + b\sum x + c\sum x^2 = \sum y$ $a\sum x + b\sum x^2 + c\sum x^3 = \sum(xy)$ $a\sum x^2 + b\sum x^3 + c\sum x^4 = \sum(x^2y)$
7	$y = a + bx + c\sqrt{x}$	$an + b\sum x + c\sum\sqrt{x} = \sum y$ $a\sum\sqrt{x} + b\sum x^2 + c\sum\sqrt{x^3} = \sum(xy)$ $a\sum\sqrt{x} + b\sum\sqrt{x^3} + c\sum x = \sum\sqrt{xy}$
8	$y = ab^x c^{x^2}$ или $lgy = lga + xlgb + x^2lgc$	$nlga + lgb\sum x + lgc\sum x^2 = \sum lgy$ $lga\sum x + lg\,b\sum x^2 + lgc\sum x^3 = \sum(xlgy)$ $lga\sum x^2 + lg\,b\sum x^3 + lgc\sum x^4 =$ $= \sum(x^2lgy)$

Таблица 15.5

Линеаризующие преобразования функций

№ п/п	Функция	Линеаризующие преобразования			
		переменных X и Y		выражения для величин a и b	
		y'	x'	a'	b'
1	$y = a + b/x$	y	$1/x$	a	b
2	$y = 1/(a + bx)$	$1/y$	x	a	b
3	$y = x/(a + bx)$	x/y	x	a	b
4	$y = ab^x$	$\lg y$	x	$\lg a$	$\lg b$
5	$y = ae^{bx}$	$\ln y$	x	$\ln a$	b
6	$y = 1/(a + be^{-x})$	$1/y$	e^{-x}	a	b
7	$y = ax^b$	$\lg y$	$\lg x$	$\lg a$	b
8	$y = a + b\lg x$	y	$\lg x$	a	b
9	$y = a/(b + x)$	$1/y$	x	b/a	$1/a$
10	$y = ax/(b + x)$	$1/y$	$1/x$	b/a	$1/a$
11	$y = ae^{b/x}$	$\ln y$	$1/x$	$\ln a$	b
12	$y = a + bx^n$	y	x^n	a	b

Пример 15.2. По совокупности сельскохозяйственных организаций имеются выборочные данные о среднегодовой стоимости основных средств на 1 га сельскохозяйственных угодий, тыс. руб. (X) и выручке от реализации продукции (работ, услуг) на 1 га сельскохозяйственных угодий, тыс. руб. (Y).

№ п/п	1	2	3	4	5	6	7	8
Y	56,0	70,1	39,2	61,9	46,7	60,1	50,7	55,3
X	49,2	80,0	22,5	60,3	32,7	44,9	54,8	68,0
№ п/п	9	10	11	12	13	14	15	16
Y	40,0	80,9	87,0	39,9	53,7	67,6	75,6	44,5
X	45,8	67,4	95,4	13,7	37,2	64,0	74,7	33,1

Требуется следующее.

1. Построить график зависимости между переменными, по которому необходимо подобрать вид уравнения регрессии.

2. Рассчитать параметры уравнения регрессии методом наименьших квадратов.

3. Оценить качество уравнения с помощью средней ошибки аппроксимации.

4. Найти коэффициент эластичности.

5. Оценить тесноту связи между переменными с помощью показателей корреляции и детерминации.

6. Оценить значимость коэффициентов корреляции и регрессии по *t*-критерию Стьюдента при уровне значимости $\alpha = 0{,}05$.

7. Охарактеризовать статистическую надежность результатов регрессионного анализа с использованием *F*-критерия Фишера при уровне значимости $\alpha = 0{,}05$.

8. Определить прогнозное значение результативного признака, если возможное значение факторного признака составит 1,5 от его среднего уровня по совокупности.

Решение. 1. График зависимости переменных *X* и *Y* строится в прямоугольной системе координат. На оси абсцисс откладываются значения факторного признака *X*, а по оси ординат — результативного признака *Y*.

На график наносятся точки, координаты которых соответствуют значениям *X* и *Y* (рис. 15.13). Характер расположения точек на графике показывает, что связь между переменными может выражаться линейным уравнением регрессии:

$$\hat{y}_x = b_0 + b_1 x.$$

2. Параметры уравнения регрессии находим методом наименьших квадратов путем составления и решения системы нормальных уравнений (15.18). Для проведения всех расчетов строится вспомогательная таблица 15.6.

В таблице все средние находятся по формуле средней арифметической простой: $\overline{X} = \sum x : n$.

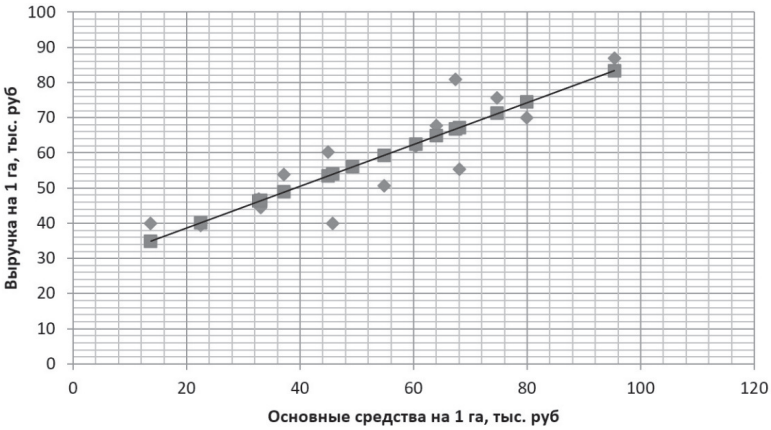


Рис. 15.13 — Зависимость выручки от реализации продукции от фондооснащенности

Таблица 15.6

Вспомогательная таблица регрессионного анализа

№ п/п	y	x	x^2	y^2	xy	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$	$\bar{A}$
1	56,0	49,2	2420,64	3136,00	2755,20	55,98	0,02	0,0005	0,04
2	70,1	80,0	6400,00	4914,01	5608,00	74,28	–4,18	17,4361	5,96
3	39,2	22,5	506,25	1536,64	882,00	40,11	–0,91	0,8359	2,33
4	61,9	60,3	3636,09	3831,61	3732,57	62,57	–0,67	0,4511	1,09
5	46,7	32,7	1069,29	2180,89	1527,09	46,17	0,53	0,2764	1,13
6	60,1	44,9	2016,01	3612,01	2698,49	53,42	6,68	44,5907	11,11
7	50,7	54,8	3003,04	2570,49	2778,36	59,30	–8,60	74,0300	16,97
8	55,3	68,0	4624,00	3058,09	3760,4	67,15	–11,85	140,3355	21,42
9	40,0	45,8	2097,64	1600,00	1832	53,96	–13,96	194,7998	34,89
10	80,9	67,4	4542,76	6544,81	5452,66	66,79	14,11	199,0960	17,44
11	87,0	95,4	9101,16	7569,00	8299,8	83,42	3,58	12,7809	4,11
12	39,9	13,7	187,69	1592,01	546,63	34,89	5,01	25,1390	12,57
13	53,7	37,2	1383,84	2883,69	1997,64	48,85	4,85	23,5446	9,04
14	67,6	64,0	4096,00	4569,76	4326,4	64,77	2,83	8,0096	4,19
15	75,6	74,7	5580,09	5715,36	5647,32	71,13	4,47	20,0089	5,92
16	44,5	33,1	1095,61	1980,25	1472,95	46,41	–1,91	3,6552	4,30
Итого	929,2	843,7	51760,11	57294,62	53317,51	929,2	–	764,9904	152,49
Среднее значение	58,08	52,73	3235,01	3581	3332,34	–	–	–	9,53

Подставим полученные суммы в систему уравнений, учитывая, что n=16.

$$\begin{cases} 929{,}2 = 16b_0 + 843{,}7b_1, \\ 53317{,}51 = 843{,}7b_0 + 51760{,}11b_1. \end{cases}$$

Решив систему, получим $b_0 = 26{,}747;\ b_1 = 0{,}594.$

Параметры уравнения регрессии также можно найти по формулам, вытекающим из системы нормальных уравнений:

$$b_1 = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - (\overline{X})^2} = \frac{3332{,}34 - 52{,}73 \cdot 58{,}08}{3235{,}01 - (52{,}73)^2} = 0{,}594,$$

$$b_0 = \overline{Y} - b_1\overline{X} = 58{,}08 - 0{,}594 \cdot 52{,}73 = 26{,}747.$$

Небольшие расхождения в результатах расчетов могут происходить за счет округления средних значений во втором случае.

Таким образом, уравнение регрессии имеет вид

$$\hat{y}_x = 26{,}747 + 0{,}594x.$$

Коэффициент регрессии показывает, что при увеличении стоимости основных средств на 1 га сельскохозяйственных угодий на 1 тыс. руб. выручка от реализации продукции на 1 га сельскохозяйственных угодий в среднем увеличивается на 594 руб.

Если в уравнение регрессии подставить фактические значения переменной X, то определяются возможные (теоретические) значения переменной $\hat{y}$, которые наносятся на график в виде уравнения прямой (рис. 15.13).

3. Качество уравнения регрессии оценивается с помощью средней ошибки аппроксимации:

$$\overline{A} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%, \quad \overline{A} = \frac{152{,}49}{16} \cdot 100\% = 9{,}53\%.$$

Значит, фактические значения выручки от реализации на 1 га сельскохозяйственных угодий от расчетных значений по уравнению регрессии в среднем различаются на 9,53%.

Качество уравнения регрессии считается хорошим, если ошибка аппроксимации не превышает 8–10%. Полученное уравнение регрессии можно оценить как вполне хорошее.

4. При линейной форме связи средний коэффициент эластичности составит

$$Э = 0{,}594\frac{52{,}7}{58{,}1} = 0{,}539.$$

Коэффициент эластичности показывает, что при увеличении фондооснащенности организации на 1% выручка от реализации продукции на 1 га сельхозугодий в среднем возрастает на 0,539%.

5. При линейной зависимости теснота связи между переменными X и Y определяется с помощью коэффициента корреляции:

$$r_{xy} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x \sigma_y},$$

где σ_x и σ_y — средние квадратические отклонения по X и Y.

Используя результаты таблицы 15.6, получим

$$\sigma_x = \sqrt{\overline{x^2} - (\overline{x})^2} = \sqrt{3235{,}007 - 52{,}731^2} = 21{,}317,$$

$$\sigma_y = \sqrt{\overline{y^2} - (\overline{y})^2} = \sqrt{3580{,}914 - 58{,}075^2} = 14{,}429,$$

$$r_{xy} = \frac{3332{,}344 - 58{,}075 \cdot 52{,}731}{21{,}317 \cdot 14{,}429} \approx 0{,}878.$$

Так как значение коэффициента корреляции близко к единице, то между признаками связь очень тесная, прямая, близкая к линейной функциональной.

Коэффициент детерминации $r^2 = 0{,}878^2 = 0{,}771$ показывает, что 77,1% различий в выручке от реализации продукции (работ, услуг) на 1 га сельскохозяйственных угодий (Y) объясняется вариацией среднегодовой стоимости основных средств на 1 га сельскохозяйственных угодий (X), а 22,9% — другими, не учтенными факторами.

6. Так как исходные данные являются выборочными, то необходимо оценить существенность или значимость величины коэффициента корреляции.

Выдвигаем нулевую гипотезу: коэффициент корреляции в генеральной совокупности равен нулю и изучаемый фактор не оказывает существенного влияния на результативный признак. $H_0: r = 0$, при $H_0: r \neq 0$.

Для проверки нулевой гипотезы применим *t*-критерий Стьюдента.

Найдем наблюдаемое значение *t*-критерия

$$t_{н.} = |r| \cdot \sqrt{\frac{n-2}{1-r^2}} = |0{,}878| \cdot \sqrt{\frac{16-2}{1-0{,}878^2}} = 6{,}86.$$

Критическое значение *t* находится по таблице *t*-распределения Стьюдента при уровне значимости α = 0,05 и числе степеней свободы $k = n - 2 = 16 - 2 = 14$ для двухсторонней критической области, $t_{кр.} = 2{,}15$. Сравниваем $t_{н.}$ с $t_{кр.}$ Так как $t_{н.} > t_{кр.}$, то нулевая гипотеза отвергается, коэффициент корреляции существенно отличен от нуля в генеральной совокупности. Значит, среднегодовая стоимость основных средств на 1 га сельскохозяйственных угодий (X) оказывает статистически существенное влияние на выручку от реализации продукции (работ, услуг) на 1 га сельскохозяйственных угодий (Y).

Статистическая значимость коэффициента регрессии также проводится с использованием *t*-критерия Стьюдента.

Находится наблюдаемое значение критерия $t_{набл} = \frac{b_1}{m_{b_1}}$, где m_{b_1} — средняя ошибка коэффициента b_1 вычисляется по формуле

$$m_{b_1} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{(n-2)\sum(x_i - \bar{X})^2}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{(n-2)\sigma_x^2 n}}; \qquad (15.40)$$

$$m_{b_1} = \sqrt{\frac{764{,}99}{14 \cdot 454{,}4221 \cdot 16}} = 0{,}0867.$$

Следовательно, $t_{н.} = \frac{0{,}594}{0{,}0867} = 6{,}85$.

Критическое значение *t* также равно 2,15. Так как $t_{н.} > t_{кр.}$, то коэффициент регрессии статистически значим. Подтверждается вывод о значимости влияния фондообеспеченности на выручку от реализации продукции на 1 га сельскохозяйственных угодий.

7. Статистическая надежность уравнения регрессии проверяется с использованием F-критерия Фишера.

Наблюдаемое значение критерия находится по формуле

$$F_{н.} = \frac{\frac{\sum(\hat{y}_i - \bar{y}_i)^2}{p}}{\frac{\sum(y_i - \hat{y}_i)^2}{n - p - 1}}, \qquad (15.41)$$

где p — число параметров при переменных X.

Если применяется линейное уравнение регрессии, то расчет $F_{н.}$ упрощается.

$$F_{н.} = \frac{r^2}{1-r^2}(n-2) = \frac{0{,}771}{1 - 0{,}771}(16-2) = 47{,}14.$$

По таблице приложения 4 находится критическое значение F-критерия при числе степеней свободы $k_1 = p = 1, k_2 = n - p - 1 = 14$ и уровне значимости α = 0,05:

$$F_{кр.} = F_{0{,}05}(k_1 = 1, k_2 = 14) = 4{,}60.$$

Так как $F_{\text{н.}} > F_{\text{кр.}}$, то уравнение регрессии статистически значимое или надежное.

При парной линейной зависимости оценка значимости всего уравнения, коэффициентов корреляции и регрессии дает одинаковые результаты, так как $t_{b_1}^2 = t_r^2 = F_\alpha(k_1, k_2)$.

8. Прогнозное значение результативного признака определяется путем подстановки в уравнение регрессии прогнозного или возможного значения факторного признака (x_p).

По условию $x_p = 52{,}731 \cdot 1{,}5 = 79{,}05$, значит, при среднегодовой стоимости основных средств на 1 га сельскохозяйственных угодий 79,05 тыс. руб. прогнозное значение выручки от реализации продукции (работ, услуг) на 1 га сельскохозяйственных угодий составит 73,7 тыс. руб., так как

$$\hat{y}_x = b_0 + b_1 x = 26{,}747 + 0{,}594 \cdot 79{,}05 = 73{,}7.$$

1. *Интервальные оценки*. Предположение о нормальном законе распределения двух переменных X и Y опирается на двумерный нормальный закон распределения вида (15.9). Как было указано ранее, выборочный коэффициент корреляции определяет ориентацию эллипса рассеяния с центром (a, b), который строится на основе предположения двумерного нормального распределения двух переменных и представляет собой предсказанный интервал для нового одиночного наблюдения при заданном числе наблюдений n.

Пусть

$$y = \beta_0 + \beta_1 x \text{ —}$$

теоретическое уравнение регрессии, а

$$\hat{y} = b_0 + b_1 x \text{ —}$$

его оценка, полученная, например, с помощью метода наименьших квадратов, тогда говорят о доверительных интервалах параметров регрессии и самого уравнения регрессии.

Доверительный интервал для коэффициента регрессии:

$$b_1 - t_\gamma \, \text{s} \, \frac{1}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \le \beta_1 \le b_1 + t_\gamma \, \text{s} \frac{1}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}, \qquad (15.42)$$

где

$$s^2 = \sum (y_i - \hat{y}_i)^2 / (n - 2) \text{ —}$$

остаточная дисперсия ошибок ($\varepsilon_i = y_i - \hat{y}_i$) на одну степень свободы, t_γ — критическое значение *t*-распределения Стьюдента с доверительной вероятностью $\gamma = 1 - \alpha$ и числом степеней свободы для двусторонней области $k = n - 2$.

При заданном значении $x = x_0$ рассматривается доверительный интервал с доверительной надежностью $\gamma = 1 - \alpha$ для пронозируемого теоретического значения $\tilde{y}$ согласно уравнению регрессии:

$$\tilde{y} \in \left[(b_0 + b_1 x_0) \pm t_\gamma \, \text{s} \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right]. \qquad (15.43)$$

По мере удаления x_0 от среднего значения ширина доверительного интервала увеличивается. Средства визуализации пакета *Statistica* позволяют с использованием матричного графика рассмотреть гистограммы всех пар переменных, рассматриваемых в анализе, регрессионные зависимости регрессии Y на X и X на Y, коэффициенты корреляции, детерминации с уровнем значимости, эллипсы рассеяния и доверительные границы уравнения регрессии. Все перечисленные результаты по данным примера 15.1 отражены на рисунке 15.14.

2. *Исследование остатков.* Остатки — это n разностей вида

$$\varepsilon_i = y_i - \hat{y}_i, \text{ где } i = 1, 2, \ldots, n.$$

Если модель корректна, то остатки ε_i — это то, что нельзя объяснить с помощью уравнения регрессии. В случае корректности модели остатки рассматриваются как наблюдаемые ошибки.

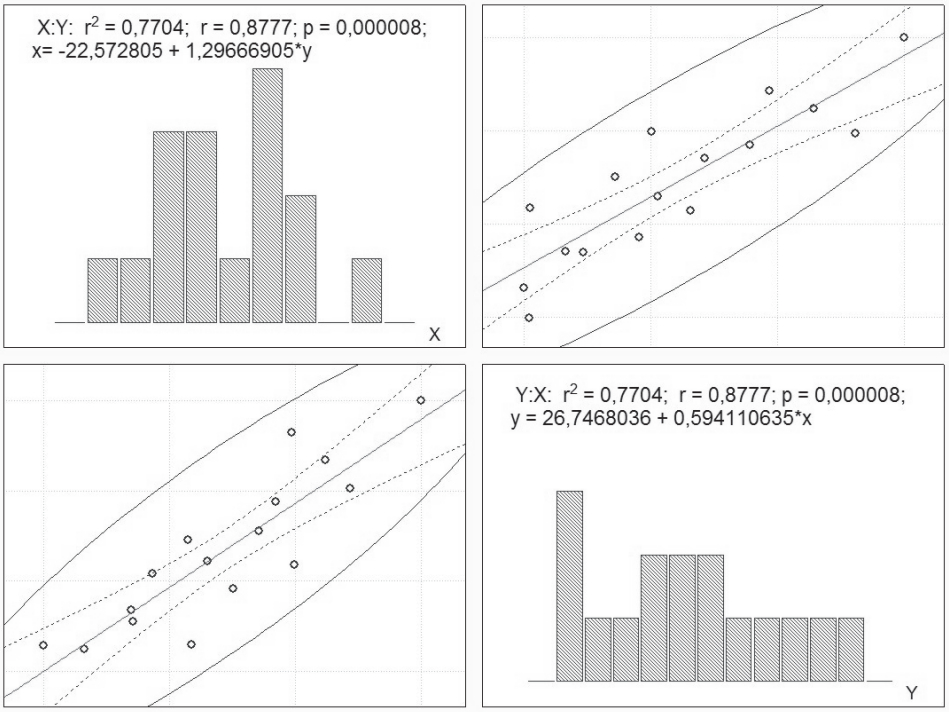


Рис. 15.14 — Матричный график

Обычные предположения исследователя относительно ошибок состоят в том, что ошибки независимы, имеют нулевые средние, постоянные дисперсии и подчиняются нормальному закону распределения. Корректность модели предполагает выполнение сделанных предположений. Если анализ остатков приводит к выводу, что предположения нарушены, то требуется проверить корректность данных, а возможно изменить вид модели, которая, однако, строится на основе теоретических предположений.

