

**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ,
МОЛОДЕЖИ И СПОРТА УКРАИНЫ**

ХАРЬКОВСКИЙ НАЦИОНАЛЬНЫЙ ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ

Малярець Л. М.

Егоршин А. А.

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

**Учебно-практическое пособие
для иностранных студентов**

Харьков. Изд. ХНЭУ, 2013

УДК 519.2(042.4)

ББК 22.17я73

М21

Рецензенты: докт. физ.-мат. наук, профессор кафедры высшей математики Национального аэрокосмического университета им. Н. Е. Жуковского «ХАИ» *Проценко В. С.*; докт. экон. наук, профессор, зав. кафедрой математики и математических методов в экономике Донецкого национального университета *Христиановский В. В.*

Рекомендовано к изданию решением ученого совета Харьковского национального экономического университета.

Протокол № 6 от 07.02.2012 г.

Малярец Л. М.

М21 Теория вероятностей и математическая статистика : учебно-практическое пособие для иностранных студентов / Л. М. Малярец, А. А. Егоршин. – Х. : Изд. ХНЭУ, 2013. – 304 с. (Русск. яз.)

Изложен теоретический материал по учебной дисциплине, сформированный по лекциям, каждая из которых сопровождается примерами и вопросами для самопроверки. Разработаны лабораторные работы для укрепления теоретических знаний и формирования практических навыков по теории вероятностей и математической статистике.

Рекомендовано для всех студентов, которые изучают теорию вероятностей и математическую статистику.

ISBN 978-966-676-569-0

УДК 519.2(042.4)

ББК 22.17я73

© Харьковский национальный
экономический университет, 2013

© Малярец Л. М.

Егоршин А. А.

2013

Введение

Учебная дисциплина «Теория вероятностей и математическая статистика» много лет преподается студентам всех специальностей Харьковского национального экономического университета преподавателями кафедры высшей математики и экономико-математических методов. В 70-е годы XX столетия многие преподаватели университета слушали курс лекций по теории вероятностей и математической статистики у прекрасного специалиста, руководителя Харьковской научной школы математиков-экономистов Воловельской Софьи Натановны. XXI век существенно отличается от прошлого. Появились персональные компьютеры и доступное программное обеспечение. Ушли в прошлое громоздкие ручные вычисления, разрешены многие прикладные проблемы многомерного статистического анализа. Появилось много новых книг и учебников по теории вероятностей. В подавляющем большинстве пособий принят следующий стиль изложения: многословно и высокопарно распространяться по поводу очевидных утверждений и в то же время пропускаются все доказательства. Однако бесспорным фактом является то, что простые вещи следует излагать просто, приводя все ключевые доказательства или мелким шрифтом, или в виде приложений, или делая ссылку на доступные источники, где имеются эти доказательства, так, как сделано в данном учебном пособии.

Более того, несмотря на требования к строгости изложения, в последнее время окончательно закрепились странные словосочетания типа «сумма событий», «произведение событий», «частотная гистограмма» или, например, определение: «Математическое ожидание – это сумма произведений значений случайной величины на вероятность их появления». Ошибкой, на взгляд авторов, в данном определении есть то, что вместо ответа на вопрос: «Что это такое?» приведено описание вычислительной формулы для частного случая дискретной случайной величины. Поэтому авторы в данном пособии попытались не следовать тенденции сокращенного изложения материала, а представили материал с объяснениями, исходя из теории и практики применения самих инструментов для решения экономических задач.

В пособии, согласно рабочей программе нормативной дисциплины «Теория вероятностей и математическая статистика», на изучение теории выделяются 17 лекций, 8 лабораторных работ и 8 практических занятий. Отличительной особенностью изложения теории в пособии является структурирование ее по лекциям, которые содержат весь необходимый материал для приобретения компетентностей теоретического содержания инструментов теории вероятностей

и математической статистики. Теоретический материал сопровождается многочисленными решенными примерами и задачами, демонстрирующими применение соответствующих теорем, определений, формул.

В пособии также изложено 8 лабораторных работ, где детально объясняется их выполнение. Для закрепления материала каждая лекция и каждая лабораторная работа заканчивается вопросами для самопроверки. Кроме того, в виде справочника написаны разделы «Теория вероятностей в вопросах и ответах», «Математическая статистика в вопросах и ответах», «Регрессионный анализ в вопросах и ответах».

После изучения дисциплины студенты должны уметь: определять вероятности случайных событий; знать основные теоремы теории вероятностей и математической статистики, их экономическую интерпретацию; применять схемы независимых испытаний для решения экономических задач; знать законы распределения и числовые характеристики одномерных и многомерных случайных величин; иметь представление о теории случайных процессов; уметь первоначально обработать статистические данные; оценивать параметры распределения; проверить статистические гипотезы, знать элементы дисперсионного анализа, теории корреляции и регрессионного анализа.

В целом, изучив материал пособия, студенты должны приобрести такие компетентности, как: знания фундаментальных основ теории вероятностей и математической статистики, умения использовать данные инструменты для решения реальных задач в экономике, ценности практической реализации методов теории вероятностей и математической статистики в деятельности экономистов, менеджеров, аналитиков.

1. Основные понятия теории вероятностей

Основы знаний об окружающем мире у человека закладываются с младенчества. Именно тогда он на практических примерах усваивает основные понятия и формулирует правила поведения, позволяющие ему выжить и избежать большинства неприятностей.

В школе начинается подробное изучение явлений природы, но для удобства усвоения все явления классифицируются по отдельным наукам. Так, выделяют определенный круг вопросов и называют его физикой, другие группы явлений – арифметикой, геометрией и т. д. Теорию вероятностей начинают изучать поздно, чаще всего уже после окончания школы, когда у человека сформировалось критическое мышление. Приступая к изучению нового раздела знаний, учащийся (студент) ожидает, что ему сначала будут даны строгие определения всех новых понятий нового этого раздела науки. И тут оказывается, что основные понятия любой науки только называются, демонстрируются на примерах и устанавливается иерархия соподчинения; строгие же определения отсутствуют. Даже в такой науке, как математика, изначально считаются известными такие понятия, как «точка», «прямая», «плоскость»; определений этих основополагающих понятий нет. Соподчиненность устанавливается фразами типа: «На прямой лежат точки». Учащийся обычно не задает лишних вопросов, ему и так кажется все понятным.

Приведем показательный пример из физики. Ни в одной книге, как в учебной, так и научной, нельзя найти определения температуры, хотя всем как будто бы понятно, что это такое. В каждом терминологическом словаре температура определяется как средняя кинетическая энергия молекул в хаотическом движении. Тут одно непонятное слово «температура» заменено другим непонятным словом «энергия». Положение еще больше запутывается, если учесть, что прибор, который называется термометром, никакую температуру не измеряет, он измеряет изменение объема тел (ртути, спирта, разных металлов) при нагревании.

Теория вероятностей – наука, основные понятия которой только называются и демонстрируются на примерах. По большому счету, они нам уже известны. В некоторых учебниках пытаются дать определения исходных понятий теории вероятностей, но этого не стоит делать: большое количество бесполезных пояснений снижает уровень понимания (то, что было ясным, становится непонятным после многословных объяснений).

Известны понятия «испытание», «случайное событие»; далее в пособии появятся «случайная величина» и «случайная функция». Соподчиненность устанавливается фразой: «При испытаниях происходят события». Примеры: при броске монеты («испытание») может выпасть герб («событие G ») или решка («событие P »); студент пришел на экзамен (на украинском языке «іспит») и получил неудовлетворительную оценку (очень неприятное событие).

Любой взрослый человек, никогда специально не изучавший теорию вероятностей, правильно ответит на вопрос: «Какова вероятность выпадения герба при броске монеты?». Более того, немного подумав, он даже обоснует, почему эта вероятность равна 0,5. Любой человек правильно объяснит разницу между вероятностью выпадения герба (0,5) и вероятностью (0,8) поражения мишени метким стрелком – при повторении испытаний стрелок чаще попадает, чем промахивается, а при броске монеты герб и решка появляются с одинаковой частотой. Можно попросить его сформулировать, что же такое «вероятность», и, в итоге, он ответит, что: **вероятность** – это число, которое показывает, как часто происходит событие (A) при испытаниях. Это число изменяется в пределах $0 \leq p_A \leq 1$, или $0 \leq p_A \leq 100\%$ (в процентах). Если $p_A = 0$, событие A является невозможным, невероятным, оно никогда не происходит, сколько бы не повторялись испытания. Если, наоборот, $p_A = 1$, то событие A обязательно произойдет в каждом испытании, иными словами, такое событие не является случайным, его называют детерминированным, достоверным.

Предложено обозначать невозможные события символом \emptyset , тогда $p_{\emptyset} = 0$. Детерминированные события следует обозначать символом Ω , тогда $p_{\Omega} = 1$. Объект Ω называют еще пространством всех событий и универсумом. Все остальные события, для которых $0 < p_A < 1$, случайные, вероятностные.

Отсюда следует универсальный способ определения (в смысле «вычисления») вероятности – **статистическое** (или стохастическое) **определение вероятности**. Испытания производятся n раз и фиксируется, сколько раз при этом появилось событие A ; число появления события называется частотой m . Относительная частота $\frac{m}{n}$ (частость) изменяется от 0 до 1 и показывает, как часто появлялось событие при n испытаниях. Ожидается, что с увеличением числа испытаний это отношение должно приближаться к некому пределу – вероятности события A :

$$p_A = \lim_{n \rightarrow \infty} \frac{m(n)}{n}.$$

Как правило, эти ожидания оправдываются, например, при испытаниях с бросками монеты, игральной кости и т. п. Однако бывают также неприятные эффекты: при длительных испытаниях изнашивается инструмент, стареет агрегат, улетучиваются легкие фракции, то есть с каждым испытанием изменяются условия проведения опыта. В этих случаях (нарушение однородности) способ статистических испытаний, естественно, не применим.

Иногда возможны иные способы определения вероятностей без проведения длительных испытаний. К ним относятся способ назначения вероятности из геометрических соображений и классический способ вычисления вероятностей, использующие наличие равновероятности некоторых исходов испытания. Оба эти способа будут рассмотрены ниже. Они не являются универсальными и имеют вполне определенную область применения.

Когда невозможен ни один из вышеупомянутых способов, применяют способ экспертных оценок. Группа экспертов обсуждает вероятности неких начальных простых событий, а вероятности более сложных последствий уже рассчитываются на основе известных теорем теории вероятностей.

Кстати, эквивалентными являются названия: «событие», «случайное событие», «случай» («казус»), «исход испытания», «результат опыта».

Рассмотрим геометрическое представление событий и *геометрический способ определения вероятностей*. На рис. 1.1 изображены два события A и B

в виде неких геометрических фигур (диаграмм Венна). Обе фигуры расположены в прямоугольнике, который представляет собой универсум (пространство всех событий, достоверное событие) Ω . Расположение и форма фигур определяются конкретными условиями задачи. Если все точки универсума равновероятны, то площади фигур будут пропорциональны вероятностям изображенных событий. Действительно, представим себе метеорологический столик (прямоугольник Ω), на котором стоят ванночки A , B . Какова вероятность того, что капли дождя попадут в ванночку A ? Поскольку предполагается, что капли дождя падают на столик равномерно, доля капель, попавших в ванночку A , будет пропорциональна только ее площади, но не ее форме и расположению на столике.

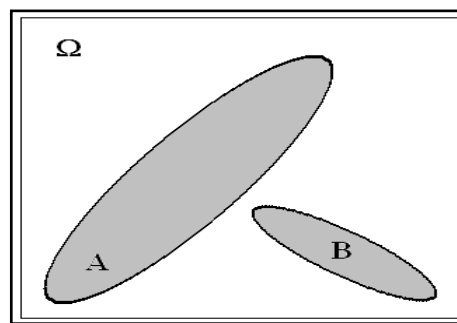


Рис. 1.1. Изображение несовместных событий

Если нам удалось построить геометрические изображения событий, то вероятности событий можно получить как отношения площадей изображенных фигур к площади универсума:

$$p_A = \frac{S_A}{S_\Omega}, \quad p_B = \frac{S_B}{S_\Omega}.$$

Пример 1. Задача о встрече. Два студента договорились встретиться в определенном месте в период от 12⁰⁰ до 13⁰⁰ и ждать друг друга не более четверти часа (15 мин.). Какова вероятность, что встреча состоится?

Обозначим через x – время прихода 1-го студента в условленное место, а через y – время прихода 2-го студента. Согласно условиям задачи, $0 \leq x \leq 1$, $0 \leq y \leq 1$, что геометрически определяет квадрат со стороной 1.

Это универсум, пространство всех равновозможных комбинаций времени прихода студентов на место встречи. Встреча состоится, если $|x - y| \leq 0,25$. Сформулированное неравенство можно переписать в виде системы:

$$\begin{cases} y \leq x + 0,25 \\ y \geq x - 0,25, \end{cases}$$

которая определяет область, изображенную на рис. 1.2.

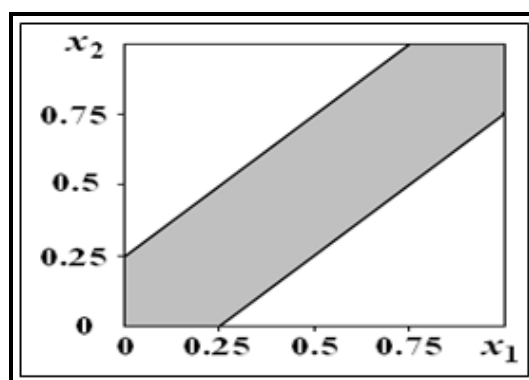


Рис. 1.2. Область встречи

Площадь этой области $S = 1 - 2 \cdot 0,5 \cdot 0,75^2 = 1 - 0,5625 = 0,4375$ равна искомой вероятности, так как площадь универсума (квадрата) равна единице. События, изображенные на рис. 1.2, несовместные, их области не пересекаются.

Несовместные события не могут появиться в одном испытании.

Иногда неважно, какое именно событие (A или B) появится, в обоих случаях получится ожидаемый результат (неважно, в какую именно ванночку попадают капли дождя, далее будет подсчитываться общий объем собранной воды). Нас интересует вероятность события (A или B), которое обозначается $A \cup B$, или $A+B$ (операция объединения множеств).

Из рис. 1.2 видно, что при вычислении вероятности $p(A+B)$ площади непересекающихся фигур надо складывать. Таким образом, геометрически проиллюстрировано утверждение: «Вероятность появления одного из **несовместных** событий (любого) равна сумме их вероятностей».

Это утверждение называется аксиомой сложения вероятностей:

$$P(A+B) = p_A + p_B,$$

где A, B – несовместные.

К сожалению, аналогия между логической операцией объединения множеств и арифметической операцией сложения чисел стала причиной появления жаргонной терминологии «сумма событий».

На рис. 1.3 изображены совместные события (их области пересекаются).

Совместные события могут появиться одновременно в одном испытании.

Операция пересечения событий A и B обозначается $A \cap B$ или просто AB (на основе последнего обозначения появилась жаргонная терминология «произведение событий»). На рис. 1.3 произведению соответствует общая зона, принадлежащая одновременно и A , и B . Этот же рисунок позволяет обосновать утверждение: «Вероятность появления одного из событий равна сумме вероятностей этих событий минус вероятность их совместного появления»: $P(A+B) = p_A + p_B - p_{AB}$. Для несовместных событий A, B $p_{AB} = 0$. Это утверждение называется теоремой сложения вероятностей. Действительно, пусть на метеорологическом столике ванночки события A и B расположены на двух уровнях, поэтому одна ванночка перекрывает другую. После испытания вода из обеих ванночек будет учитываться вместе, поэтому неважно, в какую именно ванночку попадают капли дождя. Тогда вероятность того, что капли дождя попадут в ванночки, будет пропорциональна площади общей затушеванной области, которая равна сумме площадей области A и дополнительной части области B без уже учтенной области AB .

На рис. 1.4, 1.5 изображены противоположные события A и \bar{A} (не A). Противоположные события образуют полную группу событий – их объединение дает универсум: $A + \bar{A} = \Omega$. Сумма вероятностей противоположных событий равна единице: $p_A + p_{\bar{A}} = 1$.

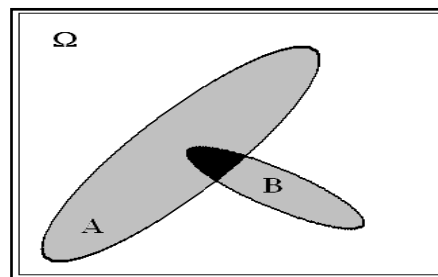


Рис. 1.3. Изображение совместных событий

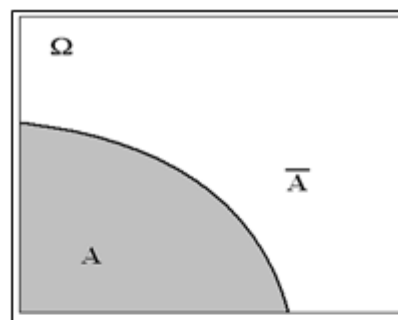


Рис. 1.4. Противоположные события

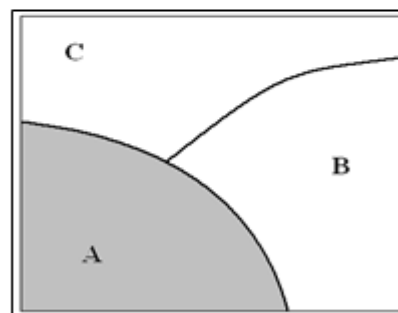


Рис. 1.5. Полная группа событий

Для обозначения вероятности противоположных событий часто используется обозначение $q_A = p_{\bar{A}}$, тогда всегда $p + q = 1$.

События A, B, C составляют **полную группу**, если при испытании одно из них обязательно произойдет: $(A + B + C) = \Omega$. Если эти события несовместные, то сумма их вероятностей будет равна единице $p_A + p_B + p_C = 1$ (см. рис. 1.5).

Успех в решении многих задач теории вероятностей зависит от умения представить исходы испытания в виде полной группы несовместных событий (этот прием будет подробно рассмотрен в следующих лекциях).

Переходим к изложению так называемого **классического способа определения вероятностей**. Этот способ имеет достаточно ограниченную область применения. Название «классический способ» закрепилось исторически, поскольку ученые – основоположники теории вероятностей – сначала изучали события, связанные с азартными играми, которые в отличие от жизненных ситуаций имеют специфические особенности – в них все определено заранее. Несмотря на внешнее разнообразие, многие игры можно свести к одной – извлечению шаров из урны.

На рис. 1.6 изображена такая урна, в которой имеется $n = 9$ шаров, одинаковых по размерам и неразличимых на ощупь. Шары помечены различными признаками, так, среди изображенных шаров имеется $m = 4$ белых (A), $k = 3$ черных (B) и $l = 2$ синих (C). Шары извлекаются из урны по одному в случайном порядке.



Рис. 1.6. Урна с шарами

Каковы вероятности извлечь черный, синий и белый шары?

На все эти вопросы даже неподготовленный человек обычно дает правильные ответы:

$$p_B = \frac{k}{n} = \frac{3}{9}; \quad p_C = \frac{l}{n} = \frac{2}{9}; \quad p_{B+C} = \frac{k+l}{n} = \frac{3+2}{9} = \frac{5}{9}.$$

Фактически используются формула $p_A = \frac{m_A}{n}$ и утверждение, что вероятность появления одного из несовместных событий равна сумме вероятностей указанных событий. Однако из этих двух утверждений одно является следствием другого. Строго доказать сразу оба утверждения невозможно, одно из них следует принять за исходный постулат (аксиому). Предлагается принять за аксиому утверждение $p_{A+B} = p_A + p_B$, если $AB = \emptyset$ (аксиому сложения вероятностей). На примере убедились в справедливости аксиомы сложения для частного случая геометрической вероятности, она представляется очевидной и для част-

ной задачи извлечения шаров из урны; теперь же считаем, что это исходное утверждение справедливо всегда и не требует доказательства.

В задачах, решаемых классическим методом, исходы испытания представлены набором «элементарных исходов» ω_i .

Элементарные исходы – несовместные $\omega_i \omega_j = \emptyset$ – составляют полную группу $\omega_1 + \omega_2 + \dots + \omega_n = \Omega$ и (внимание!) равновероятные $p(\omega_i) = p(\omega_j) = p_\omega$.

В задаче с урной каждый шар является элементарным исходом. Действительно, поскольку шары одинаковы по размерам и неразличимы на ощупь, вероятность извлечь любой из них будет одна и та же; шары извлекаются по одному, в урне нет ничего, кроме шаров.

Теперь можно легко найти вероятность каждого элементарного исхода. Применим аксиому сложения к утверждению $\omega_1 + \omega_2 + \dots + \omega_n = \Omega$ и получим $p(\omega_1 + \omega_2 + \dots + \omega_n) = p(\omega_1) + p(\omega_2) + \dots + p(\omega_n) = np_\omega = p(\Omega) = 1$, то есть $np_\omega = 1$. Отсюда имеем $p_\omega = 1/n$. Например, при броске монеты имеем два элементарных исхода (герб и решка), поэтому вероятность выпадения герба равна $p_\Gamma = 1/2$. При броске однородной кости имеем шесть элементарных исходов (I, II, III, IV, V, VI), поэтому вероятность выпадения шестерки равна $p_{VI} = 1/6$.

При некоторых элементарных исходах появляется то или иное событие (A, B, C, \dots), например, $A = \omega_1 + \omega_2 + \omega_3 + \omega_4$, $B = \omega_3 + \omega_5 + \omega_6$, $C = \omega_4$. С помощью аксиомы сложения можно найти вероятности этих событий:

$$p(A) = p(\omega_1 + \omega_2 + \omega_3 + \omega_4) = p(\omega_1) + p(\omega_2) + p(\omega_3) + p(\omega_4) = 4p_\omega = 4/n.$$

В общем случае получаем формулу $p_A = \frac{m_A}{n}$, где n – общее число элементарных исходов, а m_A – число элементарных исходов, при которых появляется событие A .

Принятие аксиомы сложения позволяет доказать некоторые утверждения, не ссылаясь ни на геометрический, ни на классический способы определения вероятностей. Так, покажем, что сумма вероятностей противоположных событий равна единице. Действительно, противоположные события составляют полную группу несовместных событий: $A + \bar{A} = \Omega$, $A\bar{A} = \emptyset$.

$$\text{Отсюда } p(A + \bar{A}) = p(A) + p(\bar{A}) = p + q = 1, \text{ то есть } p + q = 1.$$

Вернемся к задаче с урной (см. рис. 1.5). Какова вероятность при втором извлечении достать белый шар? Если после каждого извлечения шар возвращается в урну и шары перемешиваются, то вероятность появления белого шара остается одной и той же: $p_A = \frac{m}{n} = 4/9$ (вероятность события A не зависит ни от номера испытания, ни от того, какой шар был извлечен ранее). Но если после

извлечения шар в урну не возвращается, то после каждого испытания имеем дело уже с другим набором шаров в урне. Вероятность события A теперь зависит от того, какие шары были извлечены ранее. Таким образом, можно вычислить только так называемую *условную вероятность* события A , где учтено, какие события произошли до этого испытания.

Так, можно определить вероятность того, что при втором извлечении появится белый шар, если перед этим также был извлечен белый шар: $p(A|A) = \frac{3}{8}$ (общее количество шаров уменьшилось: $n_2 = 9 - 1 = 8$, уменьшилось также количество белых шаров: $m_2 = 4 - 1 = 3$).

Вероятность того, что при втором извлечении появится белый шар, если перед этим был извлечен черный шар: $p(A|B) = \frac{4}{8}$ (количество белых шаров не изменялось: $m_2 = 4$). Вероятность того, что при третьем извлечении появится синий шар, если перед этим было извлечено два синих: $p(C|CC) = \frac{0}{7} = 0$ (общее количество шаров уменьшилось на два: $n_3 = 9 - 2 = 7$, количество синих шаров также уменьшилось на два: $l_3 = 2 - 2 = 0$).

В обозначении условной вероятности условие (например, информация, что произошло в предшествующих испытаниях) записывается после вертикальной черты. При совместных событиях $AB \neq \emptyset$ иногда требуется найти вероятность появления признака B среди объектов с признаком A ; это также условная вероятность $p_{B|A} = p(B|A)$. При вычислении вероятностей классическим способом используются сведения из комбинаторики.

Пример 2. На 5-ти карточках изображены буквы А, Р, С, Т, Т. Какова вероятность, что при случайном извлечении карточек (без возвращения) появится слово СТАРТ?

Общее количество возможных комбинаций находим при помощи следующих рассуждений. На первой карточке может оказаться любая из 5-ти букв, на второй – любая из оставшихся 4-х, на третьей – из оставшихся 3-х, на четвертой – из 2-х букв, на пятой – одна последняя буква. Возможные буквы 1-й карточки сочетаются с любыми возможными буквами остальных карточек, поэтому все эти числа надо перемножить.

Таким образом, общее число равновозможных комбинаций (элементарных исходов) здесь равно $n = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5! = P_5$ – числу перестановок. Из этих комбинаций нас устраивают две (в слове СТАРТ можно взаимно переставлять одинаковые буквы Т), отсюда $m = 2$ и $p_{\text{СТАРТ}} = \frac{2}{5!} = \frac{1}{60}$.

Пример 3. Кроме 5-ти карточек с буквами А, Р, С, Т, добавлено еще 3 с буквами Б, В, Г (всего 8 карточек). Какова вероятность, что при случайном извлечении 5-ти карточек появится слово СТАРТ?

Рассуждая аналогично предыдущему примеру, получаем $n = 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4$.

Это число называется числом размещений и может быть представлено в виде отношения двух факториалов:

$$A_N^K = \frac{N!}{(N-K)!},$$

где N – общее количество объектов (карточек); K – число позиций (число извлеченных карточек).

Для данного случая:

$$n = A_8^5 = \frac{8!}{(8-5)!} = \frac{8!}{3!} = 6720, \quad m = 2 \text{ и } p_{\text{СТАРТ}} = \frac{2}{6720} = \frac{1}{3360}.$$

Пример 4. До сих пор был важен порядок расположения карточек в комбинациях. Если порядок не важен, то появляется число сочетаний.

На рис. 1.7 в большой урне находится $N = 9$ шаров, из которых $M = 4$ белых, $K = 3$ черных и $L = 2$ синих. Случайным образом отбирается $n = 4$ шара.

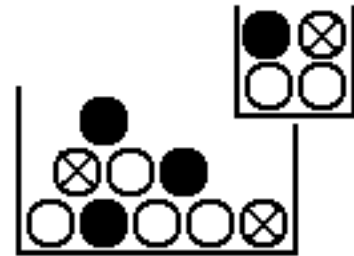


Рис. 1.7. Случайный отбор шаров

Какова вероятность, что в малой партии будет $m = 2$ белых, $k = 1$ черных и $l = 1$ синих шаров (где $m+k+l=n$)?

Поскольку при любых перестановках набор шаров в малой партии не изменяется, то общее число комбинаций при извлечении n шаров из N будет равно

$$\text{но } \frac{A_N^n}{n!} = \frac{N!}{n! (N-n)!} = C_N^n.$$

При вычислении вероятности это число будет в знаменателе дроби. В малую партию отбирается m белых шаров из M , что можно сделать C_M^m способами, k черных шаров из K (C_K^k способов) и l синих шаров из L (C_L^l способов).

Всего необходимых способов отбора шаров в малую партию будет $C_M^m C_K^k C_L^l$ – это числитель формулы классического определения вероятностей:

$$P(m, k, l) = \frac{C_M^m C_K^k C_L^l}{C_N^n} = \frac{C_M^m C_K^k C_L^l}{C_{M+K+L}^{m+k+l}}.$$

Для нашего примера:

$$P(2,1,1) = \frac{C_4^2 C_3^1 C_2^1}{C_9^4} = \frac{6 \cdot 3 \cdot 2}{126} = \frac{2}{7}.$$

Если в урне только два вида шаров ($m + k = n$), формула упрощается:

$$P(m) = \frac{C_M^m C_K^k}{C_N^n} = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n}.$$

Вопросы для самопроверки

1. Разъясните понятие «случайное событие».
2. Что такое вероятность?
3. Какие известны способы определения вероятностей?
4. Что такое невозможное событие?
5. Что такое универсум? Приведите синонимы.
6. Что такое совместные и несовместные события?
7. Перечислите основные операции с событиями.
8. Что такое противоположные события, как связаны их вероятности?
9. Что такое полная группа несовместных событий?
10. Как геометрически изображаются события с разной вероятностью?
11. Что такое элементарные исходы?
12. Приведите статистическое определение вероятностей.
13. Приведите классическое определение вероятностей.
14. Какие есть еще способы определения вероятностей?
15. Сформулируйте аксиому сложения.
16. Что такое условная вероятность?
17. Что такое независимые события?

2. Теоремы о вероятностях

Вероятности исходных простых событий определяются одним из описанных выше способов (см. лекцию 1), вероятности прочих утверждений рассчитываются с помощью теорем.

Теорема умножения вероятностей

Проще всего эта теорема формулируется и доказывается для частного случая независимых событий: «Вероятность совместного появления нескольких **независимых** событий равна произведению их вероятностей $p(AB) = p_A \cdot p_B$ ».

На студенческом жаргоне теорема формулируется так: «Вероятность произведения событий равна произведению их вероятностей» – запоминается легче, но требуется дополнительно пояснить смысл словосочетания «произведение событий» и уточнить, для всех ли событий справедливо такое утверждение.

Рассмотрим две урны с разным количеством шаров (n_1 и n_2) и разным количеством белых шаров (m_1 и m_2) среди них (рис. 2.1). Достаем по одному шару из каждой урны. Какова вероятность, что они оба белые?

Общее количество элементарных исходов при извлечении шаров одновременно из двух урн равно $n = n_1 \times n_2$ (каждый шар из 1-й урны сочетается с каждым шаром из 2-й урны).

Число элементарных исходов, при которых появляются два белых шара, равно $m = m_1 \times m_2$ (каждый белый шар из 1-й урны сочетается с каждым белым шаром из 2-й урны).

Обозначим через A появление белого шара из 1-й урны, через B – из 2-й.

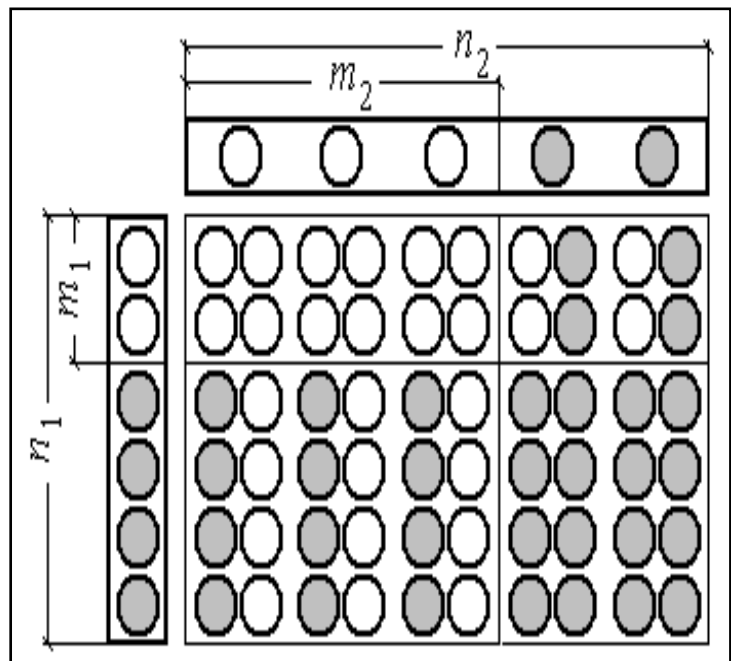


Рис. 2.1. Парные комбинации шаров из 2-х урн

Эти события независимые. Тогда получается:

$$p_{AB} = \frac{m}{n} = \frac{m_1 m_2}{n_1 n_2} = \frac{m_1}{n_1} \cdot \frac{m_2}{n_2} = p_A p_B$$

(утверждение теоремы умножения вероятностей для независимых событий).

Для зависимых событий теорема умножения формулируется так:

«**Вероятность совместного появления** двух событий равна произведению вероятности одного из них на условную вероятность другого»:

$$p(AB) = p(A) \cdot p(B|A) \text{ или } p(AB) = p(B) \cdot p(A|B).$$

Сразу следует отметить, что выражение «одно и другое события» не эквивалентно понятию «первое и второе события» – здесь порядок не важен.

Рассмотрим следующий набор элементарных исходов (рис. 2.2). Общее их количество – n , из них в m исходах появляется признак A , в k исходах – признак B , в l исходах – совместно A и B .

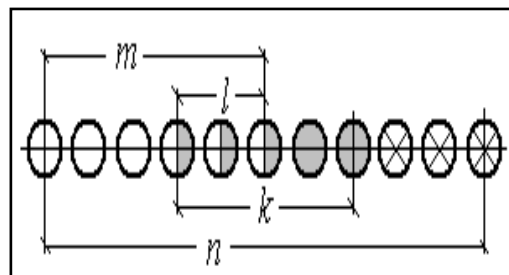


Рис. 2.2. Пример совместных событий

Вероятности этих событий: $p_A = m/n$, $p_B = k/n$, $p_{AB} = l/n$.

Что означает отношение l/m ?

В знаменателе дроби стоит количество исходов с признаком A , а в числителе – сколько из этих исходов дополнительно имеют признак B .

Следовательно, указанное отношение представляет собой условную вероятность $p_{B|A} = l/m$. Аналогично, $p_{A|B} = l/k$.

Тогда получаем: $p_{AB} = l/n = m/n \cdot l/m = p_A \cdot p_{B|A}$.

Аналогично: $p_{AB} = l/n = k/n \cdot l/k = p_B \cdot p_{A|B}$, что и требовалось доказать.

Внешний вид теоремы умножения зависит от типа событий (рис. 2.3). В краткой классификации события подразделяются на несовместные и совместные; совместные события, в свою очередь, делятся на зависимые и независимые; далее независимые события – на неоднородные, когда вероятность зависит от номера испытания (стрельба по удаляющейся мишени, одновременное извлечение шаров из разных урн, результаты экзаменов для разных студентов или по разным предметам), и однородные (несколько выстрелов одного и того же стрелка, результат многократного подбрасывания монеты, игральной кости и т. п.).

Ранее уже указывалось, что успех в решении многих задач теории вероятностей зависит от умения представить результаты испытаний в виде полной группы несовместных событий (здесь не имеются в виду элементарные исходы,

которые еще вдобавок равновероятные). Если освоить этот полезный прием, то будет достаточно одной теоремы умножения вероятностей (и аксиомы сложения вероятностей).

Краткая классификация событий

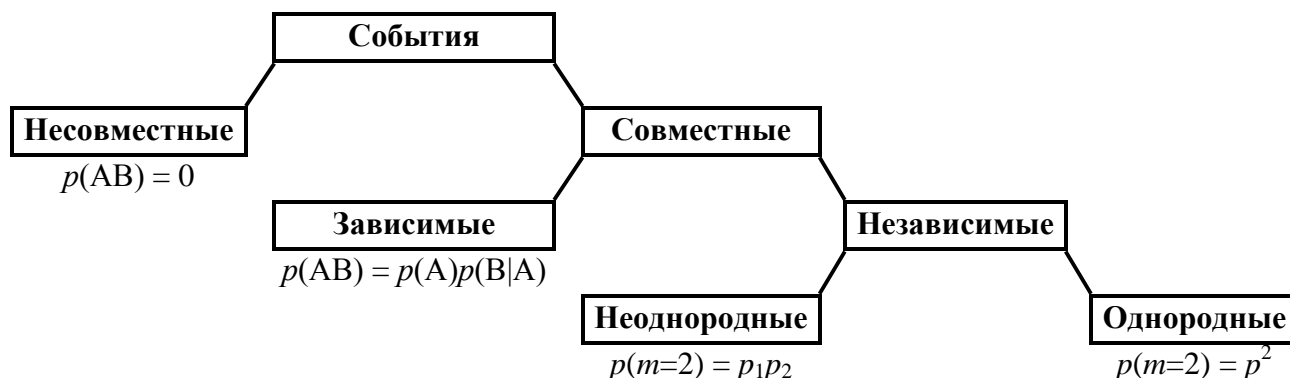


Рис. 2.3. Классификация событий

Рассмотрим следующую простую задачу. Два стрелка с вероятностями поражения мишени $p_1 = 0,8$ и $p_2 = 0,7$ одновременно стреляют по бутылке.

Какова вероятность, что бутылка будет разбита? Иными словами, какова вероятность того, что хоть один из стрелков попадет в цель?

Конечно, нельзя пользоваться аксиомой сложения $p_{A+B} = p_A + p_B = 0,8 + 0,7 = 1,5$, на что, кстати, указывает абсурдный результат (получилось, что $p_{A+B} > 1$). Немного подумав, студент находит причину недоразумения: события A и B совместные (оба стрелка могут попасть в цель), а аксиома сложения сформулирована для несовместных событий.

Составим таблицу полной группы несовместных событий этой задачи, в которой успех (попадание) будем обозначать знаком «+», неуспех (промах) – знаком «–», m – число успехов при двойном выстреле (двух испытаниях) (рис. 2.4).

В результате двойного выстрела может быть два попадания ($m = 2$), одно попадание ($m = 1$), ни одного попадания ($m = 0$). Одно попадание может

№	1	2	m	Вероятность	$P(m)$
1	+	+	2	$p_1p_2 = 0,8 \cdot 0,7 = 0,56$	0,56
2	+	–	1	$p_1q_2 = 0,8 \cdot 0,3 = 0,24$	0,38
3	–	+		$q_1p_2 = 0,2 \cdot 0,7 = 0,14$	
4	–	–	0	$q_1q_2 = 0,2 \cdot 0,3 = 0,06$	0,06

Рис. 2.4. Полная группа 2×2

появиться в двух случаях: или 1-й стрелок попал, а 2-й промахнулся (+ –); или наоборот 1-й стрелок промахнулся, а 2-й попал (– +). При выстреле у каждого стрелка имеются только два исхода (+ или –). Два возможных исхода 1-го

стрелка комбинируются с двумя исходами 2-го, в результате чего появляются $2 \times 2 = 4$ события. Эти события составляют полную группу; легко проверить, что они попарно несовместные.

В данном примере события полной группы являются независимыми неоднородными (стрелки разные). Вычисляем вероятности каждого события по теореме умножения, при этом удобно ввести особые обозначения для вероятностей промаха (вероятностей противоположных событий): $q_1 = 1 - p_1 = 1 - 0,8 = 0,2$; $q_2 = 1 - p_2 = 1 - 0,7 = 0,3$. Теперь можем правильно ответить на поставленный вопрос: «Какова вероятность хотя бы одного успеха ($m \geq 1$)?» Этому условию удовлетворяют первые три события полной группы (все события, кроме последнего). Так как все эти события несовместные, их вероятности следует сложить, то есть достаточно знать аксиому сложения:

$$P(\text{хотя бы 1 успех}) = P(m \geq 1) = 0,56 + 0,24 + 0,14 = 0,94.$$

Поскольку сумма вероятностей полной группы событий равна единице, найдем более простой и общий путь решения задачи:

$$P(\text{хотя бы 1 успех}) = 1 - P(\text{ни одного успеха}) = 1 - q_1 \cdot q_2 = 1 - 0,06 = 0,94.$$

Знание полной группы несовместных событий позволяет ответить на все возможные вопросы с условиями типа: «меньше», «не больше», «равно», «не меньше», «больше», «хотя бы» или «по крайней мере». Например, ответим на вопрос: «Какова вероятность одного успеха при двух испытаниях?» Правильный ответ: $P(m = 1) = p_1 \cdot q_2 + q_1 \cdot p_2 = 0,24 + 0,14 = 0,38$ (см. рис. 2.4).

Полная группа событий будет одинаковой как для зависимых исходов испытания, так и для независимых (однородных и неоднородных).

Рассмотрим трехкратное извлечение шаров из урны (без возвращения), первоначально содержащей $n = 9$ шаров, из которых белых – $m = 6$ и черных – $k = 3$ (рис. 2.5). Составим полную группу несовместных событий, где появление белого шара обозначено знаком «+», а черного – знаком «–» (рис. 2.6).



Рис. 2.5. Урна с 2-мя видами шаров

Полная группа несовместных событий в этой задаче должна содержать $2 \times 2 \times 2 = 8$ комбинаций. При трех испытаниях будут или $m = 3$ успеха (3 белых шара), или $m = 2$, или $m = 1$, или ни одного успеха ($m = 0$).

№	1	2	3	m	Вероятность
1	+	+	+	3	$p_A p_{A A} p_{A AA} = 6/9 \cdot 5/8 \cdot 4/7 = 20/84$
2	+	+	–	2	$p_A p_{A A} p_{B AA} = 6/9 \cdot 5/8 \cdot 3/7 = 15/84$
3	+	–	+		$p_A p_{B A} p_{A AB} = 6/9 \cdot 3/8 \cdot 5/7 = 15/84$
4	–	+	+		$p_B p_{A B} p_{A AB} = 3/9 \cdot 6/8 \cdot 5/7 = 15/84$
5	+	–	–	1	$p_A p_{B A} p_{B AB} = 6/9 \cdot 3/8 \cdot 2/7 = 6/84$
6	–	+	–		$p_B p_{A B} p_{B AB} = 3/9 \cdot 6/8 \cdot 2/7 = 6/84$
7	–	–	+		$p_B p_{B B} p_{A BB} = 3/9 \cdot 2/8 \cdot 6/7 = 6/84$
8	–	–	–	0	$p_B p_{B B} p_{B BB} = 3/9 \cdot 2/8 \cdot 1/7 = 1/84$

Рис. 2.6. Полная группа $2 \times 2 \times 2$

Два белых шара (и, значит, один черный шар) могут появиться в 3-х случаях (черный шар появляется или в первом, или во втором, или в третьем извлечениях). Один белый шар (и два черных) также может появиться в 3-х случаях (белый шар в первом, или втором, или в третьем испытаниях). Все события полной группы попарно несовместны, поэтому сумма их вероятностей должна равняться единице. Исходы испытаний здесь зависимые, поэтому в формулах теоремы умножения появятся условные вероятности (см. последнюю колонку в таблице на рис. 2.6).

Ответим на стандартные вопросы: 1. Какова вероятность появления хотя бы 1-го белого шара при трехкратном извлечении без возвращения? 2. Какова вероятность появления хотя бы одного черного шара? 3. Какова вероятность появления одного черного шара? 4. Какова вероятность появления одного белого шара? Ответы:

$$1. P(m \geq 1) = 1 - P(m = 0) = 1 - p_B p_{B|B} p_{B|BB} = 1 - 1/84 = 83/84.$$

$$2. P(m < 3) = 1 - P(m = 3) = 1 - p_A p_{A|A} p_{A|AA} = 1 - 20/84 = 64/84.$$

3. Этому условию удовлетворяют три равновероятных события полной группы № 2, 3, 4, поэтому $P(m = 2) = 3 \cdot p_A p_{A|A} p_{B|AA} = 3 \cdot 15/84 = 45/84$.

4. Этому условию удовлетворяют три равновероятных события полной группы № 5, 6, 7, поэтому $P(m = 1) = 3 \cdot p_A p_{B|A} p_{B|AB} = 3 \cdot 6/84 = 18/84$.

Проверяем правильность вычислений:

$$P(m = 3) + P(m = 2) + P(m = 1) + P(m = 0) = 20/84 + 45/84 + 18/84 + 1/84 = 84/84 = 1.$$

При большем числе испытаний или большем числе исходов в одном испытании полную группу несовместных событий надо составлять автоматически, чтобы не запутаться. Рассмотрим порядок составления полной группы $2 \times 2 \times 2 \times 2$ при 4-х испытаниях с двумя исходами в каждом. Два исхода 1-го испытания (выделены в таблице на рис. 2.7 серым цветом) повторяются на каж-

дом исходе 2-го испытания; получается группа 2×2 (выделена черной рамочкой). Все эти четыре исхода первых двух испытаний повторяются на каждом исходе 3-го испытания; получается группа $2 \times 2 \times 2$ (выделена двойной рамочкой). Все эти восемь исходов первых трех испытаний повторяются на каждом исходе 4-го испытания; получается полная группа $2 \times 2 \times 2 \times 2 = 16$ событий.

В колонке m таблицы на рис. 2.7 подсчитано число успехов в каждой комбинации, откуда виден единственный недостаток автоматического составления полной группы – события с одинаковым числом успехов не сгруппированы вместе.

В последней колонке таблицы на рис. 2.7 приведены формулы для вычисления вероятностей при повторении однородных независимых испытаний, откуда видно, что события полной группы с одинаковым числом успехов равновероятны (они не равновероятны только для неоднородных испытаний). Отсюда получается формула Бернулли: $P_n(m) = C_n^m p^m q^{n-m}$. Действительно, если есть m успехов, значит в остальных $(n - m)$ испытаниях успеха нет; согласно

№	1	2	3	4	m	P
1	+	+	+	+	4	p^4
2	–	+	+	+	3	$p^3 q$
3	+	–	+	+	3	$p^3 q$
4	–	–	+	+	2	$p^2 q^2$
5	+	+	–	+	3	$p^3 q$
6	–	+	–	+	2	$p^2 q^2$
7	+	–	–	+	2	$p^2 q^2$
8	–	–	–	+	1	$p q^3$
9	+	+	+	–	3	$p^3 q$
10	–	+	+	–	2	$p^2 q^2$
11	+	–	+	–	2	$p^2 q^2$
12	–	–	+	–	1	$p q^3$
13	+	+	–	–	2	$p^2 q^2$
14	–	+	–	–	1	$p q^3$
15	+	–	–	–	1	$p q^3$
16	–	–	–	–	0	q^4

Рис. 2.7. Группа $2 \times 2 \times 2 \times 2$

теореме умножения вероятность такого события равна $p^m q^{n-m}$; равновероятных событий в полной группе будет C_n^m , так как число сочетаний показывает, сколькими способами можно разместить m плюсов (+) на n позициях. Для данного примера ($n = 4$) вычислим и проверим по приведенной таблице на рис. 2.7 вероятность появления двух успехов: $P_4(m = 2) = C_4^2 p^2 q^{4-2} = 6 p^2 q^2$.

Число исходов при каждом испытании может быть каким угодно, поэтому в общем случае составляется полная группа $k_1 \times k_2 \times \dots \times k_n$. Для примера составим полную группу событий 3×3 , где в каждом из двух испытаний может быть по 3 исхода, которые обозначены «+», «0», «–». Три исхода 1-го испытания (выделены в таблице на рис. 2.8 серым цветом) повторены на каждом из трех исходов 2-го испытания, тем самым автоматически перечислены всевозможные комбинации.

До сих пор рассматривалась схема испытаний заданное число раз (задано n). Существует также схема испытаний до первого успеха (обе схемы предложены швейцарскими математиками, братьями Бернулли).

Составим полную группу несовместных событий для схемы испытаний до первого успеха, причем максимальное число испытаний ограничим $n \leq 5$ (рис. 2.9).

Несложно убедиться, что здесь сумма вероятностей событий полной группы действительно равна единице. Рассмотрение этой задачи с неограниченным числом испытаний приводит к так называемому экспоненциальному закону распределения.

№	1	2
1	+	+
2	0	+
3	–	+
4	+	0
5	0	0
6	–	0
7	+	–
8	0	–
9	–	–

Рис. 2.8. Группа 3×3

№	1	2	3	4	5	P
1	+					p
2	–	+				pq
3	–	–	+			pq^2
4	–	–	–	+		pq^3
5	–	–	–	–	+	pq^4
6	–	–	–	–	–	q^5

Рис. 2.9. Испытания до 1-го успеха

Теорема о полной вероятности

Данная теорема применяется для решения часто встречающейся специфической задачи, поэтому полезно запомнить конечный результат в виде отдельной теоремы. В рассматриваемой задаче событие A появляется совместно с одним из событий H_i , которые составляют полную группу несовместных событий и называются гипотезами. Даны вероятности гипотез $p(H_i)$ и условные вероятности появления события A в присутствии каждой гипотезы $p(A|H_i)$. Требуется найти вероятность события A .

Составим полную группу несовместных событий для случая 3-х гипотез H_1 , H_2 , H_3 , в присутствии которых может появиться или не появиться событие A .

Эти три гипотезы сочетаются с двумя альтернативами – A или \bar{A} .

Всего событий в полной группе будет $3 \times 2 = 6$ (рис. 2.10).

Вероятности каждой комбинации полной группы вычисляем по теореме умножения.

№	H	A	P
1	H_1	A	$p(H_1)p(A H_1)$
2	H_2	A	$p(H_2)p(A H_2)$
3	H_3	A	$p(H_3)p(A H_3)$
4	H_1	\bar{A}	
5	H_2	\bar{A}	
6	H_3	\bar{A}	

Рис. 2.10. Группа 3×2

В составленной таблице полной группы событий (см. рис. 2.10) только в первых трех случаях появляется событие A . Ввиду несовместности событий полной группы вычисленные вероятности надо сложить:

$$p(A) = p(H_1)p(A|H_1) + p(H_2)p(A|H_2) + p(H_3)p(A|H_3).$$

Полученная формула составляет суть теоремы о полной вероятности.

Пример 1. Студент знает ответы на m билетов из n . Что для него выгоднее – сдавать экзамен первым или последним?

Если он идет сдавать первым, то вероятность успешной сдачи экзамена равна $p_A = \frac{m}{n}$.

Если же студент идет сдавать экзамен вторым, то предшествующий студент уже получил один билет. Какой именно? Имеют место две гипотезы: или изъят билет «хороший» (вероятность этой гипотезы – $p_{H_1} = \frac{m}{n}$), или изъят «плохой» билет (вероятность этой гипотезы – $p_{H_2} = \frac{n-m}{n}$).

Шансы второго студента на успех существенно зависят от того, какая гипотеза будет реализована в действительности: $p_{A|H_1} = \frac{m-1}{n-1}$; $p_{A|H_2} = \frac{m}{n-1}$.

Согласно теореме о полной вероятности:

$$p_A = p_{H_1} p_{A|H_1} + p_{H_2} p_{A|H_2} = \frac{m}{n} \cdot \frac{m-1}{n-1} + \frac{n-m}{n} \cdot \frac{m}{n-1} = \frac{m}{n} \cdot \frac{(m-1) + (n-m)}{n-1} = \frac{m}{n}.$$

Шансы на успех для 2-го студента не изменились.

В качестве полезного упражнения предлагается вычислить полную вероятность успеха для 3-го студента.

Пример 2. На конвейер детали поступают от 3-х поставщиков. Первый поставщик (гипотеза H_1) поставляет m_1 деталей, второй (гипотеза H_2) – m_2 деталей, третий (гипотеза H_3) – m_3 деталей; всего $n = m_1 + m_2 + m_3$. Известны вероятности брака (доля бракованных изделий) для каждого поставщика: для 1-го поставщика она равна $p(A|H_1)$, для 2-го – $p(A|H_2)$, для 3-го – $p(A|H_3)$. Какова вероятность брака на конвейере?

Эту задачу можно решить, не обращаясь к теореме о полной вероятности (естественно, получим тот же самый результат). Найдем количество поставленных бракованных изделий от каждого поставщика: $k_1 = m_1 p(A|H_1)$, $k_2 = m_2 p(A|H_2)$, $k_3 = m_3 p(A|H_3)$, всего на конвейере будет $k = k_1 + k_2 + k_3$ бракованных деталей.

Вычисляем вероятность брака на конвейере:

$$\begin{aligned} p_A &= \frac{k_1 + k_2 + k_3}{n} = \frac{m_1}{n} \cdot p_{A|H_1} + \frac{m_2}{n} \cdot p_{A|H_2} + \frac{m_3}{n} \cdot p_{A|H_3} = \\ &= p_{H_1} p_{A|H_1} + p_{H_2} p_{A|H_2} + p_{H_3} p_{A|H_3}. \end{aligned}$$

В результате опять получена формула полной вероятности. Из вышеприведенного выражения также следует, что полная вероятность $p(A)$ есть среднее взвешенное из условных вероятностей $p(A|H_i)$ с весовыми коэффициентами m_i или $p(H_i)$.

Теорема (формула) Байеса

Формула Байеса оценивает относительные вклады каждого члена в формуле полной вероятности. Так, для случая 3-х гипотез относительные вклады будут равняться $\frac{p_{H_1} p_{A|H_1}}{p_A}$, $\frac{p_{H_2} p_{A|H_2}}{p_A}$, $\frac{p_{H_3} p_{A|H_3}}{p_A}$. Как их правильно обозначить?