Важность исследования остатков подчеркивается цитатой известного астронома С. Ф. Гершеля в классической книге «Прикладной регрессионный анализ» Н. Дрейпера и Г. Смита [38]: «Почти все величайшие открытия в астрономии были созданы путем рассмотрения того, что мы ранее называли количественными или качественными остаточными феноменами. Иначе говоря, они вытекают из анализа той части числовых или качественных результатов наблюдения, которая остается необъясненной после выделения и учета всего того, что согласуется со строгим применением известных методов».

Традиционным методом изучения остатков является построение графиков. Рассматривают гистограммы, нормальные графики. В последнем случае остатки ранжируются и рассчитываются стандартизированные значения статистики u (по формуле $u = \frac{\varepsilon_i - \bar{\varepsilon}}{s}$), исходя из предположения о нормальности распределения остатков и откладываются на вертикальной оси. На горизонтальной оси откладываются наблюдаемые значения, если остатки подчиняются нормальному закону распределения, то точки лежат близко к прямой, иллюстрирующей нормальный закон распределения, в противном случае необходимо преобразование переменных (например, логарифмирование). В случае полунормального графика на оси OY рассматриваются только положительные значения нормального закона распределения.

Для примера 15.1 нормальный вероятностный график остатков позволяет утверждать, что остатки могут подчиняться нормальному закону распределения (рис. 15.15), так как точки, иллюстрирующие остатки, лежат близко к прямой.

Можно показать, что коэффициент корреляции между остатками и $\hat{y}_i$ равен нулю (рис. 15.16), поэтому наклон графика указывает на имеющийся дефект в данных (ошибки, артефакты).

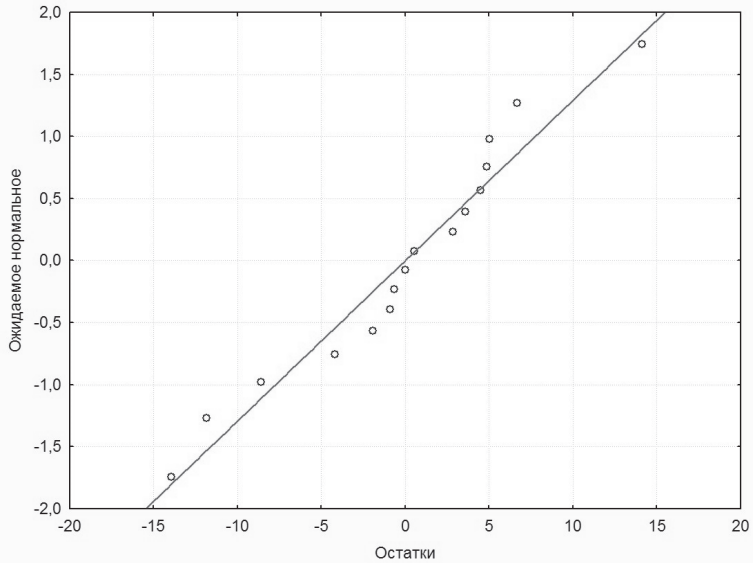


Рис. 15.15 — Нормальный вероятностный график остатков

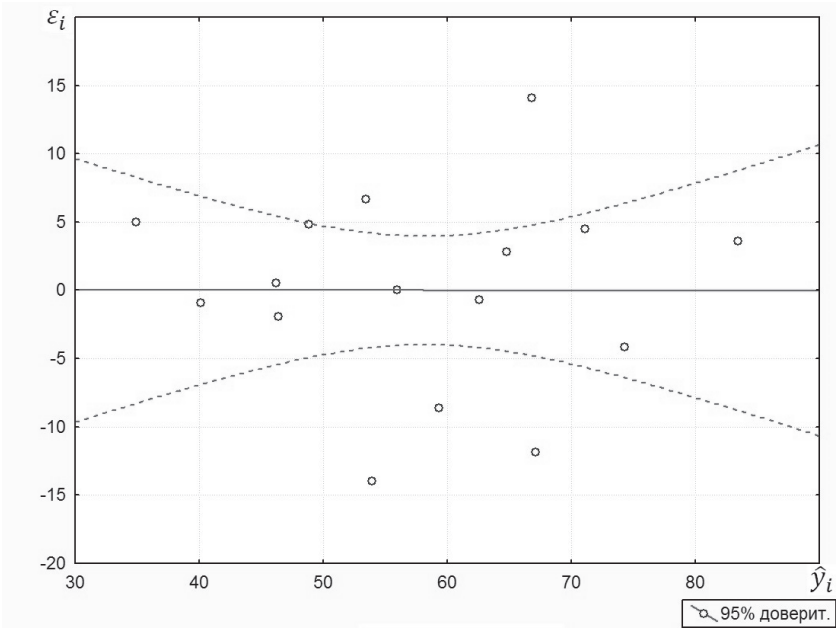


Рис. 15.16 — График предсказанных значений и остатков

График остатков также позволяет сделать вывод о несоответствии вида модели регрессии, как видно из рисунка 15.17, парабола второй степени лучше подходит к данным, чем линейная модель, что отразится на графике остатков (рис. 15.18).

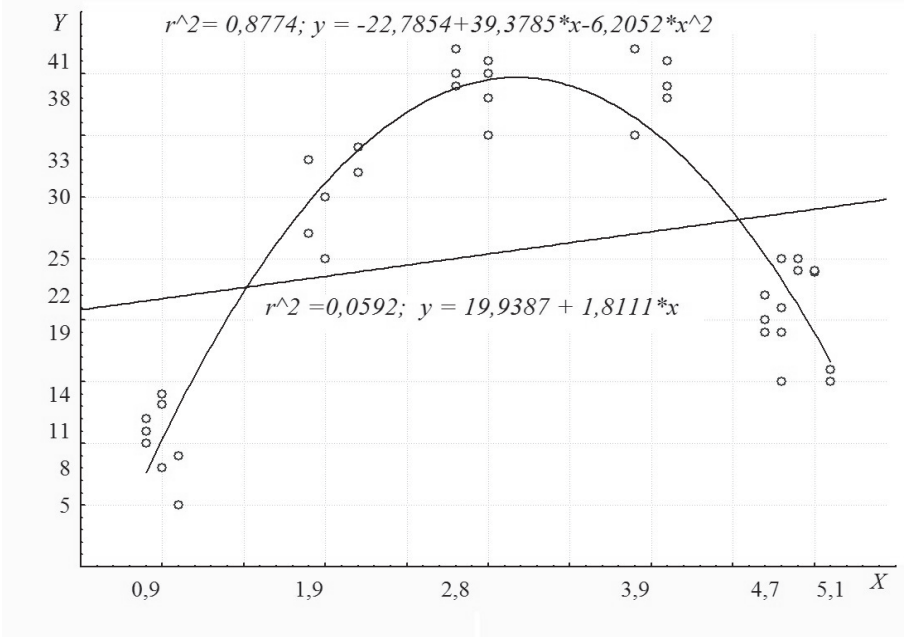


Рис. 15.17 — Пример неадекватной подгонки

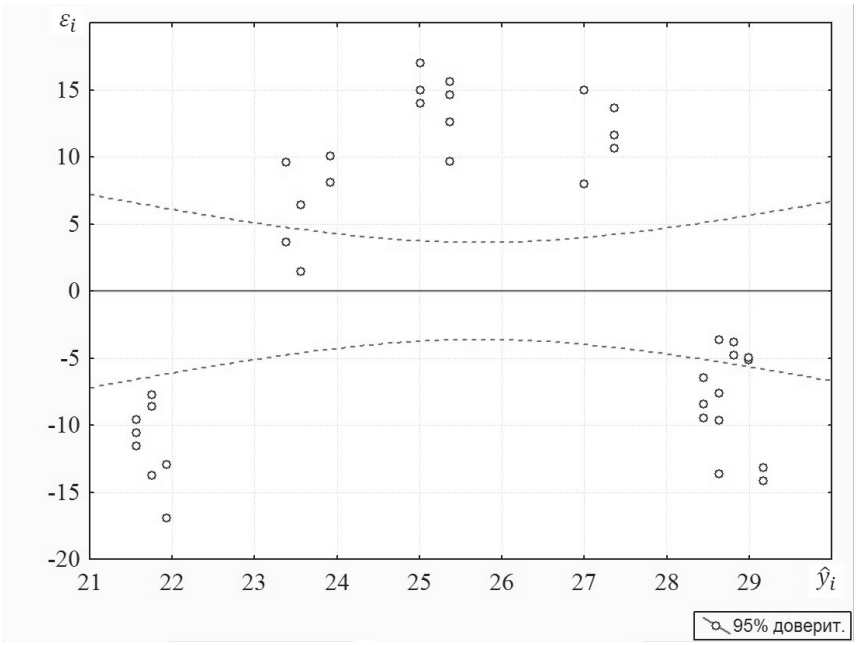


Рис. 15.18 — График предсказанных значений и остатков к линейной регрессии на рисунке 15.17

На графиках, иллюстрирующих распределение остатков, необходимо обращать внимание на *выбросы* — наблюдения, отклонения для которых резко отличаются от остальной совокупности отклонений, они могут представлять собой ошибки наблюдений или артефакты (искусственно созданные наблюдения, например, в записи второго значения результативного фактора примера 15.1, человек не поставил запятую и вместо 70,1 получилось 701 и т. п.).

Пример 15.3 [74]. На плоскости XOY задано множество точек $(x_i, y_i), i = \overline{1,n}$.

Требуется провести прямую $y = a + bx$ так, чтобы сумма квадратов расстояний точек (x_i, y_i) от этой прямой была минимальна (ортогональная регрессия).

Решение. Подобная задача возникает при описании неровностей профиля математико-статистическими методами, а также в более сложной форме она связана с методами многомерного статистического анализа, в частности, методом главных компонент, суть которого также в поиске «главных моментов инерции» системы точек. Следуя [74], в упрощенной форме интерпретируем механически задачу нахождения средней линии случайного профиля. Пусть в точках (x_i, y_i) расположены единичные массы, будем искать ось вращения, доставляющей данной системе наименьший момент инерции (наименьшую кинетическую энергию вращения при заданной угловой скорости). Из курса механики известно, что такая ось существует и проходит через центр тяжести $O'(\bar{x}, \bar{y})$, перпендикулярная ей ось будет давать для той же системы точек максимальный момент инерции. Таким образом можно указанные оси принять за новую систему координат. Найдем угол наклона α новой системы координат к старой, начало отсчета которой в точке $O'(\bar{x}, \bar{y})$ (рис. 15.19).

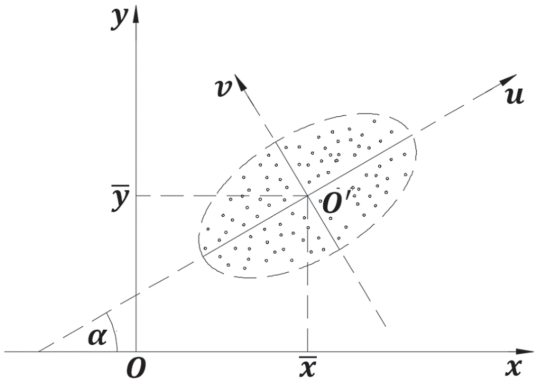


Рис. 15.19 — Ортогональная регрессия

Если (u_i, v_i) — координаты точек (x_i, y_i) в новой системе координат, то искомая прямая удовлетворяет условию

$$S(\alpha) = \sum_{i=1}^{n} v_i^2, \tag{15.44}$$

следовательно,

$$\sum_{i=1}^{n} v_i \frac{dv_i}{d\alpha} = 0. \tag{15.45}$$

Из курса аналитической геометрии известны формулы преобразования координат:

$$\begin{cases} u_i = x_i' cos\alpha + y_i' sin\alpha, \\ v_i = -x_i' sin\alpha + y_i' cos\alpha, \end{cases} \tag{15.46}$$

где $x_i' = x_i - \bar{x}, y_i' = y_i - \bar{y}$.

Отсюда

$$\frac{dv_i}{d\alpha} = (-x_i' sin\alpha + y_i' cos\alpha)_\alpha' = -(x_i' cos\alpha + y_i' sin\alpha) = -u_i,$$

значит,

$$\sum_{i=1}^{n} v_i \frac{dv_i}{d\alpha} = -\sum_{i=1}^{n} v_i u_i = 0.$$

По формулам преобразования координат (15.37) получим

$$\sum_{i=1}^{n} v_i u_i = \sum_{i=1}^{n} (x_i' cos\alpha + y_i' sin\alpha)(-x_i' sin\alpha + y_i' cos\alpha) = 0$$

или

$$(cos^2\alpha - sin^2\alpha) \sum_{i=1}^{n} (x_i' y_i') - sin\alpha\, cos\alpha \sum_{i=1}^{n} (x_i'^2 - y_i'^2) = 0.$$

Отсюда

$$tg\, 2\alpha = \frac{2\sum_{i=1}^{n} (x_i' y_i')}{\sum_{i=1}^{n} (x_i'^2 - y_i'^2)} = \frac{2\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2 - \sum_{i=1}^{n} (y_i - \bar{y})^2)}. \tag{15.47}$$

Полученные результаты позволяют построить *прямую ортогональной регрессии*, удовлетворяющую условиям задачи и проходящую через точку $O'(\bar{x}, \bar{y})$, под углом α — одним из четырех, найденных из формулы (15.47) и соответствующему направлению «наибольшей дисперсии» — наибольшему диаметру эллипса, ограничивающего корреляционное облако (рис. 15.19).

Следует отметить, что главная идея этой задачи, апеллирующая к механике, имеет далекие следствия, например, в многомерном статистическом анализе (метод главных компонент, факторный анализ).

15.4. Множественный корреляционно-регрессионный анализ

I. Многочисленные исследования показывают, что результативная переменная обычно зависит не от одной переменной *X*, а от совокупности переменных, состоящей из p величин, в виде вектора

$$X = (X_1, X_2, \dots, X_p).$$

Если уравнение нелинейное, то (если это возможно) оно вначале приводится к линейному. Параметры линейного уравнения множественной регрессии вида

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

находятся методом наименьших квадратов, для чего строится и решается следующая система нормальных уравнений:

$$\begin{cases} \sum y = nb_0 + b_1 \sum x_1 + b_2 \sum x_2 + \cdots + b_p \sum x_p, \\ \sum y x_1 = b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 \dots + b_p \sum x_1 x_p, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \sum y\, x_p = b_0 \sum x_p + b_1 \sum x_1 x_p + b_2 \sum x_2 x_p \dots + b_p \sum x_p^2. \end{cases} \quad (15.48)$$

Множественный коэффициент регрессии b_j показывает, на сколько единиц изменяется в среднем результативный признак Y, если j-й факторный признак *X* увеличить на единицу, при условии, что все другие факторы в линейной модели закреплены на постоянном, обычно среднем, уровне.

Уравнение множественной регрессии может быть построено в стандартизованном масштабе, когда единицей измерения признаков принимается их среднее квадратическое отклонение:

$$t_y = \beta_1 t_{x_1} + \beta t_{x_2} + \cdots + \beta_p\, t_{x_p} + \varepsilon,$$

где t — стандартизованные переменные:

$$t_y = \frac{y-\bar{y}}{\sigma_y},\ t_{x_1} = \frac{x_1-\bar{x}_1}{\sigma_{x_1}},\ \dots,\ t_{x_p} = \frac{x_p-\bar{x}_p}{\sigma_{x_p}},$$

β_j — стандартизованные коэффициенты регрессии (следует отличать их от параметров истинной модели), σ — среднее квадратическое отклонение.

Параметры множественного уравнения регрессии могут быть найдены на основе матрицы парных коэффициентов корреляции, тогда определяется уравнение регрессии в стандартизованном масштабе из системы нормальных уравнений:

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 r_{x_1x_2} + \beta_3 r_{x_1x_3} + \cdots + \beta_p r_{x_1x_p}, \\ r_{yx_2} = \beta_1 r_{x_2x_1} + \beta_2 + \beta_3 r_{x_2x_3} + \cdots + \beta_p r_{x_2x_p}, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ r_{yx_p} = \beta_1 r_{x_1x_p} + \beta_2 r_{x_2x_p} + \beta_3 r_{x_3x_p} + \cdots + \beta_p. \end{cases} \quad (15.49)$$

В общем случае решение системы (15.49) можно представить с использованием формул Крамера, для случая двух независимых переменных решения имеют вид

$$\beta_1 = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{1 - r_{x_1x_2}^2},\ \beta_2 = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{1 - r_{x_1x_2}^2}. \quad (15.50)$$

Стандартизованные коэффициенты регрессии показывают, на сколько средних квадратических отклонений изменяется зависимая переменная Y, при увеличении независимой переменной X_j на одно среднее квадратическое отклонение, при неизменном среднем уровне всех других факторов. β-коэффициенты можно сравнивать между собой, так как они являются величинами относительными. Чем больше величина β-коэффициента по данному фактору, тем больше его влияние по сравнению с другими факторами. Коэффициенты b и β связаны между собой формулами

$$b_j = \beta_j \frac{\sigma_y}{\sigma_{x_j}},\ \beta_j = b_j \frac{\sigma_{x_j}}{\sigma_y}. \qquad (15.51)$$

Для оценки тесноты связи между признаками применяются парные, частные и множественные коэффициенты (индексы) корреляции и детерминации. Обычно изучение зависимости проводится по выборочным данным, поэтому оценивается значимость коэффициентов регрессии, корреляции и всего уравнения множественной регрессии в целом.

Если в модели несколько факторных переменных, то доля суммы квадратов, объясняемая регрессией, называется множественным коэффициентом детерминации (квадратом множественного коэффициента корреляции R):

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2},\ 0 \le R \le 1. \qquad (15.52)$$

Значимость R^2 определяется по F −критерию:

$$F_{\text{кр.}}(k_1 = p, k_2 = n - p - 1), F_{\text{набл.}} = \frac{R^2}{(1-R^2)} : \frac{p}{n-p-1}.$$

Если $F_{\text{набл.}} > F_{\text{кр.}}$, то гипотезу H_0: $R^2 = 0$ отвергают и связь между факторными переменными и результативной переменной считают статистически значимой.

Если Y зависит только от одной переменной X, то $R = r$ — парному коэффициенту корреляции.

В экономических исследованиях корреляционно-регрессионный анализ проводится в следующей последовательности.

1. Исходя из целей и задач исследования зависимости, устанавливается результативный (y) и факторные (x_j) признаки.

2. По совокупности объектов определяются значения результативного и факторных признаков.

3. Обосновывается, для случая парной зависимости, обычно графическим методом, модель уравнения регрессии.

4. Методом наименьших квадратов определяются параметры уравнения регрессии.

5. Определяется теснота связи между изучаемыми признаками.

6. Оценивается значимость уравнения связи, его параметров и показателей тесноты связи.

Пример 15.4. Исследовать влияние продуктивности коров и производительности труда работников животноводства на себестоимость производства молока в сельскохозяйственных организациях. Результативным признаком (Y) является себестоимость производства 1 ц молока.

Факторные признаки: x_1 — среднегодовой надой молока на корову, характеризующий уровень продуктивности коров; x_2 — прямые затраты труда на 1 ц молока, выражающий уровень трудоемкости производства и являющимся обратным показателем производительности труда. Исходные данные представлены в таблице 15.7. Известны также парные коэффициенты корреляции:

$$r_{yx_1} = -0{,}7905;\ r_{yx_2} = 0{,}6270;\ r_{x_1x_2} = -0{,}5812.$$

Требуется определить:

1) параметры множественного уравнения регрессии в натуральной и стандартизованной форме;

2) средние коэффициенты эластичности для каждого фактора;

3) коэффициенты частной и множественной корреляции;

4) общий и частные *F*-критерии Фишера.

Решение. 1. Линейное уравнение множественной регрессии в натуральной форме имеет вид

$$y = b_0 + b_1x_1 + b_2x_2,$$

а в стандартизованной

$$t_y = \beta_1 t_{x_1} + \beta_1 t_{x_2}.$$

Найдем β-коэффициенты, используя парные коэффициенты корреляции.

$$\beta_1 = \frac{r_{yx_1} - r_{yx_2}r_{x_1x_2}}{1 - r^2_{x_1x_2}} = \frac{-0{,}7905 - 0{,}6270 \cdot (-0{,}5812)}{1 - 0{,}5812^2} = -0{,}6435;$$

$$\beta_2 = \frac{r_{yx_2} - r_{yx_1}r_{x_1x_2}}{1 - r^2_{x_1x_2}} = \frac{0{,}6270 - (-0{,}7905) \cdot (-0{,}5812)}{1 - 0{,}5812^2} = 0{,}2530.$$

Таблица 15.7

Средние значения и колеблемость изучаемых признаков

Признак	Обозначение переменной	Среднее значение	Среднее квадратическое отклонение
Себестоимость 1 ц молока, руб.	*Y*	1823,36	369,54
Надой молока на корову за год, ц	X_1	59,85	14,75
Прямые затраты труда на 1 ц молока, чел.-ч	X_2	2,14	1,06

Линейное уравнение множественной регрессии в стандартизованном масштабе имеет вид

$$t_y = -0{,}6435t_{x_1} + 0{,}2530t_{x_2}.$$

По абсолютной величине β-коэффициентов можно сделать вывод об относительной силе влияния факторов на изменение результативного признака. На себестоимость производства молока более сильное влияние оказывает надой молока на корову, а влияние трудоемкости производства молока оказывает значительно меньшее влияние.

Для определения параметров множественного уравнения регрессии в натуральной форме воспользуемся формулами:

$$b_j = \beta_j \frac{\sigma_y}{\sigma_{x_j}},\ b_0 = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2.$$

$$b_1 = \beta_1 \frac{\sigma_y}{\sigma_{x_1}} = -0{,}6435 \cdot \frac{369{,}54}{14{,}75} = -16{,}1220;$$

$$b_2 = \beta_2 \frac{\sigma_y}{\sigma_{x_2}} = 0{,}253 \cdot \frac{369{,}54}{1{,}06} = 88{,}2015;$$

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 = 1823{,}36 - (-16{,}122)\cdot 59{,}85 - 88{,}2015\cdot 2{,}14 = 2599{,}5105.$$

Получим линейное уравнение множественной регрессии

$$\hat{y} = 2599{,}5105 - 16{,}122\, x_1 + 88{,}2015\, x_2.$$

Коэффициенты множественной регрессии показывают, что при увеличении среднегодового надоя молока на корову на 1 ц себестоимость молока в среднем снижается на 16,12 руб., исключив влияние трудоемкости производства, а при увеличении прямых затрат труда на 1 ц молока на 1 чел.-ч себестоимость производства молока в среднем увеличивается на 88,20 руб., исключив влияние продуктивности коров.

2. Средние коэффициенты эластичности находятся по формуле:

$$Э_{yx_j} = b_j \frac{\overline{X}_j}{\overline{Y}};$$

$$Э_{yx_1} = b_1 \frac{\overline{X}_1}{\overline{Y}} = -16{,}122 \cdot \frac{59{,}85}{1823{,}36} = -0{,}529;$$

$$Э_{yx_2} = b_2 \frac{\overline{X}_2}{\overline{Y}} = 88{,}2015 \cdot \frac{2{,}14}{1823{,}36} = 0{,}104.$$

Значит, при увеличении среднегодового надоя молока на корову на 1%, себестоимость производства молока снижается в среднем на 0,529%, при исключении влияния трудоемкости. Если увеличить трудоемкость производства молока на 1%, то себестоимость молока в среднем увеличивается на 0,104%, при исключении влияния продуктивности коров.

3. Коэффициенты частной корреляции определяются через парные коэффициенты корреляции по формулам:

$$r_{yx_1\cdot x_2} = \frac{r_{yx_1} - r_{yx_2}r_{x_1x_2}}{(1-r^2_{yx_2})(1-r^2_{x_1x_2})} = \frac{-0{,}7905 - 0{,}6270\cdot(-0{,}5812)}{\sqrt{(1-0{,}627^2)(1-0{,}5812^2)}} = -0{,}6722;$$

$$r_{yx_2\cdot x_1} = \frac{r_{yx_2} - r_{yx_1}r_{x_1x_2}}{(1-r^2_{yx_1})(1-r^2_{x_1x_2})} = \frac{0{,}627 - (-0{,}7905)\cdot(-0{,}5812)}{\sqrt{(1-0{,}7905^2)(1-0{,}5812^2)}} = 0{,}3363;$$

$$r_{x_1x_2\cdot y} = \frac{r_{x_1x_2} - r_{yx_1}r_{yx_2}}{(1-r^2_{yx_1})(1-r^2_{x_1x_2})} = \frac{-0{,}5812 - (-0{,}7905)\cdot(0{,}627)}{\sqrt{(1-0{,}7905^2)(1-0{,}627^2)}} = -0{,}1804.$$

Коэффициенты частной корреляции характеризуют тесноту связи между двумя переменными, при исключении влияния третьей переменной. Значит, связь между себестоимостью производства молока и продуктивностью коров обратная и тесная, исключив влияние трудоемкости. Связь между трудоемкостью и себестоимостью молока довольно слабая при исключении влияния продуктивности. Связь между факторами X_1 и X_2 очень слабая.

Коэффициент множественной корреляции находится по формуле

$$R_{yx_1x_2} = \sqrt{\beta_1 r_{yx_1} + \beta_2 r_{yx_2}} = \sqrt{(-0{,}6435)\cdot(-0{,}7905) + 0{,}253\cdot 0{,}627} = \sqrt{0{,}5087 + 0{,}1586} = \sqrt{0{,}6673} = 0{,}8169;\ R^2 = 0{,}6673.$$

Величина коэффициента множественной корреляции показывает, что связь между *Y, X*1 и *X*2 очень тесная. Множественное уравнение регрессии объясняет

66,7% вариации себестоимости производства молока, в том числе влиянием продуктивности коров объясняется 50,9% вариации себестоимости, а трудоемкостью производства — 15,8%.

4. Оценим значимость уравнения регрессии и коэффициента R^2 с помощью критерия Фишера. Наблюдаемое или фактическое значение критерия находится по формуле

$$F_{\text{н.}} = \frac{R^2_{yx_1x_2}}{1-R^2_{yx_1x_2}} : \frac{p}{n-p-1},$$

где p — число факторов в линейном уравнении регрессии; n — число единиц наблюдения.

$$F_{\text{н.}} = \frac{0{,}6673}{1-0{,}6673} : \frac{2}{48-2-1} = 45{,}13.$$

При уровне значимости α=0,05 и числе степеней свободы $k_1 = p = 2$, $k_2 = n - p - 1 = 48 - 2 - 1 = 45$ по таблице значений F-критерия Фишера критическое значения $F_{\text{кр.}} = 3{,}23$. Сравниваем $F_{\text{н.}}$ с $F_{\text{кр.}}$.

Так как $F_{\text{н.}} > F_{\text{кр.}}$, то нулевую гипотезу о незначимости величины R^2 отклоним, т. е. уравнение множественной регрессии и R^2 статистически значимы.

В уравнении множественной регрессии не все факторы могут оказывать статистически существенное влияние на изменение результативного признака. Оценка значимости факторов в уравнении регрессии может быть дана с помощью частного F-критерия или t−критерия Стьюдента.

$$F_{\text{н}\,x_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1-R^2_{yx_1x_2}} \cdot \frac{n-p-1}{1} = \frac{0{,}6673-0{,}627^2}{1-0{,}6673} \cdot \frac{48-2-1}{1} = 37{,}08.$$

При $\alpha = 0{,}05, k_1 = 1, k_2 = 45, F_{\text{кр}} = 4{,}08$. Так как $F_{\text{н}\,x_1} > F_{\text{кр}}$, то в уравнение регрессии целесообразно включение фактора X_1 после X_2. Фактор X_1 оказывает статистически значимое влияние на Y.

$$F_{\text{н}\,x_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1-R^2_{yx_1x_2}} \cdot \frac{n-p-1}{1} = \frac{0{,}6673-0{,}7905^2}{1-0{,}6673} \cdot \frac{48-2-1}{1} = 5{,}74.$$

$F_{\text{кр.}} = 4{,}08$. Так как $F_{\text{н}\,x_2} > F_{\text{кр.}}$, то это свидетельствует о статистической значимости влияния фактора X_2 и целесообразности включения его в уравнение множественной регрессии. В данной задаче на уровень себестоимости производства молока статистически значимое влияние оказывает как продуктивность коров, так и трудоемкость производства продукции.

II. Рассмотрим модели множественного регрессионного анализа для изучения объектов, которые можно представить моделью типа «черный ящик» (рис. 15.20).

На рисунке 15.20 (вектор обозначаем «жирно»):

$X = (X', Z)$ — факторы (вектор входных переменных);

X' — управляемые, независимые переменные;

Z — контролируемые, но неуправляемые факторы, изменение которых, часто прогнозируемо;

Y — «отклик» («показатель качества управления», «выход»), $f(y)$ — закон распределения, $M(Y)$ — математическое ожидание случайной величины Y; W — помехи.

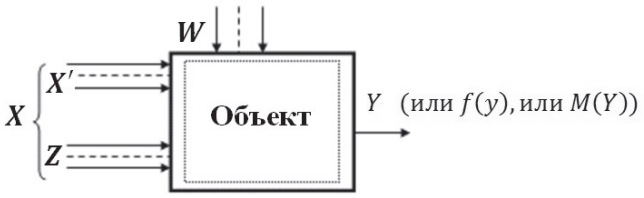


Рис. 15.20 — Модель объекта

Исследование объекта может быть пассивным, когда фиксируются «естественные» значения X и соответствующие им значения Y (пассивный эксперимент), и активным, когда осуществляются целенаправленные изменения X (активный эксперимент). В первом случае вопросы организации сбора данных не являются первостепенными и чем их больше, тем лучше. Обработка результатов пассивного эксперимента ведется методами «классического» регрессионного анализа. Во втором случае имеем дело с планированием эксперимента и соответствующими специальными формами регрессионного анализа. Планирование эксперимента, если то позволяет объект исследования, существенно эффективнее пассивного эксперимента в смысле минимизации числа опытов и точности получаемых выводов. Но для таких объектов, как, например, социально-экономические системы, активный эксперимент применим лишь в редких случаях. В основном он используется в точных науках, при изучении технологических объектов, в сельскохозяйственном эксперименте и др., т. е. только там, где изменения параметров X на границах допуска не приводят к фатальным последствиям и где физически возможно и экономически не дорого «управлять» факторами X.

Регрессионная модель представляет собою математическое выражение, связывающее входные переменные X с одним «выходом» Y. Поэтому в реальных условиях, когда результат функционирования объекта должен характеризоваться несколькими показателями Y, возникает отдельная проблема выбора единственного показателя, наиболее полно характеризующего особенности изучаемого объекта. Хотя эта задача практически очень важна, она выходит за рамки математической теории регрессионного анализа и в данном случае мы не будем ее рассматривать.

По результатам наблюдений определяется уравнение регрессии. В случае парной регрессии подобрать модель и оценить ее адекватность легко визуально (рис. 15.17).

Уравнение регрессии — это зависимость случайной величины Y от неслучайных факторов X, т. е. зависимость «следствия» Y от «причин» X:

$$Y = \eta(\mathrm{X}, \beta) + \varepsilon, \qquad (15.53)$$

где $X = (X_1, X_2, \dots, X_j, \dots, X_p)$ — вектор факторов, $j = 1, 2, \dots, p$;

$\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_j, \dots, \beta_p)$ — вектор параметров модели;

$\eta(\mathrm{X}, \beta)$ — функция регрессии (или функция отклика) случайной величины Y на неслучайные X;

$\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_i, \dots, \varepsilon_n)$ — вектор ошибок наблюдений.

При многократном однотипном воздействии X на входе получаем на выходе объекта различные значения Y. То есть определенному значению $X_i = (x_{i1}, x_{i2}, \dots, x_{ig}, \dots, x_{im})$ соответствует некоторое распределение $f(y_i), i = 1, 2, \dots, n$ случайной величины Y (рис. 15.20), где $i = 1, 2, \dots, n$ (n — число всех опытов или объем выборки), $l = 1, 2, \dots, m_i$ (m_i — число повторных или «параллельных» опытов, встречающихся в выборке). Для случая парной регрессии повторные опыты представлены на рисунке 15.17.

Уравнение регрессии вида (15.53) описывает только статику объекта, т. е. предполагается, что взаимосвязь показателя Y и факторов X, установленная в определенный момент (интервал) времени от времени не зависит (т. е. параметры модели (15.53) не зависят от времени).

Регрессионные модели, в зависимости от рассматриваемых факторов, могут быть использованы в целях: объяснения сути явления (предсказательная модель), прогнозирования (прогнозная модель), управления. Если установлена зависимость Y только от управляемых факторов X', то это уравнение теоретически может быть использовано в целях управления объектом (заметим, что при построении уравнения по результатам пассивного эксперимента ошибка в управлении может быть неприемлемой). Функциональная модель и модель для прогнозирования содержит все группы факторов X', Z (рис. 15.20):

$$Y = \eta(X', Z, \beta) + \varepsilon.$$

Обычно функциональная модель более сложная, чем предсказательная.

Чаще всего исследователь не знает вида функции отклика η, его цель — найти ее, в простейшем случае — найти параметры $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_j, \dots, \beta_p)$ (параметры истиной модели следует отличать от β_i — стандартизованных коэффициентов регрессии).

В целях построения $\eta(X, \beta)$ обычно предполагают, что это — гладкая функция в области допустимых значений: $X \in X_{\text{доп}}$. В этом случае возможно ее разложение в ряд Тейлора в окрестности некоторой точки, например, точки, соответствующей «центру» эксперимента — среднему значению $\bar{X}$. В результате получаем полином степени p вида:

$$Y = \beta_0 + \sum_j \beta_j x_j + \sum_{uj} \beta_{uj} x_u x_j + \sum_j \beta_{jj} x_j^2 + \dots + \varepsilon, \qquad (15.54)$$

где $j = 1, 2, \dots, p$, $\sum_{uj}$ — сумма парных взаимодействий $x_u\, x_j$, $u, j = 1, 2, \dots, k$, $u \neq j$, p — число факторов; β_{uj} — коэффициент парного взаимодействия, β_{jj} — коэффициент при квадрате переменной и т. д.; в формуле (15.54) степень полинома равна двум.

Обычно разложение ограничивают конечным числом членов ряда. Например:

$$Y = \beta_0 + \sum_j \beta_j x_j. \qquad (15.55)$$

Выражение (15.55) — это линейная регрессия.

Модели типа (15.54) имеют общий характер, они линейны по параметрам β_j и фактически сводятся к виду (15.55), поэтому подобные модели называют линейными.

По результатам эксперимента могут быть определены не «истинные» коэффициенты регрессии β, соответствующие генеральной совокупности, а лишь их оценки $B = (b_0, b_1, \dots, b_j, \dots, b_p)$, вычисленные по выборке объемом n.

При выводе и использовании формул регрессионного анализа пользуются векторной формой представления уравнений регрессии:

$$Y = X\beta + \varepsilon; \quad \hat{Y} = XB, \tag{15.56}$$

где Y — n-мерный вектор наблюдений результативной переменной; X — матрица размерности $n \times (p+1)$ наблюдаемых значений факторных признаков $X_1, X_2, \ldots, X_p$ значений независимых переменных, первый столбец матрицы X содержит переменную $x_{i0} = 1, i = 1, 2, \ldots, n$ и позволяет учесть b_0; β, B — векторы размерности $(p+1)$ коэффициентов и их оценок соответственно; ε — вектор ошибок размерности n, отклонений значений результативного признака y_j от значений $\hat{y}_j$, найденных по уравнению $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}$; $\hat{Y}$ — предсказанные (прогнозируемые) значения выходной величины.

Представим исходные данные в матричной форме:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1\ x_{11} x_{12}\ \ldots\ x_{1p} \\ 1\ x_{21} x_{22}\ \ldots\ x_{2p} \\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots \\ 1\ x_{n1} x_{n2}\ \ldots\ x_{np} \end{bmatrix}, \mathrm{B} = \begin{bmatrix} b_0 \\ b_1 \\ \ldots \\ b_p \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{bmatrix}. \tag{15.57}$$

Тогда функция ошибок (потерь) $S = \sum_{i=1}^{n} \varepsilon_i^2$, представляется как произведение вектора-строки $\varepsilon^{\mathrm{T}} = (\varepsilon_1, \varepsilon_2 \ldots \varepsilon_n)$ на вектор-столбец ε.

Из формулы (15.56) видно, что $\varepsilon = Y - XB$. Значит,

$$S = \varepsilon^T \varepsilon = (Y - XB)^T (Y - XB) = Y^T Y - B^T X^T Y - Y^T XB + B^T X^T XB,$$

$$S = Y^T Y - 2B^T X^T Y + B^T X^T XB. \tag{15.58}$$

Вектор столбец частных производных по всем параметрам b_j из (15.58) в матричном виде записывается в следующем виде:

$$\frac{\partial S}{\partial B} = -2X^T Y + 2(X^T X)B = 0, \tag{15.59}$$

значит, $X^T Y = X^T XB$.

Умножив обе части уравнения слева на матрицу $(X^T X)^{-1}$, обратную матрице плана эксперимента $(X^T X)$, получим

$$(X^T X)^{-1}(X^T Y) = (X^T X)^{-1}(X^T X)B,$$

где $(X^T X)^{-1}(X^T X) = E$, где E — единичная матрица размерности $(p+1)$.

Отсюда оценки вектора B по методу наименьших квадратов имеют вид

$$B = (X^T X)^{-1} X^T Y. \tag{15.60}$$

Замечание. Анализ и построение зависимостей — одна из основных проблем прикладной статистики и науки вообще. В классической постановке проблема сводится к оценке по результатам наблюдений некоторой линейной модели

$$y_i = \beta_0 + \sum_j \beta_j x_{ij}, (i=1, 2, \ldots, n), \tag{15.61}$$

где y_i — n случайных величин (наблюдаемые входные переменные), являющихся линейными комбинациями x_{ji} с p неизвестными постоянными $\beta_1, \beta_2, \ldots, \beta_p$ плюс ошибки $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$; $\{x_{ij}\}$ — известные значения наблюдений (постоянные коэффициенты).

Указанная модель имеет общий характер и в зависимости от значений $\{x_{ij}\}$ может описывать три различные схемы:

а) если $x_{ij} = \{0;1\}$, то это модель дисперсионного анализа;

б) если x_{ij} пробегает непрерывное множество значений (например, время t, температура T), то это модель регрессионного анализа;

в) если x_{ij} совмещает переменные а) и б), то это модель ковариационного анализа.

Входные переменные x_j можно разделить на ненаблюдаемые (латентные) и наблюдаемые (явные). Среди последних выделяют контролируемые и управляемые переменные. Посредством управляемых переменных проводятся активные (планируемые) эксперименты. Контролируемые переменные позволяют провести только пассивный эксперимент, однако часто их поведение возможно прогнозировать. ■

Используя процедуры регрессионного анализа, следует иметь в виду, что полученным результатам можно доверять, если выполняются постулаты регрессионного анализа, перечисленные ниже.