Рассмотрим две эквивалентные формулы для расчета вероятности совместного появления двух событий A и H_i :

$$p_{AH_i} = p_A p_{H_i|A} = p_{H_i} p_{A|H_i}.$$

Отсюда следует, что

$$\frac{p_{H_1} p_{A|H_1}}{p_A} = p_{H_1|A}, \quad \frac{p_{H_2} p_{A|H_2}}{p_A} = p_{H_2|A}, \quad \frac{p_{H_3} p_{A|H_3}}{p_A} = p_{H_3|A}.$$

Интерпретация для последней задачи с браком на конвейере:

Сборщик обнаружил на конвейере бракованную деталь (произошло событие A). Какова вероятность, что эту деталь изготовил 1-й поставщик? 2-й поставщик? 3-й? Если эти вероятности существенно различаются, то для уменьшения брака на конвейере следует, в первую очередь, принять меры к уменьшению брака (наладка оборудования) у поставщика с наибольшим значением $p_{H_i|A}$; именно этот поставщик поставляет на конвейер наибольшее количество бракованных изделий. Далеко не всегда это поставщик с максимальной долей брака $p_{A|H_i}$, так как такой поставщик может изготавливать небольшую партию деталей и в сумме давать на конвейер значительно меньше бракованных изделий по сравнению с другими поставщиками.

Теорема сложения вероятностей

Эта теорема уже была рассмотрена при изучении геометрического способа определения вероятностей: «Вероятность появления одного из событий равна сумме их вероятностей минус вероятность их совместного появления», то есть $p_{A+B} = p_A + p_B - p_{AB}$. Справедливость утверждения этой теоремы можно также установить и для классического способа определения вероятностей (рис. 2.11).

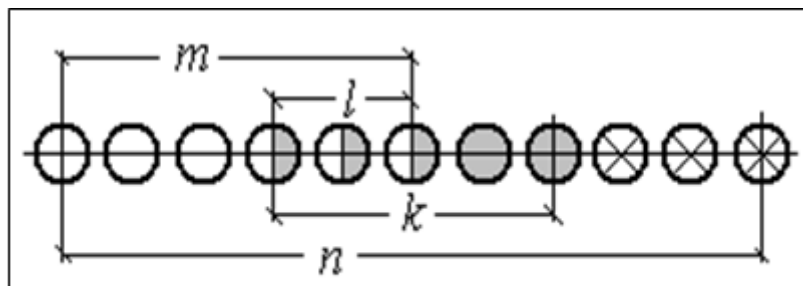


Рис. 2.11. Теорема сложения вероятностей

Здесь общее количество элементарных исходов – n , из них в m исходах появляется признак A , в k исходах – признак B , в l исходах – совместно A и B .

Вероятности этих событий: $p_A = m/n$, $p_B = k/n$, $p_{AB} = l/n$.

Нас интересует вероятность появления A или B , чему благоприятствуют m исходов с признаком A и дополнительно $(k - l)$ исходов с признаком B . Отсюда $p_{A+B} = \frac{m + k - l}{n} = p_A + p_B - p_{AB}$, что и требовалось доказать.

Пример. Два стрелка с вероятностями поражения мишени $p_1 = 0,8$ и $p_2 = 0,7$ одновременно стреляют по бутылке. Какова вероятность, что бутылка будет разбита? Иными словами, какова вероятность того, что хотя бы один из стрелков попадет в цель?

Эта задача уже рассматривалась выше. Теперь решим ее с помощью теоремы сложения (и теоремы умножения):

$$p_{A+B} = p_A + p_B - p_{AB} = 0,8 + 0,7 - 0,8 \cdot 0,7 = 1,5 - 0,56 = 0,94.$$

К сожалению, формулировка теоремы сложения усложняется с увеличением числа событий. Так, для трех событий A, B, C формула будет выглядеть так: $p_{A+B+C} = (p_A + p_B + p_C) - (p_{AB} + p_{AC} + p_{BC}) + p_{ABC}$, а для четырех событий A, B, C, D еще сложнее:

$$p_{A+B+C+D} = (p_A + p_B + p_C + p_D) - (p_{AB} + p_{AC} + p_{AD} + p_{BC} + p_{BD} + p_{CD}) + (p_{ABC} + p_{ABD} + p_{ACD} + p_{BCD}) - p_{ABCD}.$$

Освоив методику составления полной группы несовместных событий, можно никогда не пользоваться столь неудобными выражениями.

Принцип практической невозможности редких событий

Как уже указывалось выше, человек с детства усваивает некоторые правила поведения в этом мире, позволяющие ему выживать. Естественно, он может ошибаться, но разные ошибки приводят к разным последствиям. На основе многовекового опыта человек сформулировал для себя скептическое отноше-

ние к случайному появлению редких событий, которые могут произойти с вероятностью менее 5 %. Люди готовы с интересом обсуждать эти редкие события, но только когда они их не затрагивают. Однако, если неожиданное неприятное событие происходит именно с ними, они склонны искать виновника и, как правило, находят его. Так, игроки сразу выявляют шулера, у которого при бросании с виду обычной игральной кости три раза выпала шестерка, так как вероятность этого события $\frac{1}{6^3} = \frac{1}{216}$ менее 0,5 %. Считается, что при бросках монеты герб не может случайно появиться 5 раз подряд, поскольку вероятность этого события $\frac{1}{2^5} = \frac{1}{32}$ менее 3 %. Здесь многое зависит от терпимости индивида, поскольку надо выбирать между ошибкой «упустить виновного» и ошибкой «наказать невинного». Однако уже в ГОСТ-е законодательно установлена нижняя граница терпимости 1 % – если произошло событие с вероятностью менее 1 %, оно не признается случайным и требуется искать причину.

Вопросы для самопроверки

1. Сформулируйте теорему умножения вероятностей для независимых событий.
2. Сформулируйте теорему умножения вероятностей в общем виде.
3. Поясните, что такое условная вероятность.
4. Изложите краткую классификацию событий.
5. Сформулируйте теорему сложения вероятностей для случая двух событий. Сформулируйте теорему о полной вероятности.
6. Объясните смысл формулы Байеса.
7. Как составить полную группу несовместных событий $2 \times 2 \times 2$?
8. Как составить полную группу несовместных событий при испытаниях до первого успеха? Как найти вероятность хотя бы одного успеха?
9. Приведите и объясните формулу Бернулли.
10. Сформулируйте принцип невозможности редких событий.

3. Случайные величины

«Случайная величина» – очередное неопределяемое понятие, которое только демонстрируется на примерах. Так, известно, что при бросании игральной кости появляется одна из граней I, II, III, IV, V, VI. Но можно сказать и по-другому: «Число выпадающих очков есть случайная величина, которая может принимать одно из значений 1, 2, 3, 4, 5, 6».

Обычно принято обозначать случайную величину прописными рукописными буквами, например \mathcal{X} (или греческими, например ξ), а ее значения – строчными латинскими буквами x_1, x_2, x_3 и т. д.

Если все возможные значения случайной величины можно заранее перечислить, то такую случайную величину называют *дискретной*.

Если же возможные значения заполняют сплошь некоторый интервал, то такую случайную величину называют *непрерывной* (в математике есть строгое определение непрерывной функции, на основе этого можно сформулировать более строгое определение непрерывной случайной величины).

Все возможные значения случайной величины составляют полную группу несовместных событий.

Некоторые значения случайной величины появляются чаще других, некоторые – очень редко. Иными словами, разные значения случайной величины появляются с разной вероятностью.

Соответствие между отдельными значениями случайной величины и вероятностью их появления (функциональная зависимость) называют *законом распределения* данной случайной величины.

Дискретная случайная величина

Для этого частного случая можно более просто изложить некоторые понятия, связанные со случайными величинами.

Закон распределения дискретной случайной величины может быть задан в табличной форме в виде «ряда распределения» (рис. 3.1). Все значения случайной величины образуют

\mathcal{X}	x_1	x_2	x_3	\dots	x_k
$P(x)$	p_1	p_2	p_3	\dots	p_k

Рис. 3.1. Ряд распределения

полную группу несовместных событий, поэтому сумма их вероятностей должна равняться единице: $p_1 + p_2 + p_3 \dots + p_k = 1$. Чаще всего у дискретной случайной величины число возможных значений (k) конечное. Однако бывают также

дискретные случайные величины с бесконечным (но счетным) числом возможных значений.

Так, в схеме испытаний до первого успеха случайным является число испытаний, которое в принципе может быть сколь угодно велико.

В этой задаче получается бесконечный ряд распределения (рис. 3.2). Тут должна равняться единице бесконечная сумма вероятностей, убывающих по геометрической прогрессии:

\mathcal{X}	1	2	3	4	...
$P(x)$	p	pq	pq^2	pq^3	...

Рис. 3.2. Бесконечный ряд распределения

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

Закон распределения может быть представлен графически, причем условились для дискретной случайной величины соединять отдельные точки (x_i, p_i) отрезками прямых. В результате появляется полигон (многоугольник) вероятностей, который дает наглядную информацию об особенностях распределения.

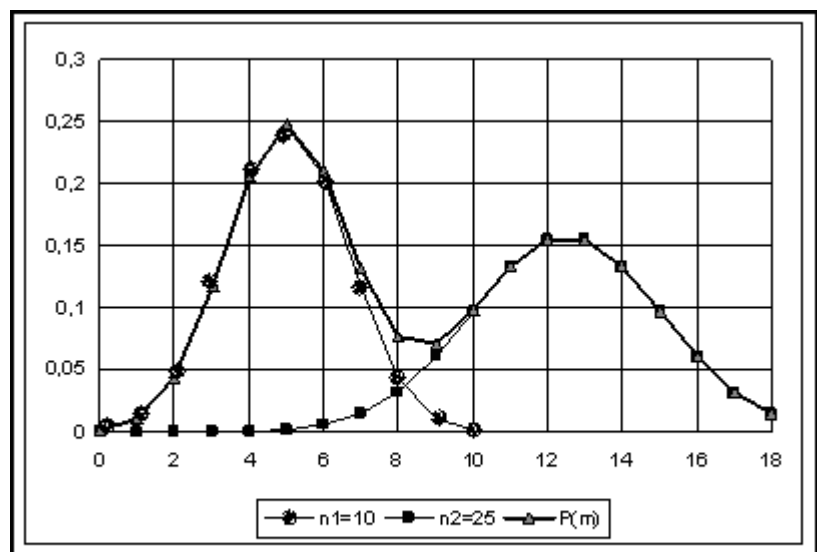


Рис. 3.3. Полигон распределения

Так, на рис. 3.3 изображено двухвершинное распределение, из чего специалист делает вывод о неоднородности совокупности — здесь явно имеет место смесь двух случайных величин с разными характеристиками.

Числовые характеристики случайных величин

Числовые характеристики предназначены для количественного описания различий между случайными величинами. Они подразделяются на следующие группы:

характеристики положения, к которым относятся мода, медиана и математическое ожидание;

характеристики разброса, к которым относятся размах, дисперсия, среднее квадратическое отклонение, коэффициент вариации;

характеристики формы, к которым относятся моменты распределения, коэффициенты асимметрии и плосковершинности.

Для непрерывной случайной величины используются еще так называемые квантили (одним из квантилей является медиана). Эти характеристики рассмотрим позже при изучении непрерывных случайных величин.

Характеристики положения

Мода (Mo) – это наивероятнейшее значение случайной величины $P(x = Mo) = \max\{p_i\}$. Для одновершинного (одномодального) распределения мода (значение x_i с наибольшей вероятностью p_i) является наглядной характеристикой положения. Как видим, эта мера не является универсальной.

Основной же характеристикой положения является математическое ожидание.

Математическое ожидание – это центр, вокруг которого группируются значения случайной величины, «среднее» ее значение. На рис. 3.3 показана смесь двух одновершинных распределений (их полигоны изображены тонкими линиями). Видно, что значения первой случайной величины группируются вокруг значения $M_1 = 5$, а второй – вокруг $M_2 = 12,5$.

Для того чтобы понять, как следует правильно вычислять эту характеристику, рассмотрим пример организации лотереи.

Пусть выпущено $n = 100$ лотерейных билетов. Разыгрываются денежные призы от 1 грн до 10 грн.

Выигрыши (значения случайной величины \mathcal{X}) и соответствующие количества выигрышных билетов (m) перечислены в таблице на рис. 3.4; в последней ее строке вычислены вероятности выигрышей. Какова должна быть цена одного билета? Каков ожидаемый средний выигрыш на один билет?

\mathcal{X}	0	1	2	5	10
m	50	30	15	4	1
p	0,5	0,3	0,15	0,04	0,01

Рис. 3.4. Пример лотереи

Для ответа на этот вопрос подсчитаем общую сумму выигрышей и разделим ее на количество билетов:

$$M(x) = \frac{m_1 x_1 + m_2 x_2 + m_3 x_3 + m_4 x_4 + m_5 x_5}{m_1 + m_2 + m_3 + m_4 + m_5} = \frac{50 \cdot 0 + 30 \cdot 1 + 15 \cdot 2 + 4 \cdot 5 + 1 \cdot 10}{100} = 0,9.$$

Стоимость одного билета (ожидаемый средний выигрыш на один билет) оказалась равной 90 коп. Такого выигрыша в таблице нет, то есть математическое ожидание – условная величина. Формула, которую мы получили для этого частного случая случайной величины, – это формула для вычисления среднего

взвешенного, где в качестве весов выступают частоты m_i или вероятности выигрышей p_i . Эта же формула применяется для определения центра тяжести системы грузов m_i , расположенных в точках с абсциссами x_i , или центра тяжести полигона.

Полученную формулу можно преобразовать к виду:

$$\begin{aligned} M(x) &= x_1 \cdot p_1 + x_2 \cdot p_2 + x_3 \cdot p_3 + x_4 \cdot p_4 + x_5 \cdot p_5 = \\ &= 0 \cdot 0,5 + 1 \cdot 0,3 + 2 \cdot 0,15 + 5 \cdot 0,04 + 10 \cdot 0,01 = 0,9. \end{aligned}$$

Для общего случая дискретной случайной величины имеем:

$$a = M(x) = \sum_{i=1}^k x_i p_i,$$

где k может быть бесконечным ($k = \infty$). Математическое ожидание – это число, а обозначение $M(x)$ похоже на обозначение функции. Предлагаем там, где это возможно, обозначать математическое ожидание буквой a .

Кстати, в англо-американской литературе математическое ожидание обозначается $E(x)$ – от слова *expect* – ожидание.

Для симметричных одномодальных распределений математическое ожидание совпадает с модой (и медианой).

Характеристики разброса

Одной характеристики положения недостаточно для представления о случайной величине. Например, чем отличается продукция мастера от продукции ученика? Они оба придерживаются одного задания и намерены изготавливать детали строго по чертежу. Но большая часть продукции ученика является браком, так как размеры его деталей часто выходят за пределы допуска; в то же время практически вся продукция мастера принимается. Таким образом, продукция мастера и ученика отличается разной изменчивостью.

Наиболее простой характеристикой изменчивости (разброса, рассеивания) является **размах** – разность между максимальным и минимальным значениями случайной величины: $d = x_{\max} - x_{\min}$. К сожалению, размах – нестабильная характеристика, она зависит только от предельных значений величины, которые имеют малую вероятность появления (и в малой партии могут не появиться). Основными же характеристиками разброса являются дисперсия и среднее квадратическое отклонение.

Для каждой детали имеется какое-то отклонение от заданного размера (центра группировки, математического ожидания): $(x_i - a)$. Большие отклонения встречаются редко, а малые – значительно чаще. Возникает идея характеризовать меру изменчивости средней величиной отклонения:

$M(x - a) = \sum (x_i - a) \cdot p_i$. Ожидается, чем больше среднее отклонение, тем будет больше случайная изменчивость, тем должен быть большим разброс данных вокруг центра a . Однако, как будет показано далее, среднее отклонение оказывалось всегда равным нулю (и для мастера, и для ученика) из-за разных знаков отклонений. Для того чтобы погасить знаки, все отклонения предварительно возводят в квадрат и находят средний квадрат отклонения: $D(x) = M(x - a)^2 = \sum (x_i - a)^2 \cdot p_i$. Эта характеристика называется **дисперсией** (в переводе – разброс). Вычислим ее на примере задачи о лотерее, где уже найдено математическое ожидание $a = 0,9$:

$$D(x) = (0-0,9)^2 \cdot 0,5 + (1-0,9)^2 \cdot 0,3 + (2-0,9)^2 \cdot 0,15 + (5-0,9)^2 \cdot 0,04 + (10-0,9)^2 \cdot 0,01 = 0,81 \cdot 0,5 + 0,01 \cdot 0,3 + 1,21 \cdot 0,15 + 16,81 \cdot 0,04 + 82,81 \cdot 0,01 = 2,09.$$

Однако не очень удобно, что размерность дисперсии равна квадрату размерности случайной величины. Поэтому извлекают корень квадратный из дисперсии и полученную характеристику называют **средним квадратическим отклонением**; по традиции ее обозначают греческой буквой сигма: $\sigma_x = \sqrt{M(x - a)^2} = \sqrt{\sum (x_i - a)^2 p_i}$. Обычное среднее отклонение (взвешенное среднее) равно нулю, поэтому используют среднее *квадратическое* (взвешенное с весами p_i). Вместо понятия «среднее квадратическое отклонение» часто употребляют выражения «стандартное отклонение» и даже «стандартная ошибка». Но надо сразу предостеречь против последнего словосочетания. За рубежом различают термины *Standard Deviation* (стандартное отклонение) и *Standard Error* (стандартная ошибка). Последняя величина равна $\frac{\sigma_x}{\sqrt{n}}$, и в отечественной литературе называется «ошибкой среднего» (о ней будет сказано позже). Для примера с лотереей $\sigma_x = \sqrt{2,09} = 1,446$.

Иногда увеличение среднего размера (математического ожидания) сопровождается одновременным увеличением разброса. Для характеристики изменчивости таких случайных величин может оказаться полезным **коэффициент вариации**, который равен $v_x = \frac{\sigma_x}{M(x)} \cdot 100\%$. Если коэффициент вариации оказывается меньшим 2 %, то такую случайную величину считают константой (ее изменчивостью можно пренебречь). Сверху коэффициент вариации не ограничен. Для примера с лотереей $v_x = \frac{1,446}{0,9} \cdot 100\% = 160,6\%$.

Характеристики формы

Полный набор характеристик дискретной случайной величины представляют **моменты распределения**: $m_1 = M(x)$, $m_2 = M(x^2)$, $m_3 = M(x^3)$, $m_4 = M(x^4)$. Заметим, что момент 1-го порядка есть основная характеристика положения – математическое ожидание $m_1 = M(x) = a$. Кроме этого, введены **центральные моменты распределения**: $\mu_1 = M(x - a)$, $\mu_2 = M(x - a)^2$, $\mu_3 = M(x - a)^3$, $\mu_4 = M(x - a)^4$. Первый центральный момент всегда равен нулю (это будет показано далее); второй центральный момент есть основная характеристика разброса – дисперсия $\mu_2 = M(x - a)^2 = (\sigma_x)^2$. Моменты 2-го и 4-го порядков имеют размерности куба и четвертой степени от размера исходной величины, поэтому применяют еще безразмерные **нормированные** или **стандартизированные моменты распределения**: $\rho_3 = \frac{\mu_3}{\sigma_x^3} = A$, $\rho_4 = \frac{\mu_4}{\sigma_x^4}$. Нормированный момент 3-го

порядка называется коэффициентом асимметрии; если $A = 0$ – распределение симметричное, при $A > 0$ – скошено влево, при $A < 0$ – скошено вправо. Нормированный момент 4-го порядка называется коэффициентом плосковершинности, или игольчатости, или же коэффициентом эксцесса. Если $\rho_4 = 3$, форма полигона распределения будет «нормальной», при $\rho_4 < 3$ – приплюснутой (плосковершинной), при $\rho_4 > 3$ – игольчатой. Заметим, что в отечественной литературе принято вычитать тройку из 4-го нормированного момента $E = \rho_4 - 3$, тогда нормальной форме соответствует $E = 0$, плосковершинности – $E < 0$, игольчатости – $E > 0$. Наличие игольчатости трактуется как результат смеси двух случайных величин с разными дисперсиями (смесь продукции мастера и ученика).

Свойства математического ожидания

Знание свойств характеристик позволяет значительно облегчить процесс их вычисления.

1. Математическое ожидание постоянной равно самой постоянной. Вариант: среднее значение постоянной равно самой постоянной.

Действительно, если $x_i = C$, то $M(x) = \sum x_i p_i = \sum C p_i = C \cdot \sum p_i = C$.

2. Постоянный множитель можно выносить за знак математического ожидания. Вариант: постоянный множитель можно выносить за знак среднего.

Действительно, $M(kx) = \sum kx_i p_i = k \cdot \sum x_i p_i = k \cdot M(x)$.

3. Математическое ожидание суммы случайных величин равно сумме математических ожиданий. Вариант: среднее суммы равно сумме средних.

Пусть x_i – значения случайной величины \mathcal{X} – появляются с вероятностями p_{x_i} , а y_j – значения случайной величины \mathcal{Y} – появляются с вероятностями p_{y_j} . Составим ряд распределения суммы случайных величин $\mathcal{Z} = \mathcal{X} + \mathcal{Y}$. Всевозможные значения сумм $z_{ij} = (x_i + y_j)$ появляются с вероятностями $p_{x_i y_j}$. Покажем, что сумма этих вероятностей равна единице.

Пусть события Y_1, Y_2, Y_3 составляют полную группу несовместных событий.

Тогда из рис. 3.5 следует $p(XY_1) + p(XY_2) + p(XY_3) = p(X)$.

Учитывая это, получаем:

$$\sum_i \sum_j p_{x_i y_j} = \sum_i \left(\sum_j p_{x_i y_j} \right) = \sum_i p_{x_i} = 1.$$

Можно изменить порядок суммирования:

$$\sum_j \sum_i p_{x_i y_j} = \sum_j \left(\sum_i p_{x_i y_j} \right) = \sum_j p_{y_j} = 1.$$

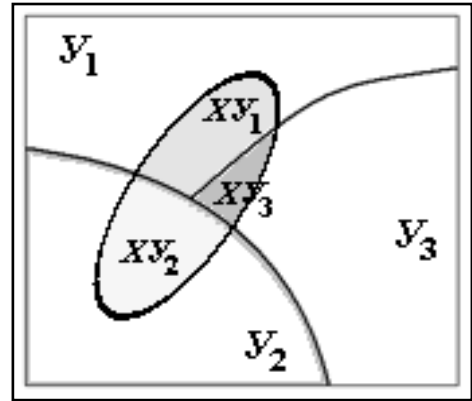


Рис. 3.5. Совместное появление X с Y_1, Y_2, Y_3

Теперь вычислим математическое ожидание суммы случайных величин и покажем, что оно равно сумме математических ожиданий:

$$\begin{aligned} \sum_i \sum_j (x_i + y_j) \cdot p_{x_i y_j} &= \sum_i x_i \left(\sum_j p_{x_i y_j} \right) + \\ &+ \sum_j y_j \left(\sum_i p_{x_i y_j} \right) = \sum_i x_i p_{x_i} + \sum_j y_j p_{y_j} = M(x) + M(y). \end{aligned}$$

4. Математическое ожидание произведения *независимых* случайных величин равно произведению их математических ожиданий.

Пусть x_i – значения случайной величины \mathcal{X} – появляются с вероятностями p_i , а y_j – значения случайной величины \mathcal{Y} – появляются с вероятностями q_j . Составим ряд распределения произведения случайных величин $\mathcal{Z} = \mathcal{X}\mathcal{Y}$. Всевозможные значения произведений $z_{ij} = x_i y_j$ *независимых* случайных величин появляются с вероятностями $p_i q_j$. Нетрудно убедиться, что сумма этих вероятностей равна единице. Вычисляем математическое ожидание произведения и получаем:

$$M(xy) = \sum_i \sum_j x_i y_j p_i q_j = \sum_i x_i p_i \sum_j y_j q_j = M(x) \cdot M(y),$$

что и требовалось доказать.

Приведем некоторые следствия доказанных выше утверждений.

Нулевое, или центральное, свойство математического ожидания: среднее значение отклонений всегда равно нулю (1-й центральный момент всегда равен нулю) $\mu_1 = M(x - a) = 0$, где $a = M(x)$.

Действительно, согласно свойствам математического ожидания, $M(x - a) = M(x) - M(a) = a - a = 0$.

Эквивалентная формула для вычисления дисперсии:

$$\begin{aligned} D(x) &= M(x - a)^2 = M(x^2 - 2ax + a^2) = M(x^2) - 2a \cdot M(x) + M(a^2) = \\ &= M(x^2) - 2a^2 + a^2 = M(x^2) - a^2 = M(x^2) - (M(x))^2. \end{aligned}$$

Эта более удобная для расчетов формула читается так: «Дисперсия равна разности математического ожидания квадрата и квадрата математического ожидания случайной величины».

Для примера с лотереей, где $M(x) = 0,9$, вычисляем:

$$M(x^2) = 0^2 \cdot 0,5 + 1^2 \cdot 0,3 + 2^2 \cdot 0,15 + 5^2 \cdot 0,04 + 10^2 \cdot 0,01 = 2,9; \quad D(x) = 2,9 + 0,9^2 = 2,09.$$

Это же значение было получено ранее более громоздким способом.

Точно так же можно выразить остальные центральные моменты через не-центральные, например, $\mu_3 = m^3 - 3m_1m^2 + 2(m_1)^3$.

Свойства дисперсии

1. Дисперсия постоянной равна нулю.

Действительно, так как $M(C) = C$, то $D(C) = M(C - C)^2 = M(0) = 0$.

2. При вынесении постоянного множителя за знак дисперсии его надо возводить в квадрат:

$$D(kx) = M(kx - ka)^2 = M(k^2(x - a)^2) = k^2 \cdot M(x - a)^2 = k^2 \cdot D(x).$$

3. Дисперсия суммы *независимых* случайных величин равна сумме их дисперсий $D(x + y) = D(x) + D(y)$. Здесь важна оговорка о «независимости» величин. При невыполнении этого условия данное свойство не имеет места. Например, при $y = x$ дисперсия суммы не равна сумме дисперсий: $D(x + x) = D(2x) = 4 \cdot D(x) \neq D(x) + D(x)$.

4. Дисперсия суммы двух случайных величин равна сумме их дисперсий и удвоенной ковариации, где ковариация – смешанный центральный момент μ_{xy} – математическое ожидание произведения отклонений случайных величин от своих центров: $\mu_{xy} = M(x - a)(y - b)$, где $a = M(x)$, $b = M(y)$. Если обозначить дисперсии и ковариацию σ_{xx} , σ_{yy} , σ_{xy} (эквивалентные обозначения), то утверждается, что $D(x + y) = \sigma_{xx} + \sigma_{yy} + 2\sigma_{xy}$.

Доказательство:

$$\begin{aligned} D(x + y) &= M((x + y) - (a + b))^2 = M((x - a) + (y - b))^2 = \\ &= M(x - a)^2 + M(y - b)^2 + 2 \cdot M(x - a)(y - b) = \sigma_{xx} + \sigma_{yy} + 2\sigma_{xy}. \end{aligned}$$

Для независимых случайных величин $\sigma_{xy} = M(x - a)(y - b) = M(x - a) \cdot M(y - b)$ (согласно 4-му свойству математического ожидания), $M(x - a) = M(y - b) = 0$ (центральное свойство математического ожидания), следовательно, для независимых случайных величин ковариация равна нулю $\sigma_{xy} = 0$. В этом случае справедливо утверждение: $D(x + y) = D(x) + D(y)$.

В другом частном случае $y = x$ (зависимые величины) ковариация равна дисперсии $\sigma_{xy} = \sigma_{xx}$ и снова получается правильный ответ, который согласуется с правилом вынесения постоянного множителя за знак дисперсии: $D(x + x) = \sigma_{xx} + \sigma_{xx} + 2\sigma_{xx} = 4\sigma_{xx} = 4 \cdot D(x)$.

Следствие: для независимых случайных величин:

$$D(\alpha x + \beta y) = \alpha^2 D(x) + \beta^2 D(y).$$

Например, $D(x - y) = D(x) + D(y)$; здесь $\beta = -1$, $\beta^2 = 1$.

Правило «3-х сигм»

Интересно, что для любой случайной величины справедливо следующее практическое правило: «Случайные отклонения от центра распределения, превышающие $3\sigma_x$, маловероятны». На практике это означает, что все данные с отклонениями, большими $3\sigma_x$, считаются неслучайными, чаще всего, выбросами, грубыми ошибками в записи чисел, или же поясняются неоднородностью совокупности (эти сомнительные данные с большими отклонениями очевидно не принадлежат изучаемой совокупности).

Вероятность появления больших отклонений оценивается с помощью неравенства Чебышева. Запишем и преобразуем формулу для дисперсии:

$$\begin{aligned} \sigma_x^2 &= \sum (x_i - a)^2 p_i = \sum_{|x-a| \leq t \cdot \sigma_x} (x_i - a)^2 p_i + \sum_{|x-a| > t \cdot \sigma_x} (x_i - a)^2 p_i \geq \sum_{|x-a| > t \cdot \sigma_x} (x_i - a)^2 p_i \geq \\ &\geq \sum_{|x-a| > t \cdot \sigma_x} (t \cdot \sigma_x)^2 p_i = t^2 \sigma_x^2 \sum_{|x-a| > t \cdot \sigma_x} p_i = t^2 \sigma_x^2 \cdot P(|x - a| > t \cdot \sigma_x). \end{aligned}$$

Отсюда получаем довольно грубую оценку вероятности появления больших отклонений:

$$P(|x - a| > t \sigma_x) \leq \frac{1}{t^2},$$

и вероятности противоположного события:

$$P(|x - a| \leq t\sigma_x) > 1 - \frac{1}{t^2}.$$

Если $t = 3$, то $P(|x - a| \leq 3\sigma_x) > 1 - 1/9 > 0,89$.

Таким образом, с уровнем доверия, не меньшим 90 %, можно утверждать, что случайные отклонения $|x - a|$ не превышают $3\sigma_x$.

В заключение рассмотрим функцию, которая будет основной в определении непрерывной случайной величины, а именно кумуляту – функцию накопленных вероятностей.

Для примера в таблице на рис. 3.6 приведен закон распределения \mathcal{X} – числа испытаний m до первого успеха для $p = 0,3$ ($q = 0,7$). Вероятности $P(m)$ здесь убывают по геометрической прогрессии $P(m) = pq^{m-1}$. В последней колонке таблицы вычислены накопленные суммы вероятностей $F(m) = \sum_{k=1}^m P(k)$; в общем случае эта функция определяется как $F(m) = P(\mathcal{X} \leq m)$.

Знание кумуляты позволяет легко находить вероятности попадания случайной величины в полуоткрытые интервалы по формуле:

$$P(m_1 < m \leq m_2) = F(m_2) - F(m_1).$$

Например,

$$\begin{aligned} P(2 \leq m \leq 8) &= P(1 < m \leq 8) = F(8) - F(1) = \\ &= 0,942 - 0,300 = 0,642. \end{aligned}$$

Этот способ значительно проще непосредственного суммирования вероятностей:

$$\begin{aligned} P(2 \leq m \leq 8) &= P(2) + P(3) + P(4) + P(5) + P(6) + P(7) + P(8) = \\ &= 0,210 + 0,147 + 0,103 + 0,072 + 0,050 + 0,035 + 0,025 = 0,642. \end{aligned}$$

Для непрерывной случайной величины функция $F(x)$ называется функцией распределения или интегральной функцией распределения. Она определена для всех значений x , в том числе и между узлами дискретной случайной величины (рис. 3.7).

Согласно определению $F(x) = P(\mathcal{X} \leq x)$, на интервалах $[x_i, x_{i+1})$ кумулята сохраняет постоянные значения, равные $F(x_i)$; $F(x < x_1) = 0$; $F(x \geq x_k) = 1$ (здесь x_k – наибольшее значение случайной величины).

m	$P(m)$	$F(m)$
1	0,3000	0,3000
2	0,2100	0,5100
3	0,1470	0,6570
4	0,1029	0,7599
5	0,0720	0,8319
6	0,0504	0,8824
7	0,0353	0,9176
8	0,0247	0,9424
9	0,0173	0,9597
10	0,0121	0,9718
11	0,0084	0,9802
12	0,0059	0,9861
13	0,0042	0,9903
14	0,0029	0,9932
15	0,0020	0,9952

Рис. 3.6. Кумулята $F(m)$

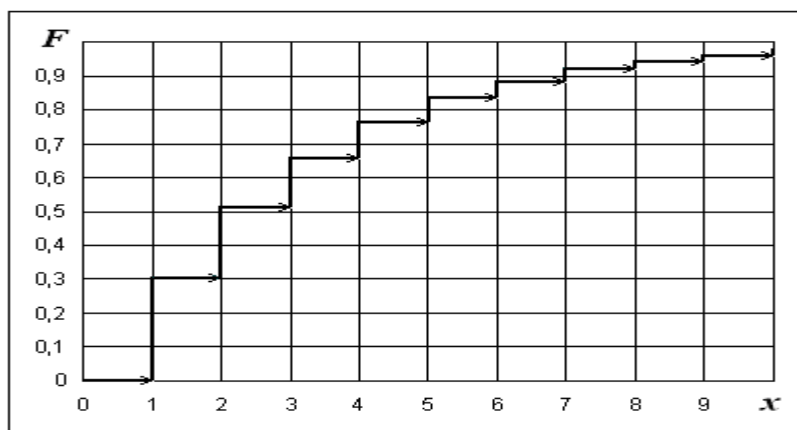


Рис. 3.7. График кумуляты

График кумуляты (функции распределения для дискретной величины) ступенчатый, значения кумуляты изменяются скачками в заданных узлах x_i случайной величины от 0 до 1.

Вопросы для самопроверки

1. Поясните, что такое случайная величина, как она связана со случайными событиями.
2. Какая разница между дискретной и непрерывной величинами?
3. Что такое закон распределения, как он задается?
4. Перечислите характеристики случайных величин.
5. Что такое математическое ожидание, как оно вычисляется?
6. Что такое дисперсия, как она вычисляется?
7. Что такое среднее квадратичное отклонение?
8. Перечислите свойства математического ожидания.
9. Перечислите свойства дисперсии.
10. Сформулируйте нулевое или центральное свойство математического ожидания.
11. Что такое ковариация, чему она равна для независимых случайных величин?
12. Что такое коэффициент асимметрии и коэффициент эксцесса?
13. Сформулируйте правило «3-х сигм».
14. Что такое кумулята, где она применяется?

4. Распределение Бернулли – Пуассона – Лапласа

Распределение Бернулли

Другие эквивалентные названия этого распределения – *биномиальное распределение, задача Бернулли*, или *задача о повторении однородных независимых испытаний*.

Швейцарские математики Иоганн и Якоб Бернулли доказали следующую формулу для расчета $P_n(m)$ – вероятности числа успехов (m) при n -кратном повторении однородных независимых испытаний, в каждом из которых событие A может появиться с вероятностью p :

$$P_n(m) = C_n^m p^m (1-p)^{n-m} = \frac{n!}{m!(n-m)!} p^m q^{n-m}.$$

Действительно, представим себе фрагмент полной группы несовместных событий при n -кратном повторении однородных независимых испытаний, где появилось m успехов. Как обычно, появление успеха (события A) будем обозначать знаком «+», а непоявление успеха (появление противоположного события \bar{A}) – знаком «-».

На рис. 4.1 изображено одно из событий полной группы, где m успехов были получены в первых же m испытаниях (знаки «+» в первых m позициях).

Значит, в оставшихся $(n-m)$ испытаниях успехов не было (знаки «-» в оставшихся $n-m$ позициях).

Поскольку вероятность успеха p в любом испытании одинакова, то, согласно теореме умножения, вероятность события, изображенного на рис. 4.1, равна произведению вероятностей $p^m q^{n-m}$, где $q = 1 - p$.

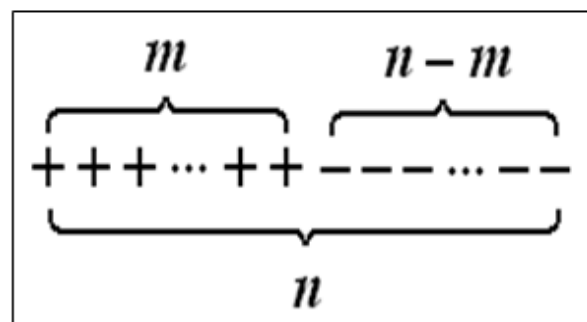


Рис. 4.1. Фрагмент полной группы несовместных событий

Порядок появления m успехов (порядок сомножителей p и q) не изменяет произведения вероятностей. При повторении однородных независимых испытаний события полной группы с одинаковым числом успехов равновероятны. Таких равновероятных событий в полной группе будет C_n^m , так как число сочетаний показывает, сколькими способами можно разместить m плюсов (+) на n позициях.

Согласно аксиоме сложения, при объединении несовместных событий их вероятности складываются, откуда и следует формула Бернулли:
 $P_n(m) = C_n^m p^m (1-p)^{n-m}$.

Можно составить ряд распределения случайной величины $\mathcal{X} = m$ – числа успехов в n однородных независимых испытаниях (рис. 4.2).

Распределение Бернулли зависит от 2-х параметров – p и n . На рис. 4.3 приведены типичные полигоны распределения Бернулли при различных значениях параметров.

$\mathcal{X} = m$	0	1	2	3	...	n
$P(m)$	$P_n(0)$	$P_n(1)$	$P_n(2)$	$P_n(3)$...	$P_n(n)$

Рис. 4.2. Ряд распределения Бернулли

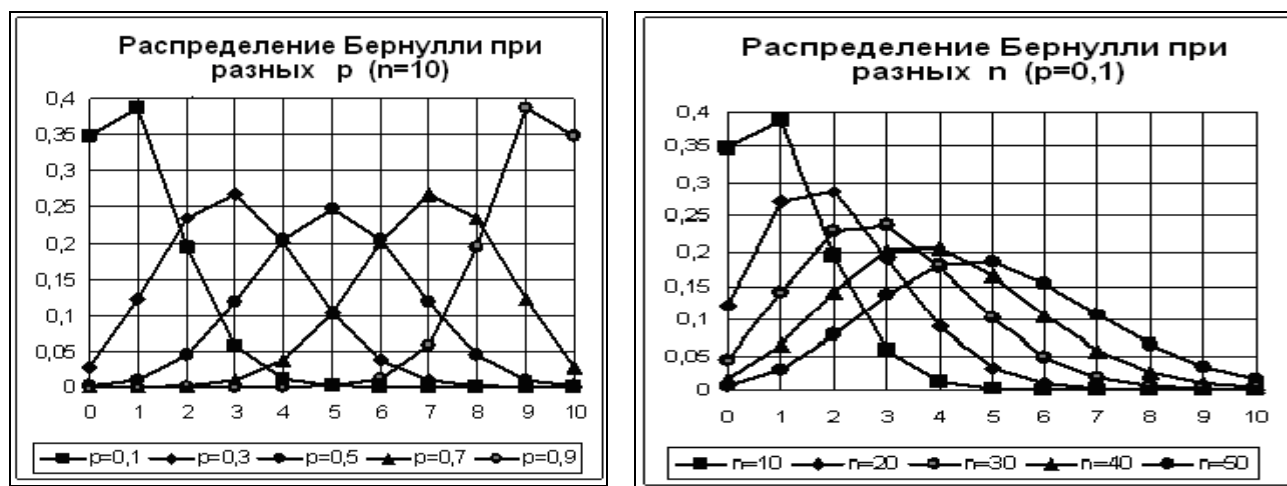


Рис. 4.3. Вид распределения Бернулли при различных значениях параметров

При $p = 0,5$ распределение симметричное, при $p < 0,5$ – скошено влево, при $p > 0,5$ – скошено вправо. При увеличении n форма распределения приближается к некоторому стандартному симметричному виду (независимо от значения параметра p).

Рассмотрим характеристики распределения Бернулли.

Прежде всего надо убедиться, что при любых значениях параметров сумма вероятностей $\sum P_n(m)$ равна единице. Действительно,

$$\sum_{m=0}^n P_n(m) = \sum_{m=0}^n C_n^m p^m (1-p)^{n-m} = \sum_{m=0}^n \frac{n!}{m!(n-m)!} p^m q^{n-m} = (p+q)^n = 1.$$

При суммировании вероятностей $\sum P_n(m)$ появилась формула бинома Ньютона $\sum_{m=0}^n \frac{n!}{m!(n-m)!} x^m y^{n-m} = (x+y)^n$, поэтому распределение Бернулли

называют также биномиальным. Коэффициенты $C_n^m = \frac{n!}{m!(n-m)!}$ в формуле Бернулли называются еще биномиальными коэффициентами, они могут быть легко рассчитаны с помощью треугольника Паскаля (рис. 4.4):

$(n=1)$					1		1					
$(n=2)$					1		2		1			
$(n=3)$				1		3		3		1		
$(n=4)$			1		4		6		4		1	
$(n=5)$		1		5		10		10		5		1
$(n=6)$	1		6		15		20		15		6	1

Рис. 4.4. Биномиальные коэффициенты

В треугольнике каждый коэффициент равен сумме двух соседних с ним коэффициентов предыдущего ряда.

Для вычисления математического ожидания и дисперсии требуется суммировать следующие выражения:

$$M(m) = \sum_{m=0}^n m \cdot P_n(m) = \sum_{m=0}^n m \cdot \frac{n!}{m!(n-m)!} p^m q^{n-m};$$

$$M(m^2) = \sum_{m=0}^n m^2 \cdot P_n(m) = \sum_{m=0}^n m^2 \cdot \frac{n!}{m!(n-m)!} p^m q^{n-m}.$$

Можно предложить студентам найти способ вычисления сумм $M(m)$, $M(m(m-1))$, $M(m(m-1)(m-2))$, но мы вычислим все суммы с помощью доказанных ранее свойств математического ожидания и дисперсии.

Напоминаем, что рассматривается задача о повторении однородных независимых испытаний, в каждом из которых случайная величина \mathcal{X}_i (число успехов в одном испытании) может принимать только два значения: 0 или 1.

Определяем первые три момента распределения для i -го испытания (рис. 4.5): $m_1 = 0 \cdot q + 1 \cdot p = p$, $m_2 = 0^2 \cdot q + 1^2 \cdot p = p$, $m_3 = 0^3 \cdot q + 1^3 \cdot p = p$.

Отсюда следует $M(\mathcal{X}_i) = m_1 = p$, $D(\mathcal{X}_i) = m_2 - (m_1)^2 = p - p^2 = pq$, $\mu_3(\mathcal{X}_i) = m_3 - 3m_1m_2 + 2(m_1)^3 = pq(q - p)$.

\mathcal{X}_i	0	1
$P(x)$	q	p

Рис. 4.5. Исходы i -го испытания

Общее число успехов при n испытаний равно сумме $\mathcal{X} = \mathcal{X}_1 + \mathcal{X}_2 + \dots + \mathcal{X}_n$. Известно, что математическое ожидание суммы случайных величин равно сумме математических ожиданий, дисперсия суммы независимых случайных величин также равна сумме диспер-

сий, можно доказать, что для независимых случайных величин такое же свойство имеет место и для момента 3-го порядка.

Отсюда получаем: $M(m) = np$, $D(m) = npq$, $\mu_3(m) = npq(q - p)$.

$$\sigma_m = \sqrt{npq}, \quad A = \frac{q - p}{\sqrt{npq}}, \quad v_m = \sqrt{\frac{q}{np}} \cdot 100 \% .$$

Можно также доказать двойное неравенство, однозначно устанавливающее местоположение моды (наивероятнейшего числа успехов при n испытаниях): $np - q \leq Mo \leq np + p$. На интервале длиной $(np + p) - (np - q) = p + q = 1$ может быть только одно целое Mo или же два соседних, если целочисленны края интервала. Предлагается студенту в качестве полезного упражнения самостоятельно вывести это двойное неравенство. *Указание:* необходимо в неравенство $P_n(Mo-1) \leq P_n(Mo) \leq P_n(Mo+1)$ подставить выражения для $P_n(m)$ и сократить общие множители.

Для примера решим типичную задачу на распределение Бернулли. Вероятность того, что электротехнический прибор потребует ремонта в гарантийный срок, равна 0,25. Найти вероятность того, что на протяжении гарантийного срока из семи приборов ремонта потребуют $m = 0, 1, 2, \dots$ приборов; потребуют ремонта не более четырех, не меньше двух, больше двух и меньше шести приборов.

Решение. Речь идет о повторении *однородных независимых* испытаний (так как $p = 0,25 = Const$). Число испытаний (число приборов) равно $n = 7$ ($n < 30$), поэтому применяем формулу Бернулли:

$$P_7(m) = C_7^m p^m q^{7-m} = \frac{7!}{m!(7-m)!} p^m q^{7-m}.$$

В этой задаче $p = 0,25$; $q = 1 - p = 0,75$; биномиальные коэффициенты можно определить также по треугольнику Паскаля.

Для $m = 0, 1, 2, \dots, 7$ расчеты удобно свести в таблицу (рис. 4.6), в последнем столбце которой приведены накопленные суммы вероятностей (значения кумуляты) $F(m)$. Напоминаем, что функция распределения (кумулята) определена как $F(m) = P(\mathcal{X} \leq m)$, поэтому $F(0) = P_n(0) = 0,133484$.

С помощью кумуляты вероятность попадания случайной величины в полуоткрытый интервал $m_1 < m \leq m_2$ вычисляется как разность значений функции $F(m)$ на краях этого интервала:

$$P(m_1 < m \leq m_2) = F(m_2) - F(m_1-1).$$

m	C_n^m	p^m	q^{n-m}	$P_n(m)$	$F(m)$
0	1	1	0,133484	0,133484	0,133484
1	7	0,25	0,177979	0,311462	0,444946
2	21	0,0625	0,237305	0,311462	0,756409
3	35	0,015625	0,316406	0,173035	0,929443
4	35	0,003906	0,421875	0,057678	0,987122
5	21	0,000977	0,5625	0,011536	0,998657
6	7	0,000244	0,75	0,001282	0,999939
7	1	0,000061	1	0,000061	1,000000

Рис. 4.6. Расчеты по формуле Бернулли

Вычисляем:

$$P(m \leq 4) = F(4) = 0,987;$$

$$P(m \geq 2) = P(2 \leq m \leq 7) = F(7) - F(1) = 1 - 0,4450 = 0,555;$$

$$P(2 < m < 6) = P(3 \leq m \leq 5) = F(5) - F(2) = 0,999 - 0,756 = 0,242.$$

Заметим, что во многих учебниках функция $F(x)$ определена немного по-другому: $F(x) = P(\mathcal{X} < x)$. Тогда для нашего примера надо было бы принять $F(0) = 0$ и изменить формулу расчета вероятности попадания случайной величины в интервал, полуоткрытый сверху: $P(m_1 \leq m < m_2) = F(m_2) - F(m_1)$ или же $P(m_1 \leq m \leq m_2) = F(m_2 + 1) - F(m_1)$. Оба определения функции $F(x)$ эквивалентны, но следует придерживаться одного стандарта.

Вычисляем характеристики распределения Бернулли:

$$M(m) = np = 7 \cdot 0,25 = 1,75; D(m) = npq = 7 \cdot 0,25 \cdot 0,75 = 1,3125;$$

$$\sigma_m = \sqrt{D(m)} = 1,1456; M(m) - q \leq Mo \leq M(m) + p;$$

$$1,75 - 0,75 \leq Mo \leq 1,75 + 0,25; 1 \leq Mo \leq 2; P(1) = P(2) = P_{max}.$$

Согласно правилу «3-х сигм», вероятные значения m не превышают:

$$M(m) + 3 \cdot \sigma_m = 1,75 + 3 \cdot 1,146 = 5,19 \approx 5.$$

На рис. 4.7 приведен полигон распределения Бернулли для данного примера.



Рис. 4.7. Полигон распределения Бернулли

Распределение Пуассона

Для $n > 30$ производить расчеты по формуле Бернулли становится затруднительным из-за слишком больших величин факториалов, поэтому используют асимптотические формулы Пуассона и Лапласа, которые становятся все более и более точными с увеличением n (именно в тех случаях, когда расчеты по исходной формуле Бернулли практически невозможны). Формула Пуассона применяется для больших $n > 30$ и малых $p < 0,05$, таких, что $np < 5$ (поэтому распределение Пуассона применяется для изучения распределения числа редких событий). Во всех остальных случаях ($n > 30$, $np \geq 5$, $nq \geq 5$) применяется асимптотическая формула Лапласа. Используем для примера партию электролампочек, из которых 2 % выходит из строя при перевозке. Какова вероятность того, что число испорченных лампочек будет не больше 5? Если перевозится $n = 10$ лампочек, то расчеты вероятностей надо производить по формуле Бернулли; если в партии $n = 100$ лампочек – по формуле Пуассона ($n = 100 > 30$; $p = 0,02 < 0,05$; $np = 2 < 5$); если же в партии $n = 1000$ лампочек – по формуле Лапласа ($n = 1000 > 30$; $np = 20 > 5$; $nq = 980 > 5$).

Закон распределения редких событий Пуассона

$$P(m) = e^{-a} \cdot \frac{a^m}{m!}$$

имеет самостоятельное значение и свою область применения – теорию массового обслуживания.

Получим вышеприведенное выражение как предельную формулу для распределения Бернулли при $n \rightarrow \infty$, $p \rightarrow 0$, $q \rightarrow 1$, но при этом $np = a = \text{Const}$.

Преобразуем формулу Бернулли:

$$\begin{aligned} P_n(m) &= \frac{n!}{m!(n-m)!} p^m q^{n-m} = \frac{n(n-1)(n-2)\dots(n-m+1)}{1 \cdot 2 \cdot 3 \dots m} \cdot \left(\frac{a}{n}\right)^m \cdot \left(1 - \frac{a}{n}\right)^{n-m} = \\ &= \frac{a^m}{m!} \cdot \frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \dots \frac{n-m+1}{n} \cdot \left(1 - \frac{a}{n}\right)^{-m} \cdot \left(1 - \frac{a}{n}\right)^n \rightarrow \frac{a^m}{m!} \cdot \lim_{n \rightarrow \infty} \left(1 - \frac{a}{n}\right)^n = \frac{a^m}{m!} \cdot e^{-a}. \end{aligned}$$

Поскольку формулу Пуассона можно рассматривать как предельный случай формулы Бернулли, то для нее сохранились все характеристики распределения Бернулли с заменой $p \approx 0$, $q \approx 1$, $np = a$:

$$(a - 1) \leq Mo \leq a; M(m) = a; D(m) = a; A = \frac{1}{\sqrt{a}}.$$

Распределение Пуассона зависит от одного параметра a . Типичный вид полигонов распределения Пуассона для разных значений параметра a показан на рис. 4.8, из которого видно, что с увеличением a вид распределения приближается к стандартной симметричной форме (распределению Лапласа).



Рис. 4.8. Вид распределения Пуассона в зависимости от параметра a

Вычисления вероятностей $P(m)$ распределения Пуассона удобно производить по рекуррентной формуле: $P(m) = P(m-1) \cdot \frac{a}{m}$, где $P(0) = e^{-a}$.

Например, для $a = 2$:

$$\begin{aligned} P(0) &= e^{-2} = 0,135; & P(1) &= P(0) \cdot \frac{2}{1} = 0,270; & P(2) &= P(1) \cdot \frac{2}{2} = 0,270; \\ P(3) &= P(2) \cdot \frac{2}{3} = 0,180; & P(4) &= P(3) \cdot \frac{2}{4} = 0,090; & P(5) &= P(4) \cdot \frac{2}{5} = 0,036; \\ P(6) &= P(5) \cdot \frac{2}{6} = 0,012; & P(7) &= P(6) \cdot \frac{2}{7} = 0,003 \text{ и т. д.} \end{aligned}$$

Обычно в задачах на распределение Пуассона задается среднее число появления некоторого события за определенный период времени: $a = \lambda t$, где λ – интенсивность появления события – среднее число событий в единицу времени.

Для примера рассмотрим следующую задачу.

В одном кубическом метре воздуха в среднем находится 1000 болезнетворных микробов. Для анализа взяли 2 литра (дм^3) воздуха. Найти вероятность того, что в пробе будет выявлено $m = 0, 1, 2, \dots$ болезнетворных микробов, не больше трех микробов, от двух до пяти микробов, хотя бы один микроб.

Решение. Это типичная задача на распределение Пуассона, где задана интенсивность $\lambda = 1000$ для $t = 1 \text{ м}^3$. Для $t = 2 \text{ л} = 0,002 \text{ м}^3$ получаем $a = 1000 \cdot 0,002 = 2$ – это среднее количество микробов в 2-х литрах воздуха.

Вычисления вероятностей распределения Пуассона $P(m) = e^{-a} \cdot \frac{a^m}{m!}$ производим по рекуррентным формулам:

$$P(m) = P(m-1) \cdot \frac{a}{m} = P(m-1) \cdot \frac{2}{m}; \quad P(0) = e^{-a} = e^{-2} = 0,135335.$$

Возможные значения m не ограничены сверху, однако, согласно правилу «3-х сигм», достаточно рассчитать вероятности $P(m)$ только для

$m < a + 3\sqrt{a} = 2 + 3 \cdot 1,414 = 6,2$. Для $m = 0, 1, 2, \dots, 10$ вычисления сведены в таблицу на рис. 4.9, рядом с ней построен график полигона (рис. 4.10).

m	$P(m)$	$F(m)$
0	0,135335	0,135335
1	0,270671	0,406006
2	0,270671	0,676676
3	0,180447	0,857123
4	0,090224	0,947347
5	0,036089	0,983436
6	0,012030	0,995466
7	0,003437	0,998903
8	0,000859	0,999763
9	0,000191	0,999954
10	0,000038	0,999992

Рис. 4.9. Расчеты по формуле Пуассона



Рис. 4.10. Полигон распределения Пуассона

С помощью функции $F(m)$ находим:

$$P(m \leq 3) = F(3) = 0,857123;$$

$$P(2 \leq m \leq 5) = F(5) - F(1) = 0,983436 - 0,406006 = 0,57743;$$

$$P(m \geq 1) = 1 - P(0) = 1 - 0,135335 = 0,864665.$$

Определяем характеристики распределения Пуассона:

$$M(m) = a = 2; D(m) = a = 2; \sigma_m = \sqrt{D(m)} = 1,414;$$

$$a - 1 \leq Mo \leq a; 1 \leq Mo \leq 2; P(1) = P(2) = P_{max}.$$

Здесь целочисленны оба края интервала длиной в единицу, поэтому самыми вероятными оказались сразу два соседних значения.

Вывод формулы для расчета вероятностей распределения Пуассона

Выше распределение Пуассона было получено как предельное из распределения Бернулли при определенных условиях: $n \rightarrow \infty$, $p \rightarrow 0$, но при этом $np = a = Const$. Здесь приведен вывод распределения Пуассона без ссылок на распределение Бернулли. Рассмотрим события, которые наступают в случайные моменты времени. Последовательность таких событий называется *поток событий*. Примерами потоков событий служат: поступление вызовов на АТС, в пункт неотложной медицинской помощи, прибытие самолетов в аэропорт, последовательности отказов элементов и многое другое. *Простейшим* (пуассоновским) называют поток событий, который обладает свойствами

стационарности, отсутствия последействия и ординарности. Свойство *стационарности* характеризуется тем, что вероятность появления событий на любом промежутке времени зависит только от длительности t промежутка, но не от начала его отсчета. Свойство *отсутствия последействия* характеризуется тем, что вероятность появления событий на любом промежутке времени не зависит от того, появлялись или не появлялись события в моменты времени, предшествующие началу рассматриваемого промежутка. Свойство *ординарности* характеризуется тем, что за малый промежуток времени вероятность появления более одного события пренебрежимо мала по сравнению с вероятностью появления только одного события.

Обозначим через $P_m(t)$ вероятность появления m событий за период времени t ; тогда $P_0(t)$ означает вероятность того, что за этот период не произойдет ни одного события; $1 - P_0(t)$ – произойдет хотя бы одно событие.

Далее обозначим $\lim_{t \rightarrow 0} \frac{1 - P_0(t)}{t} = \lambda$, тогда для малого интервала времени h

имеем $1 - P_0(h) = \lambda h + o(h)$, где $o(h)$ убывает быстрее, чем h .

Принимаем постулаты Пуассона: каково бы ни было число появления событий в период времени $(0; t)$, условная вероятность того, что в течение последующего малого интервала времени $(t; t + h)$ появится одно событие, равна $\lambda h + o(h)$, а вероятность того, что появится более одного события, есть $o(h)$. Из этих условий выводятся дифференциальные уравнения для определения $P_m(t)$.

На протяжении объединенного интервала времени $(0; t + h)$ m событий ($m \geq 1$) могут появиться 3-мя взаимоисключающими способами: 1) ни одного события за время $(t; t + h)$ и m событий за время $(0; t)$; 2) одно событие за время $(t; t + h)$ и $(m - 1)$ событий за время $(0; t)$; 3) более одного события за время $(t; t + h)$ и остальные события за время $(0; t)$. Вероятность первой возможности равна произведению $P_m(t)$ на вероятность того, что на интервале $(t; t + h)$ не произойдет ни одного события; эта последняя вероятность равна $1 - \lambda h - o(h)$. Аналогично вторая возможность имеет вероятность $P_{m-1}(t) \cdot \lambda h + o(h)$, а вероятность третьей возможности убывает быстрее, чем h . Отсюда для $m \geq 1$ имеем:

$$P_m(t + h) = P_m(t) \cdot (1 - \lambda h) + P_{m-1}(t) \cdot \lambda h + o(h);$$

$$\frac{P_m(t + h) - P_m(t)}{h} = -\lambda P_m(t) + \lambda P_{m-1}(t) + \frac{o(h)}{h}.$$

Учтем, что $\lim_{h \rightarrow 0} \frac{o(h)}{h} = 0$, тогда после предельного перехода $h \rightarrow 0$ получим:

$$P'_m(t) = -\lambda P_m(t) + \lambda P_{m-1}(t).$$

Это рекуррентное соотношение позволяет записать дифференциальное уравнение для определения $P_m(t)$, если уже известно выражение для $P_{m-1}(t)$. Для начала найдем $P_0(t)$.

При $m = 0$ вторая и третья возможности не возникают, поэтому рекуррентное уравнение упрощается:

$$P_0(t + h) = P_0(t) \cdot (1 - \lambda h) + o(h), \text{ откуда } P_0'(t) = -\lambda P_0(t).$$

С учетом естественного начального условия $P_0(0) = 1$ из последнего дифференциального уравнения получаем $P_0(t) = e^{-\lambda t}$.

При $m = 1$ имеем $P_1'(t) = -\lambda P_1(t) + \lambda e^{-\lambda t}$, откуда с учетом начального условия $P_1(0) = 0$ получаем $P_1(t) = (\lambda t) \cdot e^{-\lambda t}$.

Действуя таким же образом, для остальных m получаем $P_m(t) = e^{-\lambda t} \cdot \frac{(\lambda t)^m}{m!}$. Это распределение Пуассона, для которого $M(m) = \lambda t$, следовательно, параметр λ есть интенсивность – среднее число появления событий в единицу времени.

Вопросы для самопроверки

1. Сформулируйте задачу Бернулли, приведите ее другие названия.
2. Опишите особенности распределения Бернулли в зависимости от его параметров.
3. Приведите расчетные формулы для основных характеристик распределения Бернулли.
4. Укажите область применения асимптотических формул Пуассона и Лапласа.
5. Приведите формулу Пуассона и укажите область применения этого распределения. Что такое рекуррентная формула?
6. Опишите особенности распределения Пуассона в зависимости от параметра.
7. Приведите расчетные формулы для основных характеристик распределения Пуассона.

5. Распределение Лапласа

Для $n > 30$, $np \geq 5$, $nq \geq 5$ распределение Бернулли практически точно аппроксимируется асимптотической формулой Лапласа:

$$P_n(m) = \frac{1}{\sqrt{2\pi} \sqrt{npq}} \cdot e^{-\frac{(m-mp)^2}{2npq}} = \frac{\phi(t_m)}{\sigma_m},$$

где обозначено $t_m = \frac{m-a}{\sigma_m}$, $a = np$, $\sigma_m = \sqrt{npq}$, $\phi(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}$.

Эта аппроксимация называется *локальной теоремой Лапласа*.

На рис. 5.1 для сравнения приведены полигоны распределений Бернулли и Лапласа для $p = 0,3$; $n = 10$ и $n = 20$. Даже для таких небольших значений n соответствие очень хорошее (здесь $np = 3$ и $np = 6$).

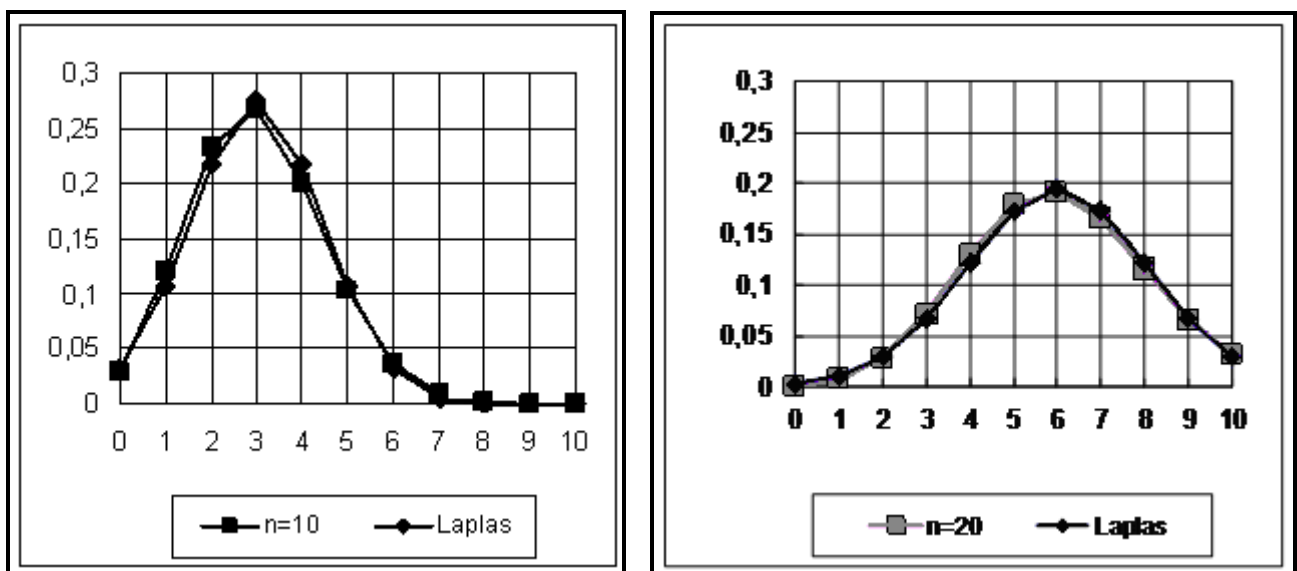


Рис. 5.1. Сравнение распределений Бернулли и Лапласа

Функция $\phi(t)$ затабулирована и называется *дифференциальной функцией Лапласа*, или же функцией *стандартизованного нормального распределения Гаусса*. Дело в том, что (как выяснилось впоследствии) Лаплас открыл частный случай более общего закона природы, который Гаусс назвал «нормальным». Известно, что в русском языке слово «нормальный» имеет совершенно иной смысл, чем в немецком, но это новое понятие не противоречит сути, поскольку нормальное распределение действительно является неким стандартным устой-

чивым законом природы, к которому при определенных условиях приближаются многие другие законы распределений.

Функция $\varphi(t)$ имеет некоторые особенности, которые учтены при составлении таблиц.

Во-первых, эта функция четная (график ее симметричен относительно оси ординат), поэтому таблицы составлены только для неотрицательных значений t ; для отрицательных t используется соотношение четности $\varphi(-t) = \varphi(t)$.

Во-вторых, функция $\varphi(t)$ неотрицательная и имеет горизонтальную асимптоту – ось абсцисс; иными словами, при увеличении t значения $\varphi(t)$ приближаются к нулю, поэтому таблицы составлены только до значений $t \leq 5$; так: $\varphi(3) = 0,00443$; $\varphi(4) = 0,00013$; $\varphi(5) = 0,00001$; для больших значений аргумента $\varphi(t) \approx 0$.

Максимальное значение функции достигается при $t = 0$ и равно $\varphi(0) = 0,3989$.

Характерные особенности дифференциальной функции Лапласа изображены на рис. 5.2.



Рис. 5.2. Дифференциальная функция Лапласа

Интегральная теорема Лапласа

Для больших n вычисление вероятностей отдельных значений m лишено особого смысла, так как даже для самого вероятного значения – моды получается $P(Mo) = \frac{\varphi(0)}{\sqrt{npq}} < \frac{0,4}{\sqrt{npq}} \sim \frac{1}{\sqrt{n}}$ – очень малое число при большом n .

В практических задачах для больших n требуется находить вероятности попадания случайной величины в некоторые интервалы $P(m_1 \leq m \leq m_2)$, то есть вычислять вероятности не одного значения m , а вероятности всех целых значений m из интервала $[m_1, m_2]$:

$$P(m_1 \leq m \leq m_2) = \sum_{m=m_1}^{m_2} P_n(m) = \sum_{t=t_{m_1}}^{t_{m_2}} \varphi(t_m) \cdot \frac{1}{\sigma_m}.$$

Рассмотрим приращение Δt_m при увеличении m на единицу:

$$\Delta t_m = t_{m+1} - t_m = \frac{(m+1) - a}{\sigma_m} - \frac{m - a}{\sigma_m} = \frac{1}{\sigma_m} = \frac{1}{\sqrt{npq}} \rightarrow 0.$$

Оказывается, $P(m_1 \leq m \leq m_2) = \sum_{t=t_{m_1}}^{i_{m_2}} \varphi(t_m) \cdot \frac{1}{\sigma_m} = \sum_{t=t_{m_1}}^{i_{m_2}} \varphi(t) \cdot \Delta t \rightarrow \int_{t_{m_1}}^{i_{m_2}} \varphi(t) \cdot dt.$

Лаплас ввел функцию $\Phi(t) = \int_0^t \varphi(s) ds = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-\frac{s^2}{2}} ds$, которая сейчас назы-

вается интегральной функцией Лапласа.

Теперь вероятность попадания случайной величины m в интервал $[m_1, m_2]$

можно записать $P(m_1 \leq m \leq m_2) \approx \int_{t_{m_1}}^{i_{m_2}} \varphi(t) \cdot dt = \Phi(t_{m_2}) - \Phi(t_{m_1})$ как разность значе-

ний интегральной функции Лапласа на краях этого интервала. Это утверждение называется **интегральной теоремой Лапласа**.

Функция $\Phi(t)$ имеет некоторые особенности, которые учтены при составлении таблиц. Во-первых, эта функция нечетная (график ее центрально симметричен относительно начала ординат), поэтому таблицы составлены только для неотрицательных значений t ; для отрицательных t используется соотношение нечетности $\Phi(-t) = -\Phi(t)$.

Во-вторых, функция $\Phi(t)$ возрастающая и имеет горизонтальные асимптоты – $\Phi_{min} = -0,5$ и $\Phi_{max} = 0,5$; иными словами, при увеличении t значения $\Phi(t)$ приближаются к 0,5, поэтому таблицы составлены только до значений $t \leq 5$; то есть: $\Phi(3) = 0,49865$; $\Phi(4) = 0,49997$; $\Phi(5) = 0,5000$; для еще больших значений аргумента $\Phi(t) = 0,5$. Значение функции при $t = 0$ равно нулю $\Phi(0) = 0$. Характерные особенности интегральной функции Лапласа изображены на рис. 5.3.



Рис. 5.3. Интегральная функция Лапласа

Три основных формы интегральной теоремы Лапласа

Для лучшего усвоения материала введем условную классификацию утверждений интегральной теоремы Лапласа и перечислим задачи, которые решаются с помощью каждой из них.

1. В качестве первой формы теоремы примем общую формулу для вычисления вероятности попадания случайной величины в заданные интервалы:
 $P(m_1 \leq m \leq m_2) \approx \Phi(t_{m_2}) - \Phi(t_{m_1})$.

2. Вторая форма предназначена для вычисления вероятности попадания случайной величины в интервалы, симметричные относительно центра:
 $P(|m - a| \leq t \cdot \sigma_m) = 2\Phi(t)$.

Действительно, для симметричных интервалов $m_2 = a + t \cdot \sigma_m$, $m_1 = a - t \cdot \sigma_m$,
 $t_{m_2} = \frac{(a + t \cdot \sigma_m) - a}{\sigma_m} = t$, $t_{m_1} = \frac{(a - t \cdot \sigma_m) - a}{\sigma_m} = -t$. Вычисляем:

$$P(|m - a| \leq t \cdot \sigma_m) = P(m_1 \leq m \leq m_2) = \Phi(t) - \Phi(-t) = \Phi(t) + \Phi(t) = 2\Phi(t).$$

В качестве примера применения выведенной формулы ответим на вопросы: «Какова вероятность того, что отклонения случайной величины m от центра $a = np$ не превысят σ_m , $2\sigma_m$, $3\sigma_m$?»

Используя таблицы интегральной функции Лапласа, получим:

$$P(|m - a| \leq \sigma_m) = 2\Phi(1) = 2 \cdot 0,3413 = 0,6826;$$

$$P(|m - a| \leq 2\sigma_m) = 2\Phi(2) = 2 \cdot 0,4772 = 0,9544;$$

$$P(|m - a| \leq 3\sigma_m) = 2\Phi(3) = 2 \cdot 0,4987 = 0,9974.$$

Только в 3-х случаях из 1000 возможны появления отклонений, превышающие три сигмы; с гарантией 95 % можно утверждать, что для распределения Лапласа (и нормального закона Гаусса) случайные отклонения не превысят две сигмы (это утверждение может быть ошибочным лишь в 5-ти случаях из 100).

3. Перепишем неравенство $|m - a| \leq t \cdot \sigma_m$ в виде $|m - np| \leq t \sqrt{npq}$ и далее (разделив обе части неравенства на n): $\left| \frac{m}{n} - p \right| \leq t \sqrt{\frac{pq}{n}}$.

Получаем следующее выражение, которое мы назовем «третьей формой интегральной теоремы Лапласа»: $P\left(\left| \frac{m}{n} - p \right| \leq t \sqrt{\frac{pq}{n}}\right) = 2\Phi(t)$ или же

$P\left(\left| \frac{m}{n} - p \right| \leq \varepsilon\right) = 2\Phi\left(\varepsilon \sqrt{\frac{n}{pq}}\right)$. В описании этой формулы два раза встречается

слово вероятность P и p , что приводит к некорректным словосочетаниям типа:

«Вероятность того, что относительная частота $\frac{m}{n}$ отклонится от вероятности p

появления события в одном испытании не более чем на ε , равна удвоенному

значению функции Лапласа с аргументом $\varepsilon \sqrt{\frac{n}{pq}}$ ». Это определение громоздко

и малопонятно. Следует избегать в одном предложении двух одинаковых слов, да еще с разным смыслом. Например, вероятность некоторого утверждения

$P\left(\left|\frac{m}{n} - p\right| \leq \varepsilon\right)$ можно назвать гарантией, надежностью, уровнем доверия. Кста-

ти, вероятность противоположного утверждения называют уровнем значимости

и обычно обозначают греческой буквой альфа: $\alpha = 1 - P$. Если потребуется, то

вероятность p появления события в одном испытании можно назвать долей в совокупности, а относительную частоту – долей в выборке. Выражение

$\varepsilon = t \sqrt{\frac{pq}{n}}$ называется погрешностью. Теперь теорему можно сформулировать

так: «С уровнем доверия $P = 2\Phi(t)$ можно утверждать, что отклонение относи-

тельной частоты $\frac{m}{n}$ от своей доли в совокупности p (от своего предельного

значения p) не превысит погрешности $\varepsilon = t \sqrt{\frac{p(1-p)}{n}}$ ».

Формула $P\left(\left|\frac{m}{n} - p\right| \leq t \sqrt{\frac{pq}{n}}\right) = 2\Phi(t)$, или $P\left(\left|\frac{m}{n} - p\right| \leq \varepsilon\right) = 2\Phi\left(\varepsilon \sqrt{\frac{n}{pq}}\right)$, свя-

зывает вместе четыре величины n, p, ε, P . Если в условиях задачи заданы лю-

бые три величины, то четвертую можно найти по вышеприведенной формуле. В

связи с этим появляется 4 типа задач.

I. Даны n, p, ε . Требуется найти P .

Заданы параметры распределения n, p . Дополнительно задана допустимая

погрешность ε . Требуется найти уровень доверия утверждения $\left|\frac{m}{n} - p\right| \leq \varepsilon$, как

часто оно будет выполняться.

Решение. Из выражения для погрешности $\varepsilon = t \sqrt{\frac{p(1-p)}{n}}$ находим t и вы-

числяем уровень доверия (гарантию) $P = 2\Phi(t)$.

Пример. В процессе статистических испытаний монету предполагается подбросить $n = 100$ раз. Можно ли утверждать, что в результате опыта отклонение относительной частоты от своего теоретического значения $p = 0,5$ (то есть погрешность ε) не превысит $0,01$?

Определим уровень доверия утверждения $\left| \frac{m}{n} - 0,5 \right| \leq 0,01$. Из выражения

$0,01 = t \sqrt{\frac{0,5 \cdot 0,5}{100}}$ получим $t = 0,2$ и далее по таблицам Лапласа найдем

$P = 2\Phi(0,2) = 2 \cdot 0,079 = 0,158$. Уровень доверия $15,8\%$ явно не достаточен, чтобы гарантировать выполнение заданного условия; поскольку только в 16 случаях из 100 возможны столь малые отклонения ($\varepsilon = 0,01$).

II. Даны n, p, P . Требуется найти ε .

Утверждения должны быть надежными со стандартным уровнем доверия 90% , 95% или 99% . Какую предельную погрешность можно ожидать с таким заданным уровнем доверия?

Решение. Из соотношения $P = 2\Phi(t)$ при заданном P по таблицам интегральной функции Лапласа находим t . Далее вычисляем предельную ожидаемую погрешность:

$$\varepsilon = t \sqrt{\frac{p(1-p)}{n}}.$$

Пример. Предполагается подбросить монету 100 раз. Какую погрешность следует ожидать с гарантией 95% ?

Из выражения $0,95 = 2\Phi(t)$ с помощью таблиц Лапласа находим $t = 1,96 \approx 2$. Далее вычисляем $\varepsilon = 2 \sqrt{\frac{0,5 \cdot 0,5}{100}} = 0,1$. Иными словами, при 100 испытаниях относительная частота может варьировать в довольно широких пределах от $0,4$ до $0,6$.

III. Даны p, ε, P . Требуется найти n .

Погрешность ε уменьшается с увеличением числа испытаний. Сколько же требуется провести испытаний, чтобы получать надежные и точные прогнозы?

Решение. Из соотношения $P = 2\Phi(t)$ при заданном P находим t . Далее записываем условие для предельной погрешности $\varepsilon \geq t \sqrt{\frac{pq}{n}}$, откуда при извест-

ных p, ε, t находим n : $n \geq pq \cdot \left(\frac{t}{\varepsilon} \right)^2$.

Пример. Сколько раз надо подкинуть монету, чтобы с гарантией 90 % снизить погрешность до 0,05? Иными словами, сколько требуется испытаний, чтобы с гарантией 90 % относительная частота появления герба при бросании монеты не выходила за пределы интервала (0,45; 0,55)?

Из соотношения $0,9 = 2\Phi(t)$ находим $t = 1,64$. Далее из неравенства

$$0,05 \geq 1,64 \cdot \sqrt{\frac{0,5 \cdot 0,5}{n}} \text{ определяем } n \geq 0,5 \cdot 0,5 \cdot \left(\frac{1,64}{0,05}\right)^2 = 16,4^2 = 269.$$

IV. Даны n , P и m/n . Найти p .

В предыдущих задачах были заданы параметры распределения и требовалось предсказать результат опыта (задача теории вероятностей). В задаче IV известен результат опыта, и надо сделать заключение о параметрах распределения (именно в этом и заключается статистический способ определения вероятностей).

Решение. Соотношение $P = 2\Phi(t)$ при заданном P определяет t . Далее записываем условие $|m/n - p| \leq \varepsilon$, которое должно выполняться с заданным уровнем доверия: $\left|\frac{m}{n} - p\right| \leq t \sqrt{\frac{p(1-p)}{n}}$. Из этого неравенства находим p .

Пример. В результате 400 бросков монеты относительная частота появления герба получилась равной 0,5. Какие значения параметра p согласуются с такими результатами опыта? Заключение должно иметь 90-процентный уровень доверия.

Из соотношения $0,9 = 2\Phi(t)$ находим $t = 1,64$. Далее выписываем неравенство $|0,5 - p| \leq 1,64 \cdot \sqrt{\frac{p(1-p)}{400}}$, возводим его в квадрат и преобразуем:

$$(0,5 - p)^2 \leq 0,082^2 \cdot p(1 - p); 1,0067 \cdot p^2 - 1,0067 \cdot p + 0,25 \leq 0;$$

отсюда $p_1 \leq p \leq p_2$,

где p_1, p_2 – корни квадратного трехчлена $p_{1,2} = 0,5 \pm 0,041$.

Итак, по результатам опыта можно сделать заключение, что с гарантией 90 % искомая вероятность p находится в пределах интервала $0,46 \leq p \leq 0,54$.

В заключение заметим, что основная асимптотическая формула $P(m_1 \leq m \leq m_2) \approx \Phi(t_{m_2}) - \Phi(t_{m_1})$ дает достаточно точные значения лишь для больших $n > 100$.

Причина этого заключается в слишком неточной замене интегральной суммы интегралом $P(m_1 \leq m \leq m_2) = \sum_{m=m_1}^{m_2} P(m) \Delta m \rightarrow \int_{m_1}^{m_2} P(m) dm = \int_{t_{m_1}}^{t_{m_2}} \varphi(t) \cdot dt$ с теми же самыми пределами интегрирования. На рис. 5.4 изображен полигон распределения Лапласа для $n = 10$, $p = 0,3$ (локальная формула Лапласа является достаточно точной аппроксимацией распределения Бернулли даже для небольших значений n). Пусть требуется вычислить $P(2 \leq m \leq 6) = P(2) + P(3) + P(4) + P(5) + P(6) = \sum P(m)$. Так как $\Delta m = 1$, то $P(2 \leq m \leq 6) = \sum P(m) \Delta m$.

Это площадь столбиковой фигуры (рис. 5.4) с границами $(m_1 - 0,5)$ и $(m_2 + 0,5)$, поэтому более правильным будет замена

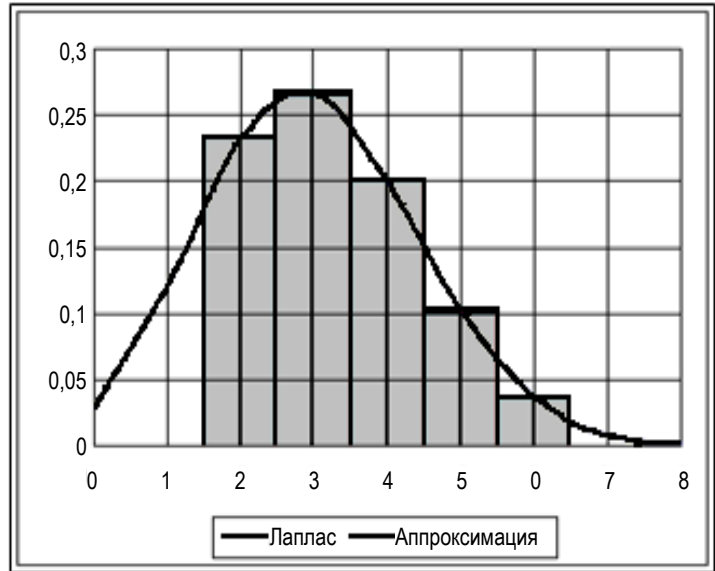


Рис. 5.4. Аппроксимация распределения Лапласа столбиковой фигурой

интегральной суммы интегралом, пределы которого несколько шире, чем в общепринятой формуле:

$$P(m_1 \leq m \leq m_2) \approx \int_{m_1-0,5}^{m_2+0,5} P(m) dm = \int_{t_{m_1-0,5}}^{t_{m_2+0,5}} \varphi(t) \cdot dt = \Phi(t_{m_2+0,5}) - \Phi(t_{m_1-0,5}).$$

Так, для данного примера $n = 10$, $p = 0,3$ точное значение $P(2 \leq m \leq 6)$, рассчитанное по формулам Бернулли, равно $\sum P(m) = 0,840$; по стандартной формуле Лапласа — $\Phi(t_6) - \Phi(t_2) = 0,736$ (погрешность 12,4 %); по уточненной формуле — $\Phi(t_{6,5}) - \Phi(t_{1,5}) = 0,842$ (погрешность всего 0,2 %).

Доказательство локальной теоремы Лапласа

При преобразовании формулы Бернулли $P_n(m) = \frac{n!}{m!(n-m)!} p^m q^{n-m}$ для больших значений $n > 30$ использована асимптотическая формула Стирлинга $n! \approx \sqrt{2\pi n} \cdot \left(\frac{n}{e}\right)^n$, которая дает тем более точные значения, чем больше n . Но даже для сравнительно небольших значений n эта формула дает уже достаточно хорошие приближения факториалов.

Например, для $n = 3$ точное значение $3! = 6$, а по формуле Стирлинга получается 5,84 (погрешность 2,7 %); для $n = 6$ точное значение $6! = 720$, а по формуле Стирлинга – 710,1 (погрешность 1,4 %); для $n = 12$ точное значение $12! = 479\,001\,600$, а по формуле Стирлинга – 475 691 882 (погрешность 0,7 %).

Запишем $m = np + t\sqrt{npq} = np\left(1 + t\sqrt{\frac{q}{np}}\right)$, тогда $(n - m) = nq - t\sqrt{npq} =$
 шем $m = np + t\sqrt{npq} = np\left(1 + t\sqrt{\frac{q}{np}}\right)$, тогда $(n - m) = nq - t\sqrt{npq} = nq\left(1 - t\sqrt{\frac{p}{nq}}\right)$.

Преобразуем формулу Бернулли для $n > 1$ ($np, nq > 5$):

$$\begin{aligned} P_n(m) &\approx \frac{\sqrt{2\pi n}}{\sqrt{2\pi m}\sqrt{2\pi(n-m)}} \cdot \left(\frac{n}{e}\right)^n \cdot \left(\frac{e}{m}\right)^m \cdot \left(\frac{e}{n-m}\right)^{n-m} \cdot p^m \cdot q^{n-m} = \\ &= \frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{n}{np\left(1 + t\sqrt{\frac{q}{np}}\right)nq\left(1 - t\sqrt{\frac{p}{nq}}\right)}} \cdot \left(\frac{np}{m}\right)^m \cdot \left(\frac{nq}{n-m}\right)^{n-m} \rightarrow \\ &\rightarrow \frac{1}{\sqrt{2\pi}\sqrt{npq}} \cdot \left(1 + t\sqrt{\frac{q}{np}}\right)^{-m} \cdot \left(1 - t\sqrt{\frac{p}{nq}}\right)^{-(n-m)} = \\ &= \frac{1}{\sqrt{2\pi}\sqrt{npq}} \cdot \exp\left\{-m \cdot \ln\left(1 + t\sqrt{\frac{q}{np}}\right)\right\} \cdot \exp\left\{-(n-m) \cdot \ln\left(1 - t\sqrt{\frac{p}{nq}}\right)\right\} = \\ &= \frac{1}{\sqrt{2\pi}\sqrt{npq}} \cdot \exp\left\{-np\left(1 + t\sqrt{\frac{q}{np}}\right)\ln\left(1 + t\sqrt{\frac{q}{np}}\right) - nq\left(1 - t\sqrt{\frac{p}{nq}}\right)\ln\left(1 - t\sqrt{\frac{p}{nq}}\right)\right\}. \end{aligned}$$

Преобразуем выражение под знаком экспоненты, используя известное разложение логарифма $\ln(1 + \gamma) = \gamma - \frac{\gamma^2}{2} + O(\gamma^3)$:

$$\begin{aligned} &-np\left(1 + t\sqrt{\frac{q}{np}}\right)\ln\left(1 + t\sqrt{\frac{q}{np}}\right) - nq\left(1 - t\sqrt{\frac{p}{nq}}\right)\ln\left(1 - t\sqrt{\frac{p}{nq}}\right) = \\ &= -np\left(1 + t\sqrt{\frac{q}{np}}\right)\left(t\sqrt{\frac{q}{np}} - \frac{1}{2}\left(t\sqrt{\frac{q}{np}}\right)^2 + \dots\right) - nq\left(1 - t\sqrt{\frac{p}{nq}}\right)\left(-t\sqrt{\frac{p}{nq}} - \frac{1}{2}\left(t\sqrt{\frac{p}{nq}}\right)^2 - \dots\right) = \\ &= \left\{-\frac{t^2}{2} + \dots\right\} \rightarrow -\frac{t^2}{2}. \end{aligned}$$

Окончательно получаем:

$$P_n(m) = \frac{n!}{m!(n-m)!} p^m q^{n-m} \rightarrow \frac{1}{\sqrt{2\pi}\sqrt{npq}} \exp\left\{-\frac{t^2}{2}\right\} = \frac{\varphi(t)}{\sigma_m},$$

$$\text{где } t = \frac{m - np}{\sqrt{npq}} = \frac{m - a}{\sigma_m}, \quad \varphi(t) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\}.$$

Вопросы для самопроверки

1. Сформулируйте локальную теорему Лапласа, укажите область применения распределения Лапласа.
2. Перечислите особенности дифференциальной функции Лапласа.
3. Перечислите особенности интегральной функции Лапласа.
4. Сформулируйте интегральную теорему Лапласа.
5. Приведите вариант интегральной теоремы Лапласа для симметричных отклонений случайной величины от своего центра.
6. Сформулируйте третью форму интегральной теоремы Лапласа.
7. Перечислите задачи, которые решаются с помощью третьей формы интегральной теоремы Лапласа.
8. Как определить потребное число испытаний, чтобы получать точные и надежные прогнозы?

6. Непрерывная случайная величина

Значения непрерывной случайной величины сплошь заполняют некоторый интервал, поэтому их невозможно перечислить и задать такую случайную величину рядом распределения.

Универсальным способом задания случайной величины является **функция распределения** (или **интегральная функция распределения**): $F(x) = P(\mathcal{X} \leq x)$ – вероятность того, что случайная величина \mathcal{X} примет значение, не меньшее заданного x . Для дискретной случайной величины эта функция называется кумулятой и представляет собой функцию накопленных вероятностей $F(x_j) = p_1 + p_2 + \dots + p_j$.

Из определения функции $F(x)$ следует, что $0 \leq F(x) \leq 1$ (так как $F(x)$ – вероятность); $F(-\infty) = P(\mathcal{X} \leq -\infty) = 0$ (невозможное событие); $F(\infty) = P(\mathcal{X} \leq \infty) = 1$ (достоверное событие). Покажем, что функция распределения неубывающая. Действительно, из рис. 6.1 следует, $F(x_2) = P(\mathcal{X} \leq x_2) = P(\mathcal{X} \leq x_1) + P(x_1 < \mathcal{X} \leq x_2) = F(x_1) + P(x_1 < \mathcal{X} \leq x_2) \geq F(x_1)$ для $x_2 > x_1$ (иными словами, большим значениям аргумента соответствуют не меньшие значения функции распределения).

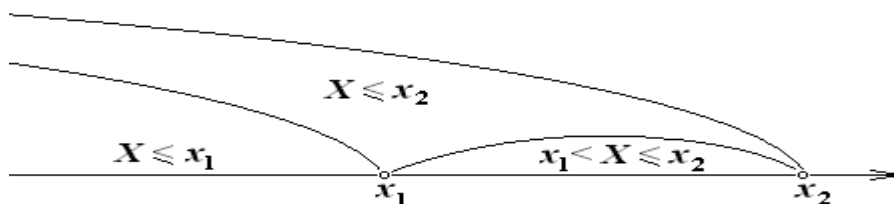


Рис. 6.1. Интервалы изменения переменной x

Фактически в общем виде доказали **интегральную теорему**:

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1):$$

вероятность попадания случайной величины в полуоткрытый интервал $x_1 < X \leq x_2$ равна разности значений функции распределения на краях этого интервала (или она равна приращению функции распределения на этом интервале).

Для дискретной случайной величины график функции распределения ступенчатый со скачками – конечными приращениями p_i (разрывами 1-го рода) в значениях x_i дискретной случайной величины. Для непрерывной случайной величины функция распределения непрерывная, ее график не имеет разрывов. Для непрерывной функции малым приращениям аргумента соответствуют малые приращения функции: $\lim_{\Delta x \rightarrow 0} \frac{\Delta F}{\Delta x} = 0$.

Отсюда следует, что для непрерывной случайной величины вероятность появления любого ее конкретного значения равна нулю:

$$P(X = x) = \lim_{\Delta x \rightarrow 0} \frac{P(x < X \leq x + \Delta x)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{F(x < X \leq x + \Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{\Delta F(x)}{\Delta x} = 0.$$

Поэтому для непрерывной случайной величины эквивалентны следующие записи интегральной теоремы:

$$P(x_1 < X \leq x_2) = P(x_1 \leq X < x_2) = P(x_1 < X < x_2) = P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1).$$

Кроме функции распределения $F(x)$, для описания непрерывной случайной величины вводят еще одну функцию – плотность вероятности – отношение вероятности попадания случайной величины в заданный интервал к длине этого интервала. В определении этой новой функции для конкретного значения x используется предельный переход:

$$f(x) = \lim_{\Delta x \rightarrow 0} \frac{P(x \leq X \leq x + \Delta x)}{\Delta x}.$$

Преобразуем это выражение:

$$f(x) = \lim_{\Delta x \rightarrow 0} \frac{P(x \leq X \leq x + \Delta x)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{\Delta F(x)}{\Delta x} = \frac{dF}{dx} = F'.$$

Функция плотности вероятности $f(x)$ оказалась равной производной от функции распределения $F(x)$, поэтому функцию $f(x)$ называют также **дифференциальной функцией распределения**.

Наоборот, функция распределения $F(x)$ выражается как интеграл от функции плотности вероятности:

$$F(x) = \int_{-\infty}^x f(s) ds,$$

поэтому функцию $F(x)$ называют также **интегральной функцией распределения**.

Замечания: в интегральном представлении функция распределения является функцией верхнего предела определенного интеграла, поэтому переменная интегрирования обозначена другой литерой. Нижний предел принят $-\infty$, чтобы не требовалось добавлять постоянную интегрирования.

Ранее была сформулирована интегральная теорема Лапласа, согласно которой (для конкретного распределения Лапласа) получено выражение:

$$P(m_1 \leq m \leq m_2) = \Phi(t_{m_2}) - \Phi(t_{m_1}),$$

которое очень напоминает общую интегральную теорему. Вопрос: является ли интегральная функция Лапласа $\Phi(t)$ интегральной функцией распределения? Ответ – нет, не является, поскольку значения $\Phi(t)$ могут быть отрицательными $\Phi(-t) = -\Phi(t)$. С помощью интегральной теоремы Лапласа можно найти функцию распределения $F(m) = P(\mathcal{X} \leq m) = P(-\infty < \mathcal{X} \leq m) = \Phi(t_m) - \Phi(-\infty) = \Phi(t_m) + \Phi(\infty) = \Phi(t_m) + 0,5$, то есть интегральная функция **распределения** Лапласа и интегральная функция Лапласа отличаются постоянным слагаемым 0,5.

Это объясняется тем, что в определении интегральной функции Лапласа нижний предел интегрирования был принят равным нулю, а не $-\infty$:

$$\Phi(t) = \int_0^t \varphi(s) ds = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-\frac{s^2}{2}} ds.$$

это было для того, чтобы можно было воспользоваться симметрией – нечетностью функции $\Phi(t)$.

Функция плотности вероятности неотрицательная $f(x) \geq 0$ как производная от неубывающей функции распределения $F(x)$. Площадь под дифференциальной кривой равна единице:

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

Это свойство используется при решении задач, так как обычно дифференциальная функция задается с точностью до постоянного множителя, который подлежит определению.

Площадь под частью дифференциальной кривой на интервале (x_1, x_2) равна вероятности попадания случайной величины в этот интервал (рис. 6.2):

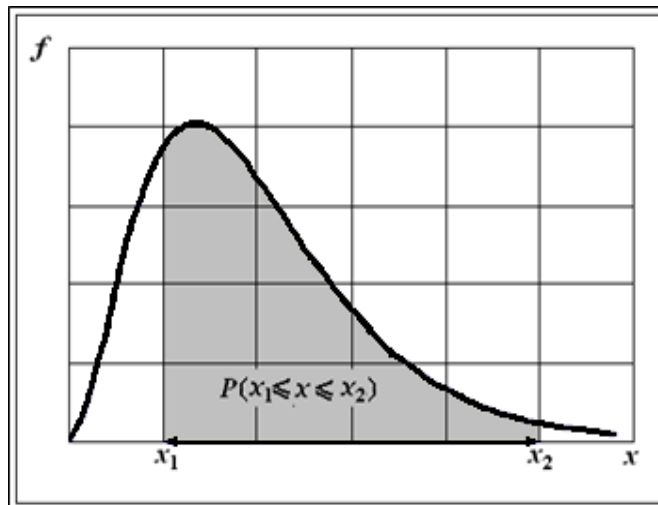


Рис. 6.2. Вероятность попадания случайной величины в интервал

$$\int_{x_1}^{x_2} f(s)ds = \int_{-\infty}^{x_2} f(s)ds - \int_{-\infty}^{x_1} f(s)ds = F(x_2) - F(x_1) = P(x_1 \leq x \leq x_2).$$

Пример. Равномерный закон распределения. По закону равномерной плотности распределены ошибки округления чисел, время ожидания транспорта, который ходит через равные интервалы времени, место остановки тела под воздействием сухого трения.

Пусть случайная величина распределена на интервале $[a, b]$ с постоянной плотностью вероятности:

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

Требуется найти константу C и составить интегральную функцию распределения $F(x)$.

Фигура под дифференциальной кривой равномерного распределения представляет собой прямоугольник с основанием $(b - a)$ и высотой C . Его площадь равна $C \cdot (b - a) = 1$, откуда $C = \frac{1}{b-a}$.

Дифференциальная функция задана тремя разными выражениями на трех интервалах, поэтому рассмотрим выражения интегральной функции на этих же интервалах.

При $x < a$ $f(x) = 0$ и $F(x) = \int_{-\infty}^x 0 ds = 0$.

При $a < x \leq b$ $f(x) = C$ и $F(x) = \int_{-\infty}^a 0 ds + \int_a^x \frac{ds}{b-a} = \frac{x-a}{b-a}$.

При $x > b$ $f(x) = 0$ и $F(x) = \int_{-\infty}^a 0 ds + \int_a^b \frac{ds}{b-a} + \int_b^x 0 ds = \frac{b-a}{b-a} = 1$.

Запишем полученные формулы в компактном виде:

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

Графики дифференциальной и интегральной функций равномерного распределения изображены на рис. 6.3.

Найдем вероятность попадания случайной величины в интервал с границами (x_1, x_2) , где $x_1 = 0,7 \cdot a + 0,3 \cdot b$, $x_2 = 0,3 \cdot a + 0,7 \cdot b$. Так как оба значения $(x_1, x_2) \in (a, b)$, то

$$P(x_1 \leq x \leq x_2) = F(x_2) - F(x_1) = \frac{(0,3a+0,7b)-a}{b-a} - \frac{(0,7a+0,3b)-a}{b-a} = 0,4.$$

Чтобы найти математическое ожидание, заметим, что произведение $f(x)\Delta x$ означает вероятность попадания случайной величины в малый интервал шириной Δx в окрестности точки x . Тогда получаем расчетную формулу математического ожидания для непрерывной случайной величины:

$$M(x) = \lim_{\Delta x \rightarrow 0} \sum x \cdot f(x) \Delta x = \int_{-\infty}^{\infty} x \cdot f(x) dx.$$

Так же рассчитываются другие моменты распределения, например математическое ожидание квадрата (момент 2-го порядка) $M(x^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x) dx$.

Дисперсию вычисляем обычным образом:

$$D(x) = M(x^2) - M^2(x).$$

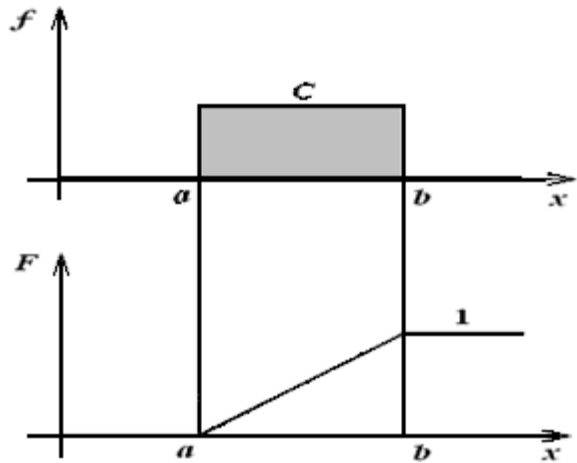


Рис. 6.3. Равномерный закон

Пример. Вычислим характеристики равномерного распределения.

$$M(x) = \int_{-\infty}^{\infty} x \cdot f(x) dx = \int_a^b x \cdot \frac{dx}{b-a} = \frac{1}{b-a} \cdot \frac{x^2}{2} \Big|_a^b = \frac{1}{2} \cdot \frac{b^2 - a^2}{b-a} = \frac{a+b}{2}.$$

Это центр тяжести прямоугольника – фигуры плотности вероятности.

$$M(x^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x) dx = \int_a^b x^2 \cdot \frac{dx}{b-a} = \frac{1}{b-a} \cdot \frac{x^3}{3} \Big|_a^b = \frac{1}{3} \cdot \frac{b^3 - a^3}{b-a} = \frac{a^2 + ab + b^2}{3}.$$

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

Распределение симметричное $A = 0$, плосковершинное $E < 0$.

Нормальный закон распределения Гаусса

Приложения нормального закона столь обширны, что их невозможно полностью перечислить. Достаточно сказать, что по нормальному закону распределены все величины в животном и растительном мире, рассеивание попаданий при стрельбе, ошибки при изготовлении деталей. Кроме того, при некоторых условиях остальные распределения приближаются к нормальному закону. Так, при увеличении числа испытаний к нормальному закону приближаются распределения Бернулли и Пуассона.

Плотность вероятности и функция распределения нормального закона выражаются через дифференциальную и интегральную функции Лапласа:

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \cdot e^{-\frac{(x-a)^2}{2\sigma_x^2}} = \frac{\varphi(t_x)}{\sigma_x},$$

$$\text{где } t_x = \frac{x-a}{\sigma_x}, \quad \varphi(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}.$$

$$F(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \int_{-\infty}^x e^{-\frac{(s-a)^2}{2\sigma_x^2}} ds = \Phi(t_x) + 0,5,$$

$$\text{где } \Phi(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-\frac{s^2}{2}} ds.$$

Графики дифференциальной и интегральной функций нормального закона изображены на рис. 6.4.

Параметры нормального закона a и σ_x совпадают с его характеристиками: $a = M(x)$ – математическое ожидание, σ_x – стандартное отклонение. Коэффициенты асимметрии и эксцесса для нормального закона равны нулю.

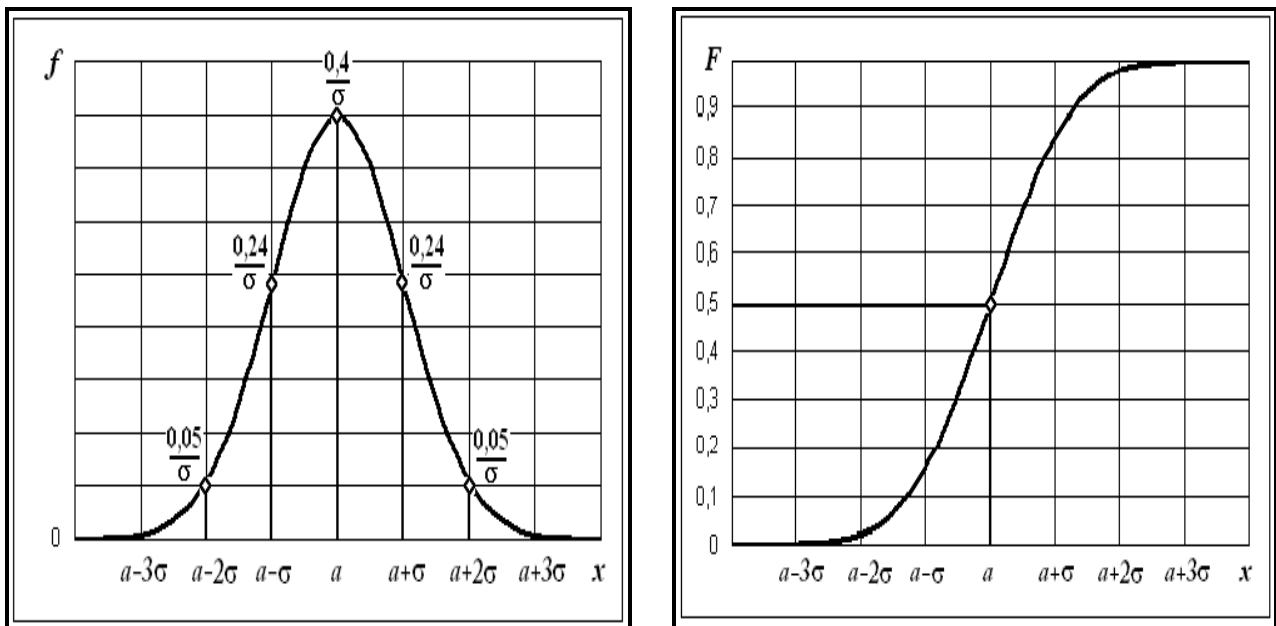


Рис. 6.4. Графики дифференциальной и интегральной функций нормального закона

На рис. 6.5 приведены графики плотности вероятности нормального распределения при различных значениях параметров.

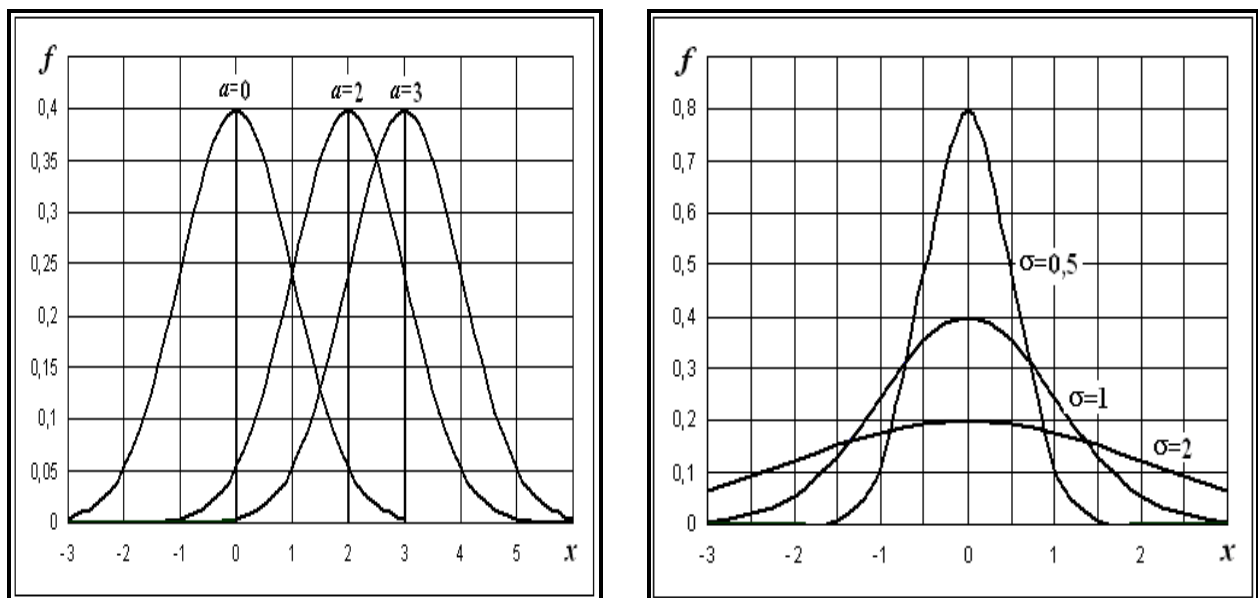


Рис. 6.5. Зависимость нормального распределения от параметров

Для того чтобы не применять стандартную фразу: «Случайная величина x распределена нормально с характеристиками a и σ_x », договорились использовать математическую запись $x \sim N(a; \sigma_x)$.