1. Результаты эксперимента должны быть свободны от систематических ошибок, т. е. математическое ожидание $M(Y)$ величины Y должно быть равно действительному значению $\tilde{Y}$ ($\tilde{Y} = \beta_0 + \sum_j \beta_j x_j$), т. е.:

$$M(Y) = \tilde{Y},$$
$$\tilde{Y} = \left(\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_i, \ldots, \tilde{Y}_n\right).$$

Следовательно, математическое ожидание ошибки ε будет равно нулю:

$$M(\varepsilon) = M\left(Y - \tilde{Y}\right) = 0,$$

или, если действительным значением считать предсказанное по уравнению регрессии значение $\hat{Y}$, то:

$$M(\varepsilon) = M\left(Y - \hat{Y}\right) = 0.$$

2. Дисперсия выходной величины Y постоянна и не зависит от величины Y_i, $i = 1, 2, \ldots, n$, т. е.

$$D(Y_i) = \sigma^2, D(\varepsilon_i) = \sigma^2, \text{для всех } i = \overline{1, n}.$$

3. Результаты наблюдений Y_i в разных точках эксперимента независимы и не коррелированны, т. е. Y_{i-1} и Y_i — не коррелированы, так что ковариации равны нулю:

$$cov(Y_{i-1}, Y_i) = M\left\{\left(Y_{i-1} - \hat{Y}_{i-1}\right)\left(Y_i - \hat{Y}_i\right)\right\} = 0;$$
$$cov(\varepsilon_{i-1}, \varepsilon_i) = M(\varepsilon_{i-1}\varepsilon_i) = 0.$$

4. Y_i, ε_i — случайные величины, подчиненные нормальному закону распределения со средними $M\ (Y) = \tilde{Y}$, $M\ (\varepsilon_i) = 0$, дисперсиями $D\left(\tilde{Y}_i\right) = D(\varepsilon_i) = \sigma^2$, т. е.

$$Y_i \to N(\tilde{Y}_i, \sigma^2),$$
$$\varepsilon_i \to N(0, \sigma^2),$$

где $N(a, \sigma^2)$ — обозначение нормальных распределений наблюдаемой величины Y и ее ошибки ε с математическим ожиданием a и дисперсией σ^2.

5. Входные переменные X_j — независимы, неслучайны, измеряются без ошибок.

Проиллюстрируем постулаты на рисунке 15.21.

Выбор структуры уравнения наилучшей регрессии (наиболее точно описывающей исследуемый процесс) можно осуществить, используя R^2 — квадрат множественного коэффициента корреляции или дисперсионный анализ. Структура уравнения регрессии усложняется (например, в полиномиальном случае повышается степень многочлена) до тех пор, пока увеличение соответствующего критерия не станет пренебрежительно малым.

Однако вывод о корректности модели по условию $R^2 \approx 1$ не всегда верен. И вот почему. Результата $R^2 \approx 1$ можно добиться, увеличивая число оцениваемых параметров β_j и в случае «насыщенности» $p = n$, коэффициент детерминации R^2 будет равен 1, но модель при этом не обязательно корректна. Особенно вывод опасен при неверном определении N — числа серий повторных опытов, когда проводятся параллельные опыты (пример повторных опытов рассмотрен на рисунке 15.17). С другой стороны, если $\hat{y}_i = \bar{y}$, то есть $b_1 = \cdots = b_p = 0$ или адекватна модель $y = \beta_0 + \varepsilon$, то $R^2 = 0$.

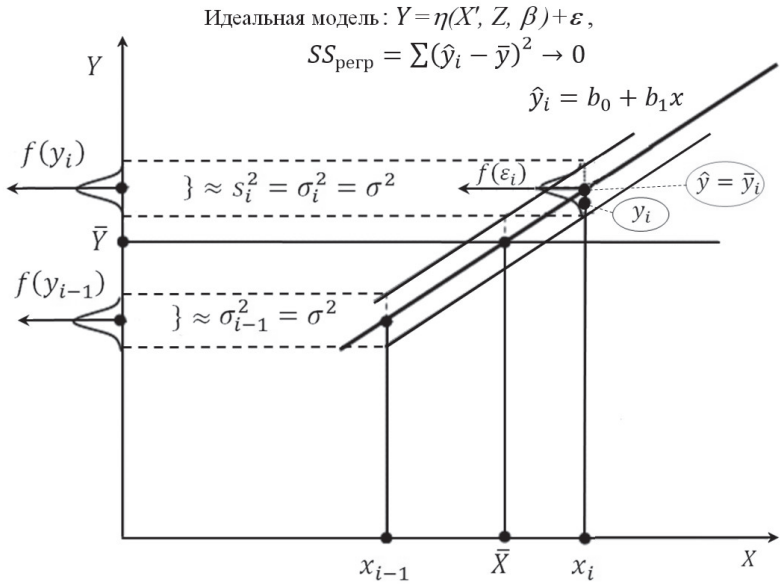


Рис. 15.21 — Графическая иллюстрация постулатов регрессионного анализа

Итак, величина R^2 и суммы квадратов не всегда дают однозначный ответ на вопрос, адекватна ли модель.

При наличии повторных опытов или наблюдений неадекватность подгонки наблюдений с помощью предполагаемой функции регрессии может быть проверена статистически. Для проверки адекватности модели применяют оценки чистой ошибки, определяемые по повторным опытам. Чистая ошибка является оценкой истинной дисперсии σ^2.

Пусть число опытов в каждой точке x_i равно $l = 1, 2, \ldots, m_i$, наблюдаемые значения результативной переменной — y_{il}. Тогда, если предполагаемая линия регрессии адекватно описывает результаты наблюдений, то среднее m_i наблюдений будет лежать близко к расчетному значению ($\bar{y}_i \to \hat{y}_i$) и, следовательно, величина $(\hat{y}_i - \bar{y}_i)^2$ будет мала и велика, в случае плохой аппроксимации. Значит, в случае «идеальной модели» величина

$$SS_{\text{регр}} = \sum_i (\hat{y}_i - \bar{y}_i)^2 \to 0.$$

В случае повторных наблюдений для проверки адекватности модели сумма квадратов ошибок ($SS_{\text{ост}}$) может быть разбита на две части:

$$SS_{\text{ост}} = SS'_{\text{ост}} + S_a. \qquad (15.62)$$

Тогда $SS'_{\text{ост}}$ — вклад в $SS_{\text{ост}}$, связанный с чистой ошибкой, при каждом x_i будет равен

$$s_i^2 = \sum_{l=1}^{m_i}(y_{li} - \bar{y}_i)^2, \tag{15.63}$$

где $\bar{y}_i = \frac{\sum_{l=1}^{m_i} y_{li}}{m_i}, k_{s_i^2} = m_i - 1.$

Общая сумма квадратов s_i^2, то есть чистая ошибка, будет равна

$$SS'_{\text{ост}} = \sum_{i=1}^{N}\sum_{l=1}^{m_i}(y_{li} - \bar{y}_i)^2, \tag{15.64}$$

где N — число серий повторных опытов.

Средний квадрат для чистой ошибки:

$$s_Z^2 = \frac{SS'_{\text{ост}}}{k_Z}, k_Z = \sum_{i=1}^{N}(m_i - 1). \tag{15.65}$$

Откуда дисперсия адекватности (MS_a) находится как разность суммы квадратов ошибок и суммы квадратов, обусловленных чистой ошибкой, деленная на соответствующее число степеней свободы с учетом того, что регрессионная модель имеет p параметров при переменных:

$$S_a = SS_{\text{ост}} - SS'_{\text{ост}} = \sum_{i=1}^{N}\sum_{l=1}^{m_i}(y_{li} - \hat{y}_i)^2 - \sum_{i=1}^{N}\sum_{l=1}^{m_i}(y_{li} - \bar{y}_i)^2, \tag{15.66}$$

$$k_a = k_{SS_{\text{ост}}} - k_{SS'_{\text{ост}}} = n - p - k_Z - 1. \tag{15.67}$$

Таблица дисперсионного анализа для уравнения регрессии имеет следующий вид (табл. 15.8).

Наблюдаемое значение F −критерия находится по формуле $F_{\text{н}} = \frac{MS_a}{s_\varepsilon^2}$.

Если $F_{\text{н}} > F_{\text{кр}}$, то модель не адекватна, где $F_{\text{кр}} = F_\alpha\ (k_1 = k_a, k_2 = k_\varepsilon)$.

Таблица 15.8

Схема дисперсионного анализа

Источник вариации		Число степеней свободы, k	Суммы квадратов, SS	Средние квадраты, MS
Обусловленный регрессией		p	$SS_{\text{регр}} = \sum(\hat{y}_i - \bar{y})^2$	$MS_R = \frac{SS_{\text{регр}}}{p}$
Остаток	неадекватность	k_a	S_a	$MS_a = \frac{S_a}{k_a}$
	чистая ошибка	k_ε	$SS'_{\text{ост}}$	$s_\varepsilon^2 = \frac{SS'_{\text{ост}}}{k_\varepsilon}$
Общий, скорректированный на среднее Y		$n - 1$	$S_o = \sum(y_i - \bar{y})^2$	

Таким образом, мера адекватности F есть оценка среднего разброса относительно линии регрессии, обусловленная случайными причинами. Индекс детерминации в этом случае равен

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{S_a + SS'_{\text{ост}}}{S_o}. \tag{15.68}$$

Из формулы (15.68) следует ряд утверждений.

1. При отсутствии чистой ошибки $S_{\text{ост}}$ уменьшается и R^2 может обманчиво увеличиться.

2. При наличии чистой ошибки и малом значении S_a величина индекса детерминации ограничена сверху $R^2 < 1 - SS'_{\text{ост}}/S_o$, так как модель не может объяснить чистую ошибку.

3. Выбросы влияют как на S_a, так и на $SS'_{\text{ост}}$, и, следовательно, занижают R^2.

4. Рекомендуется не вычислять R^2 для модели без свободного члена, а также не сравнивать его для моделей с разными факторными признаками (не являющимися подмножествами друг друга).

Замечание. 1. *Дисперсионный анализ и F-критерий.*

В подпространстве оценок (сравнить рис. 14.3 и 15.22) $\mathrm{V_r}$ выделим подпространство количественных факторов $\mathrm{X} = \{x_1, x_2, \ldots, x_p\}$ и ортогональный ему вектор 1 (см. формулы (15.56)–(15.60)). Цель — найти параметры модели (15.56) из условия наилучшей аппроксимации вектора y в пространстве натянутом на X и 1. Решение — вектор $\hat{y}$, для которого $Y\hat{Y} \perp (X \bigcup 1)$. Вектор $O\bar{Y}$ — ортогональная проекция вектора y на вектор 1. Вектор $\hat{y}$ — ортогональная проекция вектора y на подпространство $(X \bigcup 1)$. Согласно теореме о трех перпендикулярах $\mathrm{Пр}_1 y = \bar{Y}$ — проекция вектора y на вектор 1 равна $\bar{Y}$. Мы опять получаем теорему Пифагора (рис. 14.3, 15.22) и целый ряд равенств

$$OA^2 + OB^2 = O\hat{Y}^2,$$
$$(OA^2 + OB^2) + Y\hat{Y}^2 = OY^2,$$
$$(y - \bar{Y})^2 = (\hat{Y} - \bar{Y})^2 + (Y - \hat{Y})^2.$$

Последнее равенство в векторной форме иллюстрирует понятие коэффициента (индекса) детерминации R^2 (рис. 15.22):

$$R^2 = \frac{(\hat{Y} - \bar{Y})^2}{(y - \bar{Y})^2} = (cos\varphi)^2,$$

где φ — угол между векторами $(\hat{Y} - \bar{Y})$ и $(y - \bar{Y})$.

Геометрическая интерпретация *F*-статистики отражена в разделе 14.1, посвященном дисперсионному анализу.

2. Важным моментом при построении искомой зависимости является отбор факторов x_j, существенно влияющих на результативную переменную y. Известно достаточно много путей отбора, условно их можно разделить на два класса: формальные и содержательные (семантические или смысловые).

Формальные методы основываются на идее перебора различных уравнений *общей линейной модели* (например, различные модификации пошаговой регрессии с последовательным включением или исключением независимых переменных) до момента достижения некоторого критерия, например F-критерия Фишера (дисперсионный анализ), характеризующего (при заданном уровне значимости α) значимость вклада переменной в регрессию.

В случае нелинейной модели рассматривается подход *обобщенных линейных моделей*, связанный с преобразованием переменной *Y* (аналогично преобразованиям в табл. 15.4), исходя из предположения ее принадлежности распределениям: Пуассона (*лог-линейная регрессия*), биномиальному (*логит-регрессия и пробит-регрессия*), гамма (*анализ выживаемости*).

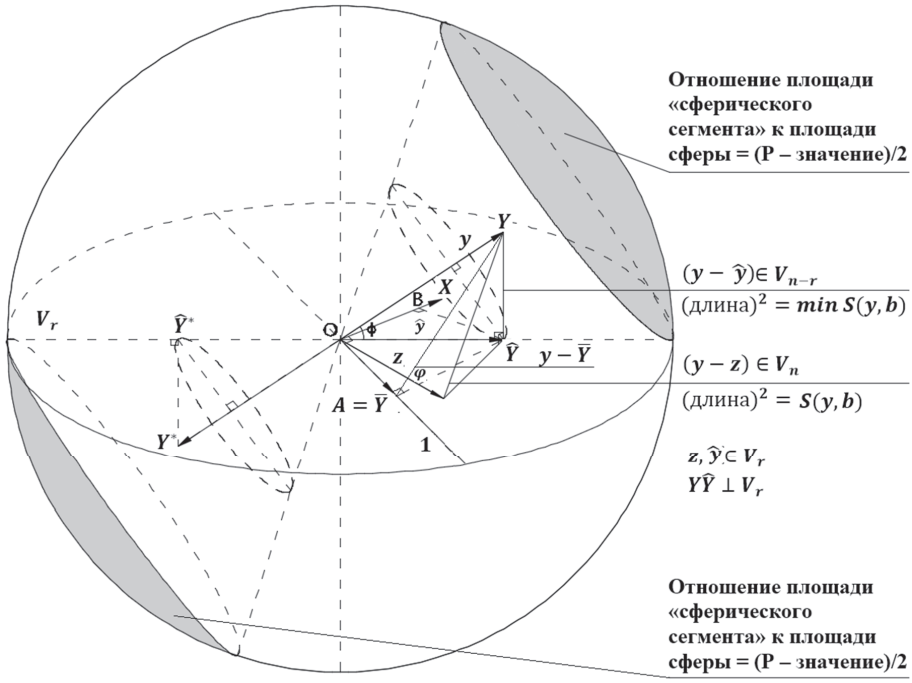


Рис. 15.22 — Геометрическая интерпретация коэффициента R^2 и статистики

Содержательные методы предполагают достижение целей моделирования, при этом, как отмечалось выше, различают виды моделей.

а) Физические модели, описывающие функциональные особенности изучаемых процессов. Построение функциональной модели — редкий случай, так как принципиально невозможно учесть все причинно-следственные связи и их взаимодействия.

б) Модель для управления процессом. Предполагается возможным для любого y_j найти такие x_{ji} (управляющие воздействия), что, задав их в модели, получим требуемое y_j.

в) Модель для предсказания. Дает возможность по заданным x_{ji} определить прогнозируемое y_i.

Наблюдения за функционированием сложного объекта можно представить в виде множества точек некоторого фазового пространства. Тогда физические модели, модели управления и предсказания — это возможные проекции объекта на различные плоскости, поэтому на практике эти модели не совпадают.

Очевидно, что практически более необходимы модели, описывающие содержательные стороны изучаемого процесса. А попытка совмещения формальных и содержательных критериев — это типичная многокритериальная задача, решение которой обычно не однозначно.

3. Альтернативным методом «выбора наилучшей регрессии» является метод группового учета аргумента (МГУА), позволяющий определить в данном классе функций с помощью компьютера оптимальную структуру искомой зави-

симости и идентифицировать параметры по внешним критериям, установленным человеком. В середине XX века Кенуй и Тьюки предложили метод складного ножа (бутстреп-метод), согласно которому вся совокупность данных разбивается на части и проводится статистический анализ всех частей для достижения несмещенности оценок. Этот метод и был положен в основу МГУА. Человек задает среду решения (список переменных и перечень опорных функций), а также критерий селекции (отбора). Для проверки адекватности последовательность наблюдений разбивается на части (последовательности). Первая часть (обучающая последовательность) используется для определения оценок коэффициентов по методу наименьших квадратов. Вторая часть (проверочная последовательность) используется для селекции моделей по внешнему критерию, например построение лучшей прогнозирующей модели.

Сегодня МГУА известен как один из первых методов «глубокого обучения» (см. раздел IV в [53]), популярного в работе с «большими данными».

Автором МГУА является А. Г. Ивахненко. МГУА хорошо проявил себя при решении многих практических задач (прогнозирования, принятия решений и т. д.), где другие методы оказывались не приемлемы, начиная с 1969 г. ■

Темы (вопросы) для самоконтроля

1. Виды и формы связей между признаками.
2. Корреляция, коэффициент корреляции и его свойства.
3. Причинная корреляция.
4. Проверка гипотезы о значимости выборочного коэффициента корреляции.
5. Честные коэффициенты корреляции.
6. Однофакторный регрессионный анализ.
7. Метод наименьших квадратов.
8. Значимость уравнения регрессии.

Глава 16
Анализ временных рядов

Статистические данные в экономике, коммерции, технике и так далее могут рассматриваться не только в пространстве, но и во времени, путем построения и анализа одного или нескольких временных рядов.

Определим *дискретный временной ряд* как последовательность измерений значений переменной (процесса) за определенный период через одинаковые промежутки или моменты времени:

$$Y_1, Y_2, Y_3, \dots, Y_n. \tag{16.1}$$

Последовательные значения результатов наблюдения во времени обычно зависимы. С детерминистской точки зрения временной ряд можно представить как

$$Y_t = f(t) + \varepsilon_t, \tag{16.2}$$

где $t = 1, 2, \dots, n$;

$f(t)$ — гладкая (непрерывная и дифференцируемая) функция, характеризующая долгосрочное движение в зависимости от времени — тренд; Y_t — уровень временного ряда; ε_t — случайный ряд возмущений.

При наличии во временном ряде тенденции и циклических колебаний значения каждого последующего уровня ряда зависят от предыдущих. Корреляционную зависимость между последовательными уровнями временного ряда называют *автокорреляцией уровней ряда*. Количественно ее можно измерить с помощью линейного коэффициента корреляции между уровнями исходного временного ряда и уровнями этого ряда, сдвинутыми на один или несколько периодов или моментов времени, называемого коэффициентом автокорреляции.

Коэффициент автокорреляции уровней ряда первого порядка, смещенных на одну единицу времени, определяется по формуле

$$r_1 = \frac{\sum_{t=2}^{n}(y_t - \bar{y}_1)(y_{t-1} - \bar{y}_2)}{\sqrt{\sum_{t=2}^{n}(y_t - \bar{y}_1)^2 \sum_{t=2}^{n}(y_{t-1} - \bar{y}_2)^2}}, \tag{16.3}$$

где $$\bar{y}_1 = \frac{\sum_{t=2}^{n} y_t}{n-1}, \; \bar{y}_2 = \frac{\sum_{t=2}^{n} y_{t-1}}{n-1}.$$

Коэффициент автокорреляции уровней ряда второго порядка:

$$r_2 = \frac{\sum_{t=3}^{n}(y_t - \bar{y}_3)(y_{t-2} - \bar{y}_4)}{\sqrt{\sum_{t=3}^{n}(y_t - \bar{y}_3)^2 \sum_{t=3}^{n}(y_{t-2} - \bar{y}_4)^2}}, \tag{16.4}$$

где $$\bar{y}_3 = \frac{\sum_{t=3}^{n} y_t}{n-2}, \quad \bar{y}_4 = \frac{\sum_{t=3}^{n} y_{t-2}}{n-2}.$$

Аналогично можно определить коэффициенты автокорреляции более высоких порядков.

Так как коэффициент автокорреляции строится по аналогии с линейным коэффициентом корреляции, то по нему можно судить о наличии линейной или близкой к линейной тенденции. Чем ближе коэффициент автокорреляции первого порядка к единице, тем более выражена линейная тенденция. Для некоторых временных рядов, имеющих сильную нелинейную тенденцию, коэффициент автокорреляции уровней исходного ряда может приближаться к нулю.

Последовательность коэффициентов автокорреляции уровней первого, второго и прочих порядков называют *автокорреляционной функцией временного ряда.* Если наиболее высоким оказался коэффициент автокорреляции первого порядка, исследуемый ряд содержит только тенденцию. Если наиболее высоким оказался коэффициент автокорреляции порядка τ, то временной ряд содержит циклические или сезонные колебания с периодичностью в τ моментов времени. Если ни один коэффициент не является значимым, можно сделать вывод о том, что либо ряд не содержит тенденции и циклических колебаний, либо содержит сильную не линейную тенденцию.