По аналогии с распределением Лапласа сформулируем три формы интегральной теоремы нормального закона:

1. Основная форма – вероятность попадания нормально распределенной величины x в заданный интервал с границами x_1, x_2 :

$$P(x_1 \leq x \leq x_2) = F(x_2) - F(x_1) = (\Phi(t_{x_2}) + 0,5) - (\Phi(t_{x_1}) + 0,5) = \Phi(t_{x_2}) - \Phi(t_{x_1})$$

2. Вторая форма предназначена для вычисления вероятности попадания случайной величины в интервал с симметричными границами:

$$P(|x - a| \leq t\sigma_x) = 2\Phi(t).$$

3. Утверждение, которое мы называем третьей формой интегральной теоремы нормального закона и приводим для полноты аналогии с тремя формами интегральной теоремы Лапласа, будет доказано немного позже:

$$P\left(|\bar{x} - a| \leq t \frac{\sigma_x}{\sqrt{n}}\right) = 2\Phi(t).$$

Фактически записана вторая форма теоремы для случайной величины $X_{cp} \sim N\left(a; \frac{\sigma_x}{\sqrt{n}}\right)$. В записи третьей формы вместо \mathcal{X}_{cp} традиционно используется \bar{x} . Это не совсем правильно, но допустимо, если понимать, о чем идет речь. Ведь \bar{x} – это одно число, константа, она не имеет изменчивости и потому не может быть как-то распределена. Но предполагается, что имеется n одинаково распределенных случайных величин \mathcal{X}_i , из них составляется среднее – тоже случайная величина – $\mathcal{X}_{cp} = (\mathcal{X}_1 + \mathcal{X}_2 + \dots + \mathcal{X}_n) / n$. Утверждается, что эта случайная величина распределена нормально (центральная предельная теорема) с указанными выше характеристиками. Значения случайной величины \mathcal{X}_{cp} могут быть сгенерированы (датчиком случайных чисел на компьютере) следующим образом: генерируются первые n значений случайной величины x , и из них составляется среднее – первое значение \mathcal{X}_{cp} ; следующие n значений x дают второе значение \mathcal{X}_{cp} и т. д. Пример: вместо того, чтобы подбрасывать одновременно 10 костей и находить среднее значение выпавших очков, можно подбросить 10 раз одну и ту же игральную кость и вычислить среднее число выпавших очков за 10 бросков.

Найдем основные характеристики \mathcal{X}_{cp} , используя свойства математического ожидания и дисперсии:

$$M(\mathcal{X}_{cp}) = M((\mathcal{X}_1 + \mathcal{X}_2 + \dots + \mathcal{X}_n) / n) = (M(\mathcal{X}_1) + M(\mathcal{X}_2) + \dots + M(\mathcal{X}_n)) / n = \\ = (a + a + \dots + a) / n = (n \cdot a) / n = a.$$

$$D(\mathcal{X}_{cp}) = D((\mathcal{X}_1 + \mathcal{X}_2 + \dots + \mathcal{X}_n) / n) = (D(\mathcal{X}_1) + D(\mathcal{X}_2) + \dots + D(\mathcal{X}_n)) / n^2 = \\ = (D_x + D_x + \dots + D_x) / n^2 = (n \cdot D_x) / n^2 = D_x / n.$$

$$\sigma_{\bar{x}} = \sqrt{D(\mathcal{X}_{cp})} = \frac{\sigma_x}{\sqrt{n}}.$$

Последнее выражение в отечественной литературе называется ошибкой среднего, а в зарубежной – стандартной ошибкой.

Как и для 3-й формы интегральной теоремы Лапласа, здесь удобно пользоваться терминологией: $P = 2\Phi(t)$ – уровень доверия утверждения о том, что отклонение среднего от математического ожидания $|\bar{x} - a| \leq \varepsilon$ не превысит по-

грешности $\varepsilon = t \frac{\sigma_x}{\sqrt{n}}$.

В связи с этим появляются три типовых задачи.

I. Известны параметры распределения a и σ_x . Дополнительно заданы n и погрешность ε . Найти уровень доверия P .

Решение. Из выражения $\varepsilon = t \frac{\sigma_x}{\sqrt{n}}$ при известных σ_x , n , ε находим t и далее вычисляем $P = 2\Phi(t)$.

II. Известны параметры распределения a и σ_x . Дополнительно заданы n и уровень доверия P . Найти погрешность ε .

Решение. Из $P = 2\Phi(t)$ по таблицам Лапласа находим t . Далее вычисляем погрешность по формуле $\varepsilon = t \frac{\sigma_x}{\sqrt{n}}$.

III. Известны параметры распределения a и σ_x . Дополнительно заданы уровень доверия P и погрешность ε . Найти нужное число n .

Решение. Из $P = 2\Phi(t)$ находим t . Далее из формулы $\varepsilon = t \frac{\sigma_x}{\sqrt{n}}$ при уже известных σ_x , t , ε определяем $n \geq \left(\frac{t \sigma_x}{\varepsilon} \right)^2$.

При изучении 3-й формы интегральной теоремы Лапласа была рассмотрена еще одна задача (IV), которая формулируется так: «Известны результаты опыта. Что можно сказать о теоретических характеристиках распределения?»

Подобные задачи составляют предмет математической статистики и будут рассмотрены в соответствующих разделах курса.

Показательный, или экспоненциальный, закон распределения

По этому закону распределено время работы оборудования до первого отказа. Его дифференциальная функция с точностью до постоянного множителя выражается формулой: $f(t) = ke^{-\lambda t}$ – для $t \geq 0$; для $t < 0$ $f(t) = 0$.

Сомножитель k находим из условия – площадь под дифференциальной кривой равна единице:

$$\int_0^{\infty} f(t) dt = k \int_0^{\infty} e^{-\lambda t} dt = -\frac{k}{\lambda} e^{-\lambda t} \Big|_0^{\infty} = \frac{k}{\lambda} (e^0 - e^{-\infty}) = \frac{k}{\lambda} = 1.$$

Отсюда $k = \lambda$.

Найдем интегральную функцию показательного распределения:

для $t < 0$ $F(t) = 0$;

$$\text{для } t \geq 0 \quad F(t) = \int_{-\infty}^0 0 ds + \int_0^t \lambda e^{-\lambda s} ds = \int_0^{\lambda t} e^{-x} dx = -e^{-x} \Big|_0^{\lambda t} = 1 - e^{-\lambda t}.$$

При вычислении интеграла была сделана замена переменной $x = \lambda s$, $dx = \lambda ds$.

Вычисляем характеристики показательного распределения:

$$M(t) = \int_0^{\infty} t \cdot f(t) dt = \int_0^{\infty} t \cdot \lambda e^{-\lambda t} dt = \frac{1}{\lambda} \int_0^{\infty} x e^{-x} dx = \frac{1}{\lambda} \left[-x e^{-x} \Big|_0^{\infty} + \int_0^{\infty} e^{-x} dx \right] = \frac{1}{\lambda} (0 + 1) = \frac{1}{\lambda}.$$

Была сделана замена переменной $x = \lambda t$ и применено правило интегрирования по частям; внеинтегральный член оказался равным нулю.

Математическое ожидание – это среднее значение случайной величины, центр тяжести фигуры под дифференциальной кривой.

Для вычисления дисперсии предварительно найдем $M(t^2)$:

$$M(t^2) = \int_0^{\infty} t^2 \cdot f(t) dt = \int_0^{\infty} t^2 \cdot \lambda e^{-\lambda t} dt = \frac{1}{\lambda^2} \int_0^{\infty} x^2 e^{-x} dx = \frac{1}{\lambda^2} \left[-x^2 e^{-x} \Big|_0^{\infty} + 2 \int_0^{\infty} x e^{-x} dx \right] = \frac{1}{\lambda^2} (0 + 2) = \frac{2}{\lambda^2}$$

Тогда

$$\sigma_t^2 = M(t^2) - (M(t))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}; \quad \sigma_t = \frac{1}{\lambda}; \quad v_t = \frac{\sigma_t}{M(t)} \cdot 100\% = 100\%.$$

Наивероятнейшее значение (мода) для показательного распределения равно нулю, то есть $M_0 = 0$ (чаще всего оборудование выходит из строя в момент включения).

Интегральная теорема для показательного закона выглядит так:

$$P(t_1 \leq t \leq t_2) = F(t_2) - F(t_1) = (1 - e^{-\lambda t_2}) - (1 - e^{-\lambda t_1}) = e^{-\lambda t_1} - e^{-\lambda t_2}.$$

Обозначим через $T = M(t) = 1/\lambda$ среднее время работы оборудования до первого отказа. Тогда:

$$P(0 \leq t \leq T) = e^0 - e^{-\lambda T} = 1 - e^{-1} = 1 - 0,368 = 0,632;$$

$$P(T \leq t \leq 2T) = e^{-\lambda T} - e^{-2\lambda T} = e^{-1} - e^{-2} = 0,368 - 0,135 = 0,233;$$

$$P(2T \leq t \leq 3T) = e^{-2\lambda T} - e^{-3\lambda T} = e^{-2} - e^{-3} = 0,135 - 0,050 = 0,085;$$

$$P(t > 3T) = 1 - P(t \leq 3T) = 1 - F(3T) = 1 - (1 - e^{-3\lambda T}) = e^{-3\lambda T} = e^{-3} = 0,050.$$

Здесь учтено, что $\lambda T = \lambda / \lambda = 1$.

Графики функций этого распределения приведены на рис. 6.6.

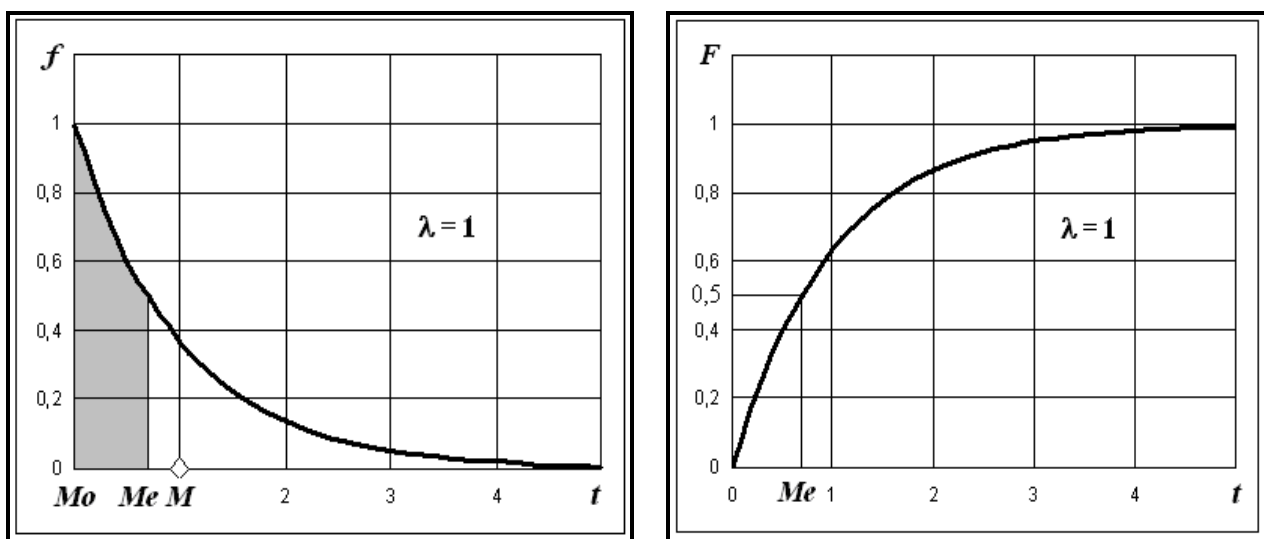


Рис. 6.6. Функции распределения показательного закона для $\lambda = 1$

Квантили распределения

Для непрерывной случайной величины в качестве дополнительных характеристик используют так называемые **квантили**, к которым относятся медиана, квартили, децили и процентиля.

Квантили делят фигуру под дифференциальной кривой на равновеликие части, или же они делят интервал изменения (варьирования) случайной величины на равновероятные части.

Медиана делит интервал варьирования на две части с вероятностью 50 % попадания случайной величины в каждую часть.

Квартили делят интервал варьирования на четыре части с вероятностью 25 % попадания случайной величины в каждую часть.

Децили делят интервал варьирования на десять частей с вероятностью 10 % попадания случайной величины в каждую часть.

Проценти делят интервал варьирования на сто частей с вероятностью 1 % попадания случайной величины в каждую часть.

Обозначения квантилей x_α , где α – вероятность того, что случайная величина примет значение, большее квантиля x_α : $P(x > x_\alpha) = \alpha$ (где α – это вероятность того, что случайная величина примет значение, большее квантиля x_α ; площадь фигуры плотности вероятности *справа* от квантиля). Выражение для вероятности противоположного события $P(x \leq x_\alpha) = F(x_\alpha) = 1 - \alpha$ приводит к вычислительной формуле $F(x_\alpha) = 1 - \alpha$. Так, для показательного закона медиану $Me = t_{0,5}$ находим из равенства $F(t_{0,5}) = 1 - 0,5 = 0,5$ (см. рис. 6.6). Это значение оказалось равным $Me = \ln 2$. Площадь фигуры под дифференциальной кривой справа от медианы равна 0,5.

Последовательные квартили (нижняя, средняя и верхняя квартили) обозначаются $x_{0,75} < x_{0,50} < x_{0,25}$.

Некоторые соображения, приводящие к нормальному закону

Рассмотрим простой пример из теории стрельбы. Пусть x – отклонения попаданий от точки прицеливания. Замечено, что вероятность этих отклонений зависит только от величины отклонения, но не от его знака. Учтем этот факт в записи:

$$\begin{aligned} f(x) &= \phi(x^2); \\ f(x, y) &= \phi(x^2 + y^2). \end{aligned}$$

Далее учтем факт независимости отклонений по оси x и по оси y .

Тогда из теоремы умножения вероятностей для независимых событий следует:

$$\begin{aligned} f(x, y) dx dy &= f(x) dx \cdot f(y) dy, \\ \phi(x^2 + y^2) &= \phi(x^2) \phi(y^2). \end{aligned}$$

Полученное функциональное уравнение имеет единственное решение:

$$f(x) = \phi(x^2) = A \cdot \exp\{kx^2\},$$

где коэффициент $k < 0$, так как большие отклонения имеют меньшую вероятность появления.

Обозначим этот коэффициент как $k = -\frac{1}{2h^2}$.

Множитель A находим из условия равенства единице площади под дифференциальной кривой: $A = \frac{1}{\sqrt{2\pi} \cdot h}$. Здесь используется интеграл Пуассона:

$$\int_0^{\infty} e^{-\frac{t^2}{2}} dt = \sqrt{\frac{\pi}{2}}.$$

Вычисляем дисперсию отклонений (при этом учтем, что $M(x) = 0$):

$$\sigma_x^2 = \int_{-\infty}^{\infty} x^2 f(x) dx = \frac{2}{\sqrt{2\pi} \cdot h} \int_0^{\infty} x^2 e^{-\frac{x^2}{2h^2}} dx = h^2.$$

Этот интеграл вычисляется по правилу «интегрирования по частям».

Окончательно получаем:

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \exp\left\{-\frac{x^2}{2\sigma_x^2}\right\},$$

то есть отклонения попаданий от точки прицеливания действительно распределены нормально.

Вопросы для самопроверки

1. Что такое функция распределения? Каковы ее свойства и график? Как она называется для дискретной случайной величины?
2. Сформулируйте общую интегральную теорему, приведите ее вариант для непрерывной случайной величины.
3. Что такое функция плотности вероятности, каковы ее свойства и график?
4. Как связаны между собой функция распределения и функция плотности вероятности?
5. Как вычисляются основные характеристики непрерывной случайной величины?
6. Сформулируйте закон равномерного распределения, опишите область его применения, приведите выражения для дифференциальной и интегральной функций.
7. Опишите характеристики равномерного закона. Выполняется ли для него правило «3-х сигм»?
8. Сформулируйте показательный закон распределения, опишите область его применения, приведите выражения для дифференциальной и интегральной функций. Опишите характеристики показательного закона.

9. Что такое нормальный закон распределения Гаусса, каковы его характерные особенности?
10. Какие отличия имеются между нормальным распределением и распределением Лапласа?
11. Каковы параметры и характеристики нормального закона?
12. Сформулируйте три формы интегральной теоремы нормального закона.
13. Какие задачи решаются с помощью 3-й формы интегральной теоремы нормального закона?
14. Что такое квантили? Перечислите их разновидности.
15. Что такое медиана и как она рассчитывается?
16. Что такое квартили и как они рассчитываются?

7. Предельные теоремы теории вероятностей

К предельным теоремам теории вероятностей относят *закон больших чисел* и *центральную предельную теорему*. Строго говоря, предельными являются также асимптотические формулы Пуассона и Лапласа, но они формально не относятся к упомянутому выше разделу теории вероятностей.

В доказательстве закона больших чисел и центральной предельной теоремы принимали участие русские математики Чебышев П. Л. и Ляпунов А. М., и, может быть, поэтому в отечественной научной литературе доказанным теоремам придается слишком большое, чуть ли не мистическое, значение. В действительности же доказывались очевидные вещи, в справедливости которых никто не сомневался. Здесь имеет место принцип: математики всегда доказывают все утверждения в отличие от представителей инженерных наук, которые никогда не доказывают утверждения, в справедливости которых уже убедились на основе опытов или правдоподобных рассуждений.

Закон больших чисел

1. *Теорема Бернулли*. Доказывается, что при увеличении числа однородных независимых испытаний относительная частота появления события A стремится к ее вероятности.

Это предположение изначально лежит в основе универсального метода статистического определения вероятностей:

$$\lim_{n \rightarrow \infty} \frac{m(n)}{n} = p.$$

Для строгого доказательства рассмотрим 3-ю форму интегральной теоремы Лапласа:

$$P\left(\left|\frac{m}{n} - p\right| \leq \varepsilon\right) = 2\Phi(t), \quad \varepsilon = t\sqrt{\frac{pq}{n}}.$$

При $t = 3$ уровень доверия (вероятность выполнения этого условия, гарантия) уже равен $P = 0,997$. Если потребуется еще больший уровень доверия, можно принять $t = 4, 5$ и даже 10 . Таким образом, с гарантией, сколь угодно близкой к единице, можно утверждать, что отклонение относительной частоты $\frac{m}{n}$ от своего предельного значения p не превысит погрешности ε , которая

стремится к нулю с увеличением n : $\left|\frac{m}{n} - p\right| \leq 10\sqrt{\frac{pq}{n}} \xrightarrow{(n \rightarrow \infty)} 0, \quad \frac{m}{n} \xrightarrow{(n \rightarrow \infty)} p.$

2. Теорема Чебышева. Доказывается, что с увеличением повторных измерений среднее значение величины приближается к своему математическому ожиданию.

В справедливости этого утверждения также никто не сомневался.

Действительно, пусть случайная величина \mathcal{X} наблюдается n раз; данные наблюдений сгруппированы (рис. 7.1), так что известны частоты m_i появления каждого возможного значения x_i , где $(m_1 + m_2 + m_3 + \dots + m_k) = n$. Тогда среднее значение случайной величины по этим n наблюдениям будет равно:

\mathcal{X}	x_1	x_1	x_1	\dots	x_k
m	m_1	m_2	m_3	\dots	m_k

Рис. 7.1. Вариационный ряд

$$\bar{x} = \frac{m_1 x_1 + m_2 x_2 + m_3 x_3 + \dots + m_k x_k}{n} = \sum x_i \cdot \frac{m_i}{n} \xrightarrow{(n \rightarrow \infty)} \sum x_i p_i = M(x).$$

Для строгого доказательства рассмотрим неравенство Чебышева, которое ранее использовали для обоснования правила «3-х сигм», но сейчас запишем это неравенство для случайного среднего \mathcal{X}_{cp} :

$$P\left(|\bar{x} - a| \leq t \frac{\sigma_x}{\sqrt{n}}\right) > 1 - \frac{1}{t^2}.$$

При $t = 10$ гарантия этого утверждения будет больше $0,99$; если требуется еще больший уровень доверия, принимаем еще большее значение t .

Таким образом, с гарантией, сколь угодно близкой к единице, можно утверждать, что отклонение среднего от математического ожидания не превысит погрешности, которая стремится к нулю при увеличении n :

$$|\bar{x} - a| \leq t \frac{\sigma_x}{\sqrt{n}} \xrightarrow{(n \rightarrow \infty)} 0, \text{ откуда } \bar{x} \xrightarrow{(n \rightarrow \infty)} a = M(x).$$

Центральная предельная теорема

Давно было замечено, что с увеличением числа слагаемых распределение суммы случайных величин приближается к нормальному.

Так, на рис. 7.2а изображены распределения суммы одного, двух и трех слагаемых ($m = 1, 2, 3$), каждое из которых имеет ошибку округления до целых значений. Известно, что ошибки округления распределены по равномерному закону на интервале $(0; 1)$. Оказывается, сумма двух таких слагаемых имеет ошибку, которая распределена по треугольному закону Симпсона. Распределение ошибки суммы трех слагаемых очень похоже на нормальное. Здесь исходное распределение было симметричным, поэтому получилась быстрая сходимость к предельному нормальному виду.

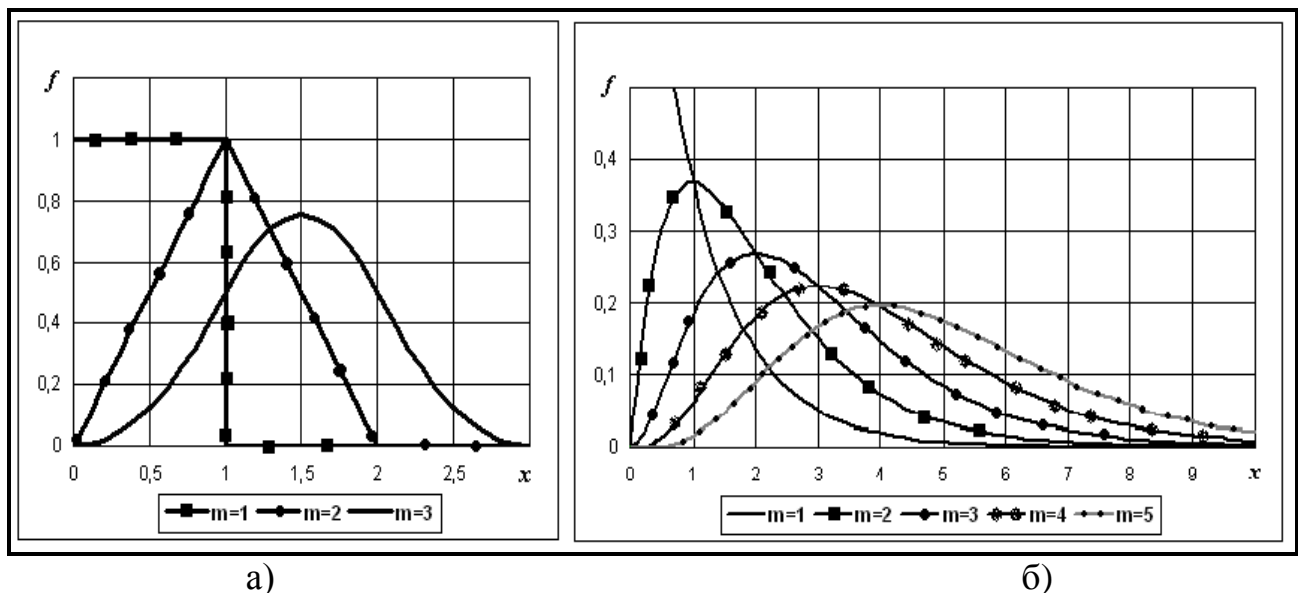


Рис. 7.2. Распределения сумм случайных величин

На рис. 7.2б изображены распределения времени работы сложного устройства с учетом резервирования – при выходе элемента из строя он заменяется на запасной (наладчиком или автоматически). Известно, что время работы любого оборудования до первого отказа распределено по экспоненциальному закону. Оказывается, что при резервировании суммарное время работы устройства распределено по законам Эрланга (частный случай гамма-распределения). Распределение времени работы каждого элемента далеко от симметричного, но чем больше в устройстве резервных элементов, тем ближе распределение суммарного времени его работы к нормальному виду. Темп сходимости к предельному симметричному распределению здесь не такой высокий по сравнению с

рис. 7.2а, но тем не менее считается, что при самых неблагоприятных условиях сумма 10 случайных величин распределена практически нормально.

Распределение Бернулли – это распределение суммы $m = \mathcal{X}_1 + \mathcal{X}_1 + \dots + \mathcal{X}_n$, где \mathcal{X}_i – число успехов в одном испытании (0 или 1). Известно, что при увеличении числа испытаний распределение Бернулли быстро приближается к распределению Лапласа (частный случай нормального распределения).

Оценим теперь с точки зрения математика-прикладника теорему, доказанную Ляпуновым А. М., которая утверждает, что всегда, когда случайная величина является суммой большого количества независимых случайных величин, дисперсия которых мала по сравнению с дисперсией их суммы, эта случайная величина распределена по закону, который приближается к нормальному; данный эффект проявляется, когда наблюдаемые явления определяются влиянием большого количества независимых случайных факторов, вклад каждого из которых в общий процесс очень малый.

У любого человека (неизучающего математику) после прочтения текста этой теоремы возникают вопросы. Что такое «большое количество»? Десять – это «большое количество» или еще не очень? Что такое «малая дисперсия»? Насколько она должна быть меньше общей дисперсии? Что значит: закон «приближается» к нормальному? Как быстро? Что значит «малый вклад»? Вообще, что нового мы узнали по сравнению с уже известными фактами?

На основе центральной предельной теоремы было сформулировано утверждение, которое мы назвали 3-й формой интегральной теоремы нормального закона:

$$P\left(|\bar{x} - a| \leq t \frac{\sigma_x}{\sqrt{n}}\right) = 2\Phi(t).$$

Уточняем: это утверждение справедливо при повторении однородных независимых испытаний не менее $n = 10$ раз; распределение исходной случайной величины x может быть каким угодно.

Кстати, в центральной предельной теореме вовсе не предполагается, что случайные слагаемые должны иметь одинаковое распределение.

Композиция распределений случайных величин

Распределение суммы независимых случайных величин называется композицией распределений.

Сначала рассмотрим композицию распределений двух дискретных независимых случайных величин \mathcal{X} , \mathcal{Y} . Дискретная величина \mathcal{X} определена на

дискретном множестве значений x , которые появляются с вероятностями $P_x(x)$; если же значение x не принадлежит заданному дискретному множеству, будем считать его вероятность равной нулю. Аналогично условимся задавать закон распределения дискретной величины \mathcal{Y} – если y не принадлежит заданному дискретному множеству значений, то его вероятность принимаем равной нулю, а если принадлежит, то $P_y(y)$. Вводим новую случайную величину $\mathcal{Z} = \mathcal{X} + \mathcal{Y}$. Вероятность совместного появления конкретных слагаемых x, y вычисляется по теореме умножения $P_x(x) \cdot P_y(y)$. Сумма $z = x + y$ может появиться не одним способом; согласно аксиоме сложения, вероятности всех таких комбинаций x, y надо сложить:

$$P_z(z) = \sum_{x+y=z} P_x(x)P_y(y).$$

Эту формулу можно переписать в виде:

$$P_z(z) = \sum_x P_x(x)P_y(z-x) \quad \text{или} \quad P_z(z) = \sum_y P_x(z-y)P_y(y).$$

Для непрерывных случайных величин можно получить очень похожие формулы, где дискретная сумма заменяется на интеграл:

$$f_z(z) = \int_{-\infty}^{\infty} f_x(x)f_y(z-x)dx \quad \text{или} \quad f_z(z) = \int_{-\infty}^{\infty} f_x(z-y)f_y(y)dy.$$

Эти интегралы иногда называются сверткой.

Пример 1. Найдем закон распределения двух случайных величин, каждая из которых распределена по показательному закону $f_1(x) = \lambda e^{-\lambda x}$, $x \geq 0$:

$$f_2(x) = \int_{-\infty}^{\infty} f_1(y) \cdot f_1(x-y)dy = \int_0^x \lambda e^{-\lambda y} \cdot \lambda e^{-\lambda(x-y)}dy = \lambda e^{-\lambda x} \int_0^x \lambda dy = \lambda(\lambda x) e^{-\lambda x}.$$

Здесь бесконечные пределы интегрирования заменены на 0 и x , так как подынтегральная функция отлична от нуля только при $y \geq 0$ и $(x-y) \geq 0$.

Теперь найдем закон распределения трех слагаемых, каждое из которых распределено по показательному закону:

$$\begin{aligned} f_3(x) &= \int_{-\infty}^{\infty} f_2(y) \cdot f_1(x-y)dy = \int_0^x \lambda(\lambda y) e^{-\lambda y} \cdot \lambda e^{-\lambda(x-y)}dy = \\ &= \lambda e^{-\lambda x} \int_0^x (\lambda y) d\lambda y = \lambda \frac{(\lambda x)^2}{2} e^{-\lambda x}. \end{aligned}$$

Аналогичными выкладками получаем:

$$f_4(x) = \int_{-\infty}^{\infty} f_3(y) \cdot f_1(x-y)dy = \int_0^x \lambda \frac{(\lambda y)^2}{2} e^{-\lambda y} \cdot \lambda e^{-\lambda(x-y)}dy = \lambda e^{-\lambda x} \int_0^x \frac{(\lambda y)^2}{2} d\lambda y = \lambda \frac{(\lambda x)^3}{2 \cdot 3} e^{-\lambda x}.$$

В общем виде получается гамма-распределение с целочисленным параметром m (распределение Эрланга):

$$f_m(x) = \lambda \frac{(\lambda x)^{m-1}}{(m-1)!} e^{-\lambda x}.$$

Графики этого распределения при различных значениях параметра m приведены на рис. 7.2б.

Гамма-распределение обладает своеобразной устойчивостью – композиция гамма-распределений с параметрами m_1 и m_2 снова приводит к гамма-распределению с параметром $m = m_1 + m_2$. С увеличением m это распределение приближается к нормальному.

Нормальное распределение также обладает устойчивостью – композиция двух нормальных распределений $N(a_1; \sigma_1)$ и $N(a_2; \sigma_2)$ снова приводит к нормальному распределению $N(a; \sigma)$ с параметрами $a = a_1 + a_2$, $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$ (складываются математические ожидания и дисперсии).

Наоборот, равномерное распределение устойчивостью не обладает.

Пример 2. Найдем закон распределения двух случайных величин, каждая из которых распределена по равномерному закону на интервале $(0; 1)$:

$$(x) = 1, \text{ если } x \in (0; 1), \text{ и } f_1(x) = 0, \text{ если } x \notin (0; 1).$$

Вычисляем интеграл свертки:

$$f_2(x) = \int_{-\infty}^{\infty} f_1(y) f_1(x-y) dy.$$

Здесь подынтегральная функция отлична от нуля (и равна единице) только для условий $0 \leq y \leq 1$ и $0 \leq (x-y) \leq 1$, которые не выполняются, если $x \notin (0; 2)$, иными словами, $f_2(x) = 0$ для $x < 0$ или $x > 2$.

На интервале $0 \leq x \leq 1$ приведенные выше условия можно записать как $0 \leq y \leq x$, а на интервале $1 \leq x \leq 2$ – как $(x-1) \leq y \leq 1$.

Определяем $f_2(x)$ для $0 \leq x \leq 1$:

$$f_2(x) = \int_0^x 1 dy = y \Big|_0^x = x.$$

Определяем $f_2(x)$ для $1 \leq x \leq 2$:

$$f_2(x) = \int_{x-1}^1 1 dy = y \Big|_{x-1}^1 = 2 - x.$$

Записываем выражения $f_2(x)$ по интервалам изменения x :

$$f_2(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x \leq 1 \\ 2 - x, & 1 \leq x \leq 2 \\ 0, & x > 2 \end{cases}.$$

Это треугольный закон распределения Симпсона.

Графики композиций равномерного закона приведены на рис. 7.2а.

Функции случайного аргумента

Рассмотрим закон распределения и характеристики функции $\mathcal{Y} = \varphi(\mathcal{X})$ случайного аргумента \mathcal{X} . Распределение \mathcal{X} считается известным.

Если \mathcal{X} – дискретная величина, заданная своим рядом распределения (рис. 7.3а), то ряд распределения функции (рис. 7.3б) составляется просто заменой x_i на $y_i = \varphi(x_i)$:

\mathcal{X}	x_1	x_2	x_3	\dots	x_k		\mathcal{Y}	$\varphi(x_1)$	$\varphi(x_2)$	$\varphi(x_3)$	\dots	$\varphi(x_k)$
$P(x)$	p_1	p_2	p_3	\dots	p_k		$P(y)$	p_1	p_2	p_3	\dots	p_k

а)

б)

Рис. 7.3. Дискретные распределения x и y

Осталось расположить значения y_i в порядке возрастания; вероятности повторяющихся значений y_i надо складывать.

Пример. Дан ряд распределения \mathcal{X} . Составим ряд распределения $\mathcal{Y} = \mathcal{X}^2$.

\mathcal{X}	-1	0	1	2
p	0,2	0,5	0,2	0,1

а)

\mathcal{Y}	1	0	1	4
p	0,2	0,5	0,2	0,1

б)

\mathcal{Y}	0	1	4
q	0,5	0,4	0,1

в)

Рис. 7.4. Ряд распределения x (а) и ряды распределения $y = x^2$ (б, в)

С вычислением характеристик функции нет никаких проблем, причем можно даже не составлять ряд распределения \mathcal{Y} : $M(y) = \sum \varphi(x_i) p_i$.

Для вышеприведенного примера можно найти математическое ожидание функции, предварительно составив ее ряд распределения:

$$M(y) = \sum y_j q_j = 0 \cdot 0,5 + 1 \cdot 0,4 + 4 \cdot 0,1 = 0,8;$$

или же непосредственно по исходному ряду для аргумента \mathcal{X} :

$$M(y) = \sum (x_i)^2 p_i = (-1)^2 \cdot 0,2 + 0^2 \cdot 0,5 + 1^2 \cdot 0,2 + 2^2 \cdot 0,1 = 0,8.$$

Дисперсия функции и моменты 3-го и 4-го порядков вычисляются по обычным формулам: $D(y) = M(y^2) - (M(y))^2$, $m_3(y) = M(y^3)$, $m_4(y) = M(y^4)$.

Пусть теперь \mathcal{X} – непрерывная величина, заданная функцией плотности вероятности $f_x(x)$. Как и для дискретного случая, характеристики функции $\mathcal{Y} = \varphi(\mathcal{X})$ можно вычислять непосредственно, не составляя функций распределения для \mathcal{Y} : $M(y) = M(\varphi(x)) = \int_{-\infty}^{\infty} \varphi(x) f_x(x) dx$.

Однако иногда требуется, зная распределение \mathcal{X} , получить в явном виде функции распределения для \mathcal{Y} . Обычно известна функция плотности вероятности $f_x(x)$ аргумента \mathcal{X} . Какой вид имеет функция плотности вероятности $f_y(y)$ для $\mathcal{Y} = \varphi(\mathcal{X})$?

Здесь понадобится также функция обратного преобразования $\mathcal{X} = \psi(\mathcal{Y})$, которая всегда существует для монотонной ветви преобразования $\mathcal{Y} = \varphi(\mathcal{X})$. Известно, что $\frac{d\varphi}{dx} \cdot \frac{d\psi}{dy} = 1$ или $\varphi'_x \psi'_y = 1$.

Запишем двумя способами формулу для вероятности попадания случайной величины в дифференциально малую окрестность:

$$dp = f_y(y) dy = f_x(\varphi(x)) dx = f_x(y) \psi'_y(y) dy.$$

Отсюда получаем $f_y(y) = f_x(y) \psi'_y(y)$.

Пример. Логнормальное распределение.

Случайная величина распределена по логарифмически нормальному закону, если ее логарифм $y = \ln x$ распределен нормально. Основная область применения логнормального закона – социологические и экономические исследования. В частности, этим законом хорошо описывается распределение таких экономических показателей, как доход, заработная плата, потребительский спрос и т. п. Логнормальное распределение внешне очень похоже на гамма-распределение и как гамма-распределение используется для описания существенно положительных величин $x > 0$.

Если случайная величина $y = \ln x$ распределена нормально с характеристиками $\mu_0 = M(y)$ и $\sigma_0^2 = D(x)$, то этот факт кратко обозначается как $y \sim N(\mu_0; \sigma_0)$. При этом величина x имеет логнормальное распределение с этими же параметрами, что кратко обозначается как $x \sim \Lambda(\mu_0; \sigma_0)$. Требование positivity $x > 0$ можно заменить на более общее условие $x > \theta_0$ и рассматривать логнормальное распределение разностей $(x - \theta_0)$: $y = \ln(x - \theta_0)$. Такое обобщен-

ное логнормальное распределение обозначается $\Lambda(\mu_0; \sigma_0; \theta_0)$ и используется в случаях, когда известно, что случайная величина x по своему смыслу не может быть меньше некоторого граничного значения θ_0 .

Функцию плотности вероятности (дифференциальную функцию) логнормального распределения получим из следующих соображений. Пусть $f_y(y)$ – дифференциальная функция нормального закона для $y = \ln(x - \theta_0)$:

$$f_y(y) = \frac{1}{\sqrt{2\pi} \cdot \sigma_0} \cdot \exp \left\{ -\frac{(y - \mu_0)^2}{2\sigma_0^2} \right\}.$$

Тогда $dp = f_y(y)dy = f_y(\ln(x - \theta_0)) \cdot \frac{dx}{x - \theta_0} = f_x(x)dx$ – вероятность попадания случайной величины y в интервал $(y, y + dy)$ или величины x в интервал $(x, x + dx)$. Отсюда получаем функцию плотности вероятности логнормального закона в виде:

$$f_x(x) = \frac{dp}{dx} = \frac{f_y(\ln(x - \theta_0))}{x - \theta_0} = \frac{1}{\sqrt{2\pi} \cdot \sigma_0 \cdot (x - \theta_0)} \cdot \exp \left\{ -\frac{(\ln(x - \theta_0) - \mu_0)^2}{2 \cdot \sigma_0^2} \right\}.$$

Графики этой функции для разных значений параметров μ_0 , σ_0 , ($\theta_0 = 0$) приведены на рис. 7.5.

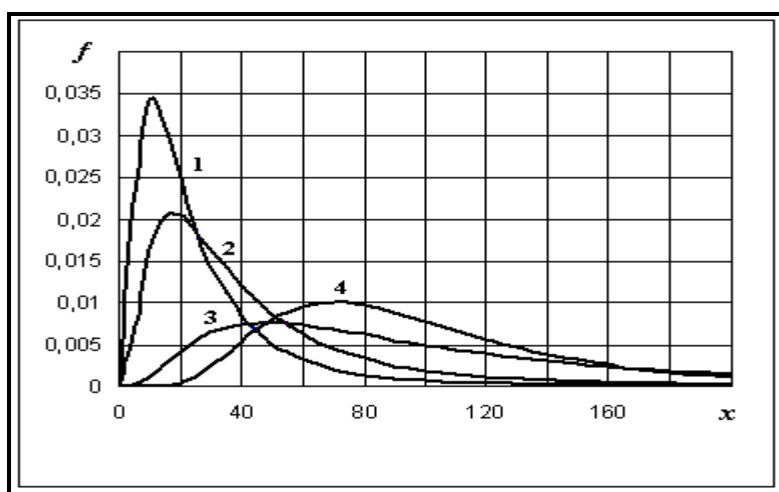


Рис. 7.5. Графики дифференциальной функции логнормального распределения

1 – ($\mu_0=3$; $\sigma_0=0,8$); 2 – ($\mu_0=3,5$; $\sigma_0=0,8$); 3 – ($\mu_0=4,5$; $\sigma_0=0,8$);

4 – ($\mu_0=4,5$; $\sigma_0=0,5$).

Характеристики $\mu_0 = M(y)$, $\sigma_0^2 = D(y)$ связаны с аналогичными характеристиками исходного показателя $\mu_x = M(x)$, $\sigma_x^2 = D(x)$ следующими соотношениями:

$$\begin{cases} \mu_x - \theta_0 = \exp\left(\mu_0 + \frac{\sigma_0^2}{2}\right) \\ \sigma_x^2 = (\mu_x - \theta_0)^2 \cdot (\exp(\sigma_0^2) - 1) \end{cases}, \text{ или наоборот } \begin{cases} \sigma_0^2 = \ln\left[1 + \left(\frac{\sigma_x}{\mu_x - \theta_0}\right)^2\right] \\ \mu_0 = \ln(\mu_x - \theta_0) - \frac{\sigma_0^2}{2} \end{cases}.$$

Вывод формулы композиции двух непрерывных величин

Ранее эта формула была получена по аналогии с формулой композиции дискретных величин. Пусть распределение непрерывных величин \mathcal{X} , \mathcal{Y} задано функциями плотности вероятности $f_x(x)$, $f_y(y)$. Известно, что при независимости случайных величин плотность их совместного распределения равна произведению $f_{xy}(x, y) = f_x(x) \cdot f_y(y)$. Запишем выражение для интегральной функции распределения суммы $\mathcal{Z} = \mathcal{X} + \mathcal{Y}$ в виде двойного интеграла: $F_z(z) = P(x + y \leq z) = \iint_{x+y \leq z} f_x(x) f_y(y) dx dy$, где область интегрирования опреде-

ляется неравенствами (для примера, рассмотрим случай неотрицательных величин): $0 \leq y \leq z - x$, $x \geq 0$. Преобразуем двойной интеграл к повторному:

$F_z(z) = \int_0^\infty f_x(x) dx \int_0^{z-x} f_y(y) dy$. Для определения плотности вероятности $f_z(z)$ дифференцируем полученное выражение по z :

$$f_z(z) = \int_0^\infty f_x(x) dx \cdot \frac{d}{dz} \int_0^{z-x} f_y(y) dy; \quad f_z(z) = \int_0^\infty f_x(x) f_y(z-x) dx,$$

что и требовалось доказать.

Вопросы для самопроверки

1. Сформулируйте закон больших чисел.
2. Сформулируйте центральную предельную теорему.
3. Как найти закон распределения суммы случайных величин?
4. Как найти закон распределения функции дискретного случайного аргумента?
5. Как вычисляются основные числовые характеристики дискретного случайного аргумента?
6. Приведите формулу для плотности вероятностей функции непрерывного случайного аргумента.
7. Как вычисляются основные числовые характеристики непрерывного случайного аргумента?
8. Что означает композиция законов распределений?

9. Что такое свертка распределений?

10. К чему приводит композиция двух нормальных распределений?

8. Система случайных величин

До сих пор в работе рассматривались случайные величины, возможные значения которых определялись одним числом. Такие величины называются одномерными. Однако часто результат опыта описывается несколькими случайными величинами, которые образуют комплекс или систему. Например, координаты точки попадания при стрельбе определяются двумя числами (абсцисса и ордината); состояние газа описывается тремя показателями (давление, температура, удельный объем). Такие комплексные случайные величины называются двумерными, трехмерными и так далее по числу компонент системы. В общем случае недостаточно изучить отдельно распределения каждой компоненты, поскольку между компонентами могут быть взаимные связи.

Геометрически систему нескольких случайных величин можно представить как случайную точку в многомерном пространстве.

Закон распределения дискретной двумерной величины

В дискретном случае двумерное распределение $(\mathcal{X}, \mathcal{Y})$ можно задать двумерной таблицей, в которой каждой паре возможных значений (x_i, y_j) поставлена в соответствие вероятность p_{ij} появления такой комбинации (рис. 8.1).

Первая строка таблицы содержит всевозможные значения компоненты \mathcal{X} , а первый столбец – компоненты \mathcal{Y} .

\mathcal{Y}	\mathcal{X}						$p(y)$
	x_1	x_1	...	x_i	...	x_n	
y_1	p_{11}	p_{12}	...	p_{i1}	...	p_{n1}	$p(y_1)$
y_2	p_{21}	p_{22}	...	p_{i2}	...	p_{n2}	$p(y_2)$
...
y_j	p_{1j}	p_{2j}	...	p_{ij}	...	p_{nj}	$p(y_j)$
...
y_k	p_{1k}	p_{2k}	...	p_{ik}	..	p_{nk}	$p(y_k)$
$p(x)$	$p(x_1)$	$p(x_2)$...	$p(x_i)$...	$p(x_n)$	1

Рис. 8.1. Двумерная случайная величина

В последней строке и последнем столбце вычислены суммы вероятностей p_{ij} по столбцам и строкам.

Так как все комбинации (x_i, y_j) образуют полную группу несовместных событий, то общая сумма вероятностей $\sum \sum p_{ij} = 1$. Поскольку отдельно все значения y_j также образуют полную группу несовместных событий, то суммы ве-

роятностей p_{ij} по столбцам равны полным вероятностям событий ($\mathcal{X} = x_i$). Следовательно, первая и последняя строки таблицы представляют собой ряд распределения \mathcal{X} . Аналогично, первый и последний столбцы таблицы представляют собой ряд распределения \mathcal{Y} . Так как, согласно теореме умножения вероятностей, $p_{ij} = P(\mathcal{X} = x_i, \mathcal{Y} = y_j) = p(x_i)p(y_j|x_i) = p(y_j)p(x_i|y_j)$, то можно составить ряды условных распределений $p(y_j|x_i) = p_{ij}/p(x_i)$ и $p(x_i|y_j) = p_{ij}/p(y_j)$.

Иными словами, столбцы таблицы пропорциональны условным вероятностям $p(y_j|x_i)$, а строки – $p(x_i|y_j)$. Для независимых случайных величин $p_{ij} = p(x_i) \cdot p(y_j)$ и распределения во всех параллельных рядах таблицы будут одинаковыми: $p(y_j|x_i) = p(y_j)$; $p(x_i|y_j) = p(x_i)$.

Характеристики дискретной двумерной случайной величины

Кроме стандартных характеристик, отдельно для каждой компоненты многомерной случайной величины:

$$\begin{aligned} m_x = M(x) &= \sum_i \sum_j x_i p_{ij} = \sum_i x_i p(x_i); & m_y = M(y) &= \sum_i \sum_j y_j p_{ij} = \sum_j y_j p(y_j); \\ \sigma_x^2 &= M(x - m_x)^2 = \sum_i \sum_j (x_i - m_x)^2 p_{ij} = \sum_i x_i^2 p(x_i) - m_x^2; \\ \sigma_y^2 &= M(y - m_y)^2 = \sum_i \sum_j (y_j - m_y)^2 p_{ij} = \sum_j y_j^2 p(y_j) - m_y^2, \end{aligned}$$

вычисляются еще смешанные центральные моменты, которые называются ковариациями (для двумерной величины имеется только одна ковариация – коэффициент совместной изменчивости x и y):

$$\begin{aligned} Cov(x, y) = \mu_{xy} = \sigma_{xy} &= M(x - m_x)(y - m_y) = \sum_i \sum_j (x_i - m_x)(y_j - m_y) p_{ij} = \\ &= \sum_i \sum_j x_i y_j p_{ij} - m_x m_y = M(xy) - m_x m_y. \end{aligned}$$

Дисперсии являются частным случаем ковариации при $x = y$:

$$\mu_{xx} = \sigma_{xx} = M(x - m_x)^2; \quad \mu_{yy} = \sigma_{yy} = M(y - m_y)^2.$$

Нормированный смешанный центральный момент (нормированная ковариация) называется коэффициентом корреляции:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

Коэффициент корреляции является мерой тесноты связи между x и y . Ранее уже было доказано, что для независимых случайных величин ковариация (и коэффициент корреляции) равны нулю.

Для любого столбца таблицы можно вычислить условные характеристики: математическое ожидание $M(y | x_i)$ и дисперсию $D(y | x_i)$:

$$M(y | x_i) = \sum_j y_j \cdot p(y_j | x_i) = \sum_j y_j p_{ij} / p(x_i).$$

Аналогично вычисляются условные характеристики для строк таблицы:

$$M(x | y_j) = \sum_i x_i \cdot p(x_i | y_j) = \sum_i x_i p_{ij} / p(y_j).$$

Если множество значений двумерной величины представить точками с координатами (x_i, y_j) и весами (весовыми коэффициентами) p_{ij} , то условные математические ожидания $M(y | x_i)$ будут представлять собой центры тяжести (средние взвешенные) в каждом вертикальном ряду таблицы для $x = x_i$, а $M(x | y_j)$ – центры тяжести в каждом горизонтальном ряду таблицы для $y = y_j$, точка с координатами $(M(x), M(y))$ – общий центр тяжести всей системы точек.

Для независимых случайных величин все условные характеристики будут одинаковы для всех рядов таблицы: $M(y | x_i) = M(y)$ и все $M(x | y_j) = M(x)$.

Введем некоторую классификацию типов связей (зависимостей).

Зависимость называется **функциональной**, если каждому значению аргумента (аргументов) соответствует единственное значение функции (каждому значению объясняющих переменных соответствует единственное значение результативного признака, случайного разброса нет).

Зависимость называется **стохастической (статистической)**, если при изменении объясняющих переменных меняется закон распределения результативной переменной (меняется его условное распределение) – меняется вид распределения или только его характеристики (условные математические ожидания, дисперсии и т. п.). Таким образом, в отличие от функциональной связи при статистической зависимости нет однозначного соответствия между множеством значений аргументов и множеством значений функции.

Статистическая зависимость называется **корреляционной**, если при изменении аргумента меняется условное математическое ожидание функции (каждому значению объясняющих переменных соответствует свое среднее значение результативной переменной). При корреляционной зависимости мы следим за изменением только одной характеристики – центра условного распределения (условного математического ожидания).

Корреляционная зависимость является частным случаем общей статистической зависимости. Естественно, существуют также иные виды статистиче-

ских зависимостей некорреляционного типа, например, когда меняется условная дисперсия.

Коэффициент корреляции является мерой тесноты корреляционной связи; когда он равен нулю, корреляционной зависимости нет (все условные математические ожидания одинаковы). Однако при $\rho_{xy} = 0$ могут быть иные виды статистической зависимости, поэтому из равенства нулю коэффициента корреляции еще нельзя утверждать, что случайные величины \mathcal{X} , \mathcal{Y} независимы; говорят, что такие величины «некоррелированы». Покажем, что максимальное значение коэффициента корреляции по абсолютной величине равно единице.

Для этого рассмотрим дисперсию линейной комбинации

$$D(x \cdot \sigma_y \pm y \cdot \sigma_x) = \sigma_x^2 \sigma_y^2 + \sigma_x^2 \sigma_y^2 \pm 2\sigma_x \sigma_y \mu_{xy} \geq 0,$$

где μ_{xy} – ковариация: $\mu_{xy} = \sigma_x \sigma_y \rho_{xy}$.

Отсюда получаем $-1 \leq \rho_{xy} \leq 1$.

При $|\rho_{xy}| = 1$ зависимость линейная функциональная (нет разброса).

Закон распределения непрерывной двумерной величины

Универсальным способом задания случайной величины является ее функция распределения: $F(x, y) = P(\mathcal{X} \leq x; \mathcal{Y} \leq y)$ – вероятность того, что каждая компонента системы не превзойдет указанных значений x, y .

Геометрически (рис. 8.2) это означает вероятность попадания случайной точки в квадрант с правой верхней вершиной в точке (x, y) .

Из определения функции распределения следует:

$$0 \leq F(x, y) \leq 1, \quad F(-\infty, \infty) = 0, \quad F(\infty, -\infty) = 0, \\ F(-\infty, -\infty) = 0, \quad F(\infty, \infty) = 1, \quad F(x, \infty) = F(x), \\ F(\infty, y) = F(y).$$

Рассмотрим вероятность попадания случайной точки $\mathcal{M}(x, y)$ в прямоугольник \mathcal{D} с диагональными вершинами $(x_1, y_1) - (x_2, y_2)$. Из рис. 8.3 находим, что эта вероятность равна:

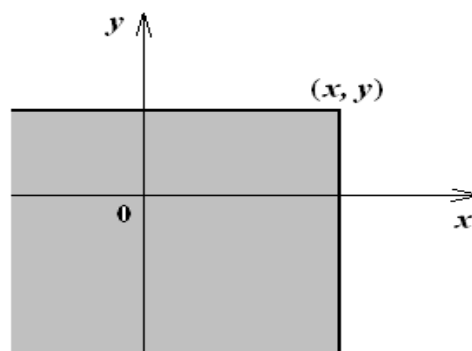


Рис. 8.2. Квадрант с вершиной (x, y)

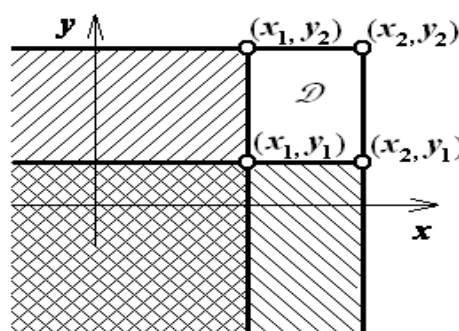


Рис. 8.3. Прямоугольник с диагональными вершинами $(x_1, y_1) - (x_2, y_2)$

$$P(\mathcal{M} \subset \mathcal{D}) = P(x_1 < \mathcal{X} \leq x_2, y_1 < \mathcal{Y} \leq y_2) = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1).$$

Вводим понятие плотность вероятности как отношение вероятности попадания случайной точки в какую-либо область к площади этой области.