Число периодов или моментов времени, по которым рассчитывается коэффициент автокорреляции, называют *лагом*.

Построение аналитической функции для моделирования тенденции (тренда) временного ряда называют аналитическим выравниванием временного ряда. Тенденция во времени может принимать разные формы, для ее формализации используют функции, рассмотренные в главе 15. Такой подход, несмотря на заслуженную критику, используется в настоящее время.

Второй подход (стохастический) заложил Эдни Юл в 1927 г. Он предложил для его объяснения пример, ставший классическим: «Если рассматривать свободное качание маятника, отклоняющегося на малый угол под действием силы тяжести, то хорошо известно, что его движение является гармоническим, т. е. оно может быть представлено синусоидальной и косинусоидальной волной с постоянными амплитудами и периодами колебаний. Но если маленький мальчик обстреливает маятник горохом нерегулярным образом, то его движение будет возмущено. Маятник будет качаться, но с нерегулярными амплитудами и периодами колебаний. Фактически вместо такого поведения, при котором расхождение между теорией и наблюдением можно отнести за счет незначительной ошибки, горох вызывает ряд толчков, *воздействующих на будущее движение системы.* Эта концепция приводит к теории *стохастических процессов*, важнейшим разделом которой является теория стохастических временных рядов» [57].

Третий подход к анализу временных рядов — это спектральный анализ в частотной области. В частном случае можно получить выравнивание по ряду Фурье, при этом обычно рассматривается не более 5 гармоник ($j = 1, 2, 3, 4, 5$):

$$y_t = a_0 + \sum_{j=1}^{k}(a_j \cos jt + b_j \sin jt). \qquad (16.5)$$

Параметры a_j и b_j находятся методом наименьших квадратов, в результате применения которого получим

$$a_0 = \frac{1}{n}\sum_{t=1}^{n} y_t,\ a_j = \frac{2}{n}\sum_{t=1}^{n} y_t \cos jt\,,\ b_j = \frac{2}{n}\sum_{t=1}^{n} y_t \sin jt\,. \qquad (16.6)$$

Анализ временных рядов преследует целый ряд целей:

1) описание поведения ряда;
2) построение модели для объяснения наблюдений;
3) расчет прогноза на краткосрочный период, исходя из предположения о сохранении тенденции развития в будущем.

Для достижения поставленных целей используют модели, основанные на детерминистском, стохастическом, спектральном и других подходах. В общем случае можно предположить в модели наличие следующих компонент:

1) тренд;

2) циклическая компонента;

3) сезонная компонента;

4) остаток или случайная компонента.

Модель, в которой временной ряд представлен как сумма перечисленных выше компонент, называется *аддитивной моделью* временного ряда. Если временной ряд представлен как произведение компонент, то она называется *мультипликативной моделью* временного ряда.

Отделить тренд и сезонность в общем случае невозможно, так как они взаимно проникают друг в друга. При выделении тренда и сезонности остается колеблющийся ряд. Удаление тренда (сглаживание временного ряда) можно осуществить с помощью скользящей средней. Скользящая средняя, в отличие от простой средней для всего временного ряда, содержит сведения о тенденциях изменения данных.

Для этого к первым $(2m+1)$ точкам ряда (16.1) подбирают полином:

$$Q_p(t) = a_p t^p + a_{p-1} t^{p-1} + \cdots + a_1 t + a_0 \qquad (16.7)$$

(для определения значения тренда в $(m+1)$ точке) и минимизируют:

$$\sum_{t=-m}^{m} (y_t - a_p t^p - a_{p-1} t^{p-1} - \cdots - a_1 t - a_0)^2. \qquad (16.8)$$

Затем подбирают полином того же порядка для второго, третьего, …, $(2m+2)$ наблюдения. Эта процедура продолжается вдоль всего ряда до последней группы из $(2m+1)$ точек. На самом деле нет необходимости подбирать полином каждый раз. Покажем, что эта процедура соответствует некоторой линейной комбинации наблюдений с постоянными коэффициентами.

Например, пусть $2m+1 = 5$ или $t: -2, -1, 0, 1, 2;\ p = 3$.

$$(Q_3(t) = a_3 t^3 + a_2 t^2 + a_1 t^1 + a_0),$$

тогда (16.8) примет вид

$$\sum_{t=-m}^{m} (y_t - a_3 t^3 - a_2 t^2 - a_1 t - a_0)^2. \qquad (16.9)$$

После дифференцирования (16.9) по a_i и преобразований (так как суммы t с нечетной степенью равны 0), получим систему уравнений:

$$\begin{cases} \sum y_t = 5a_0 + 10a_2, \\ \sum t\, y_t = 10a_1 + 34a_3, \\ \sum t^2 y_t = 10a_0 + 34a_2, \\ \sum t^3 y_t = 34a_1 + 34a_3. \end{cases} \qquad (16.10)$$

Из (1) и (3) уравнения системы (16.10) следует, что

$$a_0 = \frac{1}{35}(-3z_{-2} + 12z_{-1} + 17z_0 + 12z_1 - 3z_2). \qquad (16.11)$$

Но $y_0 = a_0$. Итак, значение тренда в какой-либо точке равно средневзвешенному значению пяти точек с данной точкой в качестве центральной и весами $\frac{1}{35}[-3, 12, 17, 12, -3]$. Для пяти точек и $p = 1$ получаем простую скользящую среднюю:

$$a_0 = \frac{1}{5}(y_{-2} + y_{-1} + y_0 + y_1 - y_2). \qquad (16.12)$$

Кроме рассмотренного подхода к выводу формул взвешенных скользящих средних существуют другие способы их определения: использование простых скользящих средних, формулы Спенсера и т. д.

При рассмотрении скользящих средних, в рамках нашего примера ($p = 3, m = 2$), следует отметить проблему крайних двух точек — они не оцениваются, хотя ее можно решить, определив из (16.10) коэффициенты a_1, a_2, a_3. Для прогноза в следующей точке следует определить Q_3 (3).

Рассмотренные выше скользящие средние (их называют иногда линейными фильтрами), являются симметрическими (т. е. коэффициенты (веса) симметричны относительно среднего).

Для прогнозирования в статистике используют асимметричные фильтры. Так в *Excel* простая *скользящая средняя (СС)* заменяет не средний, а последний уровень ряда в промежутке сглаживания, скользящая средняя используется для расчета значений в прогнозируемом периоде, на основе среднего значения переменной для указанного числа предшествующих периодов, по формуле

$$Z_{t+1} = \frac{1}{N}\sum_{h=0}^{N} Y_{t-h+1,} \qquad (16.13)$$

где N — число предшествующих периодов, входящих в СС; Y_h — фактическое значение в момент времени h; Z_h — прогнозируемое значение в момент времени h.

Асимметричные скользящие средние иногда могут учитывать степень «устаревания данных», т. е. каждое новое наблюдение будет иметь вес больше предыдущих, например при A_t в момент времени $(t + 1)$:

$$F_{t+1} = (1-\alpha)(A_{t+1} + \alpha A_t + \alpha^2 A_{t-1} + \ldots) =$$

$$= (1-\alpha)\sum_{i=0}^{\infty} \alpha^i A_{t-i+1}, \quad 0 < \alpha < 1. \qquad (16.14)$$

Рассмотренный подход к определению асимметричных СС в *Excel* носит название *экспоненциальное сглаживание* (ЭС) (или экспоненциальных средних). ЭС предназначается для предсказания значения F_{t+1} на основе прогноза для предыдущего периода F_t, скорректированного с учетом погрешностей в этом прогнозе ($A_t - F_t$), из (16.14) можно получить, что

$$F_{t+1} = F_t + \alpha\,(A_t - F_t) = \alpha A_t + (1-\alpha)F_t.$$

Существует еще целый ряд методов сглаживания и экстраполяции: модель Хольта — Уинтерса (содержит три параметра α_1, α_2, α_3 и позволяет учесть сезонность), модель Харрисона является модификацией предыдущей и выражает сезонность через гармоники. Указанные методы были разработаны для анализа экономических процессов. Широко известны также модель Бокса — Дженкинса, фильтры Калмана и Бюсси.

Практически все рассмотренные методы содержат предположения относительно исходных данных, генерирующих временной ряд. Критерием адекватности той или иной модели может служить только практическое достижение первоначальных целей анализа временных рядов (описание поведения ряда, объяснение изменения наблюдений, прогноз и т. д.).

Стационарные временные ряды. Временной ряд, не имеющий тренда (либо с исключенным трендом), или если его свойства не зависят от начала отсчета времени (механизм, генерирующий ряд, не меняется со временем, хотя и носит вероятностный характер), называется *стационарным*. Поэтому перечисленные ниже параметры для данного ряда являются постоянными:

$M(z_t) = M$ — математическое ожидание,

$M(z_t - M)^2 = \sigma^2 = D\,(z_t)$ — дисперсия,

$M[(z_t - M)(z_{t+k} - M)] = c_k$ — k-ая автоковариация,

$\rho_k = \rho_{-k} = \frac{c_k}{\sigma^2}$ — соответствующая автокорреляция.

Совокупность значений ρ_k представляется на графике и называется коррелограммой (но не каждая последовательность констант является коррелограммой).

Для стационарного процесса рассматривают 3 основных типа моделей (соответствующих определенным типам стационарных стохастических процессов).

1. Авторегрессии (АР) порядка p. В этой модели текущее значение t выражается через линейную комбинацию p предыдущих значений процесса плюс случайная компонента ε_t (*белый шум* — последовательность независимых $\varepsilon_t \to N(0, \sigma^2)$):

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_{t-p} y_{t-p} + \varepsilon_t. \qquad (16.15)$$

Важными частными случаями являются:

а) при $p = 1$, модель процесса Маркова (процесс с отсутствием последействия, то есть каждое следующее значение зависит только от предыдущего):

$$y_t = a_1 y_{t-1} + \varepsilon_t, \qquad (16.16)$$

б) при $p = 2$, модель процесса Юла:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t. \qquad (16.17)$$

2. Скользящего среднего (СС) порядка q:

$$y_t = b_1\,\varepsilon_{t-1} + b_2\,\varepsilon_{t-2} + b_3\,\varepsilon_{t-3} + \cdots + b_q\,\varepsilon_{t-q} + \varepsilon_t \qquad (16.18)$$

(термин СС не означает, что сумма весов при ε_i равна 1).

3. Авторегрессии — скользящего среднего (АРСС):

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_{t-p} y_{t-p} + b_1\,\varepsilon_{t-1} + b_2\,\varepsilon_{t-2} + b_3\,\varepsilon_{t-3} + \cdots + b_q\,\varepsilon_{t-q} + \varepsilon_t. \qquad (16.19)$$

Обычно, на практике достаточно рассматривать модели АР, СС, АРСС при p и q, не превышающих 2.

Для описания *нестационарных* процессов пользуются экспоненциально взвешенными средними, а в более общем случае моделями авторегрессии-проинтегрированного скользящего среднего (АРПСС).

Замечание. 1. Еще одним методом построения модели прогноза является метод группового учета аргумента (МГУА), рассмотренный в конце раздела 15.4.2. С точки зрения современной науки развитие сложных процессов является детерминированным только между определенными точками структурных изменений — точками бифуркаций, появление которых случайно. Поэтому можно говорить о существовании пределов предсказуемости, прогнозировать далее которые принципиально невозможно из-за случайного появления точек структурных изменений. ■

При стандартных подходах (детерминистском, стохастическом и спектральном) наличие неучтенных факторов описывается присутствием случайной составля-

ющей (ε_t), имеющей тот или иной закон распределения (обычно нормальный). Линейные модели окружающего мира, к которым сводятся перечисленные ранее методы, давно перестали удовлетворять потребностям человека. В настоящее время назрела необходимость описания сложных объектов (процессов) с помощью нелинейных методов, дающих возможность учитывать реальную жизнь (в которой нелинейные связи преобладают), а также учитывать ненаблюдаемые и неконтролируемые факторы. Для этого совокупность данных разбивается на две части: по первой строятся различные модели (обучающая последовательность), по второй отбираются лучшие модели (проверочная последовательность), дающие наилучший прогноз.

Следует отметить, что предлагаемая идеология соответствует принципу *Soft Computing* — «мягких вычислений» (терпимость к нечеткости и частичной истинности используемых данных для достижения интерпретируемости, гибкости и низкой стоимости решений), сформулированному известным математиком Лотфи Заде (1994).

Пример 16.1. Имеются данные о производстве электроэнергии в регионе за 2010–2021 гг., млрд кВт-ч.

2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
6,9	7,1	6,6	6,2	6,6	5,9	8,0	9,9	12,0	11,8	12,1	11,9

Требуется:

а) построить график временного ряда;

б) рассчитать коэффициент автокорреляции первого порядка;

в) обосновать выбор типа уравнения тренда и рассчитать его параметры;

г) дать интерпретацию параметров тренда и сделать выводы по задаче.

Решение. а) Рассмотрим систему координат $Y0t$, где Y_t — производство электроэнергии, t — порядковый номер года и нанесем в ней данные примера на график (рис. 16.1).

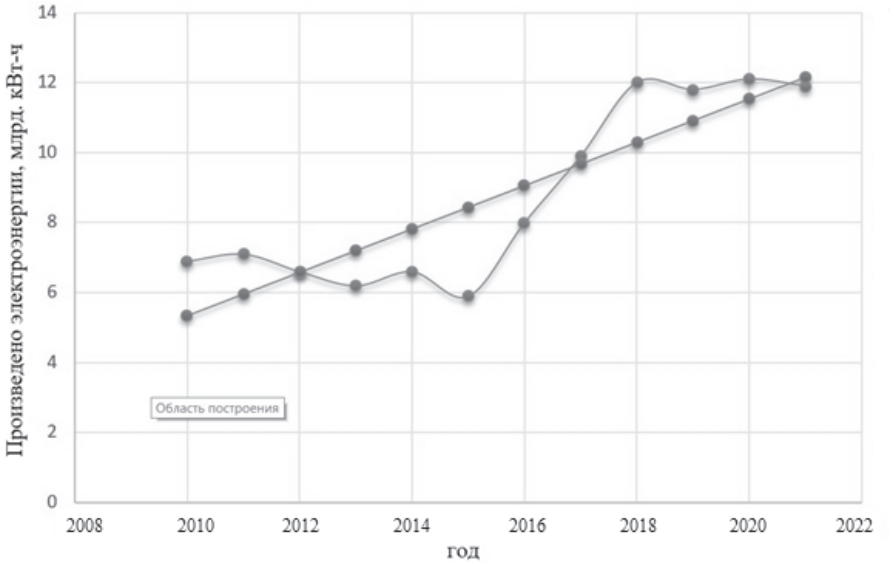


Рис. 16.1 — Динамика производства электроэнергии в регионе

б) Определим коэффициент автокорреляции первого порядка, для чего заполним вспомогательную таблицу 16.1.

$$\bar{y}_1 = \frac{\sum_{t=2}^{n} y_t}{n-1} = \frac{105-6,9}{11} = 8,92,\ \bar{y}_2 = \frac{\sum_{t=2}^{n} y_{t-1}}{n-1} = \frac{93,1}{11} = 8,46,$$

$$r_1 = \frac{\sum_{t=2}^{n}(y_t-\bar{y}_1(y_{t-1}-\bar{y}_2)}{\sqrt{\sum_{t=2}^{n}(y_t-\bar{y}_1)^2 \sum_{t=2}^{n}(y_{t-1}-\bar{y}_2)^2}} = \frac{59,9172}{\sqrt{69,1764 \cdot 62,0856}} = 0,914.$$

в) Полученное значение коэффициента автокорреляции и графическое изображение временного ряда позволяют сделать вывод о том, что временной ряд производства электроэнергии содержит тенденцию, близкую к линейной. Поэтому для моделирования его тенденции используем линейную функцию $y = a + bt$.

Для расчета параметров линейного тренда *a* и *b* используем метод наименьших квадратов, для чего составим и решим следующую систему:

$$\begin{cases} na + b\sum t = \sum y, \\ a\sum t + b\sum t^2 = \sum yt. \end{cases}$$

Заполним вспомогательную таблицу 16.2.

Воспользуемся формулами, вытекающими из системы нормальных уравнений:

$$b = \frac{\overline{yt}-\bar{y}\,\bar{t}}{\overline{t^2}-(\bar{t})^2} = \frac{64,267-8,75 \cdot 6,5}{54,167-6,5^2} = \frac{7,392}{11,917} = 0,62,$$

$$a = \bar{y} - b\bar{t} = 8,75 - 062 \cdot 6,5 = 4,72 \Rightarrow \hat{y}_t = 4,72 + 0,62t.$$

Таблица 16.1

Вспомогательная таблица для расчета коэффициента автокорреляции

t	y_t	y_{t-1}	$y_t - \bar{y}_1$	$y_{t-1} - \bar{y}_2$	$(y_t - \bar{y}_1) \cdot (y_{t-1} - \bar{y}_2)$	$(y_t - \bar{y}_{1)}{}^2$	$(y_{t-1} - \bar{y}_{2)}{}^2$
1	6,9						
2	7,1	6,9	–1,82	–1,56	2,8392	3,3124	2,4336
3	6,6	7,1	–2,32	–1,36	3,1552	5,3824	1,8496
4	6,2	6,6	–2,72	–1,86	5,0592	7,3984	3,4596
5	6,6	6,2	–2,32	–2,26	5,2432	5,3824	5,1076
6	5,9	6,6	–3,02	–1,86	5,6172	9,1204	3,4596
7	8	5,9	–0,92	–2,56	2,3552	0,8464	6,5536
8	9,9	8	0,98	–0,46	–0,4508	0,9604	0,2116
9	12	9,9	3,08	1,44	4,4352	9,4864	2,0736
10	11,8	12	2,88	3,54	10,1952	8,2944	12,5316
11	12,1	11,8	3,18	3,34	10,6212	10,1124	11,1556
12	11,9	12,1	2,98	3,64	10,8472	8,8804	13,2496
Сумма	105	93,1	-0,02	0,04	59,9172	69,1764	62,0856

Таблица 16.2

Вспомогательная таблица для расчета параметров тренда

№ п/п	y	t	yt	t^2	y_t
1	6,9	1	6,9	1	5,34
2	7,1	2	14,2	4	5,96
3	6,6	3	19,8	9	6,58
4	6,2	4	24,8	16	7,20
5	6,6	5	33,0	25	7,82
6	5,9	6	35,4	36	8,44
7	8,0	7	56,0	49	9,06
8	9,9	8	79,2	64	9,68
9	12	9	108,0	81	10,30
10	11,8	10	118,0	100	10,92
11	12,1	11	133,1	121	11,54
12	11,9	12	142,8	144	12,16
Сумма	105,0	78	771,2	650	105,0
Среднее значение	8,75	6,5	64,267	54,167	8,75

Таким образом, в среднем ежегодно производство электроэнергии в Краснодарском крае за 2010–2021 гг. увеличивалось на 0,62 млрд кВт-ч.

Темы (вопросы) для самоконтроля

1. Временной ряд.
2. Автокорреляция.
3. Основные подходы к анализу временных рядов.
4. Стационарные временные ряды.
5. Модель авторегрессии.
6. Модель скользящего среднего.
7. Модель авторегрессии-скользящего среднего.
8. Модель авторегрессии проинтегрированного скользящего среднего и ее применение.
9. Экспоненциальное сглаживание.

Однажды, когда ночь покрыла небеса невидимою своею епанчою, знаменитый французский философ Декарт, у ступенек домашней лестницы своей сидевший и на мрачный горизонт с превеликим вниманием смотрящий, некий прохожий подступил к нему с вопросом: «Скажи, мудрец, сколько звезд на сем небе?» «Мерзавец! — ответствовал сей, — никто необъятного обнять не может!» Сии, с превеликим огнем произнесенные, слова возымели на прохожего желаемое действие.

Из исторических материалов
Федота Кузьмича Пруткова (деда)

...Главный конструктор проекта уже заканчивает свой доклад:

«Новый компьютер Ультроник... более 10^{17} логических ячеек. Это больше, чем суммарное число нейронов у всех живущих в нашей стране. Уровень интеллекта невообразимо высок. Найдется в этой аудитории желающий инициировать нашу новую компьютерную систему Ультроник, задав ей первый вопрос?»

...Все в смятении, как будто чувствуя присутствие нового всемогущего разума. Адам поднимает руку.

«Ну вот, — говорит Главный конструктор, — парнишка в третьем ряду. У тебя есть вопрос к нашему... гм... новому другу?

.....

«...СЕБЯ ЧУВСТВУЕШЬ? О... весьма интересный вопрос, мой мальчик... э-э... я и сам хотел бы знать ответ», — сказал Главный конструктор. — «Давайте посмотрим, что может сказать наш друг об этом... странно... э-э... Ультроник говорит, что он не понимает, что... он не может понять, что ты имеешь в виду!»

Отдельные смешки в аудитории переросли в громовой хохот. Адам чувствовал себя крайне неловко. Они могли отреагировать как угодно, но только не смеяться.

Р. Пенроуз.
Новый ум короля: О компьютерах, мышлении и законах физики

Сократ постоянно указывал своим ученикам на то, что при правильно поставленном образовании в каждой науке надо доходить только до известного предела, который не следует переступать. По геометрии, говорил он, достаточно знать настолько, чтобы при случае быть в силах правильно измерить кусок земли, который продаешь или покупаешь, или чтобы разделить наследство, или чтоб суметь разделить работу рабочим. Но он не одобрял увлечения большими трудностями в этой науке, и хотя сам лично знал их, но говорил, что они могут занять всю жизнь человека и отвлечь его от других полезных наук, тогда как они ни к чему не нужны.

Л. Н. Толстой

Воспринимается не то, что истинно или правильно, а то, что понятно.

NN

По мере своего развития человечество придумывало все более сложные истории о самом себе ... все эти истории — вымысел. Алгоритмизация (представления информации на телевидении, в Интернете) может лишить людей возможности наблюдать собственную реальность (понимать общество и события в нем). Кто мы и что должны о себе знать, за нас будут решать алгоритмы. У нас еще есть несколько десятилетий. Это наш шанс. Приложив достаточно усилий, мы сможем понять, кто мы. Но если мы не хотим упустить этот шанс, лучше начать прямо сейчас.

Ю. Н. Харари

Важно общение и передача культуры. Воспитание. Формирование умения мыслить, анализировать. А не просто усваивать что-то, что можно найти в интернете, справочниках.

В. Н. Волкова

Заключение

Еще многое имею сказать вам,
но вы теперь не можете вместить.
Евангелие от Иоанна, 16:12

Тебе бы опыт сделать не мешало;
Ведь он для вас — источник всех наук.
Данте Алигьери
Божественная комедия

В настоящей книге рассмотрено введение в ряд классических разделов теории вероятностей и математической статистики, основывающейся на предпосылках нормальности распределения изучаемых случайных величин, которые стали основой развития методов анализа данных. Далее в прикладной статистике изучаются элементы статистики объектов нечисловой природы, категориальной статистики, непараметрической и байесовской статистики, статистического и машинного обучения.

Указанные разделы составляют основу современных методов науки о данных (*анализа структурированных, слабоструктурированных и неструктурированных данных*, логического вывода, прогнозирования, распознавания образов и других направлений, часто обобщаемых терминами «искусственный интеллект», «киберфизические системы», «цифровизация» и т. д. [76]).

Успешная практическая деятельность человека все в большей мере зависит от организации сбора и обработки информации. Совершенствование технологий записи и хранения данных — создание баз (озер) данных и знаний во всех сферах деятельности человека предъявляет новые требования к уровню подготовки специалистов. Большой объем информации, которой сопровождается деятельность практически любого предприятия и учреждения, обычно содержит полезные сведения, благодаря которым можно значительно повысить эффективность работы, совершенствуя технологию, управление и т. д. И завтра сегодняшним студентам новые идеи анализа данных будут очевидны и необходимы на практике.

Современные коммерческие организации интенсивно внедряют информационные хранилища (банки) данных и знаний, многие из которых содержат средства интеллектуального анализа данных или предполагают возможность их применения. Деловые люди осознали, что применение методов *анализа данных,* использующих вероятностную, геометрическую, логическую и так далее природу данных, может дать ощутимое преимущество в конкурентной борьбе.

Более подробно ознакомиться с современными методами прикладной статистики и *интеллектуального анализа данных* можно, например, по электронным учебникам фирмы *StatSoft*, которые в основном соответствуют описанию статистических модулей пакета *Statistica*, а также на сайте фирмы *Loginom*, создателя аналитической платформы *Loginom*. Эти направления мы предлагаем для дальнейшего самостоятельного изучения и применения на практике для анализа *структу-*

рированных данных, с помощью прикладных пакетов статистических *DATA MINING* и *KDD* (*knowledge discovery in databases*), языков программирования *Python, R* и других средств, позволяющих также анализировать *слабоструктурированные и неструктурированные данные* (*методы глубокого обучения*).

Следует отметить, что на современном этапе управление производством, фирмой, районом, регионом практически невозможно без системного подхода [21–24], разрабатывающего методики анализа целей, методы и модели совершенствования организационной структуры, управления функционированием социально-экономических объектов.

В зависимости от априорной информации об изучаемом объекте применяют следующие методы: мозговой атаки, построения сценариев, экспертной оценки, построения дерева целей, математической логики, теории множеств, теории игр, прикладной статистики, математического программирования, интеллектуального анализа данных и т. д. Разумеется, большинство методов пересекается. При этом рассматриваемые в книге вероятностные методы в рамках системного анализа являются одним из возможных подходов перевода вербального (словесного) описания модели изучаемого объекта в формальное, для решения задач управления и принятия решений.

Сегодня современные специалисты развивают системный подход в рамках необходимости изучения, кроме механических и биологических систем мультиразумных (социальных), которые характеризуются открытостью, взаимозависимыми переменными. Этот подход получил название *системного мышления*. Последние десятилетия попытка конвергенции объектов реального мира и информационных технологий посредством использования алгоритмов анализа больших данных (MCMC, частотной и байесовской статистики, нейронных сетей, машинного обучения, глубокого обучения) привела к понятию *киберфизических систем*.

Мир поступательно развивается и, безусловно, не существует абсолютных рецептов построения моделей сложных процессов. Моделирование в большей степени искусство, овладеть которым можно, только решая практические задачи. Целью настоящего изложения является не сама истина, которая со временем обычно изменяется, а возможный путь ее поиска, двигаясь по которому можно приобрести неоценимый опыт. Последнее соответствует концепциям байесовской статистики, аутопоэзиса и Кибернетики 2.0 — *процесс познания мира* (уточнение теории) *происходит в процессе наблюдения*.

Природа — древний храм, где строй живых колонн
Обрывки смутных фраз роняет временами.
Мы входим в этот храм в смятении, а за нами
Лес символов немых следит со всех сторон.

Ш. Бодлер

Приложения

Приложение 1

Значения функций

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

x	$\varphi(x)$	$\Phi(x)$	x	$\varphi(x)$	$\Phi(x)$	x	$\varphi(x)$	$\Phi(x)$
0,00	0,3989	0,0000	0,40	0,3683	0,1554	0,80	0,2897	0,2881
01	3989	0040	41	3668	1591	81	2874	2910
02	3989	0080	42	3652	1628	82	2850	2939
03	3988	0120	43	3637	1664	83	2827	2967
04	3986	0160	44	3621	1700	84	2803	2995
05	3984	0199	45	3605	1736	85	2780	3023
06	3982	0239	46	3589	1772	86	2756	3051
07	3980	0279	47	3572	1808	87	2732	3078
08	3977	0319	48	3555	1844	88	2709	3106
09	3973	0359	49	3538	1879	89	2685	3133
0,10	0,3970	0,0398	0,50	0,3521	0,1915	0,90	0,2661	0,3159
11	3965	0438	51	3503	1950	91	2637	3186
12	3961	0478	52	3485	1985	92	2613	3212
13	3956	0517	53	3467	2019	93	2589	3238
14	3951	0557	54	3448	2054	94	2565	3264
15	3945	0596	55	3429	2088	95	2541	3289
16	3939	0636	56	3410	2123	96	2516	3315
17	3932	0675	57	3391	2157	97	2492	3340
18	3925	0714	58	3372	2190	98	2468	3365
19	3918	0753	59	3352	2224	99	2444	3389
0,20	0,3910	0,0793	0,60	0,3332	0,2257	1,00	0,2420	0,3413
21	3902	0832	61	3312	2291	01	2396	3438
22	3894	0871	62	3292	2324	02	2371	3461
23	3885	0910	63	3271	2357	03	2347	3485
24	3876	0948	64	3251	2389	04	2323	3508
25	3867	0987	65	3230	2422	05	2299	3531
26	3857	1026	66	3209	2454	06	2275	3554
27	3847	1064	67	3187	2486	07	2251	3577
28	3836	1103	68	3166	2517	08	2227	3599
29	3825	1141	69	3144	2549	09	2203	3621
0,30	0,3814	0,1179	0,70	0,3123	0,2580	1,10	0,2179	0,3643
31	3802	1217	71	3101	2611	11	2155	3665
32	3790	1255	72	3079	2642	12	2131	3686
33	3778	1293	73	3056	2673	13	2107	3708
34	3765	1331	74	3034	2703	14	2083	3729
35	3752	1368	75	3011	2734	15	2059	3749
36	3739	1406	76	2989	2764	16	2036	3770
37	3726	1443	77	2966	2794	17	2012	3790
38	3712	1480	78	2943	2823	18	1989	3810
39	3697	1517	79	2920	2852	19	1965	3830

Продолжение табл.

x	$\varphi(x)$	$\Phi(x)$	x	$\varphi(x)$	$\Phi(x)$	x	$\varphi(x)$	$\Phi(x)$
1,20	0,1942	0,3849	1,70	0,0940	0,4554	2,40	0,0224	0,4918
21	1919	3869	71	0925	4564	42	0213	4922
22	1895	3883	72	0909	4573	44	0203	4927
23	1872	3907	73	0893	4582	46	0194	4931
24	1849	3925	74	0878	4591	48	0184	4934
25	1826	3944	75	0863	4599	50	0175	4938
26	1804	3962	76	0848	4608	52	0167	4941
27	1781	3980	77	0833	4616	54	0158	4945
28	1758	3997	78	0818	4625	56	0151	4948
29	1736	4015	79	0804	4633	58	0143	4951
1,30	0,1714	0,4032	1,80	0,0790	0,4641	2,60	0,0136	0,4953
31	1691	4049	81	0775	4649	62	0129	4956
32	1669	4066	82	0761	4656	64	0122	4959
33	1647	4082	83	0748	4664	66	0116	4961
34	1626	4099	84	0734	4671	68	0110	4963
35	1604	4115	85	0721	4678	70	0104	4965
36	1582	4131	86	0707	4686	72	0099	4967
37	1561	4147	87	0694	4693	74	0093	4969
38	1539	4162	88	0681	4699	76	0088	4971
39	1518	4177	89	0669	4706	78	0084	4973
1,40	0,1497	0,4192	1,90	0,0656	0,4713	2,80	0,0079	0,4974
41	1476	4207	91	0644	4719	82	0075	4976
42	1456	4222	92	0632	4726	84	0071	4977
43	1435	4236	93	0620	4732	86	0067	4979
44	1415	4251	94	0608	4738	88	0063	4980
45	1394	4265	95	0596	4744	90	0060	4981
46	1374	4279	96	0584	4750	92	0056	4982
47	1354	4292	97	0573	4756	94	0053	4984
48	1334	4306	98	0562	4761	96	0050	4985
49	1315	4319	99	0551	4767	98	0047	4986
1,50	0,1295	0,4332	2,00	0,0540	0,4772	3,00	0,00443	0,49865
51	1276	4345	02	0519	4783			
52	1257	4357	04	0498	4793	3,10	00327	49903
53	1238	4370	06	0478	4803	3,20	00238	49931
54	1219	4382	08	0459	4812			
55	1200	4394	10	0440	4821	3,30	00172	49952
56	1182	4406	12	0422	4830	3,40	00123	49966
57	1163	4418	14	0404	4838			
58	1145	4429	16	0387	4846	3,50	00087	49977
59	1127	4441	18	0371	4854			
1,60	0,1109	0,4452	2,20	0,0355	0,4861	3,60	00061	499841
61	1092	4463	22	0339	4868	3,70	00042	49989
62	1074	4474	24	0325	4875	3,80	00029	499928
63	1057	4484	26	0310	4881	3,90	00020	49995
64	1040	4495	28	0297	4887	4,00	0,0001338	499968
65	1023	4505	30	0283	4893	4,50	0000160	499997
66	1006	4515	32	0270	4898	5,00	0000015	49999997
67	0989	4525	34	0258	4904			
68	0973	4535	36	0246	4909			
69	0957	4545	38	0235	4913			

Приложение 2

Критические точки χ^2-распределения Пирсона

ν \ α	0,20	0,10	0,05	0,02	0,01	0,001
1	1,642	2,706	3,841	5,412	6,635	10,827
2	3,219	4,605	5,991	7,824	9,210	13,815
3	4,642	6,251	7,815	9,837	11,345	16,266
4	5,989	7,779	9,488	11,668	13,277	18,467
5	7,289	9,236	11,070	13,388	15,086	20,515
6	8,558	10,645	12,592	15,033	16,812	22,457
7	9,803	12,017	14,067	16,622	18,475	24,322
8	11,030	13,362	15,507	18,168	20,090	26,125
9	12,242	14,684	16,919	19,679	21,666	27,877
10	13,442	15,987	18,307	21,161	23,209	29,588
11	14,631	17,275	19,675	22,618	24,725	31,264
12	15,812	18,549	21,026	24,054	26,217	32,909
13	16,985	19,812	22,362	25,472	27,688	34,528
14	18,151	21,064	23,685	26,783	29,141	36,123
15	19,311	22,307	24,996	28,259	30,578	37,697
16	20,465	23,542	26,296	29,633	32,000	39,252
17	21,615	24,769	27,587	30,995	33,409	40,790
18	22,760	25,989	28,869	32,346	34,805	42,312
19	23,900	27,204	30,144	33,687	36,191	43,820
20	25,038	28,412	31,410	35,020	37,566	45,315
21	26,171	29,615	32,671	36,343	38,932	46,797
22	27,301	30,813	33,924	37,659	40,289	48,268
23	28,429	32,007	35,172	38,968	41,638	49,728
24	29,553	33,196	36,415	40,270	42,980	51,179
25	30,675	34,382	37,652	41,566	44,314	52,620
26	31,795	35,563	38,885	42,856	45,642	54,052
27	32,912	36,741	40,113	44,140	46,963	55,476
28	34,027	37,916	41,337	45,419	48,278	56,893
29	35,139	39,087	42,557	46,693	49,588	58,302
30	36,250	40,256	43,773	47,962	50,892	59,703

Приложение 3

Критические точки t-распределения Стьюдента

Число степеней свободы k	Уровень значимости α (двусторонняя критическая область)					
	0,10	0,05	0,02	0,01	0,002	0,001
1	6,31	12,71	31,82	63,66	318,31	636,62
2	2,92	4,30	6,97	9,93	22,33	31,60
3	2,35	3,18	4,54	5,84	10,22	12,92
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	6,87
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,90	2,37	3,00	3,50	4,79	5,41
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,06	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,15	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,02
17	1,74	2,11	2,57	2,90	3,65	3,97
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,75
25	1,71	2,06	2,49	2,79	3,45	3,73
26	1,71	2,06	2,48	2,78	3,44	3,71
27	1,70	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,47	2,76	3,41	3,67
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,16	3,37
∞	1,65	1,96	2,33	2,58	3,09	3,29
	0,05	0,025	0,01	0,005	0,001	0,0005
	Уровень значимости α (односторонняя критическая область)					

Приложение 4

Критические точки F-распределения Фишера — Снедекора

(k_1 — число степеней свободы большей дисперсии, k_2 — число степеней свободы меньшей дисперсии), уровень значимости $\alpha = 0{,}05$

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	12	16	20	24	30	50	100	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	252	253	254
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,41	19,43	19,44	19,45	19,46	19,47	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,74	8,69	8,66	8,64	8,62	8,58	8,56	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,84	5,80	5,77	5,74	5,70	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,68	4,60	4,56	4,53	4,50	4,44	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,92	3,87	3,84	3,81	3,75	3,71	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,57	3,49	3,44	3,41	3,38	3,32	3,28	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,28	3,20	3,15	3,12	3,08	3,03	2,98	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,07	2,98	2,93	2,90	2,86	2,80	2,76	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,91	2,82	2,77	2,74	2,70	2,64	2,59	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,79	2,70	2,65	2,61	2,57	2,50	2,45	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,69	2,60	2,54	2,50	2,46	2,40	2,35	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,60	2,51	2,46	2,42	2,38	2,32	2,26	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,53	2,44	2,39	2,35	2,31	2,24	2,19	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,48	2,39	2,33	2,29	2,25	2,18	2,12	2,67
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,33	2,28	2,24	2,20	2,13	2,07	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,38	2,29	2,23	2,19	2,15	2,08	2,02	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,25	2,19	2,15	2,11	2,04	1,98	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,55	2,48	2,43	2,38	2,31	2,21	2,15	2.11	2,07	2,00	1,94	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35	2,28	2,18	2,12	2,08	2,04	1,96	1,90	1,84
22	4,30	3,44	3,05	2,82	2,66	2,55	2,47	2,40	2,35	2,30	2,23	2,13	2,07	2,03	1,98	1,91	1,84	1,78
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30	2,26	2,18	2,09	2,02	1,98	1,94	1,86	1,80	1,73
26	4,22	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,05	1,99	1,95	1,90	1,82	1,76	1,69
28	4,20	3,34	2,95	2,71	2,56	2,44	2,36	2,29	2,24	2,19	2,12	2,02	1,96	1,91	1,87	1,78	1,72	1,65
32	4,15	3,30	2,90	2,67	2,51	2,40	2,32	2,25	2,19	2,14	2,07	1,97	1,91	1,86	1,82	1,74	1,67	1,59
36	4,11	3,26	2,86	2.63	2,48	2,36	2,28	2,21	2,15	2,10	2,03	1,93	1,87	1,82	1,78	1,69	1,62	1,55
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,07	2,00	1,90	1,84	1,79	1,74	1,66	1,59	1,51
60	4,00	3,15	2.76	2,52	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,81	1,75	1,70	1,65	1,56	1,48	1,39
100	3,94	3,09	2,70	2,46	2,30	2,19	2,10	2,03	1,97	1,92	1,85	1,75	1,68	1,63	1,57	1,48	1,39	1,28
200	3,89	3,04	2,65	2,41	2.26	2,14	2,05	1,98	1,92	1,87	1,80	1,69	1,62	1,57	1,52	1,42	1,32	1,19
∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,88	1,83	1,75	1,64	1,57	1,52	1,46	1,35	1,24	1,00

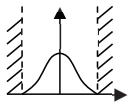
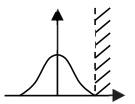
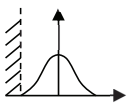
Приложение 5