Для определения функции плотности вероятности $f(x, y)$ в точке используем предельный переход:

$$f(x, y) = \lim_{\substack{\Delta x \rightarrow 0 \\ \Delta y \rightarrow 0}} \frac{F(x + \Delta x, y + \Delta y) - F(x, y + \Delta y) - F(x + \Delta x, y) + F(x, y)}{\Delta x \Delta y},$$

$$\frac{\partial^2 F(x, y)}{\partial x \partial y} = F''_{xy}(x, y).$$

Таким образом, функция плотности вероятности для двумерной случайной величины равна смешанной производной второго порядка от функции распределения.

Геометрически двумерную функцию $f(x, y)$ можно представить некоторой поверхностью – поверхностью распределения. На рис. 8.4а для примера приведена такая поверхность для двумерного нормального закона (будет рассмотрен далее). Удобно изображать двумерные распределения линиями уровня – линиями равной плотности вероятности (рис. 8.4б).

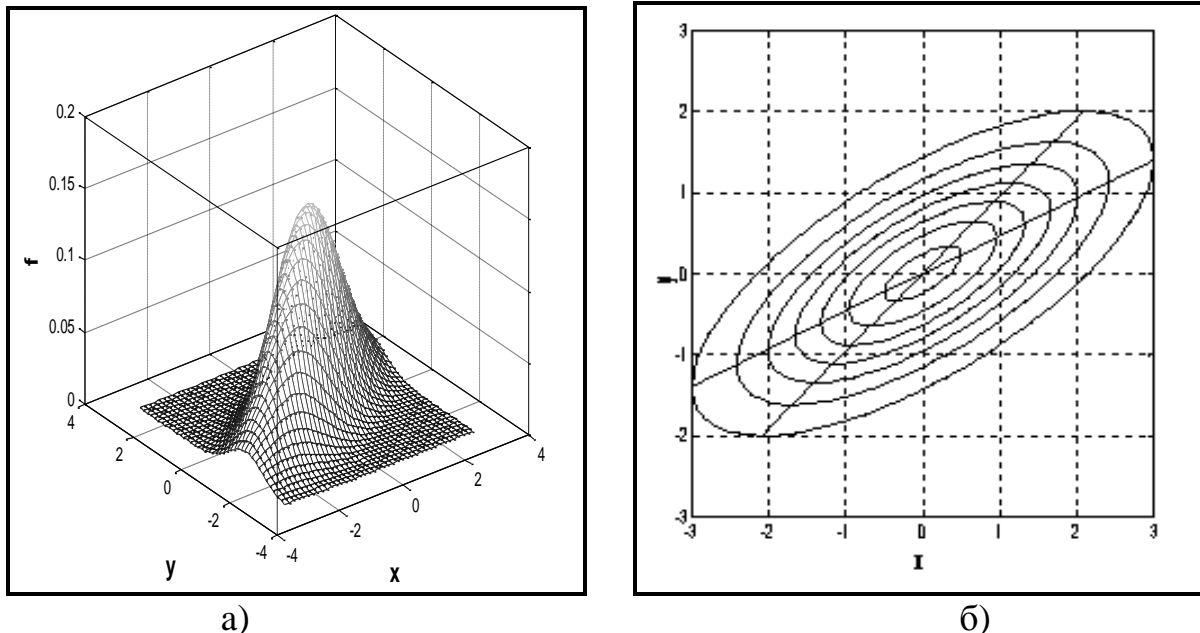


Рис. 8.4. Двумерный график плотности вероятности нормального закона (а) с параметрами ($m_x = 0; m_y = 0; \sigma_x = 1,5; \sigma_y = 1; \rho_{xy} = 0,7$) и семейство линий уровня $f = \text{Const}$ (б)

Если отдельные компоненты, входящие в систему, взаимно независимые, то по теореме умножения вероятностей имеем:

$$F(x, y) = P(\mathcal{X} \leq x; \mathcal{Y} \leq y) = P(\mathcal{X} \leq x) \cdot P(\mathcal{Y} \leq y) = F(x) \cdot F(y).$$

Дифференцируем это равенство по x и по y :

$$f(x, y) = F''_{xy}(x, y) = F'_x(x) \cdot F'_y(y) = f_1(x) \cdot f_2(y).$$

Таким образом, для независимых компонент плотность распределения системы случайных величин оказалась равной произведению плотностей распределения отдельных компонент.

Напоминаем, что случайные величины называются независимыми, если закон распределения каждой из них не зависит от того, какое значение приняла другая величина.

Зная плотность вероятности системы $f(x, y)$, всегда можно найти распределение отдельных компонент:

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy; \quad f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx.$$

Обратное утверждение в общем случае неверно – только для независимых компонент можно по их распределениям восстановить закон распределения всей системы. Согласно общей теореме умножения вероятностей записываем:

$$f(x, y) = f_1(x) \cdot f_2(y|x) = f_2(y) \cdot f_1(x|y),$$

где условные плотности вероятностей $f_2(y|x)$ и $f_1(x|y)$ можно вычислить как отношения:

$$f(y|x) = \frac{f(x, y)}{f_1(x)}; \quad f(x|y) = \frac{f(x, y)}{f_2(y)}.$$

Характеристики непрерывной двумерной величины

Все вычислительные формулы аналогичны формулам для вычисления характеристик дискретных величин, только дискретные суммы заменяются на интегралы:

$$\begin{aligned}
m_x &= M(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot f(x, y) dx dy = \int_{-\infty}^{\infty} x \cdot f_1(x) dx; \\
m_y &= M(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y \cdot f(x, y) dx dy = \int_{-\infty}^{\infty} y \cdot f_2(y) dy; \\
\sigma_x^2 &= M(x - m_x)^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - m_x)^2 f(x, y) dx dy = \int_{-\infty}^{\infty} x^2 f_1(x) dx - m_x^2; \\
\sigma_y^2 &= M(y - m_y)^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - m_y)^2 f(x, y) dx dy = \int_{-\infty}^{\infty} y^2 f_2(y) dy - m_y^2; \\
\sigma_{xy} &= M(x - m_x)(y - m_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - m_x)(y - m_y) \cdot f(x, y) dx dy = \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \cdot f(x, y) dx dy - m_x m_y.
\end{aligned}$$

Если ковариация σ_{xy} отлична от нуля, вычисляются еще условные характеристики:

$$\begin{aligned}
M(y|x) &= \int_{-\infty}^{\infty} y \cdot f_2(y|x) dy = \left(\int_{-\infty}^{\infty} y \cdot f(x, y) dy \right) / f_1(x) \\
M(x|y) &= \int_{-\infty}^{\infty} x \cdot f_1(x|y) dx = \left(\int_{-\infty}^{\infty} x \cdot f(x, y) dx \right) / f_2(y)
\end{aligned}$$

Условные характеристики являются функциями другого аргумента.

Двумерный нормальный закон

Плотность вероятности для этого закона описывается формулой:

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}} \cdot \exp\left\{-\frac{1}{2 \cdot (1-\rho_{xy}^2)} \cdot [X^2 - 2\rho_{xy}XY + Y^2]\right\},$$

где через X и Y обозначены «стандартизованные» переменные:

$$X = \frac{x - m_x}{\sigma_x}; \quad Y = \frac{y - m_y}{\sigma_y}.$$

Двумерный нормальный закон зависит от пяти параметров, которые заодно являются его характеристиками:

$$, m_y, \sigma_x, \sigma_y, \rho_{xy}.$$

При двумерном нормальном законе каждая компонента системы распределена по одномерному нормальному закону:

$$f_1(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \exp\left\{-\frac{1}{2} X^2\right\}; \quad f_2(y) = \frac{1}{\sqrt{2\pi} \cdot \sigma_y} \exp\left\{-\frac{1}{2} Y^2\right\}.$$

Поверхность нормального распределения для параметров $m_x = 0$; $m_y = 0$; $\sigma_x = 1,5$; $\sigma_y = 1$; $\rho_{xy} = 0,7$ изображена на рис. 8.4а.

Семейство линий уровня для двумерного нормального закона (см. рис. 8.4б) представляет собой семейство эллипсов $X^2 - 2\rho_{xy}XY + Y^2 = \text{Const}$.

При $\rho_{xy} = 0$ оси эллипсов параллельны координатным осям.

Условные распределения также подчиняются нормальному закону:

$$f(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{1}{\sqrt{2\pi} \cdot \sigma_y \sqrt{1-\rho_{xy}^2}} \cdot \exp\left\{-\frac{1}{2 \cdot (1-\rho_{xy}^2)} \cdot (Y - \rho_{xy}X)^2\right\};$$

$$f(x|y) = \frac{f(x,y)}{f_2(y)} = \frac{1}{\sqrt{2\pi} \cdot \sigma_x \sqrt{1-\rho_{xy}^2}} \cdot \exp\left\{-\frac{1}{2 \cdot (1-\rho_{xy}^2)} \cdot (X - \rho_{xy}Y)^2\right\}.$$

Интересно, что центры этих распределений (условные математические ожидания) линейно зависят от другой переменной:

$$M(Y|X) = \rho_{xy}X \rightarrow M(y|x) = m_y + \rho_{xy} \frac{\sigma_y}{\sigma_x} (x - m_x);$$

$$M(X|Y) = \rho_{xy}Y \rightarrow M(x|y) = m_x + \rho_{xy} \frac{\sigma_x}{\sigma_y} (y - m_y).$$

Эти линейные зависимости являются диаметрами эллипсов рассеяния, сопряженными семейству вертикальных и горизонтальных хорд соответственно.

Иными словами, $M(y|x)$ представляет собой множество середин вертикальных хорд, а $M(x|y)$ – множество середин горизонтальных хорд эллипсов. Оба диаметра при $\rho_{xy} \neq 0$ не совпадают с главными осями эллипса (см. рис. 8.4б).

Выше уже указывалось, что для независимых случайных величин $\rho_{xy} = 0$. Обратное утверждение в общем случае неверно – равенства $\rho_{xy} = 0$ недостаточно, чтобы утверждать о независимости случайных величин \mathcal{X} и \mathcal{Y} , поэтому говорят, что при $\rho_{xy} = 0$ случайные величины «не коррелированы».

Однако, если дополнительно известно, что закон совместного распределения нормальный, то из условия $\rho_{xy} = 0$ следует заключение о независимости компонент.

Действительно, для $\rho_{xy} = 0$ имеем:

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \cdot \exp\left\{-\frac{1}{2} \cdot [X^2 + Y^2]\right\};$$

$$f(x, y) = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \cdot \exp\left\{-\frac{1}{2} X^2\right\} \times \frac{1}{\sqrt{2\pi} \cdot \sigma_y} \cdot \exp\left\{-\frac{1}{2} Y^2\right\}.$$

Получили $f(x, y) = f_1(x) \cdot f_2(y)$ – условие независимости \mathcal{X}, \mathcal{Y} .

Интересно, что в случае двумерного нормального закона меняются только центры условных распределений (условные математические ожидания),

условные же дисперсии все одинаковы:

$$D(y|x) = \sigma_y^2 \cdot (1 - \rho_{xy}^2).$$

При $|\rho_{xy}| = 1$ дисперсии $D(y|x) = 0$, то есть разброса нет, зависимость функциональная.

Вопросы для самопроверки

1. Приведите определение многомерной случайной величины.
2. Как можно задать распределение двумерной случайной величины?
3. Что такое условные распределения?
4. Как, зная двумерное распределение системы, найти законы распределения ее компонентов?
5. Что такое функция распределения системы непрерывных случайных величин? Перечислите ее свойства.
6. Что такое плотность вероятностей для двумерных величин? Как эта функция связана с функцией распределения?
7. Перечислите основные характеристики двумерной случайной величины.
8. Что такое ковариация, каковы ее свойства?
9. Что такое коэффициент корреляции, каковы его свойства?
10. Чему равен коэффициент корреляции для независимых случайных величин?
11. Какие выводы можно сделать, если коэффициент корреляции равен нулю? Если он равен единице? Если он отрицательный?
12. Какие бывают типы связей?
13. Что такое корреляционная зависимость?
14. Какую зависимость характеризует коэффициент корреляции?

15. Сформулируйте двумерный нормальный закон распределения.
16. Что такое эллипсы рассеяния?
17. Каковы распределения компонент, если система величин имеет многомерное нормальное распределение?
18. Какими особенностями обладают условные распределения, если система величин имеет двумерное нормальное распределение?

9. Проблемы математической статистики

Цели теории вероятностей и математической статистики в некоторой мере противоположны. В теории вероятностей, зная теоретическое распределение случайных величин, пытаются предсказать результаты опыта (с заданной надежностью и погрешностью). В математической статистике, наоборот, по результатам эмпирического обследования пытаются сделать заключения о теоретическом распределении случайных величин.

Слово «статистика» происходит от слова *state* (государство), так как управление государством требует учета большого объема сведений из самых различных отраслей хозяйства, здравоохранения, внешнеполитической обстановки и т. п. Задача математической статистики – свести эти «простыни цифр» (выражение акад. Крылова А. Н.) до немногих понятных характеристик.

Кратко цель математической статистики можно сформулировать как разработку методов регистрации, описания и анализа данных наблюдений.

Назовем **совокупностью** или **генеральной совокупностью**, все мыслимые наблюдения изучаемой случайной величины. Количество этих наблюдений – объем совокупности – очень большое и часто бесконечное. Например, при бросках монеты объем генеральной совокупности бесконечен, монету можно подбрасывать все время без остановок. Если обследуется качество продукции, то в генеральную совокупность включаются все когда-либо произведенные изделия данного типа. Естественно, невозможно испытать все элементы генеральной совокупности, тем более, что некоторые испытания связаны с уничтожением образца (как, например, проверить вкусовые качества фруктов?). Для анализа отбирается сравнительно малая часть совокупности, которая называется **выборкой**. Обычно обозначают через N объем совокупности (если она конечная), а через n – объем выборки; естественно, объем выборки существенно меньше объема совокупности $n \ll N$.

В связи с этим возникает **проблема правильного отбора образцов** в выборку. Ведь можно в выборку отобрать (сознательно или не подозревая этого) одни бракованные элементы и на основании статистического обследования такой выборки получить совершенно неверные выводы. Выборка должна быть **репрезентативной (представительной)**, то есть правильно представлять совокупность.

Любые характеристики, вычисленные по данным выборки, называются **оценками**. Например, относительная частота m/n является оценкой вероятности p , среднее \bar{x} (центр выборки) является оценкой математического ожидания $M(x)$

(центра всей совокупности). Естественно, оценки должны быть доброкачественными и удовлетворять некоторым обязательным условиям. Правила составления доброкачественных оценок составляют **проблему статистического оценивания** числовых характеристик и параметров распределения случайных величин.

В «описательную статистику», кроме оценок числовых характеристик, входят также **эмпирические оценки функций распределения** и способы графического представления особенностей распределения данных.

На основании статистического обследования делаются некоторые выводы относительно значимости определяемых характеристик, особенностей распределения, существования или отсутствия связей, проверки однородности совокупности и т. д. Круг этих вопросов составляет **проблему проверки статистических гипотез**.

Способы составления выборочных подсовокупностей

Правила составления выборок должны обеспечивать равные шансы любым элементам совокупности быть отобранными в выборку. Предложено несколько методик составления представительных выборок от самых простых до очень сложных.

Простой случайный отбор. Так называется способ, когда элементы совокупности отбираются в выборку случайным образом, например, с помощью таблицы случайных чисел. При этом предполагается, что выборка однородная, а не представляет собой смесь нескольких подсовокупностей.

Расслоенный случайный отбор. Предварительно выясняют, из каких заметно различающихся подсовокупностей (групп, классов, слоев) состоит общая совокупность, и определяют (хотя бы ориентировочно) объемы этих групп ($N_1 + N_2 + N_3 + \dots + N_k = N$). Из каждой такой подсовокупности в выборку случайным образом отбираются элементы, количества которых пропорциональны объемам подсовокупностей $n_j = \lambda N_j$ или $n_1 : n_2 : n_3 : \dots : n_k = N_1 : N_2 : N_3 : \dots : N_k$, где $n_1 + n_2 + n_3 + \dots + n_k = n$.

На практике ориентируются на соображения по ограничению стоимости, времени и трудозатрат предварительного статистического обследования, поэтому часто выбирают упрощенные схемы отбора образцов, которые не гарантируют полной репрезентативности выборки. Однако в математической статистике данные будут анализироваться в предположении, что составленная выборка репрезентативная (представительная).

Статистическое оценивание

Доброкачественные оценки должны удовлетворять некоторым требованиям, о которых будет сказано ниже.

Сначала перечислим наиболее распространенные оценки характеристик распределения совокупности.

Для выборочных данных $\{x_i\}$ дискретной случайной величины подсчитывают частоты m_j повторяющихся значений X_j и составляют так называемый вариационный ряд

\mathcal{X}	X_1	X_2	X_3	\dots	X_k
m	m_1	m_2	m_3	\dots	m_k

Рис. 9.1. Вариационный ряд

(рис. 9.1) в виде таблицы соответствия между дискретными значениями случайной величины и частотами их появления в выборке ($\sum m_j = n$).

Относительные частоты $\frac{m_j}{n} \approx p_j$ являются стандартными оценками вероятностей p_j . Полигон относительных частот является приближением (оценкой) полигона вероятностей.

Среднее (выборочное среднее) \bar{x} является оценкой математического ожидания (генерального среднего) $M(x)$. Среднее можно подсчитать или по исходным данным: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, или по сгруппированным: $\bar{x} = \frac{1}{n} \sum_{j=1}^k m_j X_j$.

Поскольку многие характеристики выражаются через математические ожидания, для получения оценок этих характеристик надо в соответствующих формулах заменить операторы математического ожидания на операторы среднего. Таким образом, получаем оценки дисперсии, ковариации, коэффициента корреляции и многие другие (рис. 9.2).

Характеристики	Формулы	Оценки
Дисперсия	$\sigma_x^2 = M(x^2) - M^2(x)$	$s_x^2 = \overline{x^2} - (\bar{x})^2$
Ковариация	$\sigma_{xy} = M(xy) - M(x) \cdot M(y)$	$s_{xy} = \overline{xy} - \bar{x} \cdot \bar{y}$
Коэффициент корреляции	$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$	$r_{xy} = \frac{s_{xy}}{s_x s_y}$

Рис. 9.2. Некоторые характеристики и их выборочные оценки

В обозначениях оценок принято соглашение обозначать (по возможности) генеральные характеристики греческими буквами, а их выборочные оценки – соответствующими латинскими буквами или теми же символами, но с надстрочным знаком \wedge . Так, оценку дисперсии можно также обозначать как $\hat{\sigma}_x^2$. Иногда для одной и той же характеристики составляются разные оценки с разными свойствами, тогда разумно использовать разные обозначения для разных видов оценок. Далее будем обозначать через $\hat{\sigma}_x^2$ так называемую несмещенную оценку дисперсии (о которой будет сказано ниже).

Выборочные данные $\{x_i\}$ непрерывной случайной величины группируют на k интервалов с определенными границами $(s_{j-1}, s_j]$.

Ширина интервалов может быть разной $h_j = s_j - s_{j-1}$, центры интервалов обозначим $X_j = \frac{s_{j-1} + s_j}{2}$.

Подсчитываются частоты m_j попадания выборочных данных $\{x_i\}$ в эти интервалы, и строится так называемый интервальный вариационный ряд (рис. 9.3). Данные, попадающие на края интервалов, надо относить к левому (меньшему) интервалу (имеются и другие рекомендации, но только одна из них согласуется с принятым выше определением функции распределения как $F(x) = P(\mathcal{X} \leq x)$).

\mathcal{X}	X_1	X_2	X_3	\dots	X_k
	$s_0 - s_1$	$s_1 - s_2$	$s_2 - s_3$		$s_{k-1} - s_k$
m	m_1	m_2	m_3	\dots	m_k

Рис. 9.3. Интервальный ряд

Теперь формулы для вычисления среднего $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ и $\bar{x} \approx \frac{1}{n} \sum_{j=1}^k m_j X_j$

уже не эквивалентны, последняя формула (по сгруппированным данным) содержит дополнительно ошибку группировки, которая может быть существенной при малом числе интервалов. Для интервалов одинаковой ширины $h = \text{Const}$ имеются рекомендации принимать число интервалов равным корню квадратному от объема выборки $k \approx \sqrt{n}$ или же по более сложной формуле: $k \approx 1 + 3,322 \cdot \lg n$. Считается, что при $k > 10$ ошибки группировок сравнимы с другими видами ошибок, поэтому в этом случае допустимо пользоваться расчетами по сгруппированным данным, что существенно сокращает объем вычислительной работы при ручном счете.

Группировки нужны для описания закона распределения случайной величины. Ординаты **эмпирической функции плотности вероятности** (оценки дифференциальной функции распределения) вычисляются для центров интер-

валов $\hat{f}_j = \hat{f}(X_j) = \frac{\hat{p}_j}{h_j} = \frac{m_j}{nh_j}$. Поскольку все данные, попадающие в один интервал, округляются на его центр, а ошибки округления распределены по равномерному закону, то на каждом интервале $(s_{j-1} < x < s_j)$ значения эмпирической плотности вероятности считаются постоянными и равными \hat{f}_j .

График этой столбчатой функции называется **гистограммой** – это оценка графика дифференциальной функции распределения. Площадь гистограммы равна единице. При выборе слишком большого числа интервалов (очень мелкого шага) в некоторые интервалы может попасть мало наблюдений и тогда гистограмма будет иметь неоправданные «провалы». В таком случае интервалы надо укрупнять, но так, чтобы площадь гистограммы не изменялась. На рис. 9.4а приведена гистограмма с выбором слишком мелкого шага $h = 0,1$ (число интервалов $k = 20$ при объеме выборки $n = 75$). Видно много неоправданных провалов, а последние четыре наибольших наблюдения похожи на выбросы.

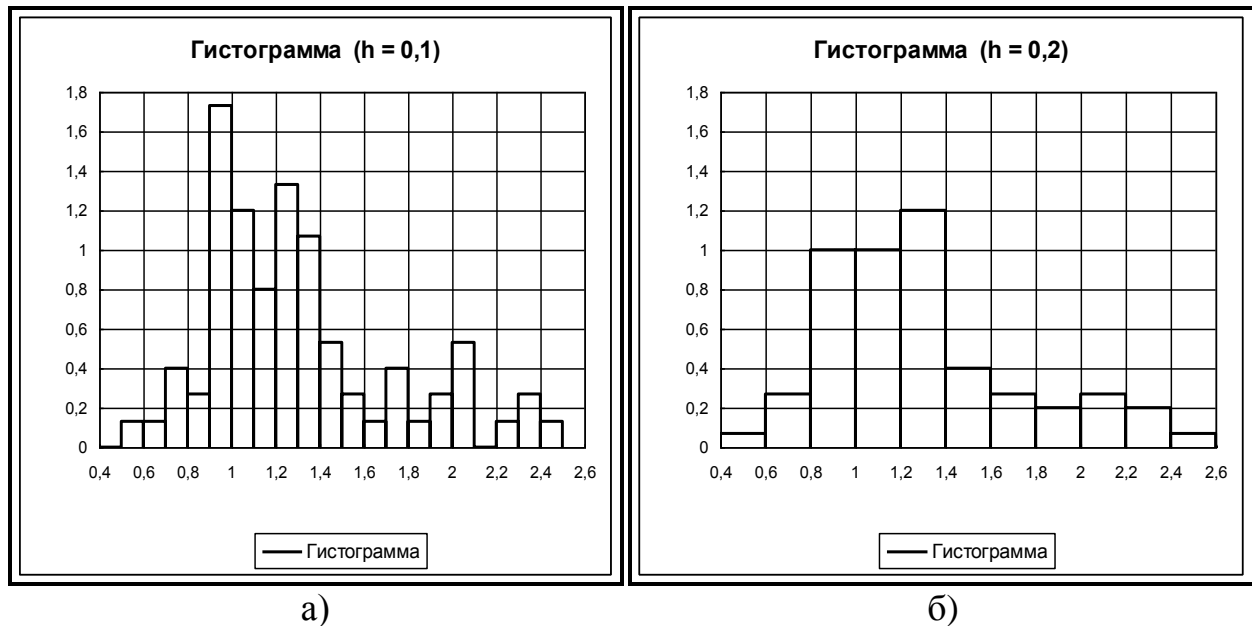


Рис. 9.4. Исходная и укрупненная гистограммы

На рис. 9.4б шаг был укрупнен вдвое $h = 0,1$ ($k = 11$), значения плотности вероятности в укрупненных интервалах вычислялись по формуле $\hat{f}_{1-2} = \frac{m_1 + m_2}{n \cdot (h_1 + h_2)}$; при укрупнении интервалов равной длины получается среднее ординат в объединяемых интервалах $\hat{f}_{1-2} = \frac{\hat{f}_1 + \hat{f}_2}{2}$. Площадь гистограммы после укрупнения не изменяется. Для объема выборки $n = 75$ рекомендуется число

интервалов ориентировочно принимать равным $k \approx \sqrt{n} = \sqrt{75} = 8,7$ ($k = 9$); по формуле $k \approx 1 + 3,322 \cdot \lg n = 1 + 3,322 \cdot \lg 75 = 15,3$ получается большее значение ($k = 15$).

Середины столбиков гистограммы при интервалах равной длины можно соединить отрезками прямых и получить полигон (рис. 9.5). Полигон производит графическое сглаживание угловатой гистограммы и целиком ей эквивалентен. Площадь такого полигона равна единице.

Обратите внимание на крайние интервалы – для построения полигона к гистограмме были добавлены слева и справа пустые интервалы.

Бытует порочная практика – вместо эмпирического графика плотности вероятности строить гистограмму частот, или гистограмму относительных частот. Для равностоящих интервалов ординаты всех



Рис. 9.5. Гистограмма и полигон

этих видов гистограмм пропорциональны, и вся разница сводится к разной градуировке оси ординат. Но чтобы не было проблем, не стоит пользоваться этими гистограммами: при необходимости укрупнять некоторые малонасыщенные интервалы вид таких гистограмм изменяется непредсказуемым образом. Далеко не всякая столбчатая диаграмма является гистограммой. В обычной столбчатой диаграмме ширина столбцов не имеет значения, она выбирается из декоративных соображений, а важны только высоты каждого столбца. Напротив, в гистограмме измеряют площади столбцов, а не их высоты – и только такой тип столбчатых диаграмм может называться гистограммой.

Эмпирическая функция распределения (оценка интегральной функции распределения) называется **кумулятой** и строится по *правым* границам интервалов группировки. Ординаты кумуляты вычисляются по формуле

$\hat{F}_j = \hat{F}(s_j) = \frac{1}{n} \sum_{i=1}^j m_i$. Для примера на рис. 9.6 приведен интервальный

вариационный ряд, который был использован ранее для построения гистограммы.

\mathcal{X}	0,2–0,4	0,4–0,6	0,6–0,8	0,8–1,0	1,0–1,2	1,2–1,4	1,4–1,6	1,6–1,8	1,8–2,0	2,0–2,2	2,2–2,4	2,4–2,6
m	0	1	4	15	15	19	6	4	3	4	3	1
Σm	0	1	5	20	35	54	60	64	67	71	74	75
\hat{F}	0	0,013	0,067	0,267	0,467	0,720	0,800	0,853	0,893	0,947	0,987	1

Рис. 9.6. Интервальный вариационный ряд

В третьей строке этой таблицы вычислены накопленные частоты Σm , а в четвертой – значения кумуляты ($\Sigma m / n$) на *правых* краях интервалов.

Поскольку все данные, попадающие в один интервал, округляются на его центр, а ошибки округления распределены по равномерному закону, то в каждом интервале ($s_{j-1} < x \leq s_j$) значения эмпирической функции распределения изменяются по линейному закону.

Для непрерывной случайной величины кумулята непрерывная кусочно-линейная (рис. 9.7).

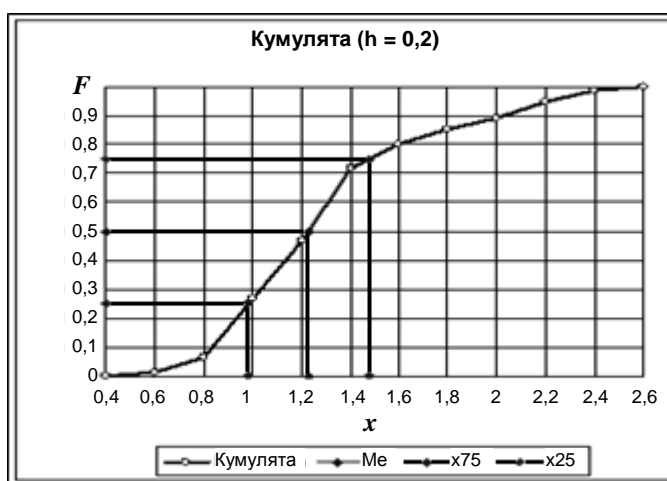


Рис. 9.7. Кумулята и квантили

Напоминаем, что для дискретной случайной величины кумулята на интервалах ($X_{j-1} < x \leq X_j$) сохраняет постоянные значения и изменяется скачками (ступеньками) только в заданных узлах $x = X_j$. Нет никакой необходимости строить такой график, он нигде не используется. Напротив, кусочно-линейный график кумуляты для непрерывной случайной величины используется для расчета оценок квантилей. На рис. 9.7 показано, как определять оценки квантилей: на оси ординат задаем $F = 0,5; 0,25; 0,75$ и в соответствии с графиком линейным интерполированием находим на оси абсцисс медиану $Me = x_{0,5} = 1,23$; нижнюю квантиль $x_{0,75} = 0,983$; верхнюю квантиль $x_{0,25} = 1,48$.

Квантили являются более надежными характеристиками положения, разброса и формы, нежели моменты распределения; они устойчивы к наличию выбросов – грубых ошибок некоторых наблюдений или описок (забыли десятичную запятую, перепутали сходные по начертанию цифры, например 3 и 8, и т. п.).



Рис. 9.8. «Усатый ящик Тьюкки»

Английский статистик Дж. Тьюкки предложил изображать распределение в виде блочной диаграммы «ящик и усы». «Ящик» (прямоугольник на рис. 9.8) ограничивает «лучшую половину наблюдений»; его границы – нижняя и верхняя квартили. Центральная линия показывает положение медианы (средней квартили). «Усы» показывают размах данных от x_{min} до x_{max} , но не более полутора межквартильного размаха от границ «ящика». Данные, которые выходят за границы «усов», считаются выбросами. В нашем примере имеется два выброса справа. Крестиком отмечено среднее \bar{x} , из-за выбросов оно сдвинуто вправо. Кстати, правило «3-х сигм» не обнаруживает этих выбросов (их отклонения от смещенного среднего не превышают 3-х сигм).

Вопросы для самопроверки

1. Что такое совокупность, генеральная совокупность, выборка?
2. Что такое репрезентативность? Приведите примеры отсутствия репрезентативности.
3. Перечислите основные проблемы математической статистики.
4. Как составляются представительные выборки данных?
5. Что такое статистические оценки? Перечислите основные статистические оценки характеристик и функций распределения.
6. Напишите сравнительные формулы для характеристик случайной величины и для их статистических оценок.
7. Как составляется интервальный вариационный ряд?
8. Что такое гистограмма? Как она строится и преобразуется при укрупнении интервалов?
9. Что такое полигон для непрерывной случайной величины? Чему равняется его площадь? Что означает часть площади полигона (гистограммы) на определенном интервале варьирования случайной величины?
10. Что такое кумулята? Как она строится для непрерывной случайной величины? Для дискретной случайной величины?
11. Как с помощью кумуляты находятся значения (оценки) квартилей?
12. Что показывает блочная диаграмма Тьюкки?

10. Свойства статистических оценок

Доброкачественные оценки должны быть *состоятельными*, *несмещенными* и *эффективными*.

Оценка b генеральной характеристики β называется *состоятельной*, если при увеличении объема выборки она приближается к своей генеральной характеристике: $\lim_{n \rightarrow N} b = \beta$. Если это свойство не выполняется, оценка является дефектной. Несостоятельными оценками пользоваться нельзя.

Все предыдущие оценки были состоятельными.

На основании закона больших чисел доказано, что относительная частота стремится к вероятности, а среднее – к математическому ожиданию $\frac{m}{n} \xrightarrow{(n \rightarrow \infty)} p$, $\bar{x} \xrightarrow{(n \rightarrow \infty)} M(x)$. Иными словами, эти оценки состоятельные.

Но тогда будут состоятельными все оценки, основанные на замене вероятностей на относительные частоты, а математических ожиданий – на средние.

Так, для оценки дисперсии s_x^2 в пределе $(n \rightarrow \infty)$ получаем:

$$s_x^2 = \overline{x^2} - (\bar{x})^2 \xrightarrow{(n \rightarrow \infty)} M(x^2) - M^2(x) = \sigma_x^2.$$

Оценка b генеральной характеристики β называется *несмещенной*, если $M(b) = \beta$. Несмещенные оценки не имеют систематических смещений.

Ранее уже было доказано (в разделе о распределении среднего \mathcal{X}_{cp}), что $M(\bar{x}) = M(x)$, $\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n}$. Первую формулу теперь можно трактовать так: выборочное среднее есть несмещенная оценка математического ожидания.

На рис. 10.1 изображено рассеяние исходных данных вокруг своего центра $M(x)$. Предполагается, что из этих данных случайным образом отбираются по n элементов в различные выборки и вычисляются их средние.

Средние различных случайных выборок рассеяны вокруг своего центра $M(\bar{x})$. Оказывается, что центр группировки выборочных средних совпадает с центром группировки исходных данных

$M(\bar{x}) = M(x)$. Систематического смещения нет.

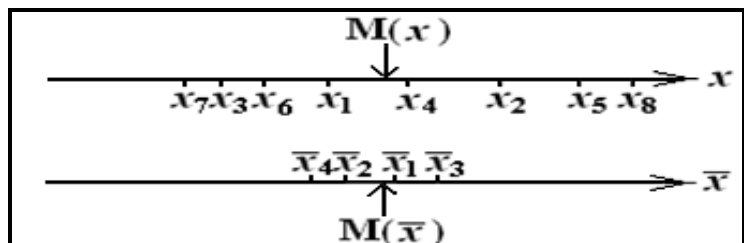


Рис. 10.1. Разбросы данных и средних

Однако выборочная оценка дисперсии уже не обладает несмещенностью, она состоятельна, но систематически занижена $M(s_x^2) < \sigma_x^2$ (рис. 10.2). Дело в том, что в генеральной дисперсии рассматриваются отклонения от центра совокупности, а в выборочной оценке – от центра выборки, а это не одно и то же:

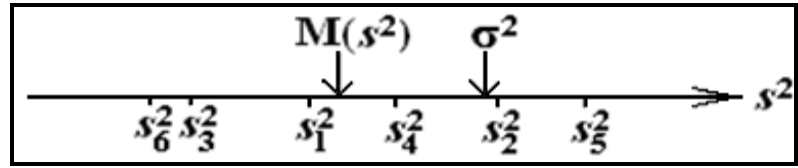


Рис. 10.2. Разброс оценок дисперсии

$$\sigma_x^2 = M(x - M(x))^2, \quad s_x^2 = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^2.$$

Выведем формулу для расчета несмещенной оценки дисперсии.

Записываем формулы для дисперсии исходной случайной величины \mathcal{X} , для дисперсии средних \mathcal{X}_{cp} и для оценки дисперсии s_x^2 :

$$\begin{aligned} \sigma_x^2 &= M(x^2) - M^2(x); \\ \sigma_{\bar{x}}^2 &= M(\bar{x}^2) - M^2(\bar{x}); \\ s_x^2 &= x^2 - (\bar{x})^2. \end{aligned}$$

Вычисляем математическое ожидание от s_x^2 (определяем центр группировки выборочных оценок дисперсии):

$$\begin{aligned} M(s_x^2) &= M(\overline{x^2}) - M(\bar{x}^2) = M(x^2) - M(\bar{x}^2) = \\ &= \{M(x^2) - M^2(x)\} - \{M(\bar{x}^2) - M^2(\bar{x})\} = \sigma_x^2 - \sigma_{\bar{x}}^2. \end{aligned}$$

При преобразованиях дважды использовали факт несмещенности оценки математического ожидания и в одном месте заменили $M(\overline{x^2})$ на $M(x^2)$, а в другом, наоборот, – $M(x)$ на $M(\bar{x})$. Получилось, что центр группировки выборочных оценок дисперсии $M(s_x^2) = \sigma_x^2 - \sigma_{\bar{x}}^2 < \sigma_x^2$ всегда меньше своего предельного значения (генеральной дисперсии). Эта систематическая ошибка уменьшается с увеличением объема выборки (так как оценка дисперсии состоятельна).

Продолжаем преобразования. Используем формулу для дисперсии среднего:

$$M(s_x^2) = \sigma_x^2 - \sigma_{\bar{x}}^2 = \sigma_x^2 - \frac{\sigma_x^2}{n} = \left(1 - \frac{1}{n}\right) \cdot \sigma_x^2 = \frac{n-1}{n} \cdot \sigma_x^2.$$

Введем поправку на несмещенность и получим несмещенную оценку дисперсии в виде:

$$\hat{\sigma}_x^2 = \frac{n}{n-1} \cdot s_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{SS_x}{df_x}.$$

Здесь $SS_x = \sum_{i=1}^n (x_i - \bar{x})^2$ – сумма квадратов n отклонений значений x_i от

центра выборки (SS – *Summa of Squares* – сумма квадратов). Однако не все n отклонений $(x_i - \bar{x})$ являются независимыми – их сумма всегда равна нулю (нулевое, или центральное, свойство среднего). Следовательно, независимых отклонений будет на единицу меньше, последнее отклонение всегда можно найти из выражения $\sum_{i=1}^n (x_i - \bar{x}) = 0$.

Введем понятие «число степеней свободы», которое равно разности количества значений случайной величины и числа наложенных на них линейных связей: $ЧСС = df_x = n - 1$ (df – *degree of freedom* – степени свободы); сейчас у нас одна связь – центральное свойство среднего, поэтому $df_x = n - 1$. Обычная оценка дисперсии равна отношению суммы квадратов (отклонений) к общему числу наблюдений, а несмещенная оценка – отношению суммы квадратов к числу степеней свободы.

Несмещенная оценка дисперсии состоятельная и не имеет систематической ошибки. Для несмещенных оценок дисперсии не выполняется важнейшее свойство дисперсии: несмещенная оценка дисперсии суммы независимых случайных величин больше не равна сумме оценок дисперсий отдельных слагаемых. Поэтому в той или иной форме используются оба вида оценок дисперсии (в англо-американской научной литературе обозначение s_x^2 закреплено за несмещенной оценкой дисперсии, но параллельно с ней оперируют с суммами квадратов SS_x , которые пропорциональны обычным оценкам дисперсии).

Оценка b генеральной характеристики β называется **эффективной**, если она имеет наименьшую дисперсию:

$$\sigma_b^2 \rightarrow \min .$$

Оценка – случайная величина, она зависит от состава случайной выборки. Если оценка неэффективная, то ее дисперсия недопустимо велика, что приводит к нестабильности результатов оценивания. Ошибка оценивания складывается из случайной ошибки и систематического смещения. Мы пытаемся использовать несмещенные оценки, для которых систематического смещения нет. Однако при этом иногда непомерно возрастает случайный разброс несмещенной оценки, из-за чего такая оценка оказывается явно неразумной (бывает даже, что оценка получается с противоположным знаком). Для пользователя слабым утешением является соображение о том, что явная ошибка оценивания является

чисто случайной. Надо искать компромисс, можно допустить небольшую систематическую ошибку, лишь бы при этом *суммарная* ошибка оценивания была небольшой. Иными словами, пусть оценка будет немного смещенной, но более эффективной.

В регрессионном анализе несмещенные оценки параметров модели (по образному выражению К. Доугерти, оценки «инструментов экономического воздействия») получают в результате решения системы уравнений с корреляционной матрицей. Теоретически эта система всегда совместна, то есть всегда имеет решение. Однако когда определитель системы линейных уравнений равен нулю (вырождение), то решение совместной системы становится неединственным, а стандартные числовые алгоритмы решения просто не работают. Еще хуже, когда определитель нулю не равен, но близок к нему. Тогда формально будет получено единственное решение, но оно оказывается нестабильным настолько, что часто не имеет разумной интерпретации. Предложено к диагональным элементам корреляционной матрицы (они равны единице) добавлять малый параметр (число порядка 0,01 – 0,05). Решение такой измененной системы уже будет стабильным, то есть существенно уменьшается дисперсия случайной ошибки (этот эффект легко заметить, произведя серию решений с последовательно увеличивающимся параметром). Однако при добавлении малого параметра система уже становится другой (пусть близкой, но другой). Появляется систематическое смещение между решениями исходной и измененной системами уравнений. Иными словами, оценки параметров модели теряют свойство несмещенности, но становятся более эффективными, в результате чего общая (суммарная) ошибка существенно снижается.

Оценка параметров распределения

До сих пор были рассмотрены оценки характеристик распределения и для самых распространенных характеристик получены готовые вычислительные формулы.

Переходим к оценкам параметров закона распределения (в будущем будем оценивать также параметры эконометрических моделей).

Любой закон распределения зависит от небольшого количества параметров. Если значения параметров известны, то можно вычислить характеристики данного закона распределения. Параметры распределения и характеристики распределения – это разные понятия.

Например, равномерный закон зависит от двух параметров a и b ; зная значения этих параметров, можно вычислить характеристики равномерного распределения, например математическое ожидание и стандартное отклонение:

$$M(x) = \frac{a+b}{2}; \quad \sigma_x = \frac{b-a}{\sqrt{12}}.$$

Чтобы получить оценки параметров, применяется один из трех методов: *моментов, максимального правдоподобия и наименьших квадратов.*

Метод моментов – самый простой из них. Согласно этому методу, надо теоретические характеристики распределения (они являются функциями неизвестных параметров) приравнять их к выборочным оценкам. В результате получаем систему уравнений для определения оценок параметров.

Например, чтобы найти два параметра равномерного закона, приравниваем две важнейшие характеристики (обычно это математическое ожидание и стандартное отклонение) к их выборочным оценкам (среднему и несмещенной оценке стандартного отклонения):

$$\begin{cases} M(x) = \bar{x} \\ \sigma_x = \hat{\sigma}_x \end{cases} \rightarrow \begin{cases} \frac{a+b}{2} = \bar{x} \\ \frac{b-a}{\sqrt{12}} = \hat{\sigma}_x \end{cases} \rightarrow \begin{cases} a = \bar{x} - \sqrt{3} \cdot \hat{\sigma}_x \\ b = \bar{x} + \sqrt{3} \cdot \hat{\sigma}_x \end{cases}.$$

Составили систему двух уравнений относительно двух неизвестных параметров равномерного распределения и получили оценки этих параметров. Резонно их обозначать \hat{a}, \hat{b} , или же, наоборот, теоретические параметры закона переобозначить греческими буквами α, β . Иногда в одном и том же тексте одновременно используются теоретическая функция распределения (например плотность вероятности) $f(x_j)$, ее эмпирическая оценка (гистограмма) \hat{f}_j и теоретическая функция, в которой параметры заменены на их оценки. Для этой последней разновидности оценки функции распределения можно ввести еще какое-нибудь свое обозначение, например $\tilde{f}(x_j)$.

У нормального закона параметры совпадают с основными характеристиками, поэтому систему уравнений не придется решать:

$$\begin{cases} M(x) = a = \bar{x} \\ \sigma_x = \hat{\sigma}_x \end{cases}.$$

Показательный закон зависит от одного параметра λ , характеристики распределения выражаются через этот параметр: $M(x) = \sigma_x = \frac{1}{\lambda}$. Если из теоретических соображений ожидается показательное распределение и $\bar{x} \approx \hat{\sigma}_x$, то параметр определяется из одного уравнения $M(x) = \bar{x} \rightarrow \frac{1}{\lambda} = \bar{x} \rightarrow \lambda = \frac{1}{\bar{x}}$.

Метод максимального правдоподобия – считается наиболее обоснованным и пропагандируется как самый модный в настоящее время способ определения параметров. Заключается он в следующем: для всей системы наблюдений $\{x_i\}$ составляется функция правдоподобия – вероятность или плотность вероятности совместного появления такой системы данных. Функция правдоподобия зависит от параметров предполагаемого закона распределения. Эти параметры следует определять из условий максимума функции правдоподобия.

Например, предполагая показательное распределение, найдем наиболее правдоподобную оценку параметра λ . Функция плотности вероятности этого распределения для наблюдения x_i имеет вид:

$$f(x_i) = \lambda \cdot \exp\{-\lambda x_i\}.$$

Для независимых наблюдений дифференциальная функция их совместного распределения получается как произведение:

$$f(x_1, x_2, \dots, x_n) = \prod (\lambda \cdot \exp\{-\lambda x_i\}) = \lambda^n \cdot \exp\{-\lambda \sum x_i\}.$$

Обычно функцию правдоподобия принимают равной логарифму натурального от этого выражения:

$$\Phi(\lambda) = n \cdot \ln(\lambda) - \lambda \sum x_i.$$

Приравниваем к нулю первую производную от функции правдоподобия $\Phi'(\lambda) = \frac{n}{\lambda} - \sum x_i = 0$, откуда получаем $\lambda = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}$. Именно такую оценку параметра получили ранее методом моментов.

Для нормального закона $f(x_i) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left\{-\frac{(x_i - a)^2}{2\sigma^2}\right\}$, тогда функция правдоподобия (логарифм от произведения функций $f(x_i)$) будет равна:

$$\Phi(a, \sigma) = -n \cdot \ln(\sigma) - \frac{1}{2\sigma^2} \cdot \sum (x_i - a)^2,$$

где постоянное слагаемое $-n \cdot \ln(\sqrt{2\pi})$ отброшено.

Приравниваем к нулю частную производную функции правдоподобия по параметру a :

$$\frac{\partial \Phi}{\partial a} = \frac{2}{2\sigma^2} \cdot \sum (x_i - a) = 0, \text{ откуда } a = \frac{1}{n} \cdot \sum x_i = \bar{x}.$$

Правдоподобная оценка параметра a совпала с оценкой по методу моментов.

Приравниваем к нулю частную производную функции правдоподобия по параметру σ :

$$\frac{\partial \Phi}{\partial \sigma} = -\frac{n}{\sigma} + \frac{2}{2\sigma^3} \cdot \sum (x_i - a)^2 = 0, \text{ откуда } \sigma^2 = \frac{1}{n} \cdot \sum (x_i - a)^2 = s_x^2.$$

Правдоподобная оценка параметра σ^2 оказалась равна обычной (смещенной) оценке дисперсии.

Метод наименьших квадратов применяется, в основном, в тех случаях, когда модель зависит от параметров *линейно*. Будем использовать этот метод для определения параметров регрессионных моделей. Он будет подробно описан в разделе «Регрессионный анализ».

Статистические критерии

Наряду с оценками характеристик и параметров вводят еще некоторые числовые комплексы K , составленные из данных наблюдений. Эти комплексы зависят от состава случайной выборки и поэтому также являются случайными величинами. Если известен (изучен) закон распределения K , то такие комплексы называются статистиками, или критериями. Известны статистики Пирсона, Стьюдента, Фишера (по именам ученых, установивших закон распределения того или иного комплекса). Многие оценки можно называть статистиками, показывая этим, что нам известны законы распределения таких оценок при изменении состава выборки.

Определим зону чисто случайного изменения критерия K , используя практический принцип невозможности редких событий. Выбираем уровень значимости α , малый настолько, что сомневаемся в случайной природе появления события с такой малой вероятностью; мы более склонны считать, что наблюдаемое редкое событие вызвано какими-то внешними воздействиями, это «что-то означает» (вот откуда терминология «уровень значимости»). Обычно принимается уровень значимости 0,1, 0,05 или 0,01. Вероятность противоположного события $P = 1 - \alpha$ называется уровнем доверия. Обычно принимают уровень доверия 0,9, 0,95 или 0,99. Зону чисто случайного изменения критерия

составляют все его значения, которые появляются с вероятностью, большей уровня значимости.

Далее все зависит от особенностей проблемы, для которой составлен тот или иной критерий. Существуют односторонние критерии, когда мы сомневаемся в случайном появлении слишком больших (или наоборот слишком малых) значений K .

Для правостороннего критерия (рис. 10.3а), зная распределение статистики K , находим квантиль K_α из условия $P(\mathcal{X} > K_\alpha) = \alpha$, или $F(K_\alpha) = P(\mathcal{X} \leq K_\alpha) = 1 - \alpha$. Площадь под дифференциальной кривой f_K справа от K_α равна α .

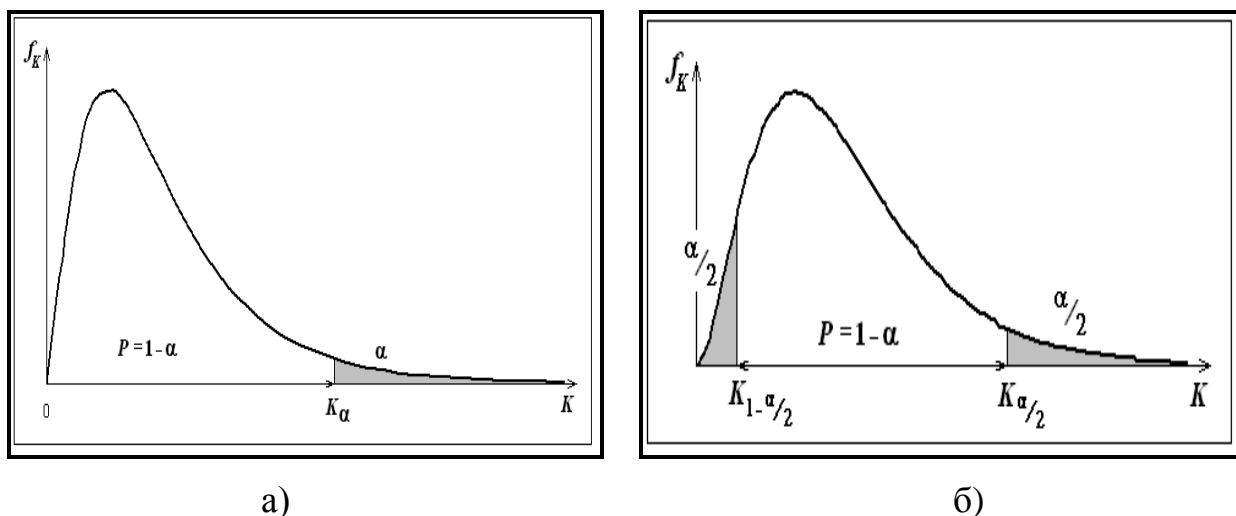


Рис. 10.3. Критическая зона для одностороннего (а) и двустороннего (б) критериев

Если для наших данных окажется, что вычисленное значение K больше критического K_α , нуль-гипотеза о случайности изменения K отвергается и принимается противоположная (альтернативная) гипотеза о неслучайном появлении столь большого K .

Если для наших данных окажется, что вычисленное значение K меньше критического K_α , «нуль-гипотеза» о случайности изменения K не может быть отвергнута.

Для двустороннего критерия (когда сомнительны большие случайные отклонения K и вправо и влево) вычисляются квантили $K_{1-\alpha/2}$ и $K_{\alpha/2}$. Площади под дифференциальной кривой f_K слева от $K_{1-\alpha/2}$ и справа от $K_{\alpha/2}$ одинаковы и равны $\alpha/2$ (рис. 10.3б). Нуль-гипотеза о случайности изменения K не может быть отвергнута, если вычисленное значение критерия попадает в интервал $K_{1-\alpha/2} < K < K_{\alpha/2}$.

Так должно быть и для симметричных распределений статистики K , но в этом случае приняты не совсем правильные обозначения. Критическое значение K_α теперь определяется из условия $P(|\mathcal{X}| > K_\alpha) = \alpha$.

Заметим, что при этих обозначениях площадь под дифференциальной кривой f_K справа от K_α равна $\alpha/2$ (рис. 10.4). Это не совсем правильно, но общепринято.

Границы между областью принятия и областью отбрасывания нуль-гипотезы несколько размыты (рис. 10.5). Любой статистический критерий имеет некоторую область неопределенности, поэтому рекомендуется использовать сразу два уровня значимости (один для принятия, другой для отбрасывания нуль-гипотезы).