Сравнение дисперсий

<table>
<tr><td colspan="3">1 выборка</td><td colspan="2">2 выборки</td><td colspan="2">m выборок</td></tr>
<tr><td colspan="3">H_0:$\sigma^2 = \sigma_0^2$ (σ_0^2 генеральная дисперсия)
H_1: $\sigma^2 < \sigma_0^2$</td><td colspan="2">H_0: $\sigma_1^2 = \sigma_2^2$
H_1: $\sigma_1^2 > \sigma_2^2$</td><td>H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$
H_1: $\sigma_1^2 > \sigma_2^2 > \ldots > \sigma_k^2$</td><td rowspan="3">$H_0$: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_m^2$
H_1: $\sigma_1^2 > \sigma_2^2 > \ldots > \sigma_m^2$
Дано: $n_1 \neq n_2 \neq \ldots \neq n_m$, $n_i \geq 4$
$K_{набл} = B_{набл} = \dfrac{V}{C}$,
$V = 2{,}303[\lg S^2 - \sum_{i=1}^{k} k_i \lg S_i^2]$,
$C = 1 + \dfrac{1}{3(m-1)}[\sum \dfrac{1}{k_i} - \dfrac{1}{k}]$,
где $k_i = n_i - 1$, $k = \sum k_i$,
$S_m^2 = \dfrac{\sum k_i S_i^2}{k}$,
$\sum = \sum_{i=1}^{k}$</td></tr>
<tr><td colspan="3">Дано: n, S^2, σ_0^2</td><td colspan="2">Дано: n_1, n_2, S_1^2, S_2^2</td><td>Дано: $n_1 = n_2 = \ldots = n_k$</td></tr>
<tr><td colspan="3">$K_{набл} = \chi^2_{набл} = \dfrac{(n-1)S^2}{\sigma_0^2}$, k=$n$–1.
а) n<30, $\chi^2_{кр.}$ при α, k — по таблице распределения Пирсона;
б) n>30, $\chi^2_{кр.}$ при α, k: $\chi^2_{кр.} = k\left(1 - \dfrac{2}{9k} + u_\alpha \sqrt{\dfrac{2}{9k}}\right)^3$.
u_α — по таблице распределения Лапласа</td><td colspan="2">$K_{набл} = F_{набл} = \dfrac{S_1^2}{S_2^2}$,
при k_1, k_2.
$S_1^2 > S_2^2$,
$k_1 = n_1 - 1$,
$k_2 = n_2 - 1$</td><td>$K_{набл} = \dfrac{S^2 \max}{\sum_{i=1}^{k} S_i^2}$,
$S^2 max = max\{S_i^2\}$</td></tr>
<tr><td>Двусторонний критерий</td><td colspan="2">Односторонний критерий</td><td>Двусторонний критерий</td><td>Односторонний критерий</td><td rowspan="2">Односторонний критерий</td></tr>
<tr><td>H_1: $\sigma^2 \neq \sigma_0^2$</td><td>H_1: $\sigma^2 > \sigma_0^2$</td><td>H_1: $\sigma^2 < \sigma_0^2$</td><td>H_1: $\sigma_1^2 \neq \sigma_2^2$</td><td>$H_1$: $\sigma_1^2 > \sigma_2^2$</td></tr>
<tr><td rowspan="2">$\chi^2_{кр.\,лев.} \leq \chi^2_{набл} \leq \chi^2_{кр.прав.}$
Тогда H_0 принимается.
$\chi^2_{кр.\,лев.}$ при $(1 - \alpha/2)$, k
$\chi^2_{кр.\,прав.}$ при $\alpha/2$, k
$\Phi(u_{\alpha/2}) = (1 - \alpha)/2$</td><td>$\chi^2_{набл} < \chi^2_{кр.прав.\,\alpha,\,k}$</td><td>$\chi^2_{набл} > \chi^2_{кр.лев.1-\alpha,\,k}$</td><td>$F_{набл} < F_{кр.}$ при k_1, k_2, $\alpha/2$</td><td>$F_{набл} < F_{кр.\,прав.}$ при k_1, k_2, α</td><td rowspan="2">По критерию Кокрена. $K_{кр.}$, α, k, m.
H_0 — принимается, если $K_{набл} < K_{кр.\alpha,k,m}$. Если H_0, принята то в качестве оценки дисперсии генеральной совокупности принимают
$S^2 = \dfrac{\sum S_i^2}{m}$</td><td>Односторонний критерий</td></tr>
<tr><td colspan="2">$\Phi(u_\alpha) = (1 - 2\alpha)/2$</td><td colspan="2">$F_{кр.}$ определяется по таблице распределения Фишера — Снедекора</td><td>H_0 принимается, если $B_{набл} < \chi^2_{кр.прав.}$при α, k. По таблице распределения Пирсона</td></tr>
</table>

Приложение 6

1 выборка. Сравнение выборочной средней с генеральной средней (гипотетической) H_0: $\overline{X} = \overline{X}_0$ или H_0: $\overline{X} - \overline{X}_0 = 0$

<table>
<tr><td></td><td>Большой объем выборки, n</td><td>Малый объем выборки, n</td></tr>
<tr><td rowspan="5">Генеральная дисперсия известна $D[X]=\sigma^2$</td><td colspan="2">Задача № 1
$K_{набл} = \dfrac{(\overline{X} - \overline{X}_0)\cdot\sqrt{n}}{\sigma}$, H_0: $\overline{X} = \overline{X}_0$; H_0 принимается если:</td></tr>
<tr><td>Двусторонний критерий</td><td>Односторонний критерий</td></tr>
<tr><td>$H_1: \overline{X} \neq \overline{X}_0$</td><td>$H_1: \overline{X} > \overline{X}_0$ | $H_1: \overline{X} < \overline{X}_0$</td></tr>
<tr><td>$\lvert K_{набл}\rvert < K_{кр,\alpha/2}$</td><td>$K_{набл} < K_{кр,\alpha}$ | $K_{набл} > -K_{кр,\alpha}$</td></tr>
<tr><td>$\Phi(K_{кр,\alpha/2}) = (1-\alpha)/2$</td><td>$\Phi(K_{кр,\alpha}) = (1-2\alpha)/2$</td></tr>
<tr><td colspan="2">$K_{кр}$ — по таблице значений интеграла Лапласа, $\Phi(x)$</td></tr>
<tr><td>Генеральная дисперсия не известна</td><td></td><td>Задача № 2
$K_{набл} = t_{набл} = \dfrac{(\overline{X} - \overline{X}_0)\cdot\sqrt{n}}{S}$
$S^2 = \dfrac{\sum X_j^2 \cdot n_j - \dfrac{(\sum X_j n_j)^2}{n}}{n-1} = \dfrac{\sum(\overline{X}_j - \overline{X})^2}{n-1}$
H_0 принимается, если:</td></tr>
<tr><td></td><td>
<table>
<tr><td>Двусторонний критерий</td><td colspan="2">Односторонний критерий</td></tr>
<tr><td>$H_1: \overline{X} \neq \overline{X}_0$</td><td>$H_1: \overline{X} > \overline{X}_0$</td><td>$H_1: \overline{X} < \overline{X}_0$</td></tr>
<tr><td>$\lvert t_{набл}\rvert < t_{кр,k,\alpha/2}$</td><td>$t_{набл} < t_{кр,k,\alpha}$</td><td>$t_{набл} > -t_{кр,k,\alpha}$</td></tr>
<tr><td> 
$-t_{кр} < t_{набл} < +t_{кр}$</td><td> 
$t_{набл} < +t_{кр}$</td><td> 
$t_{набл} > -t_{кр}$</td></tr>
<tr><td colspan="3">$t_{кр}$ по таблице t-распределения Стьюдента</td></tr>
</table>
</td></tr>
</table>

Приложение 7

2 выборки. Сравнение двух средних 2 генеральных совокупностей. H_0: $M[X_1]=M[X_2]$ или H_0: $M[X_1]-M[X_2]=0$

<table>
<tr><td></td><td colspan="3">2 выборки большого объема</td><td colspan="4">2 выборки малого объема</td></tr>
<tr><td></td><td colspan="3">Независимые n_1 и n_2</td><td colspan="2">Независимые n_1 и n_2</td><td colspan="2">Зависимые: $n_1=n_2=n$</td></tr>
<tr><td rowspan="6">Генеральные дисперсии известны $D[X_1]$ и $D[X_2]$</td><td colspan="3">Задача №3
$K_{\text{набл}}=\dfrac{\bar{X}_1-\bar{X}_2}{\sqrt{\dfrac{D[X_1]}{n_1}+\dfrac{D[X_2]}{n_2}}}$
H_0: $M[X_1]=M[X_2]$, H_0 принимается, если:</td><td colspan="4" rowspan="5"></td></tr>
<tr><td>Двустор. критерий</td><td colspan="2">Односторонний критерий</td></tr>
<tr><td>H_1: $M[X_1]\neq M[X_2]$</td><td>H_1: $M[X_1]>M[X_2]$</td><td>H_1: $M[X_1]<M[X_2]$</td></tr>
<tr><td>$\lvert K_{набл}\rvert < K_{кр,\alpha/2}$</td><td>$K_{набл} < K_{кр,\alpha}$</td><td>$K_{набл} > -K_{кр,\alpha}$</td></tr>
<tr><td>$\Phi(K_{кр,\alpha/2})=(1-\alpha)/2$</td><td colspan="2">$\Phi(K_{кр,\alpha})=(1-2\alpha)/2$</td></tr>
<tr><td colspan="3">$K_{кр}$ по таблице интеграла Лапласа, $\Phi(x)$</td></tr>
<tr><td rowspan="6">Генеральные дисперсии не известны</td><td colspan="3" rowspan="6"></td><td colspan="2">Задача №4
$D[X_1]\approx D[X_2]$
$K_{\text{набл}}=t_{\text{набл}}=$
$=\dfrac{\bar{X}_1-\bar{X}_2}{\sqrt{(n_1-1)\cdot S_1^2+(n_2-1)\cdot S_2^2}}\cdot\sqrt{\dfrac{n_1n_2(n_1+n_2-2)}{n_1+n_2}}$,
$k=n_1+n_2-2$.
H_0: $M[X_1]=M[X_2]$
H_0 принимает если:</td><td colspan="2">Задача №5
$K_{\text{набл}}=t_{\text{набл}}=\dfrac{\bar{d}\cdot\sqrt{n}}{S_d}$, где
$\bar{d}=\dfrac{\sum d_j}{n}$; $d_j=X_{1j}-X_{2j}$;
$S_d=\sqrt{\dfrac{\sum d_j^2-\dfrac{(\sum d_j)^2}{n}}{n-1}}$; $k=n-1$
H_0 принимается если:</td></tr>
<tr><td>Двустор. критерий</td><td colspan="2">Односторонний критерий</td><td>Двустор. критерий</td><td colspan="2">Односторонний критерий</td></tr>
<tr><td>H_1:$M[X_1]\neq M[X_2]$</td><td>H_1:$M[X_1]>M[X_2]$</td><td>H_1:$M[X_1]<M[X_2]$</td><td>H_1:$M[X_1]\neq M[X_2]$:
M|X,^M|X:^</td><td>H_1:$M[X_1]>M[X_2]$</td><td>H_1:$M[X_1]<M[X_2]$</td></tr>
<tr><td>$\lvert t_{набл}\rvert<t_{кр,\alpha/2,k}$</td><td>$t_{набл}<t_{кр,\alpha,k}$</td><td>$t_{набл}>-t_{кр,\alpha,k}$</td><td>$\lvert t_{набл}\rvert<t_{кр,\alpha/2,k}$</td><td>$t_{набл}<t_{кр,\alpha,k}$</td><td>$t_{набл}>-t_{кр,\alpha,k}$</td></tr>
<tr><td colspan="3">$t_{кр}$ по таблице распределения Стьюдента</td><td colspan="3">$t_{кр}$ по таблице распределения Стьюдента</td></tr>
</table>

Литература

1. *Айвазян*, *С. А.* Прикладная статистика. Исследование зависимостей / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. — М. : Финансы и статистика, 1985. — 488 с.

2. *Айвазян*, *С. А.* Прикладная статистика. Классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков и др. — М. : Финансы и статистика, 1989. — 607 с.

3. *Айвазян*, *С. А.* Прикладная статистика. Основы моделирования и первичная обработка данных / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. — М. : Финансы и статистика, 1985. — 472 с.

4. *Андрухаев*, *Х. М.* Сборник задач по теории вероятностей / Х. М. Андрухаев. — М. : Просвещение, 1985. — 160 с.

5. *Беклемишев*, *Д. В.* Курс аналитической геометрии и линейной алгебры: учебник. — 16-е изд., стер. — СПб. : Лань, 2019. — 448 с.

6. *Беклемишев*, *Д. В.* Дополнительные главы линейной алгебры : учеб. пособие / Д. В. Беклемишев. — 2-е изд., перераб. и доп. — СПб. : Лань, 2008. — 496 с.

7. *Большев*, *Л. Н.* Таблицы математической статистики / Л. Н. Большев, Н. В. Смирнов. — М. : Наука, 1983. — 416 с.

8. *Бондаренко*, *П. С.* Теория вероятностей и математическая статистика: учеб. пособие для бакалавров / П. С. Бондаренко, Г. В. Горелова, И. А. Кацко. — Краснодар : КубГАУ, 2013. — 340 с. : ил. — (Серия: Вероятность, статистика и прикладные исследования в аграрном университете).

9. *Бондаренко*, *П. С.* Теория вероятностей и математическая статистика: учебное пособие / П. С. Бондаренко, Г. В. Горелова, И. А. Кацко; под ред. И. А. Кацко, А. И. Трубилина. — М. : КНОРУС, 2017. — 390 с.

10. *Борзых*, *Д. А.* Теория вероятностей и математическая статистика в задачах. Более 360 задач и упражнений. — М. : УРСС, 2018. — 240 с.

11. *Борзых*, *Д. А.* Эконометрика в задачах и упражнениях / Д. А. Борзых, Б. Б. Демешев. — 2-е изд. — М. : УРСС, 2017. — 304 с.

12. *Боровков*, *А. А.* Математическая статистика : учебник. — М. : Наука, 1984. — 472 с.

13. *Боровков*, *А. А.* Теория вероятностей. — 4-е изд. — М. : Едиториал УРСС, 2003. — 472 с.

14. *Варден*, *Б. Л.* Математическая статистика. — М. : Изд. иностранной литературы, 1960. — 435 с.

15. *Венецкий*, *Н. Г.* Теория вероятностей и математическая статистика / Н. Г. Венецкий, Г. Е. Кильдишев. — 3-е изд. — М. : Статистика, 1975. — 264 с.

16. *Вентцель*, *Е. С.* Задачи и упражнения по теории вероятностей : учеб. пособие для вузов / Е. С. Вентцель, Л. А. Овчаров. — 3-е изд. — М. : Высшая школа, 2000. — 366 с.

17. *Вентцель*, *Е. С.* Исследование операций. — М. : Советское радио, 1972. — 552 с.

18. *Вентцель*, *Е. С.* Теория вероятностей и ее инженерные приложения / Е. С. Вентцель, Л. А. Овчаров. — М. : Высшая школа, 2000. — 480 с.

19. *Вентцель*, *Е. С.* Теория вероятностей : учебник для вузов. — 8-е изд., стер. — М. : Высш. школа, 2002. — 575 с.

20. Возможно да, возможно нет. Фишер. Статистический вывод // Наука. Величайшие теории: выпуск 47 ; пер. с итал. — М. : Де Агостини, 2015. — 176 с.

21. *Волкова*, *В. Н.* Основы теории систем и системного анализа / В. Н. Волкова, А. А. Денисов. — СПб. : Изд-во СПбГТУ, 1997. — 510 с.

22. *Волкова*, *В. Н.* Постепенная формализация моделей принятия решений. — СПб. : Изд-во Политехн. ун-та, 2006. — 120 с.

23. *Волкова*, *В. Н.* Открытые системы. Переосмысливая Л. фон Берталанфи : монография. — СПб. : ПОЛИТЕХ-ПРЕСС, 2019. — 440 с.

24. *Волкова*, *В. Н.* Моделирование систем и процессов : учебник для академического бакалавриата / В. Н. Волкова [и др.]; под редакцией В. Н. Волковой, В. Н. Козлова. — М. : Юрайт, 2016. — 450 с.

25. *Герасимович*, *А. И.* Математическая статистика / А. И. Герасимович, Я. И. Матвеева. — М. : Высшая школа, 1978. — 200 с.

26. *Гласс*, *Д.* Статистические методы в педагогике и психологии / Д. Гласс, Д. Стенли. — М. : Прогресс, 1976. — 496 с.

27. *Гливенко*, *В. И.* Теория вероятностей : учебник для высших педагогических учебных заведений. — 2-е изд. — М. : ЛЕНАНД, 2019. — 138 с.

28. Гмурман В. Е. Теория вероятностей и математическая статистика : учебник для вузов / В. Е. Гмурман. — 12-е изд. — М. : Юрайт, 2020. — 479 с.

29. *Гмурман*, *В. Е.* Руководство к решению задач по теории вероятностей и математической статистике : учебное пособие для бакалавриата и специалитета. — 11-е изд., перераб. и доп. — М. : Юрайт, 2019. — 406 с.

30. *Гнеденко*, *Б. В.* Курс теории вероятностей : учебник. — 6-е изд., перераб. и доп. — М. : Наука, 1988. — 448 с.

31. *Горелова*, *Г. В.* Теория вероятностей и математическая статистика в примерах и задачах с применением *Excel* : учеб. пособие для вузов / Г. В. Горелова, И. А. Кацко. — 4-е изд., испр. и доп. — Ростов н/Д. : Феникс, 2006. — 475 с.

32. *Грэхэм*, *Р.* Конкретная математика. Основания информатики / Р. Грэхэм, Д. Кнут, О. Паташник ; пер с англ. — М. : Мир, 1998. — 703 с.

33. *Гумбель*, *Э.* Статистика экстремальных значений. — М. : Мир, 1965. — 450 с.

34. *Давенпорт*, *В. Б.* Введение в теорию случайных сигналов и шумов / В. Б. Давенпорт, В. Л. Рут. — М. : ИЛ, 1960. — 468 с.

35. *Дауни*, *А. Б.* Байесовские модели. — М. : ДМК Пресс, 2018. — 182 с.

36. *Де Гроот*, *М.* Оптимальные статистические решения / М. Де Гроот ; пер. с англ. — М. : Мир, 1974. — 496 с.

37. *Демиденко*, *Е. З.* Линейная и нелинейная регрессия. — М. : Финансы и статистика, 1981. — 302 с.

38. *Дрейпер*, *И.* Прикладной регрессионный анализ / И. Дрейпер, Г. Смит ; пер. с англ. — 3-е изд. — М. : Вильямс, 2007. — 912 с.

39. *Доросинский*, *Л. Г.* Введение в теорию обработки сигналов от пространственно-распределенных целей в РСА. — Ульяновск : Зебра, 2016. — 145 с.

40. *Дружинин*, *Н. К.* Логика оценки статистических гипотез. — М. : Статистика, 1973. — 212 с.

41. *Дынкин*, *Е. Б.* Управляемые марковские процессы и их приложения / Е. Б. Дынкин, А. А. Юшкевич. — М. : Наука, 1975. — 341 с.

42. *Дюге*, *Д.* Теоретическая и прикладная статистика. — М. : Наука, 1972. — 384 с.

43. *Езикиэл*, *М.* Методы анализа корреляций и регрессий: линейных и криволинейных / М. Езикиэл, К. Фокс. — М. : Статистика, 1966. — 557 с.

44. *Елисеева*, *И. И.* Теория статистики с основами теории вероятностей : учеб. пособие для вузов / И. И. Елисеева, В. С. Князевский, Л. И. Ниворожкина [и др.] ; под ред. И. И. Елисеевой. — М. : ЮНИТИ-ДАНА, 2001. — 446 с.

45. *Елисеева*, *И. И.* Эконометрика: учебник / И. И. Елисеева, С. В. Курышева, Т. В. Костеева [и др.] ; под ред. И. И. Елисеевой. — 2-е изд., перераб. и доп. — М. : Финансы и статистика, 2005. — 576 с.

46. *Закс*, *Л.* Статистическое оценивание / Л. Закс ; пер. с нем. В. Н. Варыгина, под ред. Ю. П. Адлера, В. Г. Горского. — М. : Статистика, 1976. — 598 с.

47. *Закс*, *Л.* Теория статистических выводов / Л. Закс ; пер. с англ. Е. В. Чепурина, под ред. Ю. К. Беляева. — М. : Мир, 1975. — 776 с.

48. *Зорич*, *В. А.* Математический анализ. — 8-е изд., испр. — М. : МЦНМО, 2017. — 564 с. (Часть I), 676 с. (Часть II).

49. *Зорич*, *В. А.* Математический анализ задач естествознания. — М. : МЦНМО, 2017. — 160 с.

50. *Ивченко*, *Г. И.* Введение в математическую статистику / Г. И. Ивченко, Ю. И. Медведев. — 2-е изд., испр. и доп. — М. : Ленанд, 2017. — 608 с.

51. *Йейтс*, *Ф.* Выборочный метод в переписях и обследованиях / Ф. Йетс ; пер. с англ. Е. И. Арона, под ред. А. Г. Волкова. — М. : Статистика, 1965. — 434 с.

52. *Каханер*, *Д.* Численные методы и программное обеспечение / Д. Каханер, К. Моулер, С. Нэш ; пер. с англ., под ред. Х. Д. Икрамова. — 2-е изд., стер. — М. : Мир, 2001. — 575 с.

53. *Кацко*, *И. А.* Теория вероятностей и математическая статистика : учебник / И. А. Кацко, П. С. Бондаренко, Г. В. Горелова. — 2-е изд., перераб. и доп. — М. : КНОРУС, 2020. — 800 с.

54. *Кельберт*, *М. Я.* Вероятность и статистика в примерах и задачах. Т. 1: Основные понятия теории вероятностей и математической статистики / М. Я. Кельберт, Ю. М. Сухов. — М. : МЦНМО, 2007. — 456 с.

55. *Кельберт*, *М. Я.* Вероятность и статистика в примерах и задачах. Т. 2: Марковские цепи как отправная точка теории случайных процессов и их приложения / М. Я. Кельберт, Ю. М. Сухов. — М. : МЦНМО, 2010. — 560 с.

56. *Кемени*, *Дж.* Введение в конечную математику / Дж. Кемени, Дж. Снелл, Дж. Томпсон ; пер. с англ., под ред. И. М. Яглома. — 2-е изд. стер. — М. : Мир, 1965. — 488 с.

57. *Кендэл*, *М.* Временные ряды / М. Кендэл ; пер. с англ. Ю. П. Лукашина. — М. : Финансы и статистика, 1981. — 199 с.: ил.

58. *Кимбл*, *Г.* Как правильно пользоваться статистикой. — М. : Финансы и статистика, 1982. — 294 с.

59. *Кнут*, *Д. Э.* Искусство программирования. Т. 1: Основные алгоритмы / Д. Э. Кнут ; пер. с англ. — 3-е изд. — М. : Вильямс, 2001. — 720 с.

60. *Кнут*, *Д. Э.* Искусство программирования. Т. 2: Получисленные алгоритмы / Д. Э. Кнут ; пер. с англ. — 3-е изд. — М. : Вильямс, 2001. — 832 с.

61. *Коваленко*, *И. Н.* Теория вероятностей и математическая статистика : учеб. пособие / И. Н. Коваленко, А. А. Филиппова. — 2-е изд., перераб. и доп. — М. : Высшая школа, 1982. — 256 с.

62. *Кокрен*, *У.* Методы выборочного исследования / У. Кокрен ; пер. с англ. И. М. Сонина, под ред. А. Г. Волкова. — М. : Статистика, 1976. — 440 с.

63. *Колемаев*, *В. А.* Теория вероятностей и математическая статистика : учеб. пособие / В. А. Колемаев, О. Е. Староверов, В. Б. Турундаевский ; под ред. В. А. Колемаева. — М. : Высшая Школа, 1991. — 400 с.

64. *Колемаев*, *В. А.* Теория вероятностей и математическая статистика: учебник / В. А. Колемаев, В. Н. Калинина ; под ред. В. А. Колемаева. — М. : ИНФРА-М, 1997. — 302 с.

65. *Колмогоров*, *А. Н.* Введение в теорию вероятностей / А. Н. Колмогоров, И. Г. Журбенко, А. В. Прохоров. — М. : Наука, 1982. — 160 с.

66. *Колмогоров*, *А. Н.* Основные понятия теории вероятностей / А. Н. Колмогоров. — М. : ЛЕНАНД, 2019. — 120 с.

67. *Крамер*, *Г.* Математические методы статистики / Г. Крамер ; пер. с англ. А. С. Монина, А. А. Петрова, под. ред. академика А. Н. Колмогорова. — 2-е изд. стер. — М. : Мир, 1975. — 648 с.

68. *Кремер*, *Н. Ш.* Теория вероятностей и математическая статистика : учебник для вузов / Н. Ш. Кремер. — 2-е изд., перераб. и доп. — М. : ЮНИТИ-ДАНА, 2004. — 573 с.

69. *Лагутин*, *М. Б.* Наглядная математическая статистика : учеб. пособие / М. Б. Лагутин. — 7-е изд. — М. : Лаборатория знаний, 2019. — 472 с.

70. *Ларичев*, *О. И.* Теория и методы принятия решений, а также Хроника событий в Волшебных Странах. — М. : Логос, 2008. — 392 с.

71. *Левин*, *Б. Р.* Теоретические основы статистической радиотехники Т. 2 / Б. Р. Левин. — М. : Советское радио, 1968. — 503 с.

72. *Левин*, *Б. Р.* Теоретические основы статистической радиотехники. — 3-е изд., перераб. и доп. — М. : Радио и связь, 1989. — 656 с.

73. *Леман*, *Э.* Проверка статистических гипотез. — 2-е изд., испр. — М. : Наука, 1979. — 408 с.

74. *Линник*, *Ю. В.* Метод наименьших квадратов и основы теории обработки наблюдений. — М. : Физматгиз, 1962. — 352 с.

75. *Лоэв*, *М.* Теория вероятностей / М. Лоэв ; пер. с англ. Б. А. Севостьянова, под ред. Ю. В. Прохорова. — М. : ИЛ, 1962. — 720 с.

76. *Луценко*, *Е. В*. Интеллектуальные информационные системы : учеб. пособие. — 2-е изд., испр. и доп. — Краснодар : КубГАУ, 2006. — 645 с.

77. *Львовский*, *Е. Н*. Статистические методы построения эмпирических формул : учеб. пособие. — 2-е изд. — М. : Высшая школа, 1988. — 239 с.

78. *Ляховецкий*, *А. М*. Статистика : учеб. пособие для бакалавров / А. М. Ляховецкий, Н. В. Климова, Е. В. Кремянская [и др.]; под ред. В. И. Нечаева. — Краснодар : КубГАУ, 2013. — 359 с. — (Серия: Вероятность, статистика и прикладные исследования в аграрном университете).

79. *Ляховецкий*, *А. М*. Статистика : учеб. пособие / А. М. Ляховецкий, Е. В. Кремянская, Н. В. Климова [и др.]; под ред. В. И. Нечаева. — М. : КНОРУС, 2018. — 362 с.

80. *Мизес*, *Р*. Вероятность и статистика / Р. Мизес ; пер. с нем., под ред. и с предисл. А. Я. Хинчина. — 2-е изд., стер. — М. : КомКнига, 2006. — 264 с.

81. *Митропольский*, *А. К*. Техника статистических вычислений. — М. : Наука, 1971. — 576 с.

82. *Мостеллер*, *Ч. Ф*. Анализ данных и регрессия: в 2-х вып. / Ч. Ф. Мостеллер, Дж. Тьюки ; пер. с анг. Б. Л. Розовского, под. ред. Ю. П. Адлера. — М. : Финансы и статистика, 1982. — 317 с. (Вып. 1), 239 с. (Вып. 2).

83. *Мостеллер*, *Ч. Ф*. Пятьдесят занимательных вероятностных задач с решениями. — М. : Наука, 1975. — 112 с.

84. *Наследов*, *А. Д*. Математические методы психологического исследования. Анализ и интерпретация данных. — 3-е изд. — СПб. : Речь, 2007. — 392 с.

85. *Нейман*, *Ю*. Вводный курс теории вероятностей и математической статистики / Ю. Нейман ; пер. с англ. Н. Митрофанова, А. Хусу, под ред. Ю. В. Линника. — М. : Наука, 1968. — 448 с.

86. *Ниворожкина*, *Л. И*. Математическая статистика с элементами теории вероятностей в задачах с решениями : учеб. пособие для бакалавров / Л. И. Ниворожкина, З. А. Морозова, И. Э. Гурьянова ; под ред. проф. Л. И. Ниворожкиной. — 2-е изд., перераб. и доп. — М. : Дашков и К°, 2015. — 480 с.

87. *Орлов*, *А. И*. Современная прикладная статистика (обобщающая статья) / А. И. Орлов. — М. : Заводская лаборатория, 1998. — Т. 64, № 3 — С. 52–60.

88. *Орлов*, *А. И*. Прикладная статистика : учебник. — М. : Экзамен, 2004. — 656 с.

89. *Орлов*, *А. И*. Организационно-экономическое моделирование. Ч. 1. Нечисловая статистика. В 3 ч. : учебник. — М. : Изд-во МГТУ им. Н. Э. Баумана, 2009. — 541 с.

90. *Орлов*, *А. И*. Организационно-экономическое моделирование. Ч. 2. Экспертные оценки. В 3 ч. : учебник. — М. : Изд-во МГТУ им. Н. Э. Баумана, 2011. — 486 с.

91. *Орлов*, *А. И*. Организационно-экономическое моделирование. Ч. 3. Статистические методы анализа данных. В 3 ч. : учебник / А. И. Орлов. — М. : Изд-во МГТУ им. Н. Э. Баумана, 2012. — 623 с.

92. *Прохоров*, *Ю. В*. Лекции по теории вероятностей и математической статистике: учебник / Ю. В. Прохоров, Л. С. Пономаренко. — 2-е изд., испр. и доп. — М. : МГУ, 2012. — 256 с.

93. *Пугачев*, *В. С*. Теории вероятностей и математическая статистика : учеб. пособие / В. С. Пугачев. — 2-е изд., испр. и доп. — М. : ФИЗМАТЛИТ, 2002. — 496 с.

94. *Пустыльник*, *Е. И*. Статистические методы анализа и обработки наблюдений. — М. : Наука, 1968. — 288 с.

95. *Реньи*, *А*. Трилогия о математике (Диалоги о математике. Письма о вероятности. Дневник. Записки студента по теории информации.) / А. Реньи ; пер. с венгер. — М. : Мир, 1980. — 375 с.

96. *Розанов*, *Ю. А*. Теория вероятностей, случайные процессы и математическая статистика: учебник для вузов. — М. : Наука, 1985. — 320 с.

97. Сборник задач по теории вероятностей, математической статистике и теории случайных функций : учеб. пособие / под ред. А. А. Свешникова. — М. : Наука, 1970. — 656 с.

98. *Севастьянов*, *Б. А*. Курс теории вероятностей и математической статистики. — М. : Наука, 1982. — 256 с.

99. *Секей*, *Г*. Парадоксы в теории вероятностей и математической статистике / Г. Секей ; пер. с англ. — М. : Мир, 1990. — 240 с.

100. *Смирнов*, *Н. В*. Курс теории вероятностей и математической статистики для технических приложений / Н. В. Смирнов, И. В. Дунин-Барковский. — М. : Наука, 1969. — 511 с.

101. *Соколов*, *Г. А*. Теория вероятностей. Управляемые цепи Маркова в экономике / Г. А. Соколов, Н. А. Чистякова. — М.: ФИЗМАТЛИТ, 2005. — 248 с.

102. *Соколов*, *Г. А*. Управляемые цепи Маркова в экономике (дискретные цепи Маркова с доходами) : учебник / Г. А. Соколов. — 2-е изд. — М. : ИНФРА-М, 2015. — 158 с.

103. *Сосулин*, *Ю. Г*. Теория обнаружения и оценивания стохастических сигналов / Ю. Г. Сосулин. — М. : Советское радио, 1978. — 320 с.

104. *Сошникова*, *Л. А*. Многомерный статистичекий анализ в экономике: учеб. пособие для вузов / Л. А. Сошникова, В. Н. Тамашевич, Г. Уебе [и др.] ; под ред. проф. В. Н. Тамашевича. — М. : ЮНИТИ-ДАНА, 1999. — 598 с.

105. *Стивенс*, *Р*. Visual Basic. Готовые алгоритмы / Р. Стивенс ; пер. с англ. — М. : ДМК Пресс, 2000. — 384 с.

106. *Стренг*, *Г*. Линейная алгебра и ее применения / Г. Стренг. — М. : МИР, 1980. — 456 с.

107. *Таха*, *Х. А*. Введение в исследование операций / Х. А. Таха ; пер. с англ. — 6-е изд. — М. : Вильямс, 2001. — 912 с.

108. Теория выбора и принятия решений : учеб. пособие / И. М. Макаров, Т. М. Виноградская, А. А. Рубчинский [и др.]. — М. : Наука, 1982. — 328 с.

109. *Тутубалин*, *В. Н*. Теория вероятностей. — М. : МГУ, 1972. — 232 с.

110. *Тутубалин*, *В. Н*. Границы применимости (вероятностно-статистические методы и их возможности). — М. : Знание, 1977. — 64 с.

111. *Уилкс*, *С*. Математическая статистика / С. Уилкс ; пер с англ., под ред. Ю. В. Линника. — М. : Наука, 1967. — 632 с.

112. *Феллер*, *В*. Введение в теорию вероятностей и ее приложения: в 2-х томах. Т. 1 / В. Феллер ; пер. с англ. — М. : Мир, 1984. — 528 с.

113. *Фихтенгольц*, *Г. М.* Курс дифференциального и интегрального исчисления: Т. 1 / Г. М. Фихтенгольц. — М. : Наука, 1966. — 608 с.

114. *Фишер*, *Р*. Статистические методы для исследователей / Р. Фишер ; пер с 12 англ. изд. — М. : Госстатиздат, 1958. — 267 с.

115. *Френкс*, *Л*. Теория сигналов. — М. : Советское радио, 1974. — 344 с.

116. *Хабаров*, *С*. Экспертные системы : учеб. пособие. — СПб. : ЛТА: Кафедра информатики и информационных систем, ФЭУ. — URL: http://www.habarov.spb.ru/new_es/index.htm.

117. *Хальд*, *А*. Математическая статистика с техническими приложениями / А. Хальд; пер. с англ. Н. Н. Воробьева, В. В. Петрова, А. П. Хусу, под ред. Ю. В. Линника. — М. : Изд-во иностранной литературы, 1956. — 664 с.

118. *Хей*, *Д*. Введение в методы байесовского статистического вывода / Д. Хей ; пер. с англ. — М. : Финансы и статистика, 1987. — 335 с.

119. *Хинчин*, *А. Я*. Об аналитическом аппарате физической статистики: избранные работы по математической физике / А. Я. Хинчин. — М. : ЛЕНАНД, 2018. — 176 с.

120. *Хинчин*, *А. Я*. Основные законы теории вероятностей: Теорема Лапласа. Закон больших чисел. Закон повторного логарифма. — 2-е изд. — М. : ЛЕНАНД, 2017. — 88 с.

121. *Ховард*, *Р*. Динамическое программирование и марковские процессы. — М. : Советское радио, 1964. — 192 с.

122. *Хьюстон*, *А*. Дисперсионный анализ / А. Хьюстон ; пер с англ. А. Г. Кругликова. — М. : Статистика, 1971. — 88 с.

123. *Цейтлин*, *Н. А*. Из опыта аналитического статистика. — М. : Солар, 2007. — 906 с.

124. *Чернова*, *Н. И*. Теория вероятностей : учеб. пособие. — Новосибирск: Новосибирский гос. ун-т, 2007. — 160 с.

125. *Чернова*, *Н. И*. Математическая статистика : учеб. пособие. — 2-е изд. — Новосибирск : Новосибирский гос. ун-т, 2014. — 150 с.

126. *Четыркин*, *Е. М*. Вероятность и статистика / Е. М. Четыркин, И. Л. Калихман. — М. : Финансы и статистика, 1982. — 319 с.

127. *Чжун*, *К. Л*. Элементарный курс теории вероятностей. Стохастические процессы и финансовая математика / К. Л. Чжун, Ф. АитСахлиа ; пер. с англ. — М. : БИНОМ. Лаборатория знаний, 2007. — 455 с.

128. *Чистяков*, *В. П*. Курс теории вероятностей. — 4-е изд. — М. : Агар, 1996. — 256 с.

129. *Шведов*, *А. С*. Теория вероятностей и математическая статистика : учеб. пособие / А. С. Шведов. — М. : Высшая школа экономики, 2016. — 280 с.

130. *Шеффе*, *Г*. Дисперсионный анализ. — 2-е изд. — М. : Наука, 1980. — 512 с.

131. *Ширяев*, *А. Н.* Вероятность : учеб. пособие. — 2-е изд. — М. : Наука, 1989. — 640 с.

132. *Шуровьески*, *Д.* Мудрость толпы / Д. Шуровьески. — М. : Вильямс, 2007. — 304 с.

133. *Шурыгин*, *А. М.* Прикладная стохастика: робастность, оценивание, прогноз. — М. : Финансы и статистика, 2000. — 224 с.

134. Эконометрика. Практикум : учебно-практическое пособие / коллектив авторов ; под ред. И. А. Кацко. — М. : КНОРУС, 2019. — 218 с.

135. *Юденков*, *В. А.* Дисперсионный анализ : учеб. пособие. — Минск : Вышэйш. шк, 1982. — 95 с.

136. *Krauth*, *W.* Introduction to Monte Carlo algorithms. summer school in Beg-Rohu (France) and Budapest1996, 2006. cel-00092936. — URL: https://cel.archives-ouvertes.fr/cel-00092936/document.

Предметный указатель

Игорь Александрович КАЦКО,
Петр Сергеевич БОНДАРЕНКО,
Галина Викторовна ГОРЕЛОВА

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебник

Издание третье, исправленное и дополненное

Зав. редакцией
физико-математической литературы *О. Е. Гайнутдинова*
Ответственный редактор *В. В. Яески*
Корректор *Е. С. Шумская*
Выпускающий *В. А. Плотникова*

ЛР № 065466 от 21.10.97
Гигиенический сертификат 78.01.10.953.П.1028
от 14.04.2016 г., выдан ЦГСЭН в СПб

Издательство «ЛАНЬ»
lan@lanbook.ru; www.lanbook.com
196105, Санкт-Петербург, пр. Юрия Гагарина, д. 1, лит. А
Тел./факс: (812) 336-25-09, 412-92-72
Бесплатный звонок по России: 8-800-700-40-71

Подписано в печать 17.01.23.
Бумага офсетная. Гарнитура Школьная. Формат 70×100 $^{1}/_{16}$.
Печать офсетная/цифровая. Усл. п. л. 35,43. Тираж 30 экз.

Заказ № 108-23.

Отпечатано в полном соответствии с качеством
предоставленного оригинал-макета в АО «Т8 Издательские Технологии».
109316, г. Москва, Волгоградский пр., д. 42, к. 5.

ГДЕ КУПИТЬ

ДЛЯ ОРГАНИЗАЦИЙ:

Для того, чтобы заказать необходимые Вам книги, достаточно обратиться в любую из торговых компаний Издательского Дома «ЛАНЬ»:

по России и зарубежью

«ЛАНЬ-ТРЕЙД»
РФ, 196105, Санкт-Петербург, пр. Ю. Гагарина, 1
тел.: (812) 412-85-78, 412-14-45, 412-85-82
тел./факс: (812) 412-54-93
e-mail: trade@lanbook.ru
ICQ: 446-869-967

www.lanbook.com
пункт меню **«Где купить»**
раздел **«Прайс-листы, каталоги»**

в Москве и в Московской области

«ЛАНЬ-ПРЕСС»
109387, Москва, ул. Летняя, д. 6
тел.: (499) 722-72-30, (495) 647-40-77
e-mail: lanpress@lanbook.ru

в Краснодаре и в Краснодарском крае

«ЛАНЬ-ЮГ»
350901, Краснодар, ул. Жлобы, д. 1/1
тел.: (861) 274-10-35
e-mail:lankrd98@mail.ru

ДЛЯ РОЗНИЧНЫХ ПОКУПАТЕЛЕЙ:

интернет-магазин

Издательство «Лань»: http://www.lanbook.com

магазин электронных книг

Global F5

http://globalf5.com/

Издательство «ЛАНЬ»

Мы будем благодарны Вам
за пожелания по издаваемой нами литературе,
а также за предложения по изданию книг
новых авторов или переизданию
уже существующих трудов.

Мы заинтересованы в сотрудничестве
с высшими учебными заведениями
и открыты для Ваших предложений
по улучшению нашего взаимодействия.

Теперь Вы можете звонить нам бесплатно
из любых городов России по телефону

8-800-700-40-71

Дополнительную информацию
и ответы на вопросы Вы также можете получить,
обратившись по электронной почте:

cs@lanbook.ru