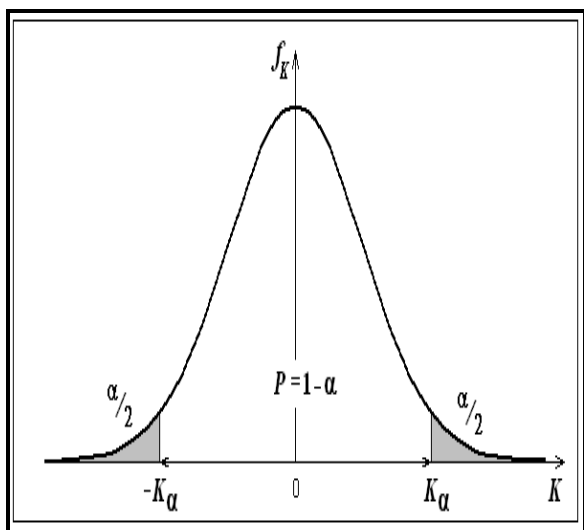


Рис. 10.4. Случай симметричного распределения статистики K

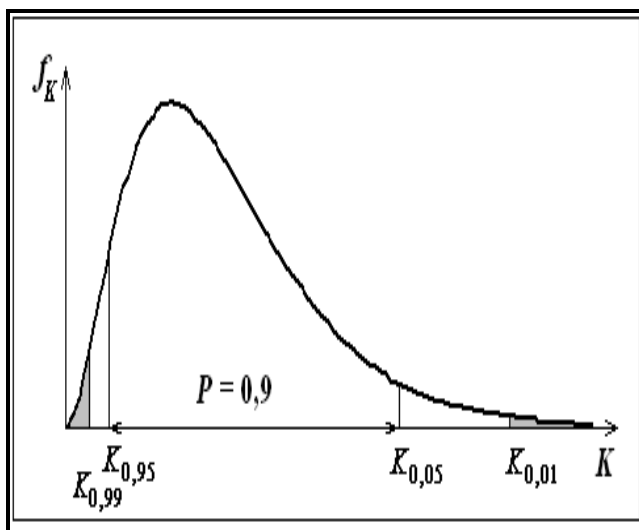


Рис. 10.5. Два уровня значимости (для принятия и отбрасывания нуль-гипотезы)

Уверены, что события с вероятностью $\alpha \leq 0,01$ случайно не происходят, поэтому принимаем такой уровень значимости для отбрасывания нуль-гипотезы. В то же время вероятность $\alpha > 0,05$ уже не может считаться малой, поэтому этот уровень значимости используется для принятия нуль-гипотезы.

Нас могут интересовать самые различные вещи, например, есть ли существенные различия между урожайностью разных сортов пшеницы, между эффективностью различных лекарственных препаратов, однотипной продукцией различных предприятий; нас может интересовать, есть ли значимое воздействие некоторых наших мероприятий на повышение качества и количества производимого продукта, нас крайне интересует надежность и безопасность, здоровье и сохранение среды обитания. Чтобы получить внятные ответы на

наши вопросы, формулируется соответствующая нуль-гипотеза, которая скептически утверждает, что никакого систематического воздействия нет, вся изменчивость определяется чисто случайными флуктуациями, нет никакого значимого различия между сравниваемыми сортами, продукцией разных предприятий, наши лекарства и наши мероприятия не приносят никакого эффекта.

Мы должны оценить вероятность появления наших данных при справедливости нуль-гипотезы и, если эта вероятность не окажется достаточно малой, вынуждены будем сделать огорчительное заключение: «Нуль-гипотеза не может быть отвергнута»; данных мало, чтобы надежно заявить противоположное; такие эффекты могут появляться чисто случайно.

Мы вовсе не утверждаем, что лекарства действительно неэффективны, что сравниваемая продукция действительно эквивалентна и т. п., мы расписываемся в собственной беспомощности – по имеющимся данным ничего определенного сказать нельзя.

Но если вероятность появления данных при справедливости нуль-гипотезы окажется меньше определенного уровня, то нуль-гипотеза отвергается и принимается противоположное утверждение, которое называется альтернативной гипотезой.

При правильно поставленных вопросах альтернативная гипотеза может утверждать, что между подсовкупностями имеются значимые различия (в любую сторону), либо более определенно утверждать, что альтернативное значение параметра больше (или наоборот меньше) того, которое свойственно при нуль-гипотезе.

Все истины, установленные экспериментально, получены в опытах, где нуль-гипотеза была отвергнута (найлены контрпримеры). Тонкий знаток и ценитель природы, писатель М. Пришвин заметил: «Да» природы условное и еле слышимое. «Нет» природы ясное и категоричное».

Государственными стандартами установлено, какую вероятность можно и нужно считать малой. Это уровень значимости $\alpha \leq 0,01$, который является вероятностью «ошибки 1-го рода» – вероятности ошибочно отвергнуть правильную нуль-гипотезу.

В то же время вероятность $\alpha > 0,05$ уже не может считаться малой, иначе мы допустим «ошибку 2-го рода» – ошибочно примем неверную альтернативную гипотезу (в юстиции также различают ошибки «наказать невиновного» и «упустить виновника»; в приемочном контроле различают «риск производителя», когда на основе недостаточного выборочного обследования бракуют всю

партию пригодной продукции, и «риск потребителя» – принимается партия некондиционной продукции).

Поэтому, если вероятность чисто случайного появления наших данных больше 5 % , делается стандартное заключение: нуль-гипотеза не может быть отвергнута (иногда говорят, нуль-гипотеза принимается); если вероятность оказалась меньше 1 % , то «нуль-гипотеза отвергается»; но если эта вероятность больше 1 % и меньше 5 % , делается более осторожное заключение: нуль-гипотеза принимается (или отвергается) *при 5-процентном уровне значимости*.

Наши заключения могут задевать чьи-то интересы, и в последнем спорном случае на нас могут оказывать определенное давление в пользу того или иного вывода; именно поэтому необходима оговорка о 5-процентном уровне значимости.

Вопросы для самопроверки

1. Что такое состоятельность оценок?
2. Что такое несмещенность оценок?
3. Что такое сумма квадратов (отклонений)?
4. Что такое число степеней свободы?
5. Приведите формулу для несмещенной оценки дисперсии.
6. Что такое эффективность оценок?
7. Как оцениваются параметры распределения? Какие для этого существуют методы?
8. Что такое статистики?
9. Как выбирается зона чисто случайного изменения статистики?
10. Как обозначаются критические значения статистик для несимметричных распределений?
11. Как обозначаются критические значения статистик для симметричных распределений?
12. Что такое зона неопределенности критерия?

11. Критерии согласия

Критерий согласия Пирсона

С помощью критериев согласия проверяют гипотезу о соответствии эмпирического распределения предполагаемому теоретическому закону, например, наиболее часто проверяют, можно ли считать наблюдаемое распределение нормальным.

Самый распространенный критерий согласия предложил К. Пирсон, который доказал, что если величины x_i распределены по стандартному нормальному закону $x_i \sim N(0, 1)$ с характеристиками $M(x_i) = 0$ и $\sigma(x_i) = 1$, то сумма их квадратов $\chi^2 = \sum_{i=1}^n x_i^2$ имеет гамма-распределение с вполне определенными па-

раметрами: $f_\nu(x) = \frac{1}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})} \cdot x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}$. Этот частный случай гамма-распределения

называется распределением Пирсона «Хи-квадрат». В общем виде гамма-

распределение $f_\alpha(x) = \lambda \frac{(\lambda x)^{\alpha-1}}{(\alpha-1)!} e^{-\lambda x}$ зависит от двух параметров (λ и α). Ранее

(при изучении композиций случайных величин) мы уже встречались с одним частным случаем гамма-распределения – распределением Эрланга

$f_m(x) = \lambda \frac{(\lambda x)^{m-1}}{(m-1)!} e^{-\lambda x}$ с целочисленным параметром $\alpha = m$. В распределении

Пирсона оба параметра полуцелые: $\lambda = 1/2$, $\alpha = \nu/2$, где $\nu = df$ – число степеней свободы (ЧСС) системы случайных величин $\{x_i\}$ (для независимых величин $df = n$, а для зависимых $df = n - \text{число связей}$). Факториал $(\alpha - 1)!$ для дробных значений α в отечественной научной литературе обозначается как $\Gamma(\alpha)$, где

$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$ – гамма-функция, для которой выполняется соотношение

$\Gamma(\alpha+1) = \alpha \Gamma(\alpha)$. В частности, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, $\Gamma(\frac{3}{2}) = \frac{1}{2} \Gamma(\frac{1}{2})$. Закон распределения

Пирсона однопараметрический (зависит только от параметра $\nu = df$). Типичный график дифференциальной функции распределения показан на рис. 11.1, где $K = \chi^2$. Характеристики закона $M(\chi^2) = \nu$, $D(\chi^2) = 2\nu$.

Для каждого значения $\nu = df$ составлены таблицы квантилей $\chi^2_{0,99}, \chi^2_{0,95}, \chi^2_{0,05}, \chi^2_{0,01}$. Зоной случайного изменения χ^2 является интервал $\chi^2_{0,95} < \chi^2 < \chi^2_{0,05}$ (так называемый 90-процентный доверительный интервал). При увеличении df распределение Пирсона приближается к нормальному, поэтому таблицы квантилей составлены только для $df \leq 30$.

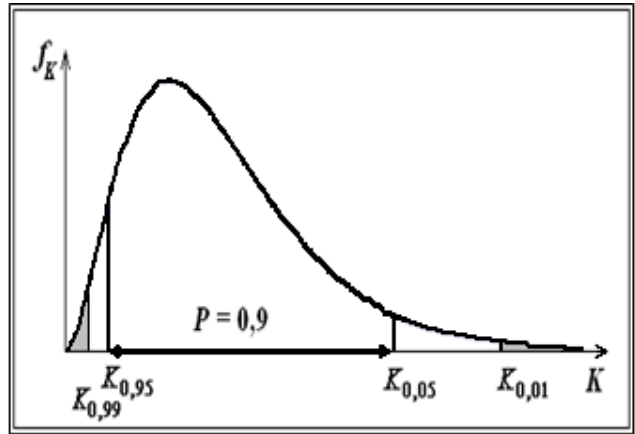


Рис. 11.1. Распределение Пирсона

Для проверки гипотезы о согласии эмпирического распределения предполагаемому теоретическому закону Пирсон составил статистику (критерий), которая обозначается χ^2 :

$$\chi^2 = \sum_{j=1}^k \frac{(m_j - \tilde{m}_j)^2}{\tilde{m}_j},$$

где m_j – наблюдаемые частоты попадания случайной величины в интервалы $s_{j-1} < X \leq s_j$; \tilde{m}_j – ожидаемые частоты по предполагаемому теоретическому закону, в котором неизвестные параметры заменены на их эмпирические оценки. Структура статистики Пирсона – это сумма квадратов отклонений частот от их ожидаемых значений с весами, обратными к \tilde{m}_j (отклонения в одну-две единицы существенны для малых \tilde{m}_j и не существенны для больших \tilde{m}_j).

Покажем, что при выполнении некоторых условий статистика Пирсона распределена по закону χ^2 . Существует некоторая вероятность p_j попадания наблюдений в интервал $(s_{j-1}, s_j]$. Количество таких наблюдений (частота m_j) распределено по закону Бернулли с характеристиками $M(m_j) = np_j$ и $D(m_j) = np_j q_j = np_j(1 - p_j)$. При $n \geq 30$, $np_j \geq 5$ распределение Бернулли уже можно считать нормальным (распределением Лапласа) и тогда величина:

$$\chi^2 = \sum_{j=1}^k \frac{(m_j - np_j)^2}{np_j q_j} = \sum_{j=1}^k \frac{(m_j - np_j)^2}{np_j(1 - p_j)}$$

будет распределена по закону χ^2 . Если интервалы $(s_{j-1}, s_j]$ достаточно узкие, настолько, чтобы можно было пренебречь малыми вероятностями $p_j < 0,1$ по сравнению с единицей, то получаем комплекс:

$$\chi^2 = \sum_{j=1}^k \frac{(m_j - np_j)^2}{np_j},$$

который распределен по закону χ^2 при $m_j \geq 5$, но при $m_j < 0,1 \cdot n$ (в каждый интервал должно попасть не менее 5-ти наблюдений, но меньше 10 % от объема выборки). Эти два несколько противоречивых требования могут быть выполнены одновременно только для выборок большого объема $n > 200$.

Теоретические вероятности попадания наблюдений в заданные интервалы вычисляются с помощью интегральной функции предполагаемого закона $\tilde{p}_i = \tilde{F}(s_j) - \tilde{F}(s_{j-1})$, а ожидаемые частоты – по формуле $\tilde{m}_j = n\tilde{p}_j$.

Отсюда получаем статистику Пирсона в стандартном виде $\chi^2 = \sum_{j=1}^k \frac{(m_j - \tilde{m}_j)^2}{\tilde{m}_j}$. Если снять обременительное требование $m_j < 0,1 \cdot n$, то ста-

тистика слегка усложняется $\chi^2 = \sum_{j=1}^k \frac{(m_j - \tilde{m}_j)^2}{\tilde{m}_j \left(1 - \frac{\tilde{m}_j}{n}\right)}$, но теперь ее можно применять

для выборок умеренного объема $30 \leq n < 200$.

Замена p_j на \tilde{p}_j приводит к тому, что отклонения частот $(m_j - \tilde{m}_j)$ больше не будут независимыми, на них будут наложены две или три связи. Действительно, поскольку $\sum \tilde{p}_j = 1$ (если это не так, следует расширить крайние интервалы – еще одно условие правильного применения критерия Пирсона), то получается, что $\sum (m_j - \tilde{m}_j) = n - n = 0$ – сумма всех отклонений равна нулю. При оценке параметров предполагаемого закона методом моментов приравниваем теоретические характеристики к их выборочным оценкам. Если закон однопараметрический (Пуассона или показательный), один параметр закона оценивается из равенства $M(x) = \bar{x}$, откуда получаем еще одну связь $\sum (m_j - \tilde{m}_j) X_j = n(\bar{x} - M(x)) = 0$, где X_j – центры интервалов. При проверке согласия распределения с однопараметрическим законом число степеней свободы равно $df = k - 2$. Большинство теоретических законов распределения двухпараметрические (Бернулли, нормальный, логнормальный, равномерный, гамма), для них оценку второго параметра получаем, приравнявая дисперсии $\sigma_x^2 = s_x^2$, что приводит еще к одной связи $\sum (m_j - \tilde{m}_j) X_j^2 = 0$, откуда для двухпараметрических законов $df = k - 3$.

Для данного числа степеней свободы по таблицам Пирсона находят квантили $\chi^2_{0,99}, \chi^2_{0,95}, \chi^2_{0,05}, \chi^2_{0,01}$. Если окажется, что вычисленное значение статистики Хи-квадрат находится в пределах $\chi^2_{0,95} < \chi^2 < \chi^2_{0,05}$, нуль-гипотеза о случайности расхождений между наблюдаемыми и ожидаемыми частотами не может быть отвергнута; предполагаемый теоретический закон не противоречит данным; можно считать, что он именно такой и можно использовать его для дальнейших вычислений. Уровень доверия нашего заключения $P = 0,9$ (90 %). Если окажется, что вычисленное значение статистики Хи-квадрат больше большей границы $\chi^2 > \chi^2_{0,01}$, нуль-гипотеза отвергается; предполагаемый теоретический закон не согласуется с данными, расхождения между наблюдаемыми и ожидаемыми частотами слишком велики, распределение неудовлетворительно описывается этим законом. Однако нуль-гипотеза отвергается также при слишком хорошем соответствии, когда вычисленное значение статистики Хи-квадрат оказывается меньше меньшей границы $\chi^2 < \chi^2_{0,99}$; в этом случае сомневаемся в достоверности данных, по всей видимости здесь имеется какая-то фальсификация; вероятность такого полного соответствия при справедливости нуль-гипотезы меньше 1 %, а такое событие является практически невероятным (невозможным).

Рассмотрим пример применения критерия согласия Пирсона.

На рис. 11.2 изображены гистограмма, полигон (графически сглаженная гистограмма) и кривая нормального распределения, параметры которой оценены методом моментов. Можно ли считать, что эмпирическое распределение нормальное?

На рис. 11.3 приведен интервальный вариационный ряд с шагом группировки $h = 0,2$. Сумма наблюдаемых частот равна $n = \sum m_j = 75$, оценки характеристик: $\bar{x} = 1,297$, $s_x = 0,433$.

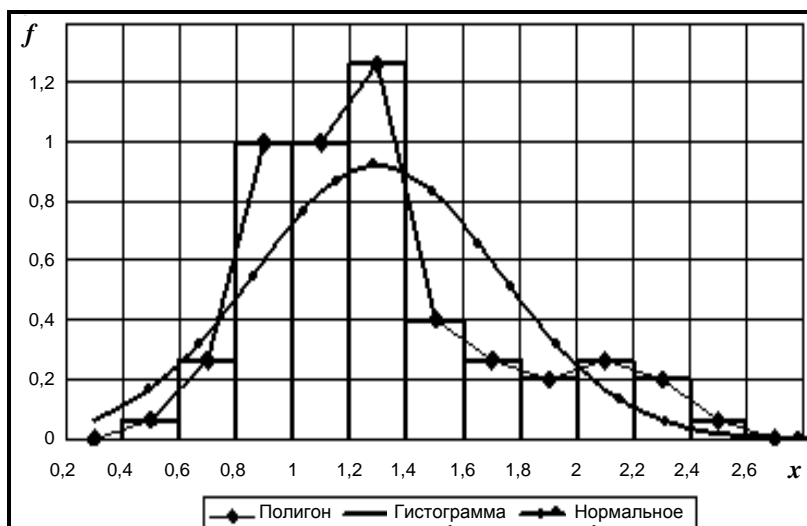


Рис. 11.2. Дифференциальные функции распределения

X	0,5	0,7	0,9	1,1	1,3	1,5	1,7	1,9	2,1	2,3	2,5
	0,4–0,6	0,6–0,8	0,8–1,0	1,0–1,2	1,2–1,4	1,4–1,6	1,6–1,8	1,8–2,0	2,0–2,2	2,2–2,4	2,4–2,6
m	1	4	15	15	19	6	4	3	4	3	1
\tilde{p}	0,034	0,072	0,121	0,165	0,183	0,164	0,119	0,070	0,034	0,013	0,004
	0,053	0,072	0,121	0,165	0,183	0,164	0,119	0,070	0,034	0,013	0,005
\tilde{m}	4,0	5,4	9,0	12,4	13,7	12,3	9,0	5,3	2,5	1,0	0,4
		9,4	9,0	12,4	13,7	12,3	9,0	9,2			
m		5	15	15	19	6	4	11			
Хи-кв		2,06	4,00	0,55	2,05	3,23	2,78	0,35			
Исправ		2,36	4,55	0,65	2,51	3,86	3,16	0,40			

Рис. 11.3. Расчеты по критерию Пирсона

В строке \tilde{p} (верхний ряд цифр) вычислены вероятности попадания наблюдений в каждый интервал по формуле:

$$\tilde{p} = \Phi(t_{X+h/2}) - \Phi(t_{X-h/2}),$$

где Φ – интегральная функция Лапласа, $t_x = \frac{x-\bar{x}}{s_x}$.

Сумма этих вероятностей оказалась равной 0,979, то есть меньше единицы. Расширяем крайние интервалы и для первого интервала с центром $X = 0,5$ вычисляем $\tilde{p} = \Phi(t_{X+h/2}) - \Phi(-\infty) = \Phi(t_{X+h/2}) + 0,5$, а для последнего с центром $X = 2,5$ – $\tilde{p} = \Phi(\infty) - \Phi(t_{X-h/2}) = 0,5 - \Phi(t_{X-h/2})$. Исправленные значения приведены в строке \tilde{p} (нижний ряд цифр). Сумма исправленных вероятностей равна единице.

В строке $\tilde{m} = n\tilde{p}$ (верхний ряд цифр) вычислены теоретические частоты, которые ожидаются согласно нормальному распределению. Сумма этих частот равна $n = 75$.

Для правильного применения критерия Пирсона малонасыщенные интервалы следует объединить с соседними так, чтобы в каждый укрупненный интервал попало не менее 5-ти наблюдений.

Укрупняем первые два интервала и последние четыре (при укрупнении частоты складываются). Укрупненные теоретические частоты записаны в строке \tilde{m} (нижний ряд цифр).

Для сравнения в следующей строке приведены укрупненные наблюдаемые частоты m .

Далее в строке «*Хи-кв*» вычислены отдельные слагаемые $\frac{(m - \tilde{m})^2}{\tilde{m}}$, а в последней строке «*Исправ*» – с поправкой на малый объем выборки $\frac{(m - \tilde{m})^2}{\tilde{m}\left(1 - \frac{\tilde{m}}{n}\right)}$.

Стандартное значение критерия получилось равным $\chi^2 = 15,01$. Исправленное значение оказалось несколько большим $\chi^2 = 17,48$. У нас было $k = 7$ укрупненных интервалов (7 пар частот для сравнения). Нормальный закон двухпараметрический, поэтому число степеней свободы равно $df = 7 - 3 = 4$. Для этого значения числа степеней свободы из таблицы Пирсона выписываем критические значения: $\chi_{0,99}^2 = 0,30$; $\chi_{0,95}^2 = 0,71$; $\chi_{0,05}^2 = 9,49$; $\chi_{0,01}^2 = 13,28$. Оба вычисленных значения статистики Пирсона χ^2 (стандартная – 15,01 и исправленная – 17,48) оказались больше большей критической границы (13,28), следовательно, гипотеза о нормальности распределения отвергается.

Критерий согласия Колмогорова – Смирнова

Это самый простой и самый ненадежный критерий. Основан он на сравнении интегральных функций распределения – теоретической $F(s_j)$ и эмпирической \hat{F}_j (кумуляты). Статистика Колмогорова – Смирнова имеет вид: $KS = \max |F(s_j) - \hat{F}_j| \cdot \sqrt{n}$. Если вычисленное значение KS оказывается меньше 1,36, теоретический закон принимается; если же KS оказывается больше 1,63 – отвергается.

Очень простой критерий, но есть одна маленькая деталь: теоретический закон должен быть известен полностью, включая знание значений его параметров.

При применении критерия Пирсона неизвестные параметры теоретического закона заменялись на свои выборочные оценки и вместо теоретической интегральной функции $F(s_j)$ использовалась интегральная функция $\tilde{F}(s_j)$, которая «подогнана под данные выборки». Из-за этого сравниваемые законы получались более согласованными, чем это есть в действительности, но этот нежелательный эффект нейтрализовывался соответствующим уменьшением числа степеней свободы.

Если же, применяя критерий Колмогорова – Смирнова, заменяем параметры распределения на его оценки, то верить критерию KS можно, когда он

отвергает предполагаемый закон. Критерий KS слишком «либеральный», и поэтому часто принимаются не совсем верные теоретические законы. Так, для примера, который был рассмотрен выше, критерий KS не нашел существенных различий в эмпирическом \hat{F}_j и нормальном $\tilde{F}(s_j)$ распределениях.

Интервальные оценки характеристик и параметров

Оценки, которыми мы пользовались до сих пор, представляли собой одно число, поэтому они называются точечными. Выше уже указывалось, что многие оценки являются статистиками, подтверждая этим, что известен закон их распределения.

В таком случае можно определить границы случайного изменения оценки K (границы доверительного интервала с заданным уровнем доверия $P = 1 - \alpha$) как квантили $K_{\alpha/2}$ и $K_{1-\alpha/2}$ (рис. 11.4).

Доверительный интервал $K_{1-\alpha/2} < K < K_{\alpha/2}$ со случайными границами (выборочные квантили) накрывает неизвестное генеральное значение статистики с заданным уровнем доверия P и не накрывает – с уровнем значимости α (то есть с малой вероятностью α наши интервальные оценки могут быть ошибочными).

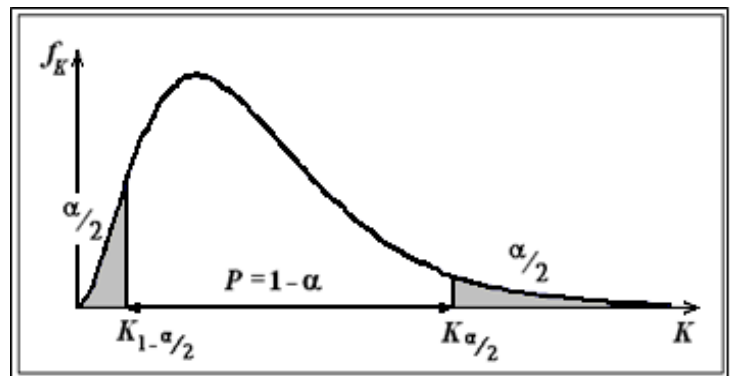


Рис. 11.4. Доверительный интервал с уровнем доверия $P=1 - \alpha$

Впервые интервальную оценку (для дисперсии) построил К. Пирсон.

Рассмотрим случайные величины x_i , распределенные по нормальному закону с одинаковыми характеристиками: $x_i \sim N(a; \sigma_x)$. Известно, что среднее этих случайных величин \bar{x} также имеет нормальное распределение с тем же центром $M(\bar{x}) = M(x)$. Тогда случайные величины $\frac{x_i - \bar{x}}{\sigma_x} \sim N(0; 1)$ будут распределены по стандартному нормальному закону с нулевым математическим ожиданием и единичной дисперсией, а сумма их квадратов – по закону Пирсона:

$$\chi^2 = \frac{\sum (x_i - \bar{x})^2}{\sigma_x^2} = \frac{n \cdot s_x^2}{\sigma_x^2}.$$

Строим доверительный 90-процентный интервал на эту статистику $\chi^2_{0,95} < \frac{n \cdot s_x^2}{\sigma_x^2} < \chi^2_{0,05}$, откуда получаем 90-процентный доверительный интервал для дисперсии:

$$\frac{n}{\chi^2_{0,05}} \cdot s_x^2 < \sigma_x^2 < \frac{n}{\chi^2_{0,95}} \cdot s_x^2.$$

С уровнем доверия $P = 0,9$ этот интервал со случайными границами накрывает неизвестное нам значение генеральной дисперсии.

Математическое ожидание статистики χ^2 равно df_x – числу степеней свободы разностей $(x_i - \bar{x})$:

$$M(\chi^2) = \frac{n \cdot M(s_x^2)}{\sigma_x^2} = df_x.$$

Отсюда получаем несмещенную оценку дисперсии в виде:

$$\hat{\sigma}_x^2 = \frac{n}{df_x} \cdot s_x^2 = \frac{SS_x}{df_x}, \text{ так как } M(\hat{\sigma}_x^2) = \frac{n}{df_x} \cdot M(s_x^2) = \sigma_x^2.$$

Осталось определить число степеней свободы, от которого зависят границы доверительного интервала. Если случайные величины взаимно независимые, то для разностей $(x_i - \bar{x})$ имеется только одна связь $\sum (x_i - \bar{x}) = 0$ (центральное свойство среднего). В этом частном случае $df_x = n - 1$. Если же на разности $(x_i - \bar{x})$ наложены еще какие-то связи (их количество обозначим через l), то $df_x = n - l - 1$.

Вывод дифференциальной функции распределения Пирсона

Сначала рассмотрим распределение суммы квадратов $\chi^2 = \sum_{i=1}^n x_i^2$ независимых случайных величин, каждая из которых распределена по стандартному нормальному закону с плотностью вероятности $f(x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}}$.

Совместную плотность вероятности величин x_1, x_2, \dots, x_n получаем в виде произведения: $f(x_1)f(x_2) \cdots f(x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{\sum x_i^2}{2}}$. Запишем выражение для интегральной функции распределения Пирсона в виде многомерного интеграла

$F(x) = P(\chi^2 \leq x) = \int \cdots \int_{\sum x_i^2 \leq x} C_1 \exp\left\{-\frac{1}{2} \sum x_i^2\right\} \cdot dx_1 dx_2 \cdots dx_n$, где областью интегрирования является внутренность многомерной сферы радиуса \sqrt{x} с центром в начале координат.

Для определения дифференциальной функции $f(x) = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \rightarrow 0} \frac{\Delta F(x)}{h}$ преобразуем числитель этого выражения:

$\Delta F(x) = \int \cdots \int_{x < \sum x_i^2 \leq x+h} C_1 \exp\left\{-\frac{1}{2} \sum x_i^2\right\} \cdot dx_1 dx_2 \cdots dx_n$. Согласно теореме о среднем,

$$\Delta F(x) = C_1 \exp\left\{-\frac{1}{2}(x + \theta h)\right\} \cdot \int \cdots \int_{x < \sum x_i^2 \leq x+h} dx_1 dx_2 \cdots dx_n \quad (\text{где } 0 < \theta < 1).$$

В интеграле $I(z) = \int \cdots \int_{\sum x_i^2 \leq z} dx_1 dx_2 \cdots dx_n$ сделаем замену переменных

$$x_i = y_i \sqrt{z} \quad \text{и преобразуем его к виду } I(z) = \sqrt{z}^n \cdot \int \cdots \int_{\sum y_i^2 \leq 1} dy_1 dy_2 \cdots dy_n = C_2 \cdot \sqrt{z}^n,$$

где константа C_2 от z не зависит. Таким образом, для малых h имеем: $\Delta F(x) = C_1 \exp\left\{-\frac{1}{2}(x + \theta h)\right\} \cdot [I(x+h) - I(x)] = C_1 C_2 \exp\left\{-\frac{1}{2}(x + \theta h)\right\} \cdot \left[\sqrt{(x+h)^n} - \sqrt{x^n}\right]$.

Отсюда следуют искомая формула для дифференциальной функции (плотности вероятности):

$$f(x) = \lim_{h \rightarrow 0} \frac{\Delta F(x)}{h} = C_1 C_2 \exp\left\{-\frac{x}{2}\right\} \cdot \frac{d\sqrt{x^n}}{dx} = C_0 x^{\frac{n}{2}-1} e^{-\frac{x}{2}}.$$

Константа C_0 определяется из условия равенства единице площади под дифференциальной кривой.

Таким образом, вывели закон распределения суммы квадратов $\chi^2 = \sum_{i=1}^n x_i^2$

для частного случая *независимых* случайных величин $x_i \sim N(0, 1)$.

Теперь рассмотрим сумму квадратов $\chi^2 = \sum_{i=1}^n x_i^2$ случайных величин, рас-

пределенных по стандартному нормальному закону, но на которые наложены l линейных связей (так что *независимых* случайных величин будет $\nu = n - l$). Всегда возможно систему n зависимых величин x_i заменить на систему меньшего числа ($\nu = n - l$) *независимых* (некоррелированных) величин u_i так, чтобы сум-

ма их квадратов χ^2 не изменилась $\sum_{i=1}^n x_i^2 = \sum_{i=1}^{\nu} u_i^2$.

В действительности эти преобразования («ортогональные преобразования») делать не придется, достаточно знать, что они возможны. Продемонстрируем вышесказанное на примере. Пусть $n = 4$ величины (случайных или неслучайных) связаны одним соотношением $x_1 + x_2 + x_3 + x_4 = 0$.

Введем новые величины следующим образом:

$$\begin{aligned} u_1 &= (x_1 - x_2 - x_3 + x_4) / 2; \\ u_2 &= (x_1 - x_2 + x_3 - x_4) / 2; \\ u_3 &= (x_1 + x_2 - x_3 - x_4) / 2; \\ u_4 &= (x_1 + x_2 + x_3 + x_4) / 2. \end{aligned}$$

Последняя величина в новой системе равна нулю $u_4 \equiv 0$.

Убедимся, что после указанного преобразования сумма квадратов не изменилась:

$$\begin{aligned} &u_1^2 + u_2^2 + u_3^2 + u_4^2 = \\ &= 0,25 \cdot (x_1^2 + x_2^2 + x_3^2 + x_4^2 - 2x_1x_2 - 2x_1x_3 + 2x_1x_4 + 2x_2x_3 - 2x_2x_4 - 2x_3x_4 + \\ &\quad x_1^2 + x_2^2 + x_3^2 + x_4^2 - 2x_1x_2 + 2x_1x_3 - 2x_1x_4 - 2x_2x_3 + 2x_2x_4 - 2x_3x_4 + \\ &\quad x_1^2 + x_2^2 + x_3^2 + x_4^2 + 2x_1x_2 - 2x_1x_3 - 2x_1x_4 - 2x_2x_3 - 2x_2x_4 + 2x_3x_4 + \\ &\quad x_1^2 + x_2^2 + x_3^2 + x_4^2 + 2x_1x_2 + 2x_1x_3 + 2x_1x_4 + 2x_2x_3 + 2x_2x_4 + 2x_3x_4) = \\ &= x_1^2 + x_2^2 + x_3^2 + x_4^2. \end{aligned}$$

Если случайные величины x_i распределены по стандартному нормальному закону $x_i \sim N(0, 1)$, то также будут распределены новые случайные величины u_i (за исключением $u_4 = \text{Const} = 0$).

Таким образом, при наличии связей, наложенных на x_i , параметр n в распределении Пирсона надо заменять на число степеней свободы ν .

Вопросы для самопроверки

1. Что такое критерий согласия?
2. Перечислите требования к применению критерия Пирсона.
3. Что такое распределение Пирсона?
4. Сформулируйте критерий согласия Колмогорова – Смирнова.
5. Что такое точечные и интервальные оценки?
6. Приведите интервальную оценку дисперсии.

12. Проверка статистических гипотез

Распределение Стьюдента

Пусть z и V – независимые случайные величины, где z распределено по стандартному нормальному закону $z \sim N(0; 1)$, а V – по закону χ^2 с числом степеней свободы df . Английский статистик Госсет (псевдоним – Стьюдент) изучил распределение комплекса $t = \frac{z}{\sqrt{V/df}}$. В частности, распределению Стьюдента подчиняется статистика $t_b = \frac{b - \beta}{\hat{\sigma}_b}$, где величина b распределена нормально с математическим ожиданием β : $b \sim N(\beta; \sigma_b)$. Поскольку выборочное среднее распределено нормально (на основании центральной предельной теоремы) $\bar{x} \sim N\left(a; \frac{\sigma_x}{\sqrt{n}}\right)$, то по Стьюденту также распределена статистика $t_{\bar{x}} = \frac{\bar{x} - a}{\frac{\hat{\sigma}_x}{\sqrt{n}}}$. Заметим, что ранее (при изучении распределений Лапласа и Гаусса)

мы через t_x обозначали стандартизованную величину $t_x = \frac{x - a}{\sigma_x}$, где в знаменателе стоит генеральное стандартное отклонение σ_x . Теперь предлагается переобозначить эту величину на z , а обозначение t_x закрепить за комплексом $t_x = \frac{x - a}{\hat{\sigma}_x}$, где в знаменателе стоит несмещенная оценка стандартного отклонения.

Распределение Стьюдента зависит только от числа степеней свободы $k = df$. Ее функция плотности вероятности (дифференциальная функция распределения) имеет вид: $f_k(x) = B_k \cdot \left(1 + \frac{x^2}{\nu}\right)^{-\frac{k+1}{2}}$.

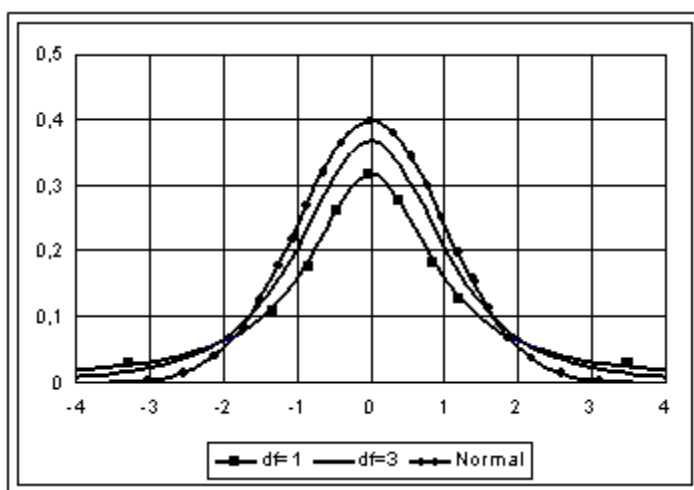


Рис. 12.1. Распределение Стьюдента

При увеличении ЧСС ($k = df$) распределение быстро приближается к нормальному (рис. 12.1).

Напомним характерные особенности нормального распределения: оно симметричное, одномодальное, для него выполняется правило «2-х сигм», а именно: с уровнем доверия $P = 0,95$ случайные отклонения от центра не превосходят $2\sigma_x$ (вернее, $1,96\sigma_x$). Распределение Стьюдента симметричное и одномодальное, но правило «2-х сигм» выполняется только для $df > 30$; для меньших значений ЧСС с уровнем доверия $P = 0,95$ случайные отклонения от центра не превосходят $t_{0,05} \cdot \sigma_x$, где квантиль $t_{0,05}$ надо определять по таблицам Стьюдента (рис. 12.2) в зависимости от df :

df	1	2	3	4	5	6	7	8	9	10	15	20	25	30	∞
$t_{0,05}$	12,7	4,3	3,2	2,8	2,6	2,4	2,4	2,3	2,3	2,2	2,1	2,1	2,1	2,0	1,96

Рис. 12.2. Фрагмент таблицы распределения Стьюдента

Как видно из таблицы на рис. 12.2, нарушение правила «2-х сигм» в распределении Стьюдента весьма существенно только для малых ЧСС.

Напоминаем общепринятые обозначения критических значений для симметричных распределений: с уровнем доверия $P = 1 - \alpha$ выполняется условие $|t| \leq t_\alpha$, то есть общепринятое здесь обозначение t_α не соответствует стандартному обозначению квантиля с уровнем значимости α (из-за модуля в неравенстве $|t| \leq t_\alpha$. Площади под дифференциальной кривой симметричного распределения справа от t_α и слева от $-t_\alpha$ одинаковы и равны $\alpha/2$.

Интервальная оценка для математического ожидания

Как уже указывалось выше, статистика $t_{\bar{x}} = \frac{\bar{x} - a}{\frac{\hat{\sigma}_x}{\sqrt{n}}}$ распределена по закону

Стьюдента с $df = n - 1$. Тогда с уровнем доверия $P = 1 - \alpha$ выполняется условие $|\bar{x} - a| \leq t_\alpha \frac{\hat{\sigma}_x}{\sqrt{n}}$. Это неравенство можно разрешить относительно a : $\bar{x} - t_\alpha \frac{\hat{\sigma}_x}{\sqrt{n}} \leq a \leq \bar{x} + t_\alpha \frac{\hat{\sigma}_x}{\sqrt{n}}$. Получили доверительный интервал со случайными границами, который с вероятностью $P = 1 - \alpha$ накрывает неизвестное значение математического ожидания (центра всей совокупности). По результатам обследования относительно малой выборки сделано заключение о важнейшей характеристике генеральной совокупности.

Математическое ожидание оценивается с уровнем доверия $P = 1 - \alpha$ и погрешностью $\varepsilon = t_\alpha \frac{\hat{\sigma}_x}{\sqrt{n}}$. Погрешность можно выразить в процентах от \bar{x} (относительная погрешность):

$$\delta = \frac{\varepsilon}{\bar{x}} \cdot 100\% = t_\alpha \frac{v_x}{\sqrt{n}}, \text{ где } v_x = \frac{\hat{\sigma}_x}{\bar{x}} \cdot 100\% - \text{коэффициент вариации.}$$

Теперь можно сформулировать три вида стандартных задач:

1. При заданном уровне доверия P можно определить погрешность оценки математического ожидания ε .
2. При заданной погрешности ε можно найти уровень доверия P .
3. Можно определить объем выборки, для которого с заданной надежностью (уровнем доверия $P = 1 - \alpha$) погрешность в оценке математического ожидания не превзойдет некоторого заданного значения; предельное значение погрешности обычно задается в процентах $\delta \leq q\%$.

Рассмотрим решение этой задачи. Из выражения для относительной погрешности имеем:

$$\delta = \frac{\varepsilon}{\bar{x}} \cdot 100\% = t_\alpha \frac{v_x}{\sqrt{n}} \leq q, \text{ откуда } n \geq t_\alpha^2 \cdot \left(\frac{v_x}{q}\right)^2.$$

Полученное неравенство еще предстоит решать итерациями, так как табличное значение квантиля t_α зависит от ЧСС, которое здесь равно $df = n - 1$.

Пример. Пусть принят уровень доверия $P = 0,95$ (соответственно, уровень значимости $\alpha = 0,05$). Предельная относительная погрешность принята равной $q = 5\%$. Тогда $n \geq t_{0,05}^2 \cdot \left(\frac{v_x}{5}\right)^2$.

$$\text{Если } v_x = 20\%, \text{ то } n \geq t_{0,05}^2 \cdot \left(\frac{20}{5}\right)^2 = 16 \cdot t_{0,05}^2.$$

Принимаем $t_{0,05} = 2$ (предельное значение для больших n) и получаем $n \geq 16 \cdot 2^2 = 64$. Проверяем: $df = n - 1 = 63$, далее по таблице Стьюдента находим $t_{0,05}(63) = 2$. Процесс закончился за одну итерацию: $n = 64$.

$$\text{Если } v_x = 10\%, \text{ то } n \geq t_{0,05}^2 \cdot \left(\frac{10}{5}\right)^2 = 4 \cdot t_{0,05}^2.$$

Принимаем $t_{0,05} = 2$ и получаем $n \geq 4 \cdot 2^2 = 16$. Проверяем: $df = n - 1 = 15$, далее по таблице Стьюдента находим $t_{0,05}(15) = 2,1$. Новое значение $n \geq 4 \cdot 2,1^2 = 17,6$ дает верхнюю границу $n = 18$. Проверяем среднее значение $n = 17$, $df = 17 - 1 = 16$; далее по таблице Стьюдента находим $t_{0,05}(16) = 2,1$. Процесс закончился за две итерации: $n = 17$.

$$\text{Если } v_x = 5\%, \text{ то } n \geq t_{0,05}^2 \cdot \left(\frac{5}{5}\right)^2 = t_{0,05}^2.$$

Принимаем $t_{0,05} = 2$ и получаем $n \geq 2^2 = 4$. Проверяем: $df = 4 - 1 = 3$, далее по таблице Стьюдента находим $t_{0,05}(3) = 3,2$. Новое значение $n \geq 3,2^2 = 10,2$ дает верхнюю границу $n = 11$. Проверяем среднее значение $n = 8$, $df = 8 - 1 = 7$; далее по таблице Стьюдента находим $t_{0,05}(7) = 2,4$. Новое значение $n \geq 2,4^2 = 5,8$ дает границу $n = 6$. Проверим эту границу: $n = 6$, $df = 6 - 1 = 5$, $t_{0,05}(5) = 2,6$, $n \geq 2,6^2 = 6,8$, откуда получаем $n = 7$. Процесс закончился за пять итераций: $n = 7$.

Если $v_x = 2\%$, то $n \geq t_{0,05}^2 \cdot \left(\frac{2}{5}\right)^2 = 0,16 \cdot t_{0,05}^2$. Принимаем $t_{0,05} = 2$ и получаем $n \geq 0,16 \cdot 2^2 = 0,64$. Принимаем $n = 2$, $df = 2 - 1 = 1$, $t_{0,05}(1) = 12,7$, $n \geq 0,16 \cdot 12,7^2 = 25,8$. Это явно завышенная оценка, поэтому проверяем следующее значение $n = 3$, $df = 3 - 1 = 2$, $t_{0,05}(2) = 4,3$, $n \geq 0,16 \cdot 4,3^2 = 3,0$. Процесс закончился. Достаточно $n = 3$.

Для еще меньших значений коэффициента вариации $v_x < 2\%$ случайную величину можно считать постоянной (ее изменчивостью можно пренебречь).

Проверка гипотезы о равенстве центров двух совокупностей

Пусть две совокупности представлены выборками (рис. 12.3) с известными средними и оценками дисперсий. Оценки дисперсий могут быть обычными s_k^2 или несмещенными $\hat{\sigma}_k^2$; для любого вида оценок дисперсии вычисляются

суммы квадратов отклонений $SS_k = \sum_{i=1}^{n_k} (x_i - \bar{x}_k)^2$ по формулам:

$$SS_k = n_k s_k^2 \text{ или } SS_k = (n_k - 1) \hat{\sigma}_k^2.$$

Характеристики	Выборка 1	Выборка 2
Объемы выборок	n_1	n_2
Средние	\bar{x}_1	\bar{x}_2
Оценки дисперсий	s_1^2	s_2^2
Несмещенные оценки	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
Суммы квадратов	$SS_1 = n_1 s_1^2 = (n_1 - 1) \hat{\sigma}_1^2$	$SS_2 = n_2 s_2^2 = (n_2 - 1) \hat{\sigma}_2^2$

Рис. 12.3. Характеристики двух выборок

Совокупности предполагаются независимыми с примерно одинаковой изменчивостью, иными словами, считается, что совокупности могут различаться только центрами (математическими ожиданиями a_1, a_2).

Последнее предположение о равенстве дисперсий совокупностей можно проверить по критерию Фишера, который будет рассмотрен позже.

Для малых выборок оценки дисперсий могут оказаться малонадежными, поэтому вычисляем объединенную дисперсию:

$$\hat{\sigma}_0^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2}.$$

Число степеней свободы для объединенной дисперсии равно $(n_1 + n_2 - 2)$, так как на отклонения $(x_i - \bar{x}_k)$ наложено 2 связи – сумма первых n_1 отклонений равна нулю и сумма следующих n_2 отклонений также равна нулю (центральное свойство средних).

Согласно центральной предельной теореме, выборочные средние распределены асимптотически нормально:

$$\bar{x}_k \sim N\left(a_k; \frac{\sigma_0}{\sqrt{n_k}}\right).$$

Разность выборочных средних $\Delta = \bar{x}_1 - \bar{x}_2$ также распределена нормально с характеристиками $M(\Delta) = a_1 - a_2$, $D(\Delta) = \sigma_0^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$, так как дисперсия суммы (и разности) *независимых* случайных величин равна сумме дисперсий. В таком случае статистика

$$t_\Delta = \frac{\Delta - M(\Delta)}{\hat{\sigma}_\Delta} = \frac{(\bar{x}_1 - \bar{x}_2) - (a_1 - a_2)}{\sqrt{\frac{SS_1 + SS_2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

имеет t -распределение Стьюдента с числом степеней свободы $df = (n_1 + n_2 - 2)$.

Нуль-гипотеза заключается в предположении $a_1 = a_2$.

Если окажется, что $|t_\Delta| = \frac{|\Delta|}{\hat{\sigma}_\Delta} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 n_2}}} \leq t_{0,05}(n_1 + n_2 - 2)$, нуль-

гипотеза не может быть отвергнута и с уровнем доверия $P = 0,95$ принимаем гипотезу о равенстве центров совокупностей. Нуль-гипотеза отвергается при $|t_\Delta| \geq t_{0,01}$. Остальные значения $t_{0,05} < t_\Delta < t_{0,01}$ попадают в область неопределенности критерия.

Для примера сравним средние 4-х выборок, характеристики которых приведены в таблице на рис. 12.4.

Выборки	Объемы, n_k	Средние, \bar{x}_k	Несмещенные оценки, $\hat{\sigma}_k^2$	Суммы квадратов
1	103	32,368	31,004	3162,4
2	3	27,367	67,704	135,4
3	30	22,243	24,566	712,4
4	17	19,247	8,253	132,0

Рис. 12.4. Характеристики 4-х выборок

Для анализа m групп понадобится сравнить $C_m^2 = \frac{m(m-1)}{2}$ пар; в данном случае получается 6 пар (рис. 12.5).

Группы $i-j$	n_i	n_j	df_{ij}	Δ_{ij}	$\hat{\sigma}_{\Delta}^2$	t_{Δ}	$t_{0,05}$	$t_{0,01}$
1 – 2	103	3	104	5,001	9,537	1,62	1,98	2,62
1 – 3	103	30	131	10,125	1,197	9,26	1,98	2,61
1 – 4	103	17	118	13,121	1,905	9,51	1,98	2,62
2 – 3	3	30	31	5,123	10,193	1,60	2,04	2,74
2 – 4	3	17	18	8,120	10,902	2,46	2,10	2,88
3 – 4	30	17	45	2,996	2,562	1,87	2,01	2,69

Рис. 12.5. Сравнение 6-ти пар выборок

В таблице на рис. 12.5 вычислены ЧСС (df_{ij}), разности $\Delta_{ij} = |\bar{x}_i - \bar{x}_j|$, несмещенные оценки дисперсий этих разностей $\hat{\sigma}_{\Delta}^2$, статистики Стьюдента t_{Δ} и табличные значения $t_{0,05}(df)$, $t_{0,01}(df)$. Значимые разности, для которых $t_{\Delta} > t_{0,05}$, выделены в таблице, откуда видно, что 1-я группа значимо отличается от 3-й и 4-й, 2-я группа значимо отличается от 4-й. Этим выводам явно не хватает наглядности. Трудно представить себе общую картину, особенно при сравнении большого количества групп (например, для 10 групп будет уже 45 сравнений).

Более наглядным является графическое сравнение интервальных оценок центров подсовокупностей (выборок), для чего на рис. 12.6 для каждой группы вычислены границы 95-процентных доверительных интервалов $\bar{x}_i - HCP_i < a_i < \bar{x}_i + HCP_i$, где $HCP_i = t_{0,05}(df_i) \cdot \frac{\hat{\sigma}_i}{\sqrt{n_i}}$ – так называемая наименьшая существенная разность; a_i – центр подсовокупности.

Группы	n_i	Средние	НСР _i	Нижние	Верхние
1	103	32,368	1,012	31,356	33,380
2	3	27,367	5,928	21,4389	33,295
3	30	22,243	1,875	20,369	24,118
4	17	19,247	2,490	16,757	21,738

Рис. 12.6. Интервальные оценки центров каждой выборки

В таблице на рис. 12.6 определены нижние и верхние границы доверительных интервалов. На рис. 12.7 эти интервалы изображены графически в одном масштабе. Картина во многом прояснилась. Группа 2 содержала всего 3 наблюдения, и доверительный интервал для оценки центра этой подсовкупности получился очень широким – он перекрывает доверительные интервалы для групп 1 и 3. Чем больше наблюдений в группе, тем уже доверительный интервал, и тем более определенные заключения можно сделать. По данному обследованию можно заключить, что группа 1 значительно отличается от групп 3 и 4. Что касается групп 3 и 4, то для них нуль-гипотеза об отсутствии значимых различий не может быть отвергнута – их доверительные интервалы перекрываются (однако здесь для сравнения 2-й и 4-й групп потребуется дополнительное исследование, так как для этих групп нарушается предпосылка о равенстве дисперсий).

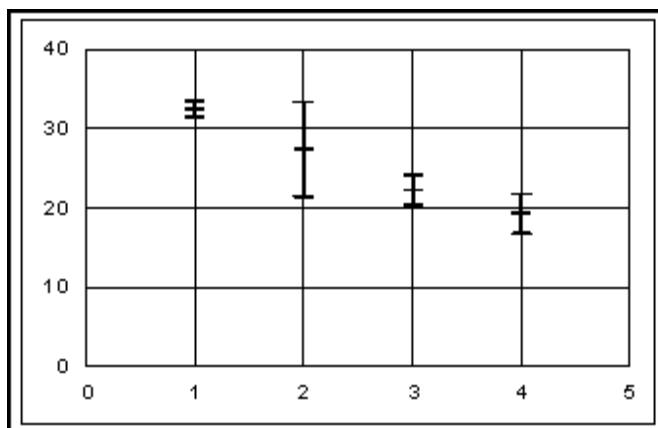


Рис. 12.7. Доверительные интервалы для каждой группы

В данном примере ясность графического анализа была обусловлена малым количеством сравниваемых групп и тем, что эти группы были расположены регулярно в порядке убывания результативного признака.

Академик Доспехов Б. А. предложил еще более наглядное представление результатов анализа (рис. 12.8). Группы должны быть отсортированы в порядке возрастания (или убывания) результативного признака (средних по группам). Далее строится символьная диаграмма (это проще, чем рисунок) из нескольких рядов звездочек. Если между некоторыми

Академик Доспехов Б. А. предложил еще более наглядное представление результатов анализа (рис. 12.8). Группы должны быть отсортированы в порядке возрастания (или убывания) результативного признака (средних по группам). Далее строится символьная диаграмма (это проще, чем рисунок) из нескольких рядов звездочек. Если между некоторыми

Группы	n_i	Средние	Однородные группы		
4	17	19,247	*		
3	30	22,243	*	*	
2	3	27,367		*	*
1	103	32,368			*

Рис. 12.8. Символьная диаграмма

группами нет значимых различий, они помечаются звездочками в одном вертикальном ряду (однородные группы).

Сравнение двух дисперсий

Английский статистик Р. Фишер изучил распределение отношения несмещенных оценок дисперсии для двух выборок, взятых из одной и той же нормально распределенной совокупности: $F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{SS_1}{SS_2} \cdot \frac{df_2}{df_1}$, где в числителе стоит большая из двух оценок дисперсий, а в знаменателе – меньшая оценка. Распределение Фишера зависит только от чисел степеней свободы числителя и знаменателя $F(df_1, df_2)$. Внешне график дифференциальной функции распределения Фишера похож на аналогичный график распределения Пирсона. Составлены таблицы с двумя входами (df_1, df_2) для определения критических значений $F_{0,05}$ и $F_{0,01}$ (определяются границы только для правостороннего критерия).

Если вычисленное значение F не превосходит табличного значения нижней границы $F \leq F_{0,05}$, нуль-гипотеза о равенстве дисперсий $\sigma_1^2 = \sigma_2^2$ не может быть отвергнута. Нуль-гипотеза отвергается, если вычисленное значение F превосходит табличное значение для верхней границы $F > F_{0,01}$. Остальные значения F попадают в область неопределенности критерия.

Для предыдущего примера (4 группы) сравним все 6 пар дисперсий. В таблице на рис. 12.9 для каждой пары выборок выписаны их объемы (n_i) и несмещенные оценки дисперсий ($\hat{\sigma}_i^2$). Вычислены отношения F большей оценки к меньшей и по таблицам для заданной пары чисел степеней свободы числителя и знаменателя найдены критические значения $F_{0,05}$ и $F_{0,01}$.

Группы $i-j$	n_i	n_j	$\hat{\sigma}_i^2$	$\hat{\sigma}_j^2$	F	$F_{0,05}$	$F_{0,01}$
1 – 2	103	3	31,004	67,704	2,184	3,09	4,82
1 – 3	103	30	31,004	24,566	1,262	1,71	2,15
1 – 4	103	17	31,004	8,253	3,757	2,07	2,86
2 – 3	3	30	67,704	24,566	2,756	3,33	5,42
2 – 4	3	17	67,704	8,253	8,204	3,63	6,23
3 – 4	30	17	24,566	8,253	2,977	2,20	3,10

Рис. 12.9. Сравнение дисперсий каждой пары выборок

Значимые дисперсионные отношения, для которых $F > F_{0,01}$, выделены в таблице, откуда видно, что изменчивость данных в 4-й группе значимо отличается от изменчивости в других группах.

При сравнении 3-й и 4-й групп дисперсионное отношение $F = 2,977$ попало в область неопределенности критерия Фишера.

Если при оценке значимости разности средних по критерию Стьюдента оказалось, что нельзя пользоваться объединенной оценкой дисперсии, дисперсию разности $\Delta = \bar{x}_1 - \bar{x}_2$ вычисляют по другой формуле:

$$\hat{\sigma}_{\Delta}^2 = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}.$$

Однако при этом возникают проблемы определения соответствующего числа степеней свободы df_{Δ} . Согласно методике Уэлча, ЧСС надо определять так:

$$df_{\Delta} = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2} \right)^2}{\frac{\hat{\sigma}_1^4}{n_1^2(n_1-1)} + \frac{\hat{\sigma}_2^4}{n_2^2(n_2-1)}}.$$

Например, для спорного случая при сравнении 3-й и 4-й групп имеем:

$$\Delta = \bar{x}_3 - \bar{x}_4 = 22,243 - 19,247 = 2,996;$$

$$\hat{\sigma}_{\Delta}^2 = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2} = \frac{24,566}{30} + \frac{8,253}{17} = 1,304;$$

$$df_{\Delta} = \frac{1,304^2}{\frac{24,566^2}{30^2 \cdot 29} + \frac{8,253^2}{17^2 \cdot 16}} = 44,9.$$

Все равно получилось то же самое значение ЧСС:

$$df_{\Delta} = (n_3 + n_4 - 2) = 45;$$

$$t_{\Delta} = \frac{\Delta}{\hat{\sigma}_{\Delta}} = \frac{2,996}{\sqrt{1,304}} = 2,62;$$

$$t_{0,05}(44) = 2,01;$$

$$t_{0,01}(44) = 2,69.$$

Вычисленное t_{Δ} попало в область неопределенности критерия.

Ранее (при объединении дисперсий) был сделан вывод об отсутствии значимых различий между выборками 3 и 4.

Вывод дифференциальной функции распределения Стьюдента

Рассмотрим распределение статистики Стьюдента $t = \frac{z}{\sqrt{V/k}}$,

где z и V – независимые случайные величины, z распределено по стандартному нормальному закону $f_z(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}$, а V – по закону Пирсона (χ^2)

$f_V(V) = C_0 \exp\left\{-\frac{V}{2}\right\} \cdot V^{\frac{k}{2}-1}$ с числом степеней свободы k .

Вследствие независимости z и V плотность вероятности их совместного распределения равна произведению $f_z(z)f_V(V) = C_1 \exp\left\{-\frac{z^2}{2} - \frac{V}{2}\right\} \cdot V^{\frac{k}{2}-1}$.

Интегральная функция распределения Стьюдента выражается через двойной интеграл:

$$F(x) = P\left(\frac{z\sqrt{k}}{\sqrt{V}} \leq x\right) = \iint_{z \leq \frac{x}{\sqrt{k}}\sqrt{V}} C_1 \exp\left\{-\frac{z^2}{2} - \frac{V}{2}\right\} \cdot V^{\frac{k}{2}-1} dz dV,$$

где область интегрирования определена неравенствами $-\infty < z \leq \frac{x}{\sqrt{k}}\sqrt{V}$ и $V \geq 0$.

Преобразуем этот интеграл:

$$F(x) = C_1 \int_0^\infty \exp\left\{-\frac{V}{2}\right\} \cdot V^{\frac{k}{2}-1} dV \cdot \int_{-\infty}^{\frac{x}{\sqrt{k}}\sqrt{V}} \exp\left\{-\frac{z^2}{2}\right\} dz.$$

Для определения плотности вероятности $f(x)$ дифференцируем полученное выражение по x :

$$\begin{aligned} f(x) &= C_1 \int_0^\infty \exp\left\{-\frac{V}{2}\right\} \cdot V^{\frac{k}{2}-1} dV \cdot \frac{d}{dx} \int_{-\infty}^{\frac{x}{\sqrt{k}}\sqrt{V}} \exp\left\{-\frac{z^2}{2}\right\} dz, \\ f(x) &= C_1 \int_0^\infty \exp\left\{-\frac{V}{2}\right\} \cdot V^{\frac{k}{2}-1} \cdot \exp\left\{-\frac{x^2 V}{2k}\right\} \frac{\sqrt{V}}{\sqrt{k}} dV = C_2 \int_0^\infty \exp\left\{-\frac{V}{2} \left(1 + \frac{x^2}{k}\right)\right\} \cdot V^{\frac{k-1}{2}} dV. \end{aligned}$$

В последнем интеграле делаем замену переменной:

$$\begin{aligned} \frac{V}{2} \left(1 + \frac{x^2}{k}\right) &= u; \quad V = \frac{2u}{1 + \frac{x^2}{k}}; \quad dV = \frac{2du}{1 + \frac{x^2}{k}}; \\ f(x) &= C_3 \left(1 + \frac{x^2}{k}\right)^{-\frac{k-1}{2}-1} \int_0^\infty \exp\{-u\} \cdot u^{\frac{k-1}{2}} du = B_k \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}. \end{aligned}$$

Константа B_k определяется из условия равенства единице площади под дифференциальной кривой.

Вывод дифференциальной функции распределения Фишера

Рассмотрим распределение статистики Фишера:

$$v = \frac{U/k_1}{V/k_2} = \frac{Uk_2}{Vk_1},$$

где U и V – независимые случайные величины, распределенные по закону Пирсона с числами степеней свободы k_1 и k_2 .

Вследствие независимости U и V плотность вероятности их совместного распределения равна произведению $f_U(U)f_V(V) = C \exp\left\{-\frac{U}{2} - \frac{V}{2}\right\} \cdot U^{\frac{k_1}{2}-1} \cdot V^{\frac{k_2}{2}-1}$.

Интегральная функция распределения Фишера выражается через двойной интеграл, где область интегрирования определена неравенствами $0 \leq U \leq \frac{xk_1}{k_2}V$, $V \geq 0$:

$$F(x) = P(v \leq x) = P\left(\frac{Uk_2}{Vk_1} \leq x\right) = \iint_{U \leq \frac{xk_1}{k_2}V} C \exp\left\{-\frac{U}{2} - \frac{V}{2}\right\} \cdot U^{\frac{k_1}{2}-1} \cdot V^{\frac{k_2}{2}-1} dU dV.$$

Преобразуем этот интеграл к виду:

$$F(x) = C \int_0^\infty e^{-\frac{V}{2}} \cdot V^{\frac{k_2}{2}-1} dV \int_0^{\frac{xk_1}{k_2}V} e^{-\frac{U}{2}} \cdot U^{\frac{k_1}{2}-1} dU.$$

Для определения плотности вероятности дифференцируем полученное выражение по x :

$$f(x) = C \int_0^\infty e^{-\frac{V}{2}} \cdot V^{\frac{k_2}{2}-1} dV \frac{d}{dx} \int_0^{\frac{xk_1}{k_2}V} e^{-\frac{U}{2}} \cdot U^{\frac{k_1}{2}-1} dU;$$

$$f(x) = C \int_0^\infty e^{-\frac{V}{2}} \cdot V^{\frac{k_2}{2}-1} \cdot e^{-\frac{xk_1}{2k_2}V} \cdot \left(\frac{xk_1}{k_2}V\right)^{\frac{k_1}{2}-1} \cdot \frac{k_1}{k_2} V dV;$$

$$f(x) = C_1 \left(\frac{xk_1}{k_2}\right)^{\frac{k_1}{2}-1} \int_0^\infty e^{-\frac{V}{2}\left(1+\frac{xk_1}{k_2}\right)} \cdot V^{\frac{k_1+k_2}{2}-1} dV.$$

В последнем интеграле делаем замену переменной:

$$V\left(1 + \frac{xk_1}{k_2}\right) = w; \quad V = \left(1 + \frac{xk_1}{k_2}\right); \quad dV = \left(\frac{dw}{1 + \frac{xk_1}{k_2}}\right);$$

$$f(x) = \frac{C_1 \left(\frac{xk_1}{k_2}\right)^{\frac{k_1-1}{2}}}{\left(1 + \frac{xk_1}{k_2}\right)^{\frac{k_1+k_2}{2}}} \int_0^\infty e^{-\frac{w}{2}} \cdot w^{\frac{k_1+k_2-2}{2}} dw = \frac{C_2 \left(\frac{xk_1}{k_2}\right)^{\frac{k_1-2}{2}}}{\left(1 + \frac{xk_1}{k_2}\right)^{\frac{k_1+k_2}{2}}}.$$

Константа C_2 определяется из условия равенства единице площади под дифференциальной кривой.

Вопросы для самопроверки

1. Чем отличается распределение Стьюдента от нормального распределения?
2. Как строится интервальная оценка математического ожидания?
3. Какова относительная погрешность интервальной оценки?
4. Как определяется объем выборки, необходимый для получения заключений с заданными надежностью и погрешностью?
5. Как сравниваются центры двух подсовкупностей?
6. Изложите графическую форму представления результатов анализа сравнения выборок.
7. Как сравниваются дисперсии?
8. Как находится дисперсия разности средних при одинаковых дисперсиях выборок?
9. Как находится дисперсия разности средних при разных дисперсиях выборок?

13. Дисперсионный анализ

Сравнение групп

Дисперсионный анализ – математический аппарат для сравнения средних нескольких популяций (групп, слоев, классов), которые определяются уровнями некоторых величин (факторов), положенных в основу классификации. Международным обозначением дисперсионного анализа является аббревиатура ANOVA (analysis of variance).

В дисперсионном анализе предполагается, что группы (популяции) различаются только средним уровнем результативной переменной, данные в каждой группе распределены нормально с одинаковой дисперсией.

В предыдущей лекции мы познакомились с методикой Стьюдента сравнения средних двух групп. Это частный случай «однофакторного» дисперсионного анализа, когда группы наблюдений определяются уровнями одного классификационного показателя (фактора).

Рассмотрим общий случай однофакторного дисперсионного анализа. Пусть имеется p групп ($i = 1, 2, \dots, p$), соответствующие уровням x_i классификационного фактора \mathcal{X} . В каждой группе имеется k_i наблюдений результативной переменной y_{ij} ($j = 1, 2, \dots, k_i$), всего наблюдений $n = \sum_{i=1}^p k_i$.

Модель однофакторного дисперсионного анализа $y_{ij} = u_i + \varepsilon_{ij}$ представляет разложение полного сигнала y_{ij} на две компоненты – на полезный сигнал u_i и помеху ε_{ij} (на детерминированную часть u_i , которая характеризует определенную группу, и случайный разброс ε_{ij} в этой группе).

Методом наименьших квадратов из условия минимума суммы квадратов ошибок $\sum_{i=1}^p \sum_{j=1}^{k_i} \varepsilon_{ij}^2$ найдено, что наилучшими оценками для u_i являются средние

значения $u_i = \bar{y}_{x_i} = \frac{1}{k_i} \sum_{j=1}^{k_i} y_{ij}$ по данным каждой группы (средние групповые).

Интересно, что точно такое же разложение имеет сумма квадратов отклонений $SSy = SSu + SS\varepsilon$ и число степеней свободы $dfy = dfu + df\varepsilon$.

Покажем это. Прежде всего убедимся, что среднее из средних групповых равно общему среднему: $\bar{u} = \bar{y} = \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^{k_i} y_{ij}$.

Действительно,

$$\bar{u} = \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^{k_i} u_i = \frac{1}{n} \sum_{i=1}^p k_i u_i = \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^{k_i} y_{ij} = \bar{y}.$$

Преобразуем полную (общую) сумму квадратов отклонений:

$$\begin{aligned} SSy &= \sum_{i=1}^p \sum_{j=1}^{k_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^p \sum_{j=1}^{k_i} ((y_{ij} - u_i) - (u_i - \bar{y}))^2 = \\ &= \sum_{i=1}^p \sum_{j=1}^{k_i} (y_{ij} - u_i)^2 + \sum_{i=1}^p \sum_{j=1}^{k_i} (u_i - \bar{y})^2 - 2 \sum_{i=1}^p \sum_{j=1}^{k_i} (y_{ij} - u_i)(u_i - \bar{y}) = \\ &= SS\varepsilon + SSu - 2 \sum_{i=1}^p (u_i - \bar{y}) \sum_{j=1}^{k_i} (y_{ij} - u_i) = SS\varepsilon + SSu. \end{aligned}$$

\downarrow
0

Последняя двойная сумма в этом выражении равна нулю из-за центрального свойства средних – алгебраическая сумма отклонений от своего среднего равна

нулю $\sum_{j=1}^{k_i} \varepsilon_{ij} = \sum_{j=1}^{k_i} (y_{ij} - u_i) = 0.$

Итак, $SSy = SSu + SS\varepsilon.$

Докажем аналогичное соотношение для чисел степеней свободы $dfy = dfu + df\varepsilon:$

$dfy = n - 1$, так как на n отклонений $(y_{ij} - \bar{y})$ наложена одна связь – сумма всех этих отклонений равна нулю;

$dfu = p - 1$, так как на p отклонений $(u_i - \bar{y})$ наложена одна связь – взвешенная сумма этих отклонений равна нулю $\sum_{i=1}^p \sum_{j=1}^{k_i} (u_i - \bar{y}) = \sum_{i=1}^p k_i (u_i - \bar{y}) = 0;$

$df\varepsilon = n - p$, так как в каждой группе сумма отклонений равна нулю $\sum_{j=1}^{k_i} \varepsilon_{ij} = 0.$

Итак, $dfy = dfu + df\varepsilon.$

Р. Фишер предложил все выкладки дисперсионного анализа оформлять в виде стандартной таблицы (рис. 13.1).

Изменчивость	Суммы квадратов	ЧСС	Средние квадраты	Дисперсионное отношение
Между группами	$SSy = \sum k u^2 - n \cdot (\bar{y})^2$	$dfu = p - 1$	$MSu = SSu / dfu$	$F = MSu / MS\varepsilon$
Внутри групп	$SS\varepsilon = \sum \sum y^2 - \sum k u^2$	$df\varepsilon = n - p$	$MS\varepsilon = SS\varepsilon / df\varepsilon$	
Общая	$SSy = \sum \sum y^2 - n \cdot (\bar{y})^2$	$dfy = n - 1$		

Рис. 13.1. Шаблон таблицы дисперсионного анализа

Изменчивость (Source of variable): *Общая* (Total) изменчивость данных y_{ij} разлагается на детерминированную часть – изменчивость средних групповых u_i (традиционное название – изменчивость *Между группами* (Between of groups)) и случайную изменчивость ε_{ij} (изменчивость *Внутри групп* (Within of groups)).

Суммы квадратов отклонений (Suma of squares) – SSy , SSu , $SS\varepsilon$.

ЧСС – числа степеней свободы (Degree of freedom) – dfy , dfu , $df\varepsilon$.

Средние квадраты (Mean of squares) – несмещенные оценки дисперсий, равные отношению сумм квадратов отклонений к своему числу степеней свободы: $MSu = SSu / dfu$, $MS\varepsilon = SS\varepsilon / df\varepsilon$.

Дисперсионное отношение Фишера F показывает, во сколько раз изменчивость между группами (изменчивость "полезного сигнала") превышает изменчивость внутри групп (случайную изменчивость): $F = MSu / MS\varepsilon$.

Статистика F распределена по закону Фишера с числами степеней свободы (dfu , $df\varepsilon$). По таблицам Фишера – Снедекора находим критические значения $F_{0,05}(dfu, df\varepsilon)$ и $F_{0,01}(dfu, df\varepsilon)$.

Если вычисленное значение F меньше нижней границы $F < F_{0,05}$, нуль-гипотеза об отсутствии значимых различий между группами не может быть отвергнута. Если вычисленное значение F больше верхней границы $F > F_{0,01}$, нуль-гипотеза отвергается, между группами имеются значимые различия. Какие именно различия – устанавливается по парным сравнением групп с помощью

критерия Стьюдента $t_{\Delta} = \frac{\Delta}{\hat{\sigma}_{\Delta}} = \frac{|u_1 - u_2|}{\sqrt{MS\varepsilon \cdot \left(\frac{1}{k_1} + \frac{1}{k_2}\right)}}$, где за оценку объединенной

дисперсии принимается $MS\varepsilon$. Для групп малого объема оценка $MS\varepsilon$ более надежна, чем объединенная дисперсия только двух групп. Критическое значение $t_{0,05}$ находится по таблицам Стьюдента для ЧСС = $df\varepsilon$.

Для $p = 2$ (сравнение двух групп) дисперсионный анализ полностью эквивалентен методике сравнения групп по критерию Стьюдента. Действительно, для двух групп $df_{\varepsilon} = k_1 + k_2 - 2$, $SS_{\varepsilon} = SS_1 + SS_2$, $MS_{\varepsilon} = \frac{SS_1 + SS_2}{k_1 + k_2 - 2}$, $\bar{y} = \frac{k_1 u_1 + k_2 u_2}{k_1 + k_2}$,

$$SSu = k_1(u_1 - \bar{y})^2 + k_2(u_2 - \bar{y})^2 = \frac{k_1 k_2}{k_1 + k_2} (u_1 - u_2)^2 = \frac{(u_1 - u_2)^2}{\left(\frac{1}{k_1} + \frac{1}{k_2}\right)},$$

$$F = \frac{MSu}{MS_{\varepsilon}} = \frac{(u_1 - u_2)^2}{\frac{SS_1 + SS_2}{k_1 + k_2 - 2} \left(\frac{1}{k_1} + \frac{1}{k_2}\right)} = t_{\Delta}^2, \text{ то есть статистика Фишера равна квадрату}$$

статистики Стьюдента; так же связаны и табличные значения $F_{0,05}(1; df) = t_{0,05}^2(df)$.

Рассмотрим полученные результаты дисперсионного анализа с другой стороны. Можно сказать, что при $F > F_{0,01}$ результативная переменная \mathcal{Y} и классификационный показатель \mathcal{X} не являются независимыми, они связаны между собой, между ними имеется статистическая (корреляционная) зависимость. Введем меру тесноты этой связи.

Имеем разложение общей суммы квадратов отклонений на детерминированную и случайную компоненты:

$$SSy = SSu + SS_{\varepsilon};$$

или в относительных единицах:

$$1 = \frac{SSu}{SSy} + \frac{SS_{\varepsilon}}{SSy}.$$

Относительный вклад детерминированной части называется индексом детерминации:

$$\eta^2 = \frac{SSu}{SSy} = 1 - \frac{SS_{\varepsilon}}{SSy}.$$

Индекс детерминации изменяется от 0 до 1 ($0 \leq \eta^2 \leq 1$).

Действительно, индекс детерминации $\eta^2 = \frac{SSu}{SSy}$ есть отношение сумм квадратов, которые не могут быть отрицательными; с другой стороны, индекс детерминации не может быть больше единицы, так как $\eta^2 = 1 - \frac{SS_{\varepsilon}}{SSy}$.

Если индекс детерминации равен нулю, то равна нулю сумма квадратов $SSu = \sum \sum (u_i - \bar{y})^2 = \sum k_i (u_i - \bar{y})^2$, следовательно, равны нулю все ее члены, откуда для любой группы будет $u_i = \bar{y}_{x_i} = \bar{y}$ – все средние групповые равны общему среднему, между средними групповыми нет различий, результативная переменная \mathcal{Y} не зависит от классификационного показателя \mathcal{X} .

Если индекс детерминации равен единице, то равна нулю сумма квадратов $SS\varepsilon = \sum \sum \varepsilon_{ij}^2$, следовательно, равны нулю все ее члены, то есть никакого случайного разброса нет, каждой группе (каждому уровню классификационного показателя \mathcal{X}) соответствует единственное значение результативной переменной u_i . Однозначное соответствие между множеством значений объясняющей переменной \mathcal{X} и множеством значений результативной переменной \mathcal{Y} называется функциональной зависимостью.

Чем ближе индекс детерминации к единице, тем ближе корреляционная зависимость к функциональной.

Индекс детерминации η^2 показывает, какая часть полной изменчивости определяется классификационным фактором (различиями между группами наблюдений). Для совместимости с мерами тесноты связей другой природы принято извлекать корень квадратный из индекса детерминации $\eta = \sqrt{\eta^2}$. Характеристика η называется корреляционным отношением.

Рассмотрим пример сравнения 4-х выборок, сведения о которых приведены в таблице на рис. 13.2.

Выборки	Объемы, k_i	Средние, $u_i = \bar{y}_{x_i}$	Несмещенные оценки, $\hat{\sigma}_i^2$	Суммы квадратов
1	103	32,368	31,004	3162,4
2	3	27,367	67,704	135,4
3	30	22,243	24,566	712,4
4	17	19,247	8,253	132,0
Всего	153	28,827	28,471	4242,2

Рис. 13.2. Исходные данные для дисперсионного анализа

Здесь в каждой группе найдены средние $u_i = \bar{y}_{x_i} = \frac{1}{k_i} \sum_{j=1}^{k_i} y_{ij}$, в итоговой строке приведено общее среднее \bar{y} при объединении всех 4-х групп наблюдений.

В последнем столбце вычислены суммы квадратов отклонений от средних групповых $SS_i = \sum_{j=1}^{k_i} (y_{ij} - u_i)^2 = \sum_{j=1}^{k_i} y_{ij}^2 - k_i u_i^2$, в итоговой строке приведена объединенная сумма квадратов случайных отклонений данных от средних групповых $SS\varepsilon = SS_1 + SS_2 + SS_3 + SS_4 = 4242,2$.

Число степеней свободы случайных отклонений в каждой группе равно $df_i = (k_i - 1)$, общее число степеней свободы случайной изменчивости получаем как их сумму $df_{\varepsilon} = \sum df_i = n - p = 153 - 4 = 149$.

Несмещенные оценки дисперсии случайной изменчивости по группам $\hat{\sigma}_i^2$ и общая MS_{ε} (в итоговой строке) получены как отношения сумм квадратов к соответствующим числам степеней свободы.

Необходимо выяснить, имеются ли значимые различия между выделенными группами наблюдений, и, если они есть, определить величину вклада классификационного фактора в общую изменчивость данных (значимые различия между группами указывают на существование связи между уровнями классификационного фактора и значениями результативной переменной; необходимо оценить тесноту этой связи).

Вычисляем сумму квадратов отклонений средних групповых от общего среднего:

$$\begin{aligned} SSu &= \sum k_i u_i^2 - n(\bar{y})^2 = \\ &= (103 \cdot 32,368^2 + 3 \cdot 27,367^2 + 30 \cdot 22,243^2 + 17 \cdot 19,247^2) - 153 \cdot 28,827^2 = \\ &= 4156,4. \end{aligned}$$

Общую сумму квадратов получаем как сумму двух вкладов: $SSy = SSu + SS_{\varepsilon} = 4156,4 + 4242,2 = 8398,6$.

Заполняем таблицу дисперсионного анализа (рис. 13.3).

Изменчивость	Суммы квадратов	ЧСС	Средние квадраты	Дисперсионное отношение	Табличные значения
Между группами	$SSu = 4156$	$dfu = 3$	$MSu = 1386$	$F = 48,86$	$F_{0,05} = 2,67$
Внутри групп	$SS_{\varepsilon} = 4242$	$df_{\varepsilon} = 149$	$MS_{\varepsilon} = 28,5$		$F_{0,01} = 3,92$
Общая	$SSy = 8398$	$dfy = 152$			$\alpha = 0,000$

Рис. 13.3. Таблица дисперсионного анализа

Вычисленное дисперсионное отношение Фишера $F = 48,86$ показывает, что изменчивость данных между группами в 48,9 раз превышает изменчивость внутри групп («мощность полезного сигнала» превышает «мощность помехи» в 48,9 раз).

Далее все зависит от наличия статистических таблиц. Если имеется таблица квантилей, то выписываем критические значения $F_{0,05}(3; 149) = 2,67$ и $F_{0,01}(3; 149) = 3,92$. Так как вычисленное значение $F = 48,86$ больше большей

границы $F > F_{0,01}$, нуль-гипотеза отвергается – между группами имеются значимые различия. Часто вместо таблиц квантилей используется таблица уровней значимости – вероятностей того, что вычисленное значение F получено чисто случайно при справедливости нуль-гипотезы. Обычно эта вероятность (уровень значимости) обозначается греческой буквой α . Нуль-гипотеза не может быть отвергнута, если $\alpha > 0,05$, и отвергается при $\alpha < 0,01$. Для данного примера получилось $\alpha = 5,5 \cdot 10^{-22} < 0,01$, следовательно, между группами имеются значимые различия.

Вычисляем индекс детерминации $\eta^2 = \frac{SSu}{SSy} = \frac{4156,4}{8398,6} = 0,495$. Получилось, что около 50 % изменчивости данных определяются различными уровнями классификационного фактора (различиями между группами наблюдений). Корреляционное отношение равно $\eta = \sqrt{0,495} = 0,703$.

Дисперсионный анализ является одним из инструментов для изучения связей между статистическими показателями. Поэтому имеет смысл напомнить терминологию классификации типов связей и ввести классификацию типов переменных.

Итак, напомним введенную раньше (в лекции 8 «Системы случайных величин») классификацию связей.

Зависимость называется **функциональной**, если каждому значению аргумента (аргументов) соответствует единственное значение функции (каждому значению объясняющих переменных соответствует единственное значение результативного признака; случайного разброса нет).

Зависимость называется **стохастической (статистической)**, если при изменении объясняющих переменных меняется вид распределения результативного показателя (меняется его условное распределение) – изменяются вид распределения или только его характеристики (условные математические ожидания, дисперсии и т. п.). Таким образом, в отличие от функциональной связи, при статистической зависимости нет однозначного соответствия между множеством значений аргументов и множеством значений функции.

Статистическая зависимость называется **корреляционной**, если при изменении аргумента изменяется только условное математическое ожидание функции (каждому значению объясняющих переменных соответствует свое среднее значение результативного показателя). При корреляционной зависимости мы следим за изменением только одной характеристики – центра условного распределения (условного математического ожидания).

Корреляционная зависимость является частным случаем общей статистической зависимости. Естественно, существуют также иные виды статистических зависимостей некорреляционного типа, например, когда при изменении аргументов меняется условная дисперсия.

В дисперсионном анализе предполагается, что группы (популяции) различаются только средним уровнем результативного признака, дисперсия в каждой группе считается одинаковой. Следовательно, с точки зрения введенной выше классификации дисперсионный анализ изучает корреляционные связи.

Рассмотрим классификацию переменных.

Выше уже неоднократно использовались термины *«объясняющие переменные»* и *«результативный показатель»*, «аргументы» и «функция»; в математике (но не в статистике!) допустимы также словосочетания «независимая и зависимая переменные» – дело в том, что в статистике аргументы (объясняющие переменные) далеко не всегда являются независимыми. Мы встречались с числовыми *непрерывными* и *дискретными* величинами (последние могли принимать только определенные фиксированные значения и не могли принимать никаких промежуточных значений между ними).

С развитием науки появилась необходимость измерять нечисловые объекты. Например, как сравнивать разные почвы, состояние экономики в разных странах, климатические условия в разные годы?

Поэтому, кроме количественных шкал, были введены качественные шкалы измерения. Отдельные значения на качественных шкалах называются категориями. Наиболее общая качественная шкала – это шкала имен (продукция разных предприятий, стадии развития растений, пол – мужской, женский, сезон – зимний, весенний, летний, осенний и т. д.). В шкале имен допустимы лишь операции сравнения «равно – не равно»: $x_i = x_j$ или $x_i \neq x_j$. К сожалению, из-за плохого перевода в русском языке появилось название «номинальная шкала» (от *name* – имя).

Следующими по информативности являются порядковые шкалы (из-за плохого перевода в русском языке такие шкалы часто называют ординальными). Порядковыми являются шкалы рангов или баллов, где допустимы любые логические операции сравнения $<, \leq, \neq, =, \geq, >$. Никакие арифметические операции в порядковых шкалах недопустимы.

Если один ученик получил оценку «пять», а другой «четыре», то нельзя сказать, что первый ученик знает предмет на единицу больше другого, так как для уровня знаний нет числовой меры.

Традиционное использование так называемого среднего балла иногда приводит к серьезным ошибкам.

Отметим разницу между шкалой рангов и другими порядковыми шкалами. Под операцией «ранжирование» понимается «сортировка» данных в порядке возрастания какого-либо показателя (количественного или порядкового) и *присвоение каждому наблюдению номера по порядку* (ранга). Если несколько данных по выбранному признаку не различимы, то им всем присваивается одинаковый ранг, средний из их порядковых номеров.

Количественные шкалы (непрерывные или дискретные) подразделяются еще на *интервальные шкалы* и *шкалы отношений*. На первый взгляд, измерение какого-либо показателя подобно измерению длины мерной линейкой (рулеткой) и с этими размерами можно производить любые арифметические операции (складывать их, перемножать и т. д.). Однако оказывается, что в некоторых количественных шкалах (интервальных), кроме логических операций сравнения, допустима всего лишь одна арифметическая операция – вычитание. К таким шкалам относятся шкала температур и шкала времени. Действительно, в шкале времени разность дат имеет ясный смысл, как длительность некоего процесса, временной промежуток между какими-либо событиями; но нет никакого смысла в произведении дат или в сумме дат. Нельзя также сказать, что температура 10 °С вдвое больше температуры 5 °С. Дело в том, что в интервальных шкалах (шкала температур и шкала времени) нет естественного начала. В шкале времени принято начинать отсчет от какого-либо события (реального или мифического) – от «Рождества Христова» или от «Сотворения мира». В шкале температур по Цельсию за начало 0 °С принята температура таяния льда. В других температурных шкалах (по Фаренгейту или по Реомюру) приняты другие начальные точки отсчета.

В шкале отношений допустимы любые арифметические операции.

Теперь мы можем определить место дисперсионного анализа в ряду других методов анализа связей. В дисперсионном анализе все объясняющие переменные (факторы) качественные, измеренные в наиболее общей шкале имен, а результирующий признак количественный, измеренный в шкале отношений.

Ранговый дисперсионный анализ Краскала – Уоллиса

Любую шкалу измерения можно всегда понизить до более простой, причем выводы, справедливые в простейших шкалах, будут более общими и надежными, чем в высших шкалах. Поэтому, кроме обычного дисперсионного анализа, используют также ранговые дисперсионные анализы Фридмана или

Краскала – Уоллиса. Так, в стандартном дисперсионном анализе требуется, чтобы данные в каждой группе были распределены нормально с одинаковой дисперсией. Если эти предпосылки не выполняются, выводы дисперсионного анализа становятся сомнительными. Наличие выбросов (далеко отклоняющихся значений) также способно исказить результаты анализа. После перехода к рангам некоторая часть информации будет потеряна, однако снимаются все вышеперечисленные обременительные предположения.

Например, в таблице на рис. 13.4 приведены данные о времени появления реакции в 4-х группах, которые отличаются условиями проведения опыта. В последних строках таблицы вычислены средние и дисперсии в каждой группе, откуда видна нежелательная особенность: большим значениям средних групповых соответствуют большие значения дисперсии.

№	I	II	III	IV
1	0,5	1,1	0,9	0,4
2	0,7	1,6	2,1	1,9
3	1,0	3,7	3,0	2,4
4	1,2	4,3	4,7	2,8
5	1,7	4,7	6,4	3,9
6	2,3	5,1	6,6	5,4
7	2,4	6,6	8,5	11,4
8	3,1	8,8	10,0	20,4
Средние	1,6	4,5	5,3	6,1
Дисперсии	0,741	5,494	8,809	39,077

Рис. 13.4. Данные 4-х выборок

Как правило, время появления какого-то события имеет экспоненциальное или гамма-распределение, которые существенно отличаются от нормального. Кроме того, последнее наблюдение в 4-й группе очень похоже на выброс (такие отклонения допустимы для экспоненциального закона, но нетипичны для нормального распределения).

По методу Краскала – Уоллиса необходимо все данные ($n = 4 \times 8 = 32$) ранжировать и для каждой группы найти средние ранги v_i .

Доказано, что статистика

$$H = \frac{12}{n(n+1)} \sum k_i \left(v_i - \frac{n+1}{2} \right)^2 = \frac{12}{n(n+1)} \sum k_i v_i^2 - 3(n+1)$$

имеет асимптотическое χ^2 -распределение с ЧСС = $p - 1$, где p – число групп.

Если нуль-гипотеза отклоняется, то для выявления значимых различий необходимо сделать $\frac{p(p-1)}{2}$ парных сравнений по критерию Стьюдента

$$t_{ij} = \frac{v_i - v_j}{\sqrt{\frac{n(n+1)}{12} \left(\frac{1}{k_i} + \frac{1}{k_j} \right)}}$$

с числом степеней свободы $df_{ij} = k_i + k_j - 2$.

Итак, ранжируем данные предыдущей таблицы и подсчитываем средние ранги в каждой группе (рис. 13.5).

№	I	II	III	IV
1	2	6	4	1
2	3	8	11	10
3	5	18	16	13,5
4	7	20	21,5	15
5	9	21,5	25	19
6	12	24	26,5	23
7	13,5	26,5	28	31
8	17	29	30	32
Суммы	68,5	153	162	144,5
Средние	8,563	19,125	20,250	18,063

Рис. 13.5. Данные после ранжирования

Расположенные в порядке возрастания наблюдения 13 – 14 оказались одинаковыми, поэтому присваиваем им одинаковый средний ранг 13,5; одинаковыми оказались также пары наблюдения 21 – 22 и 26 – 27, присваиваем этим парам средние ранги 21,5 и 26,5.

В последней строке таблицы подсчитаны средние ранги по группам.

Вычисляем статистику Краскала – Уоллиса ($p = 4$, $k_i = 8$, $n = 32$):

$$H = \frac{12}{32(32+1)} \left(8 \cdot 8,563^2 + 8 \cdot 19,125^2 + 8 \cdot 20,250^2 + 8 \cdot 18,063^2 \right) - 3(32+1) = 7,85.$$

Это значение сравниваем с табличным $H_{0,05} = \chi_{0,05}^2(4-1) = 7,81$.

Поскольку $H = 7,85 > H_{0,05}$, то нуль-гипотеза отклоняется с уровнем значимости 5 %, то есть считаем, что между группами имеются значимые различия (вспомните, что означает оговорка «с уровнем значимости 5 %»).

Теперь необходимо выяснить, какие именно группы значимо отличаются от остальных. Вычисляем разность средних рангов для 1-й и 3-й групп (максимальная разница): $\Delta_{13} = 20,250 - 8,563 = 11,687$.

Статистика Стьюдента для этих групп

$$t_{\Delta_{13}} = \frac{|v_3 - v_1|}{\sqrt{\frac{32(32+1)}{12} \left(\frac{1}{8} + \frac{1}{8}\right)}} = \frac{11,687}{\sqrt{22}} = 2,49$$

оказалась больше табличного значения $t_{0,05}(8 + 8 - 2) = 2,14$, то есть можно считать, что между группами 1 – 3 есть значимые различия (с уровнем значимости 5 %). Остальные разности незначимы.

Приведем некоторые соображения для вывода статистики Краскала – Уоллиса.

Напоминаем, что если величины x_i распределены нормально $x_i \sim N(a_i, \sigma_i)$, то сумма квадратов стандартизованных величин $\sum \frac{(x_i - a_i)^2}{\sigma_i^2}$ распределена по закону χ^2 .

Краскал и Уоллис рассматривали средние ранги v_i в каждой из p групп объема k_i . Нуль-гипотеза заключается в утверждении, что элементы в каждую группу отбираются случайным образом, поэтому ожидается (математическое ожидание), что все a_i одинаковы и равны общему среднему рангу всех n наблюдений $a_i = \frac{n+1}{2}$ (ранги – последовательные номера от 1 до n). Известны вероятности попадания элемента в ту или иную группы – они пропорциональны объемам выборок $q_i = \frac{k_i}{n}$. Этого достаточно, чтобы вывести формулу для дисперсий средних рангов:

$$\sigma_{v_i}^2 = \left(1 - \frac{k_i}{n}\right) \frac{n(n+1)}{12k_i}.$$

Согласно центральной предельной теореме, средние ранги случайных выборок объема $k_i > 5$ распределены практически нормально.

Составляем стандартную статистику Пирсона:

$$\chi^2 \approx \sum \frac{\left(v_i - \frac{n+1}{2}\right)^2}{\left(1 - \frac{k_i}{n}\right) \frac{n(n+1)}{12k_i}},$$

где для больших n можно пренебречь сомножителями $\left(1 - \frac{k_i}{n}\right)$.

Новая статистика $H = \frac{12}{n(n+1)} \sum k_i \left(v_i - \frac{n+1}{2}\right)^2$ будет иметь асимптотическое χ^2 -распределение. Число степеней свободы здесь на единицу меньше числа групп, так как общая сумма рангов известна $\sum k_i v_i = \frac{n(n+1)}{2}$ – это связь, наложенная на отклонения $\left(v_i - \frac{n+1}{2}\right)$.

В новой статистике Стьюдента для сравнения средних рангов двух групп $(v_i - v_j)$ также пренебрегаем сомножителями $\left(1 - \frac{k_i}{n}\right)$:

$$t_{ij} = \frac{v_i - v_j}{\sqrt{\sigma_{v_i}^2 + \sigma_{v_j}^2}} \approx \frac{v_i - v_j}{\sqrt{\frac{n(n+1)}{12} \left(\frac{1}{k_i} + \frac{1}{k_j}\right)}}.$$

Дополнение к выводу формул Краскала – Уоллиса

Математика – это не поваренная книга рецептов, где все надо запоминать без каких-либо доказательств.

Конечно, для нематематиков сложные выводы можно не приводить, а ограничиться ссылками на соответствующие источники. Но если математические преобразования посильны и поучительны, может быть стоит их привести как дополнительные (необязательные) материалы к изучению данной темы. Разве неинтересно узнать, каким образом была выведена формула для дисперсии средних рангов?

Итак, ранги u_i – это последовательные номера от 1 до n , переставленные в каком-то другом порядке. Поэтому можно найти сумму рангов $\sum u_i = \sum i = \frac{n(n+1)}{2}$, сумму квадратов рангов $\sum u_i^2 = \sum i^2 = \frac{n(n+1)(2n+1)}{6}$ (эту формулу можно найти в справочниках) и сумму произведений рангов $2 \sum_{i < j} u_i u_j = (\sum u_i)^2 - \sum u_i^2 = \frac{(n-1)n(n+1)(3n+2)}{12}$.

Сумму рангов в случайной выборке можно представить в виде $z = \sum \alpha_i u_i$, где $\alpha_i = 1$, если элемент u_i попал в выборку, и $\alpha_i = 0$, если не попал.

Вероятность того, что элемент попадает в выборку, равна $q_1 = \frac{k}{n}$; два элемента попадают в одну выборку с вероятностью $q_2 = \frac{k}{n} \cdot \frac{k-1}{n-1}$ (выборка без возвращения).

Вычисляем математические ожидания $M(\alpha_i) = \frac{k}{n}$, $M(\alpha_i^2) = \frac{k}{n}$, $M(\alpha_i \alpha_j) = \frac{k}{n} \cdot \frac{k-1}{n-1}$.

Отсюда получаем дисперсию:

$$D(\alpha_i) = M(\alpha_i^2) - M^2(\alpha_i) = \frac{k}{n} \cdot \left(1 - \frac{k}{n}\right)$$

и ковариацию:

$$\text{Cov}(\alpha_i, \alpha_j) = M(\alpha_i \alpha_j) - M(\alpha_i)M(\alpha_j) = -\frac{k}{n(n-1)} \cdot \left(1 - \frac{k}{n}\right).$$

Теперь вычислим математическое ожидание суммы рангов $z = \sum \alpha_i u_i$ и среднего ранга $v = z/k$ в случайной выборке:

$$M(z) = \sum u_i M(\alpha_i) = \sum u_i \frac{k}{n} = \frac{k}{n} \sum u_i = \frac{k}{n} \cdot \frac{n(n+1)}{2} = \frac{k(n+1)}{2}; \quad M(v) = \frac{n+1}{2}.$$

Этот результат мы установили ранее простейшими рассуждениями.

Вычислим дисперсию суммы рангов $z = \sum \alpha_i u_i$ в случайной выборке:

$$\begin{aligned} D(z) &= \sum u_i^2 D(\alpha_i) + 2 \sum_{i < j} u_i u_j \text{Cov}(\alpha_i, \alpha_j) = \frac{k}{n} \cdot \left(1 - \frac{k}{n}\right) \cdot \sum u_i^2 - \frac{k}{n(n-1)} \cdot \left(1 - \frac{k}{n}\right) \cdot 2 \sum_{i < j} u_i u_j = \\ &= \frac{k}{n} \cdot \left(1 - \frac{k}{n}\right) \cdot \frac{n(n+1)(2n+1)}{6} - \frac{k}{n(n-1)} \cdot \left(1 - \frac{k}{n}\right) \cdot \frac{(n-1)n(n+1)(3n+2)}{12} = k \cdot \left(1 - \frac{k}{n}\right) \cdot \frac{n(n+1)}{12}. \end{aligned}$$

Искомая дисперсия среднего ранга $v = z/k$ в случайной выборке будет в k^2 раз меньше: $D(v) = \left(1 - \frac{k}{n}\right) \cdot \frac{n(n+1)}{12k}$, что и требовалось доказать.

Вопросы для самопроверки

1. Сформулируйте задачу дисперсионного анализа.
2. Что такое суммы квадратов? Напишите разложение общей суммы квадратов на компоненты.
3. Что такое число степеней свободы? Напишите разложение общего числа степеней свободы на компоненты.
4. Что такое средние квадраты?
5. Что означает дисперсионное отношение Фишера?
6. Как определяется значимость различий между группами наблюдений?
7. Как оценить тесноту связи между классификационной и результативной переменными?
8. Как уточнить, между какими именно группами имеются значимые различия?
9. Что такое ранги?
10. Какие преимущества имеет ранговый дисперсионный анализ?

14. Регрессионный анализ

Регрессионный анализ предназначен для изучения корреляционных связей между количественными переменными, причем результативная переменная должна быть измерена в количественной непрерывной шкале. Несколько странное название «регрессионный анализ» (почему «регресс», а не «прогресс»?) закрепилось исторически. Дело в том, что характерной особенностью прогрессивно возрастающих зависимостей является увеличение темпа возрастания результативной переменной $\Delta y / \Delta x$ при увеличении абсолютного уровня объясняющей переменной x . Такими зависимостями обычно описываются нестабильные процессы типа взрыва (цепная реакция, демографический взрыв, удвоение числа публикаций через определенный период и т. п.). Большинство же интересующих нас биометрических («био» – жизнь) зависимостей регрессивны, для них характерно насыщение – постепенное снижение эффективности управляющих воздействий (объясняющих переменных), с возрастанием их абсолютных значений темпы прироста результативной переменной снижаются (рис. 14.1).

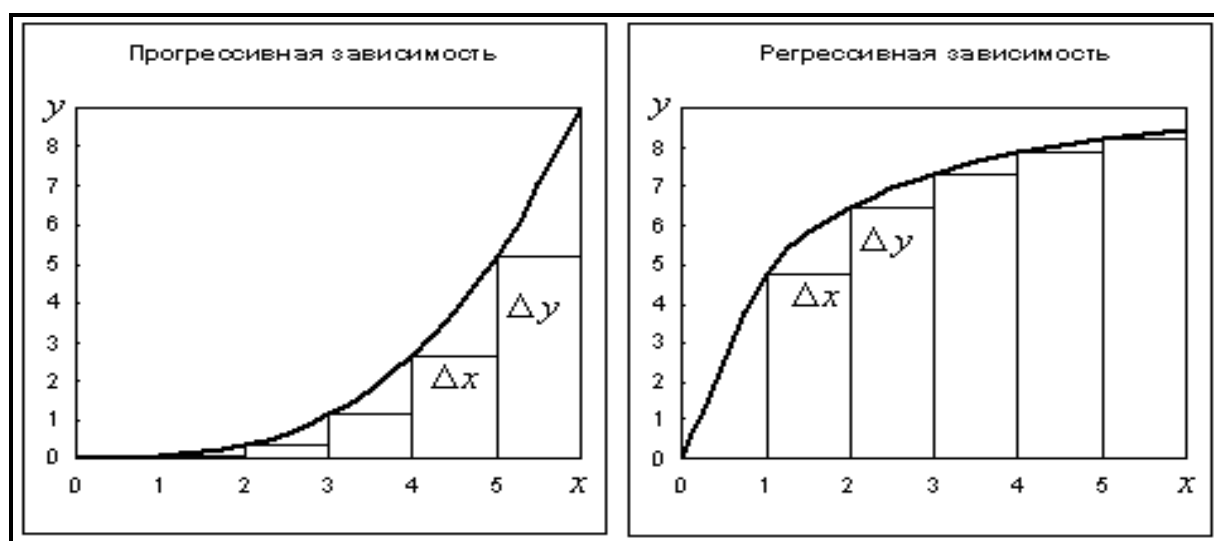


Рис. 14.1. Прогрессивная и регрессивная зависимости

В регрессионном анализе (так же, как и в дисперсионном) изучается поведение только одной характеристики распределения результативной переменной – центра группировки \bar{y} при каждом значении объясняющих переменных x_1, x_2, x_3 и т. д. Следовательно, в регрессионном анализе изучаются корреляционные зависимости. По традиции уравнения этих зависимостей называются уравнениями регрессии, а их графики – линиями регрессии. Корреляционно-регрессионный анализ не решает вопрос о направлении причинно-

следственных связей; специалист должен сам указать, какую именно переменную надо считать результативной (остальные переменные тогда будут считаться объясняющими). В регрессионном анализе все ошибки будут отнесены только к результативной переменной, а объясняющие переменные будут считаться неслучайными, измеренными точно. Из-за этой особенности одним и тем же данным будет соответствовать не одно, а несколько так называемых сопряженных уравнений регрессии в зависимости от того, какая именно переменная объявлена результативной. На рис. 14.2 изображены линии сопряженных регрессий для двух случайных величин (\mathcal{X} , \mathcal{Y}), причем на рис. 14.2а за результативную переменную принята переменная Y и найдены ее средние значения $M(y|x)$ для каждого значения x , а на рис. 14.2б за результативную переменную принята переменная X и найдены ее средние значения $M(x|y)$ для каждого значения y .

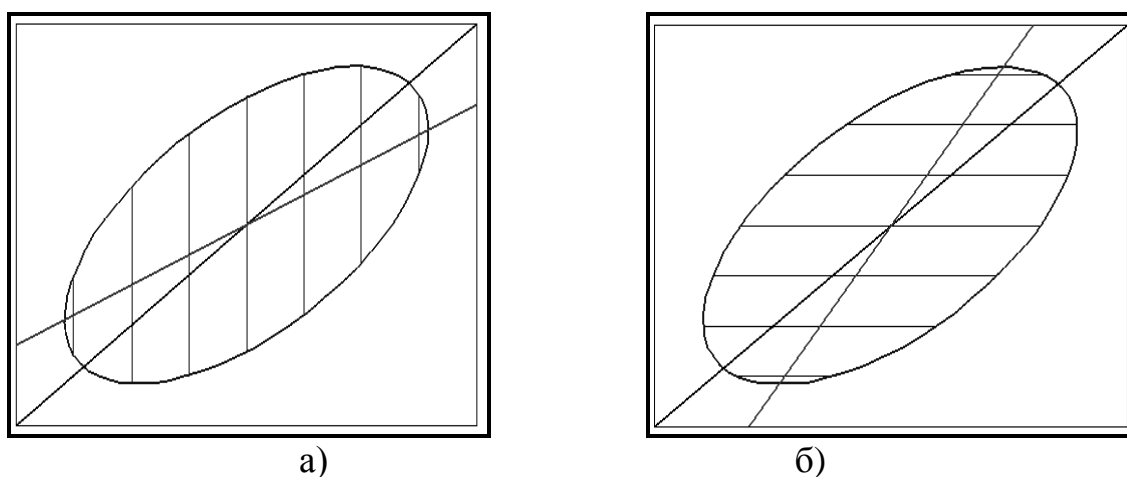


Рис. 14.2. Сопряженные линии регрессии и диагональная регрессия

Данный пример соответствует случаю совместного нормального распределения двух случайных величин (\mathcal{X} , \mathcal{Y}). С уровнем доверия $P = 0,95$ практически все точки (95 %) попадают в эллиптическую область, поскольку для двумерного нормального закона область рассеивания точек (x_i, y_i) имеет такую форму.

Если \mathcal{X} – причина, а \mathcal{Y} – следствие (рис. 14.2а), то линия регрессии $M(y|x) = f_1(x)$ будет совпадать с диаметром эллипса, сопряженным семейству вертикальных хорд (то есть серединами вертикальных хорд эллипса).

Если \mathcal{Y} – причина, а \mathcal{X} – следствие (рис. 14.2б), то линия регрессии $M(x|y) = f_2(y)$ будет совпадать с диаметром эллипса, сопряженным семейству горизонтальных хорд (серединами горизонтальных хорд эллипса).

Это совсем разные диаметры. Обе сопряженные линии регрессии не совпадают с главной осью эллипса рассеивания – это следствие выбора только

одной переменной в качестве случайной (результативной) величины, которой приписываются все ошибки (случайные и неслучайные). Чаще всего специалист не сомневается в правильности выбора направления причинно-следственных связей. Тогда разумным выглядит предположение, что одна из переменных определяет другую.

Однако бывает, что обе переменные (\mathcal{X} , \mathcal{Y}) являются разными следствиями одной и той же общей причины, что и порождает наблюдаемую связь между ними (например, обе переменные возрастают со временем). В такой задаче нет оснований принимать какую-либо переменную в качестве результативной и считать, что одна переменная определяет другую; наилучший график зависимости между равноправными переменными \mathcal{X} , \mathcal{Y} должен совпадать с главной осью рассеивания данных. По традиции этот особый вид зависимости называется диагональной регрессией. Сразу же отметим, что диагональная регрессия не есть регрессия (корреляционная зависимость) по определению, так как точки диагонали эллипса не являются средними значениями одной из переменных при заданных значениях остальных.

Метод наименьших квадратов (МНК)

Данные обычно имеют вид таблицы значений показателей (x_1, x_2, y) , один из которых является результативным (y) и выражается через оставшиеся объясняющие переменные (x_1, x_2) , которые иногда называются факторами.

Предполагается, что форма связи нам известна с точностью до параметров, наилучшие значения которых надо найти по опытным данным, то есть найти МНК-оценки параметров. Для применения метода наименьших квадратов крайне желательно, чтобы параметры входили в форму связи *линейным образом*, например, так:

линейная двухфакторная зависимость: $y = b_0 + b_1 x_1 + b_2 x_2 + e$;

квадратичная однофакторная зависимость: $y = b_0 + b_1 x + b_2 x^2 + e$;

нелинейная зависимость: $\ln y = b_0 + b_1 \ln x_1 + b_2 \ln x_2 + e$.

Здесь b_0, b_1, b_2 – параметры модели, которые подлежат определению; e – ошибки (остатки модели).

Далее будем рассматривать базовую линейную зависимость, к которой могут быть сведены многие другие зависимости соответствующими заменами переменных: $y = y_p + e$, где $y_p = b_0 + b_1 x_1 + b_2 x_2$.

Условимся суммирование по всем наблюдениям обозначать квадратными скобками (обозначения Гаусса): $[y] = \sum_{i=1}^n y_i$; $[xy] = \sum_{i=1}^n x_i y_i$.

По методу наименьших квадратов (МНК) параметры модели b_0, b_1, b_2 следует определять из условия минимума суммы квадратов ошибок по всем наблюдениям $[e^2] \rightarrow \min$.

Согласно необходимым условиям экстремума, приравняем к нулю частные производные суммы квадратов ошибок по каждому параметру модели b_0, b_1, b_2 .

В результате получим такую систему нормальных уравнений:

$$[e] = 0; [ex_1] = 0; [ex_2] = 0.$$

При преобразованиях были использованы правила:

$$\frac{\partial}{\partial b_j} [e^2] = \left[\frac{\partial e^2}{\partial b_j} \right] = 2 \cdot \left[e \frac{\partial e}{\partial b_j} \right]; e = y - b_0 - b_1 x_1 - b_2 x_2; \frac{\partial e}{\partial b_j} = -x_j; \frac{\partial e}{\partial b_0} = -1.$$

Название «система нормальных уравнений» объясняется терминологией векторного исчисления.

Таблицы значения любых переменных (y, x_0, x_2, e) представляют собой n -мерные векторы:

$$y = (y_1, y_2, \dots, y_n);$$

$$x_0 = (1, 1, \dots, 1);$$

$$x_j = (x_{j1}, x_{j2}, \dots, x_{jn});$$

$$e = (e_1, e_2, \dots, e_n).$$

Два вектора перпендикулярны (ортогональны, нормальны), если их скалярное произведение (сумма значений одноименных компонент) равно нулю

$$(e, x) = [ex] = \sum_{i=1}^n e_i x_i = 0.$$

Таким образом, система нормальных уравнений действительно представляет собой запись условий ортогональности (нормальности) вектора ошибок (e) к каждому члену модели $(1, x_1, x_2)$.

Помножим равенство $y = a_0 x_0 + a_1 x_1 + a_2 x_2 + e$ (где $x_0 = 1$) на каждую переменную, которые входят в это равенство, и вычислим средние полученных выражений по всем наблюдениям.

При этом учтем требование нормальности (ортогональности) ошибок к каждому члену модели $\bar{e} = 0, \overline{ex_1} = 0, \overline{ex_2} = 0$.

Получим:

$$\begin{cases} \bar{y} = b_0 + b_1 \bar{x}_1 + b_2 \bar{x}_2 \\ \overline{yx_1} = b_0 \bar{x}_1 + b_1 \overline{x_1^2} + b_2 \overline{x_1 x_2} ; \\ \overline{yx_2} = b_0 \bar{x}_2 + b_1 \overline{x_1 x_2} + b_2 \overline{x_2^2} \\ \overline{y^2} = b_0 \bar{y} + b_1 \overline{yx_1} + b_2 \overline{yx_2} + \overline{ye}; \\ \overline{ye} = \overline{e^2}. \end{cases}$$

Первые три равенства (объединенные фигурной скобкой) представляют собой систему нормальных уравнений в развернутой форме, а из последних двух равенств получаем выражение для оценки дисперсии остатка модели $s_e^2 = \overline{e^2} = \overline{y^2} - (b_0 \bar{y} + b_1 \overline{yx_1} + b_2 \overline{yx_2})$. Аналогичную формулу имеем для расчета суммы квадратов ошибок: $[e^2] = [y^2] - b_0 [y] - b_1 [yx_1] - b_2 [yx_2]$. Таким образом, выразили сумму квадратов ошибок через уже найденные суммы. Эта формула понадобится в дальнейшем.

Пример расчета МНК-оценок параметров

Расчеты по методу наименьших квадратов продемонстрируем на оценке параметров квадратичной модели $y = b_0 + b_1 x + b_2 x^2 + e$, которая формально сводится к предыдущей двухфакторной линейной модели заменой переменных $x_1 = x$, $x_2 = x^2$. При этом выясняется, что аргументы x_1 , x_2 не являются «независимыми» переменными в общепринятом понимании, они могут быть связаны между собой, лишь бы определитель системы нормальных уравнений был отличен от нуля. Кроме того, оказывается, что одной объясняющей переменной в нелинейной модели может соответствовать не один, а сразу несколько членов, необходимых для описания нелинейностей.

Данные на рис. 14.3 (эмпирические точки) явно уклоняются от прямой, видно явное наличие максимума зависимости; поэтому сочтено, что квадратичная модель $y = b_0 + b_1 x + b_2 x^2 + e$ будет более адекватно описывать эту нелинейную зависимость, чем линейная зависимость $y = b_0 + b_1 x + e$.

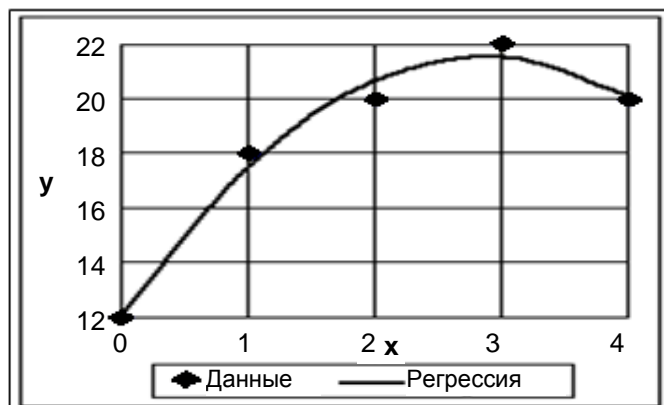


Рис. 14.3. Квадратичная зависимость

Условия ортогональности ошибок к каждому члену квадратичной модели $[e]=0$, $[ex]=0$, $[ex^2]=0$ приводят к такой системе нормальных уравнений:

$$\begin{cases} [y] = b_0 n + b_1 [x] + b_2 [x^2] \\ [yx] = b_0 [x] + b_1 [x^2] + b_2 [x^3] \\ [yx^2] = b_0 [x^2] + b_1 [x^3] + b_2 [x^4] \end{cases}.$$

Все необходимые суммы подсчитаны в таблице на рис. 14.4.

№	Данные		Расчет сумм						Расчетные значения	
	x	y	x^2	x^3	x^4	yx	yx^2	y^2	y_p	e
1	0	12	0	0	0	0	0	144	12,114	-0,114
2	1	18	1	1	1	18	18	324	17,547	0,457
3	2	20	4	8	16	40	80	400	20,696	-0,686
4	3	22	9	27	81	66	198	484	21,543	0,457
5	4	20	16	64	256	80	320	400	20,114	-0,114
Суммы	10	92	30	100	354	204	616	1752		0

Рис. 14.4. Расчет сумм для квадратичной модели

Вычисленные суммы подставляем в систему нормальных уравнений:

$$\begin{cases} 92 = b_0 \cdot 5 + b_1 \cdot 10 + b_2 \cdot 30 \\ 204 = b_0 \cdot 10 + b_1 \cdot 30 + b_2 \cdot 100 \\ 616 = b_0 \cdot 30 + b_1 \cdot 100 + b_2 \cdot 354 \end{cases}$$

и находим ее решение: $b_0 = 12,114$; $b_1 = 6,571$; $b_2 = -1,143$.

Расчетные значения $y_p = 12,114 + 6,571x - 1,143x^2$ приведены в той же таблице вместе с ошибками $e = y - y_p$. График найденной квадратичной зависимости изображен на рис. 14.3, при этом наблюдается хорошее сглаживание исходных данных.

Убеждаемся, что сумма всех ошибок равняется нулю: $[e] = 0$.

Данных немного, поэтому подсчитаем сумму квадратов ошибок непосредственно: $[e^2] = (-0,114)^2 + (0,457)^2 + (-0,686)^2 + (0,457)^2 + (-0,114)^2 = 0,914$.

Для проверки вычислим эту же сумму квадратов по формуле:

$$\begin{aligned} [e^2] &= [y^2] - b_0 [y] - b_1 [yx] - b_2 [yx^2] = \\ &= 1752 - 12,114 \cdot 92 - 6,571 \cdot 204 + 1,143 \cdot 616 = 1,116. \end{aligned}$$

Расхождение в результатах расчета двумя способами объясняется погрешностями в вычислении параметров модели с 3-мя десятичными знаками.

Если вычислить эти параметры с 4-мя десятичными знаками, то для суммы квадратов ошибок получим значение $[e^2] = 0,945$, а с 5-ю знаками – уже $[e^2] = 0,915$.

Оценка тесноты принятой формы связи

Ввиду ортогональности ошибок к каждому члену модели $[e] = 0$; $[ex_1] = 0$; $[ex_2] = 0$ ошибки будут также ортогональны к расчетным значениям $[ey_p] = 0$, где $y_p = b_0 + b_1 x_1 + b_2 x_2$.

Выражение $y = y_p + e$ представляет собой разложение полного сигнала (y) на две компоненты – детерминированную часть (y_p), которая определяется моделью (уравнением регрессии; в конечном итоге, – объясняющими переменными x_1, x_2), и ошибку (e), которая моделью не определяется. Оказывается, что точно такое же разложение имеет сумма квадратов отклонений $SSy = SS_p + SS_e$.

Покажем это. Ввиду ортогональности ошибок к свободному члену модели $[e] = 0$ получается, что $\bar{y}_p = \bar{y}$. Преобразуем полную (общую) сумму квадратов отклонений:

$$\begin{aligned} SSy &= \left[(y - \bar{y})^2 \right] = \left[(y_p - \bar{y} + e)^2 \right] = \\ &= \left[(y_p - \bar{y})^2 \right] + \left[e^2 \right] + 2 \left[(y_p - \bar{y}) \cdot e \right] = SS_p + SS_e. \end{aligned}$$

\downarrow
 0

Удвоенная сумма равна нулю ввиду ортогональности ошибок к расчетным значениям $[ey_p] = 0$ и свободному члену модели $[e] = 0$.

Получили разложение общей суммы квадратов отклонений (SSy) на две компоненты, одна из которых определяется моделью (SS_p), а другая (SS_e) моделью не определяется $SSy = SS_p + SS_e$, или в относительных величинах:

$1 = \frac{SS_p}{SSy} + \frac{SS_e}{SSy}$. Относительный вклад детерминированной части называется коэффициентом детерминации:

$$R^2 = \frac{SS_p}{SSy} = 1 - \frac{SS_e}{SSy}.$$

Коэффициент детерминации изменяется от 0 до 1 ($0 \leq R^2 \leq 1$).

Действительно, коэффициент детерминации $R^2 = \frac{SS_p}{SSy}$ есть отношение сумм квадратов, которые не могут быть отрицательными; с другой стороны, коэффициент детерминации не может быть больше единицы, так как $R^2 = 1 - \frac{SS_e}{SSy}$.

Если коэффициент детерминации равен нулю, то равна нулю сумма квадратов $SSp = \left[(y_p - \bar{y})^2 \right]$, следовательно, равны нулю все ее члены, откуда для любых значений аргументов x_1, x_2 расчетные значения будут одинаковыми $y_p = \bar{y}$, следовательно, отсутствует корреляционная зависимость выбранной формы связи.

Если коэффициент детерминации равен единице, то равна нулю сумма квадратов $SSe = [e^2]$, следовательно, равны нулю все ее члены, иными словами, никаких ошибок нет, каждому значению аргументов x_1, x_2 соответствует единственное расчетное значение y_p .

Однозначное соответствие между множеством значений объясняющих переменных и множеством значений результативной переменной является функциональной зависимостью.

Чем ближе коэффициент детерминации к единице, тем ближе найденная корреляционная зависимость к функциональной.

Коэффициент детерминации R^2 показывает, какая часть полной изменчивости определяется выбранной регрессионной моделью. Принято извлекать корень квадратный из коэффициента детерминации $R = \sqrt{R^2}$. Характеристика R называется коэффициентом корреляции: коэффициентом парной корреляции, если модель линейная однофакторная, или коэффициентом множественной корреляции – во всех остальных случаях.

Для рассмотренного выше примера аппроксимации данных квадратичной моделью была вычислена сумма квадратов ошибок $SSe = 0,914$; из таблицы расчета сумм выписываем также $n = 5$, $[y] = 92$, $[y^2] = 1752$, откуда получаем значение общей суммы квадратов $SSy = [y^2] - \frac{[y]^2}{n} = 1752 - \frac{92^2}{5} = 59,5$.

Вычисляем коэффициент детерминации: $R^2 = 1 - \frac{SSe}{SSy} = 1 - \frac{0,914}{59,5} = 0,985$, то

есть в данном примере 98,5 % изменчивости y объясняется квадратичной зависимостью от x .

Коэффициент (множественной) корреляции здесь равен $R = \sqrt{0,985} = 0,992$.

В отличие от индекса детерминации (другой меры тесноты корреляционной связи, введенной в предыдущей лекции о дисперсионном анализе) при равенстве нулю коэффициента детерминации еще нельзя утверждать, что корреляционной связи нет вообще.

На рис. 14.5 изображена функциональная (то есть наиболее тесная) квадратичная зависимость, которую ошибочно попытались аппроксимировать линейной моделью.

Ввиду симметрии расположения заданных точек наилучшая линейная модель получилась в виде $y_p = \bar{y}$, для которой $R^2 = 0$.

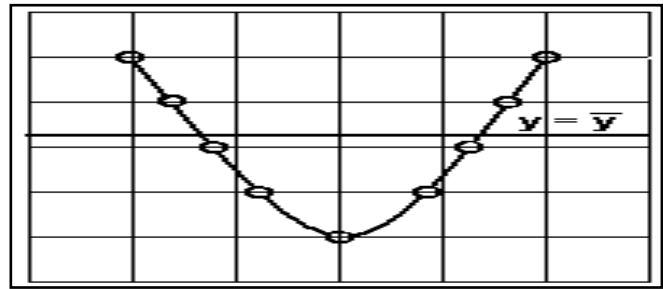


Рис. 14.5. Линейная аппроксимация нелинейной зависимости

Однофакторная линейная зависимость

Для этого частного случая можно получить готовые формулы для МНК-оценок параметров модели и коэффициента корреляции.

Систему нормальных уравнений для линейной модели $y = b_0 + b_1 x + e$

$$\begin{cases} \bar{y} = b_0 + b_1 \bar{x} \\ \overline{xy} = b_0 \bar{x} + b_1 \overline{x^2} \end{cases}$$

можно решить в общем виде и получить формулы для расчета коэффициента регрессии $b_1 = \frac{s_{xy}}{s_x^2}$ и свободного члена $b_0 = \bar{y} - b_1 \bar{x}$.

Из формулы для дисперсии остатка модели:

$$s_e^2 = \overline{y^2} - (b_0 \bar{y} + b_1 \overline{yx}) = (\overline{y^2} - \bar{y}^2) - b_1 (\overline{yx} - \bar{y} \cdot \bar{x}) = s_y^2 - b_1 s_{xy} = s_y^2 - \left(\frac{s_{xy}}{s_x} \right)^2$$

получаем коэффициент детерминации в виде:

$$R^2 = 1 - \frac{s_e^2}{s_y^2} = \left(\frac{s_{xy}}{s_x s_y} \right)^2.$$

Извлекаем корень квадратный из коэффициента детерминации и получаем коэффициент парной корреляции (Пирсона):

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

Таким образом, коэффициент множественной корреляции R для случая однофакторной линейной зависимости совпадает (по модулю) с коэффициентом парной корреляции $R = |r_{xy}|$. В общем же случае он совпадает с коэффициентом парной корреляции между расчетными и наблюдаемыми значениями $R = r_{yy_p}$.

Коэффициент парной корреляции r_{xy} и его квадрат (коэффициент детерминации R^2) оценивает тесноту линейной связи. Если $r_{xy} = 0$, линейной связи нет; при $r_{xy} = \pm 1$ имеем точную линейную зависимость.

На рис. 14.6 изображены возможные ситуации при разных значениях r_{xy} .

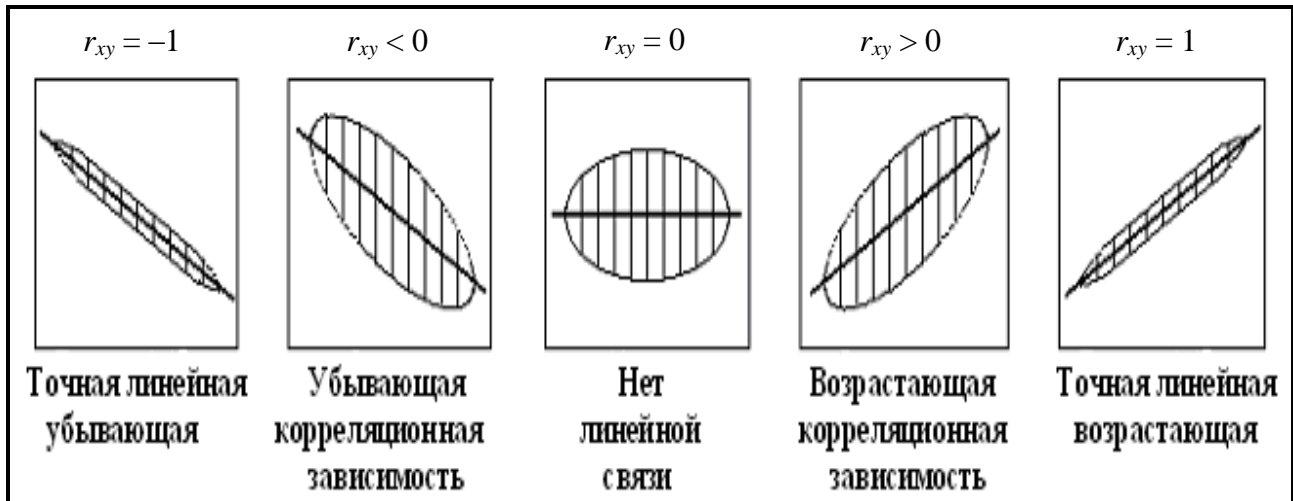


Рис. 14.6. Различные случаи тесноты связи

С использованием коэффициента парной корреляции ранее найденные формулы можно записать несколько в иной форме:

$$b_1 = r_{xy} \frac{s_y}{s_x};$$

$$s_e^2 = s_y^2 (1 - r_{xy}^2).$$

Уравнение регрессии удобно записать в стандартизованных переменных:

$$\frac{y - \bar{y}}{s_y} = r_{xy} \cdot \frac{x - \bar{x}}{s_x}.$$

Если изменить направление причинно-следственных связей, то получим очень похожее уравнение сопряженной регрессии:

$$\frac{x - \bar{x}}{s_x} = r_{xy} \cdot \frac{y - \bar{y}}{s_y}.$$

Если же оба показателя (x и y) являются следствиями одной и той же общей причины, то наилучшим описанием связи будет диагональная регрессия (Фриша):

$$\frac{y - \bar{y}}{s_y} = \pm \frac{x - \bar{x}}{s_x},$$

где знак \pm соответствует знаку ковариации s_{xy} или знаку коэффициента корреляции r_{xy} .

Нелинейные двухпараметрические модели

Как уже было сказано выше, для успешного применения МНК желательно, чтобы форма связи была линейной относительно параметров модели. Для двухпараметрических зависимостей, которые линейно зависят от параметров или могут быть приведенными к такой форме функциональными преобразованиями, существует графический способ проверки их пригодности для описания данных (иными словами, существует графический способ проверки адекватности модели).

Пусть $Y = F(x, y)$ и $X = \Phi(x, y)$ – такие функциональные преобразования, после которых форма связи формально приводится к линейному виду:

$$Y = a + bX.$$

Чаще всего используется или логарифмирование, или переход к обратным величинам. На рис. 14.7 приведены справочные сведения о нелинейных двухпараметрических зависимостях, которые могут быть сведены к линейным указанными функциональными преобразованиями.

Для наглядности на рис. 14.7 в клетках таблицы приведены эскизы графиков типовых зависимостей в исходных координатах (x, y).

Графиком линейной зависимости является прямая, а прямую человек уверенно выделяет среди множества других кривых. Отсюда следует такое правило: если в преобразованных координатах (X, Y) эмпирические точки не группируются вокруг какой-либо прямой, принятая форма связи не является адекватной.

С использованием современной вычислительной техники любые графики легко строятся и преобразуются, поэтому описанный способ идентификации формы связи является достаточно эффективным.

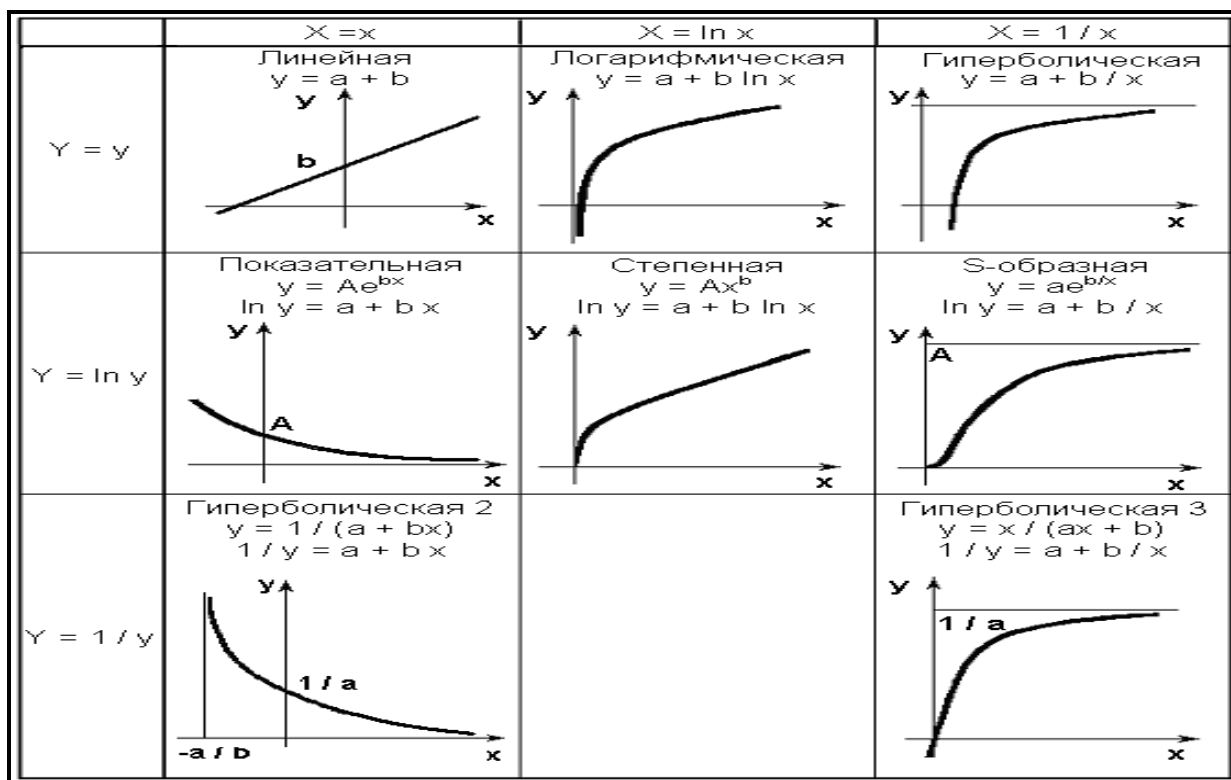


Рис. 14.7. Двухпараметрические зависимости $Y = a + bX$

Ни одна из двухпараметрических зависимостей, приведенных на рис. 14.7, не допускает существования оптимума (максимума или минимума). Если по смыслу задачи ожидается наличие экстремума (оптимума), то следует применять трехпараметрические формы связи, например квадратичную модель. Квадратичная модель с функциональными преобразованиями переменных способна описывать довольно широкий класс зависимостей с экстремумами.

Вопросы для самопроверки

1. Сформулируйте задачу регрессионного анализа.
2. Что такое линия регрессии и уравнение регрессии?
3. Что такое сопряженные уравнения и линии регрессии?
4. В чем заключается принцип наименьших квадратов?
5. Как составляется система нормальных уравнений?
6. Что такое коэффициент детерминации?
7. Чем коэффициент детерминации отличается от индекса детерминации?

Перечислите их свойства.

8. Опишите наиболее распространенные двухпараметрические нелинейные зависимости.

9. Как графически проверить правильность выбора формы связи?

15. Проблема значимости и адекватности регрессионной модели

Оценка значимости регрессионной модели

В регрессионной модели «полный сигнал» – наблюдаемые значения y – разлагается на две компоненты: «полезный сигнал» – расчетные значения y_p , которые определяются моделью (значениями аргументов x_1, x_2), и «помеху» – ошибки модели e :

$$y = y_p + e,$$

где, например, $y_p = b_0 + b_1x_1 + b_2x_2$ для двухфакторной линейной модели.

В лекции 14 об основах регрессионного анализа было показано, что точно такое же разложение имеет общая сумма квадратов отклонений $SSy = SS_p + SS_e$.

Покажем, что такое же разложение имеет также число степеней свободы $dfy = dfp + dfe$:

$dfy = n - 1$, так как на n отклонений $(y_i - \bar{y})$ наложена одна связь – сумма всех этих отклонений равна нулю $[y - \bar{y}] = 0$ (центральное свойство среднего);

$dfe = n - 1 - m$, где m – число объясняющих переменных. Для определения параметров модели принимаются условия ортогональности ошибок к каждому члену модели $[e] = 0$, $[ex_1] = 0$, $[ex_2] = 0$ – это связи, наложенные на отклонения ошибок от их среднего значения. Обычно в модели число определяемых параметров на единицу превышает число аргументов из-за обязательного наличия в модели свободного члена b_0 (кстати, наличие в модели свободного члена приводит к равенству нулю среднего значения ошибки $\bar{e} = 0$ и равенству средних $\bar{y}_p = \bar{y}$).

Для числа степеней свободы расчетных значений должно получиться: $dfp = dfy - dfe = (n - 1) - (n - 1 - m) = m$.

Рассмотрим отклонения расчетных значений от среднего значения: $y_p - \bar{y} = (b_0 + \sum b_j x_j) - \bar{y} = \sum b_j (x_j - \bar{x}_j)$. При преобразовании было использовано первое уравнение нормальной системы (см. лекцию 14) $\bar{y} = b_0 + \sum b_j \bar{x}_j$ – следствие условия $[e] = 0$.

Напоминаем, что в регрессионном анализе все объясняющие переменные x_j считаются неслучайными, поэтому оказалось, что все отклонения расчетных значений от своего среднего $(y_p - \bar{y})$ являются разными линейными комбинациями m случайных величин b_j с неслучайными коэффициентами $(x_j - \bar{x}_j)$.

Отсюда следует, что независимыми могут быть только m таких комбинаций, то есть $dfp = m$.

Для проверки значимости модели заполним на рис. 15.1 таблицу дисперсионного анализа 1, причем выразим суммы квадратов $SSp = R^2 \cdot SSy$ и $SSe = (1 - R^2) \cdot SSy$ через общую сумму квадратов SSy и коэффициент детерминации R^2 .

Источник изменчивости	Суммы квадратов	ЧСС	Средние квадраты	Дисперсионное отношение
Регрессия	$SSp = R^2 \cdot SSy$	$dfp = m$	$MSp = SSp / dfp$	$F_p = MSp / MSe$
Остаток модели	$SSe = (1 - R^2) \cdot SSy$	$dfe = n - 1 - m$	$MSe = SSe / dfe$	
Общая	$SSy = ns_x^2$	$dfy = n - 1$		

Рис. 15.1. Таблица дисперсионного анализа для оценки значимости модели

Получено следующее выражение для дисперсионного отношения Фишера

$$F_p = \frac{SSp}{SSe} \cdot \frac{dfe}{dfp} = \frac{R^2}{1-R^2} \cdot \frac{n-1-m}{m},$$

которое надо сравнивать с табличными значениями $F_{0,05}(dfp; dfe)$ и $F_{0,01}(dfp; dfe)$.

Для одномерного случая ($m = 1$) ЧСС $dfp = 1$ и дисперсионное отношение

$$F_p = \frac{r_{xy}^2}{1-r_{xy}^2} \cdot \frac{n-2}{1}$$

надо сравнивать с табличными значениями $F_\alpha(1; n-2) = t_\alpha^2(n-2)$, где $\alpha = 0,05$ и $0,01$. Интересно, что для линейной однофакторной зависимости мера тесноты связи r_{xy} и характеристика ее значимости F_p получаются одинаковыми для обеих сопряженных моделей.

Регрессионная модель считается значимой, если вычисленное значение дисперсионного отношения будет больше верхней границы $F_p > F_{0,01}$; модель признается незначимой, если $F_p < F_{0,05}$.

Оценка значимости корреляционной связи

Коэффициент детерминации (и коэффициент корреляции) представляет собой меру тесноты связи выбранной формы. Ошибка неверного выбора вида

уравнения регрессии (ошибка спецификации модели) может привести к совершенно неверным выводам относительно оценки тесноты реально существующей связи. В некоторых случаях, когда данные опыта даны в нескольких повторениях, можно найти меру чисто случайной изменчивости s_{ε}^2 (дисперсию данных по повторениям – дисперсию «внутри групп»); тогда вычисляют более объективную меру тесноты связи – индекс детерминации (и корреляционное отношение). В отличие от коэффициента детерминации $R^2 = 1 - \frac{SSe}{SSy}$ при вычислении индекса детерминации $\eta^2 = 1 - \frac{SSe}{SSy}$ не используются никакие предположения о форме корреляционной связи.

Однако параллельные наблюдения (повторения) имеют место только для планируемых опытов (активных экспериментов), что характерно для опытов физических, химических, биологических, там, где исследователь может контролировать условия опыта. В экономике же данные представляют собой наблюдения неконтролируемого процесса (пассивный эксперимент), поэтому варианты опыта почти никогда не повторяются.

Выше уже говорилось, что при понижении шкал измерения теряется какая-то часть информации, но выводы анализа становятся более общими, более объективными. При анализе парных зависимостей полезно перейти к дискретным шкалам измерения обеих переменных, то есть произвести двойную группировку данных на несколько небольших интервалов по осям \mathcal{X}, \mathcal{Y} .

Если обозначить через X_i и Y_j центры интервалов, то для каждой клетки таблицы размером $p \times q$ можно подсчитать частоты m_{ij} – количество наблюдений, попадающих в данную клетку. Все данные, попадающие в одну клетку таблицы с центром (X_i, Y_j) , считаются одинаковыми (это вносит в расчеты некоторую ошибку группировки). Сумма всех частот равна общему количеству данных $n = \sum \sum m_{ij}$. Часто такую таблицу называют корреляционной (рис. 15.2).

	X_1	X_2	X_3	\dots	X_p	$l = \sum m$	$v = \bar{x}_y$
Y_1	m_{11}	m_{21}	m_{31}	\dots	m_{p1}	l_1	v_1
Y_2	m_{12}	m_{22}	m_{32}	\dots	m_{p2}	l_2	v_2
\dots	\dots	\dots	\dots	\dots	\dots	\dots	\dots
Y_q	m_{1q}	m_{2q}	m_{3q}	\dots	m_{pq}	l_q	v_q
$k = \sum m$	k_1	k_2	k_3	\dots	k_p	n	
$u = \bar{y}_x$	u_1	u_2	u_3	\dots	u_p		

Рис. 15.2. Корреляционная таблица

Теперь суммирование по всем наблюдениям должно учитывать частоты повторения одинаковых данных, например $[xy] \approx \sum \sum m_{ij} X_i Y_j$. Сравнительные расчеты коэффициента корреляции по исходным r_{xy} и по сгруппированным r_{xy} данным дают представление о величине ошибок группировок.

Переход к сгруппированным данным позволяет получить дополнительную информацию о форме связи, более объективную меру тесноты существующей корреляционной связи и даже скорректировать наши предположения о возможном направлении причинно-следственных связей. Имея таблицу сгруппированных данных, можно для каждого значения X_i вычислить средние групповые

$u_i = \bar{y}_{x_i} = \frac{1}{k_i} \sum_{j=1}^q m_{ij} Y_j$, где $k_i = \sum_{j=1}^q m_{ij}$ – суммы частот по столбцам таблицы. Ана-

логично, для каждого значения Y_j можно вычислить средние групповые

$v_j = \bar{x}_{y_j} = \frac{1}{l_j} \sum_{i=1}^p m_{ij} X_i$, где $l_j = \sum_{i=1}^p m_{ij}$ – суммы частот по строкам таблицы.

Теперь появилась возможность для каждой из сопряженных зависимостей вычислить индексы детерминации:

$$\eta_{y/x}^2 = \frac{SSU}{SSY}; \quad \eta_{x/y}^2 = \frac{SSV}{SSX},$$

которые показывают, какая часть полной изменчивости результативной переменной объясняется наличием корреляционной связи (произвольного типа, не обязательно линейного). Оба корреляционных отношения превышают абсолютную величину коэффициента корреляции (вычисленного по сгруппированным данным):

$$\eta_{y/x}, \eta_{x/y} > |r_{XY}|.$$

Если одно из корреляционных отношений существенно превышает другое, то это является доводом в пользу выбора соответствующего направления причинно-следственных связей.

Кусочно-линейные графики средних групповых (X_i, u_i) и (v_j, Y_j) называются эмпирическими линиями регрессии. Эти графики дают возможность визуально определить вид нелинейности и выбрать более подходящую форму связи, чем традиционная линейная форма, которая часто принимается по умолчанию.

С помощью дисперсионного анализа проверяется значимость наиболее тесной корреляционной связи. Если в результате дисперсионного анализа окажется, что корреляционная связь незначимая, то незачем проводить регрессионный анализ связи заданной формы, она также будет незначимой.

На рис. 15.3 приведена заполненная таблица дисперсионного анализа 2 для проверки значимости корреляционной связи y/x , причем суммы квадратов $SSU = \eta^2 \cdot SSY$ и $SS\varepsilon = (1 - \eta^2) \cdot SSY$ выражены через общую сумму квадратов SSY и индекс детерминации $\eta^2 = \eta_{y/x}^2$.

Изменчивость	Суммы квадратов	ЧСС	Средние квадраты	Дисперсионное отношение
Средние групповые	$SSU = \eta^2 \cdot SSy$	$dfU = p - 1$	$MSU = SSU / dfU$	$F_\eta = MSU / MS\varepsilon$
Случайность	$SS\varepsilon = (1 - \eta^2) \cdot SSy$	$df\varepsilon = n - p$	$MS\varepsilon = SS\varepsilon / df\varepsilon$	
Общая	$SSY = ns_Y^2$	$dfy = n - 1$		

Рис. 15.3. Таблица дисперсионного анализа для оценки значимости корреляционной связи

Получено следующее выражение для дисперсионного отношения Фишера:

$$F_\eta = \frac{SSU}{SS\varepsilon} \cdot \frac{df\varepsilon}{dfU} = \frac{\eta^2}{1-\eta^2} \cdot \frac{n-p}{p-1},$$

которое надо сравнивать с табличными значениями $F_{0,05}(dfU; df\varepsilon)$ и $F_{0,01}(dfU; df\varepsilon)$.

Если окажется, что $F_\eta < F_{0,05}$, делаем вывод об отсутствии корреляционной связи (какой-либо формы).

Проверка адекватности модели

Недаром ошибки модели e называются остатками модели, поскольку кроме случайных ошибок, в них включаются систематические ошибки выбора неверной формы связи (ошибки спецификации модели).

Если у нас есть мера чисто случайной изменчивости (дисперсия данных по повторениям опыта), то остатки модели e можно разложить на две компоненты – случайную ε и систематическую Ω : $e = \Omega + \varepsilon$.

Точно так же разлагается сумма квадратов отклонений $SSe = SS\Omega + SS\varepsilon$ и число степеней свободы $dfe = df\Omega + df\varepsilon$.

Для проверки значимости систематической ошибки Ω (ошибки неадекватности модели) заполняем таблицу дисперсионного анализа 3 (рис. 15.4).

Две строки этой таблицы («Остаток модели» и «Случайность») дублируют соответствующие строки таблиц дисперсионных анализов 1 и 2.

Получено следующее выражение для дисперсионного отношения Фишера:

$$F_A = \frac{SS\Omega}{SS\varepsilon} \cdot \frac{df\varepsilon}{df\Omega} = \frac{\eta^2 - R^2}{1 - \eta^2} \cdot \frac{n-p}{p-1-m}.$$

Изменчивость	Суммы квадратов	ЧСС	Средние квадраты	Дисперсионное отношение
Неадекватность	$SS\Omega =$ $= (\eta^2 - R^2) \cdot SSy$	$df\Omega =$ $= p - 1 - m$	$MS\Omega =$ $= SS\Omega / df\Omega$	$F_A =$ $= MS\Omega / MS\varepsilon$
Случайность	$SS\varepsilon =$ $= (1 - \eta^2) \cdot SSy$	$df\varepsilon = n - p$	$MS\varepsilon =$ $= SS\varepsilon / df\varepsilon$	
Остаток модели	$SSe =$ $= (1 - R^2) \cdot SSy$	$dfe =$ $= n - 1 - m$	$MSe =$ $= SSe / dfe$	

Рис. 15.4. Таблица дисперсионного анализа для проверки адекватности модели

Вычисленное дисперсионное отношение Фишера F_A надо сравнивать с табличными значениями $F_{0,05}(df\Omega; df\varepsilon)$ и $F_{0,01}(df\Omega; df\varepsilon)$. Если окажется, что $F_A < F_{0,05}$, систематической ошибкой можно пренебречь и считать модель адекватной. Но если окажется, что $F_A > F_{0,01}$, то систематической ошибкой пренебречь нельзя, придется искать более подходящую форму связи.

Пример. На рис. 15.5 в корреляционной таблице приведены результаты двойной группировки данных $n = 154$ наблюдений; принято $p = 7$ интервалов равной ширины по переменной X и $q = 6$ интервалов равной ширины по переменной Y .

$\begin{matrix} X \\ Y \end{matrix}$	1	2	3	4	5	6	7	$l=\Sigma m$	ΣmX	V
6	3	0	0	0	0	0	0	3	3	1
5	2	8	3	0	0	0	0	13	27	2,077
4	2	17	4	3	0	1	0	27	66	2,444
3	0	7	20	13	4	0	0	44	146	3,318
2	0	0	4	12	20	5	4	45	218	4,844
1	0	0	0	5	9	6	2	22	115	5,227
$k=\Sigma m$	7	32	31	33	33	12	6	154		
ΣmY	36	129	99	80	61	20	10			
U	5,143	4,031	3,194	2,424	1,848	1,667	1,667			

Рис. 15.5. Корреляционная таблица размером 6×7

При группировках на интервалы равной ширины всегда можно перейти к условным переменным (линейные преобразования переменных) так, чтобы в новых переменных центры интервалов выражались последовательными целыми числами (номерах интервалов). Эти линейные преобразования переменных не изменяют ни последующих выводов анализа, ни вида графиков, у которых будет только другая разметка осей.

В последней строке и последнем столбце таблицы вычислены средние групповые U_i и V_j .

Вычисляем параметры линейной модели.

$$X_c = \sum k_i X_i / n = 3,734; (X^2)_{cp} = \sum k_i (X_i)^2 / n = 16,188; (s_X)^2 = (X^2)_{cp} - (X_{cp})^2 = 2,247;$$

$$Y_{cp} = \sum l_j Y_j / n = 2,825; (Y^2)_{cp} = \sum l_j (Y_j)^2 / n = 9,500; (s_Y)^2 = (Y^2)_{cp} - (Y_{cp})^2 = 1,521;$$

$$(XY)_{cp} = \sum m_{ij} X_i Y_j / n = 9,130; s_{XY} = (XY)_{cp} - X_{cp} \cdot Y_{cp} = -1,417;$$

$$r_{XY} = s_{XY} / (s_X \cdot s_Y) = -0,766; (r_{XY})^2 = 0,587;$$

$$b_1 = r_{XY} \cdot (s_Y / s_X) = -0,630; b_0 = Y_{cp} - b_1 \cdot X_{cp} = 5,179.$$

Линейной моделью $Y_p = 5,179 - 0,630 \cdot X$ объясняется 58,7 % общей изменчивости данных. Эта модель значима, так как дисперсионное отношение

$$F_p = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot \frac{n-2}{1} = 216,2 \text{ превышает табличное значение } F_{0,01}(1; 152) = 6,80.$$

Вычисляем индексы детерминации.

$$p = 7; U_{cp} = Y_{cp} = 2,825; (U^2)_{cp} = \sum k_i (U_i)^2 / n = 8,948;$$

$$(s_U)^2 = (U^2)_{cp} - (U_{cp})^2 = 0,969;$$

$$q = 6; V_{cp} = X_{cp} = 3,734; (V^2)_{cp} = \sum l_j (V_j)^2 / n = 15,338;$$

$$(s_V)^2 = (V^2)_{cp} - (V_{cp})^2 = 1,397;$$

$$\eta_{y/x}^2 = \frac{s_U^2}{s_Y^2} = \frac{0,969}{1,521} = 0,637; \quad \eta_{x/y}^2 = \frac{s_V^2}{s_X^2} = \frac{1,397}{2,247} = 0,622.$$

Корреляционной зависимостью y/x объясняется 63,7 % общей изменчивости данных (сопряженной зависимостью x/y объясняется 62,2 %). Корреляционная зависимость значима, так как дисперсионное отношение

$$F_\eta = \frac{\eta^2}{1 - \eta^2} \cdot \frac{n-p}{p-1} = 43,05 \text{ превышает табличное значение } F_{0,01}(6; 147) = 2,93.$$

Проверяем адекватность линейной модели. Вычисляем дисперсионное отношение $F_A = \frac{\eta^2 - r_{XY}^2}{1 - \eta^2} \cdot \frac{n-p}{p-2} = 4,06$ и сравниваем его с табличными значениями $F_{0,05}(5; 147) = 2,28$ и $F_{0,01}(5; 147) = 3,14$. Так как вычисленное значение $F_A = 4,06 > F_{0,01}$, систематической ошибкой пренебречь нельзя, линейная модель неадекватная, требуется найти более подходящую нелинейную форму связи.

На рис. 15.6а изображены графики эмпирической и теоретической регрессии, откуда видно, что, действительно, зависимость нелинейная, узлы эмпирической линии регрессии закономерно уклоняются от графика линейной регрессии.

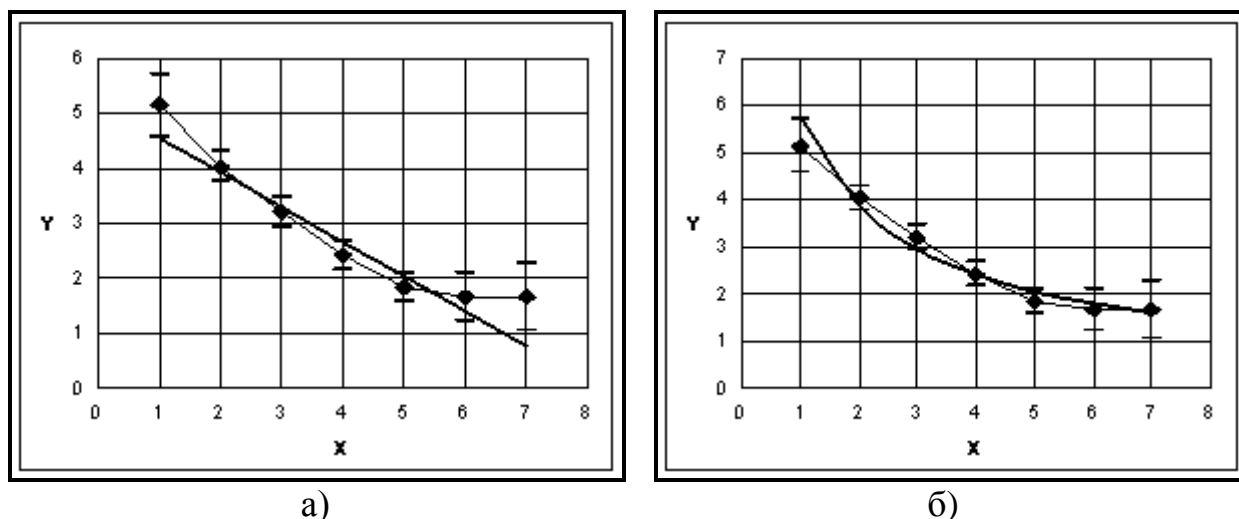


Рис. 15.6. Соответствие между эмпирической и теоретической регрессиями

(*a* – линейная модель $Y_p = 5,179 - 0,630 \cdot X$; *б* – нелинейная модель

$$Y_p = 0,212 + \frac{11,03}{X+1})$$

Для каждого узла (среднего группового) построены 95-процентные доверительные интервалы шириною $\pm HCP_{0,05}$, где $HCP_{0,05} = t_{0,05} \sqrt{\frac{ns_Y^2}{df\varepsilon} \cdot \frac{1-\eta^2}{k_i}} = \frac{1,50}{\sqrt{k_i}}$.

Крайние узлы на рис. 15.6а существенно уклоняются от линейной регрессии, ее график не пересекает крайних доверительных интервалов.

На рис. 15.6б построен график нелинейной зависимости $Y_p = a + \frac{b}{X+1}$, который пересекает доверительные интервалы для всех узлов эмпирической линии регрессии. Коэффициент детерминации возрос до $R^2 = 0,611$; дисперсионное отношение $F_A = 2,12$ понизилось и стало уже меньше табличного $F_A < F_{0,05}$. Найденная нелинейная модель адекватная.

Таблицы сопряженности и коэффициенты контингенции

Если обе переменные качественные, измеренные в наиболее общей шкале имен, то таблицу частот m_{ij} совместного появления категорий (X_i , Y_j) разных переменных называют таблицей сопряженности. В этой таблице X_i и Y_j – имена категорий (не числа), поэтому никакие арифметические операции с ними невозможны. Как и для корреляционной таблицы, подсчитывается общая сумма частот n , а также суммы частот по столбцам k_i и строкам l_j таблицы.

Относительные частоты $\frac{k_i}{n}$, $\frac{l_j}{n}$ есть оценки вероятностей появления категорий X_i и Y_j . Проверяется гипотеза о независимости качественных переменных \mathcal{X} , \mathcal{Y} (нуль-гипотеза). Имеется возможность определить теоретические

частоты \tilde{m}_{ij} совместного появления любой комбинации категорий (X_i, Y_j) , которые ожидаются при справедливости нуль-гипотезы. Действительно, при взаимной независимости категорий (X_i, Y_j) вероятность совместного появления такой комбинации равна произведению их вероятностей $\frac{k_i}{n} \cdot \frac{l_j}{n}$, откуда получаем ожидаемые частоты в виде $\tilde{m}_{ij} = \frac{k_i l_j}{n}$.

Наблюдаемые и ожидаемые частоты сравниваем по критерию Пирсона:

$$\chi^2 = \sum \sum \frac{(m_{ij} - \tilde{m}_{ij})^2}{\tilde{m}_{ij}} = \sum \sum \frac{m_{ij}^2}{\tilde{m}_{ij}} - n = n \cdot \left(\sum \sum \frac{m_{ij}^2}{k_i l_j} - 1 \right).$$

Табличные значения χ_α^2 находим для ЧСС = $(p-1)(q-1)$, где p, q – число категорий для \mathcal{X}, \mathcal{Y} .

Если окажется, что $\chi^2 > \chi_{0,01}^2$, нуль-гипотеза отклоняется и делается вывод о том, что переменные \mathcal{X}, \mathcal{Y} связаны между собой. Тогда появляется проблема оценки тесноты этой связи. Предложено несколько мер тесноты связи между качественными переменными, из которых мы рассмотрим две – коэффициент контингенции Крамера $C = \sqrt{\frac{\chi^2}{\chi_{\max}^2}}$ и коэффициент контингенции Кендала

$K = \sqrt{\frac{\chi^2}{\chi_{\max}^2 + n}}$. При абсолютном совпадении наблюдаемых и ожидаемых частот статистика Пирсона χ^2 равна нулю и равны нулю оба коэффициента контингенции.

Максимальное значение χ_{\max}^2 получается при наиболее тесной связи, когда каждой категории одной переменной соответствует только одна категория другой переменной (функциональное соответствие).

Так как категории можно переставлять, при наиболее тесной связи таблица сопряженности приобретает блочно-диагональный вид (рис. 15.7). Пусть $p > q$ (например, $p = 4, q = 3$). Вычисляем для этого случая статистику Пирсона:

	X_1	X_2	X_3	X_4	l_j
Y_1	m_{11}	m_{12}			$m_{11}+m_{12}$
Y_2			m_{22}		m_{22}
Y_3				m_{33}	m_{33}
k_i	m_{11}	m_{12}	m_{22}	m_{33}	n

Рис. 15.7. Функциональная связь между категориями

$$\begin{aligned}\chi_{\max}^2 &= n \cdot \left(\sum \sum \frac{m_{ij}^2}{k_i l_j} - 1 \right) = n \cdot \left(\frac{m_{11}^2}{m_{11}(m_{11}+m_{12})} + \frac{m_{12}^2}{m_{12}(m_{11}+m_{12})} + \frac{m_{22}^2}{m_{22}m_{22}} + \frac{m_{33}^2}{m_{33}m_{33}} - 1 \right) = \\ &= n \cdot \left(\frac{m_{11}}{m_{11}+m_{12}} + \frac{m_{12}}{m_{11}+m_{12}} + 1 + 1 - 1 \right) = n \cdot (1 + 1 + 1 - 1) = n \cdot (3 - 1) = n \cdot (q - 1).\end{aligned}$$

Таким образом, коэффициент контингенции Крамера можно записать в виде:

$$C = \sqrt{\frac{\chi^2}{\chi_{\max}^2}} = \sqrt{\frac{\chi^2}{n \cdot (d-1)}},$$

где $d = \min\{p, q\}$.

Коэффициент контингенции Кендала изменяется от 0 до $\sqrt{\frac{d-1}{d}} < 1$.

Скорректируем его: $KK = \sqrt{\frac{\chi^2}{\chi^2 + n} \cdot \frac{d}{d-1}}$.

Как правило, оказывается, что $C \leq K \leq KK$.

Пример. Рассмотрим корреляционную таблицу на рис. 15.5 размером 6×7 с числом наблюдений $n = 154$ и будем считать значения переменных X , Y именами различных категорий. Суммы частот по столбцам и строкам таблицы уже найдены.

Ниже в таблице такого же размера (рис. 15.8) подсчитаны отношения $\frac{m_{ij}^2}{k_i l_j}$.

$\begin{smallmatrix} X \\ Y \end{smallmatrix}$	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Y_6	0,429	0	0	0	0	0	0
Y_5	0,044	0,154	0,022	0	0	0	0
Y_4	0,021	0,334	0,019	0,010	0	0,003	0
Y_3	0	0,035	0,293	0,116	0,011	0	0
Y_2	0	0	0,011	0,097	0,269	0,046	0,059
Y_1	0	0	0	0,034	0,112	0,136	0,030

Рис. 15.8. Расчет статистики Пирсона

Вычисляем их сумму (2,29) и статистику Пирсона $\chi^2 = 154 \cdot (2,29 - 1) = 198,7$, которую сравниваем с табличным значением $\chi_{0,01}^2(6 \cdot 5) = 15,0$.

Так как $\chi^2 > \chi_{0,01}^2$, делаем вывод о существовании значимой связи между \mathcal{X} и \mathcal{Y} .

Коэффициенты контингенции Крамера, Кендала и скорректированный коэффициент KK равны соответственно:

$$C = \sqrt{\frac{198,7}{154 \cdot 5}} = 0,508, \quad K = \sqrt{\frac{198,7}{198,7+154}} = 0,751, \quad KK = K \cdot \sqrt{\frac{6}{5}} = 0,822.$$

Сравним эти меры с коэффициентом корреляции $|r_{XY}| = 0,766$ и с корреляционным отношением $\eta = \sqrt{0,637} = 0,798$.

Соответствие между скорректированным коэффициентом контингенции Кендала и корреляционным отношением – самой объективной мерой тесноты корреляционной связи между количественными переменными – очень хорошее ($KK \approx \eta$).

Коэффициент ранговой корреляции Спирмена

Если X, Y – порядковые переменные, то с ними не допустимы никакие арифметические операции, например, разность двух значений $(x_j - x_i)$ ничего не означает, так как из сравнения $x_j > x_i$ следует только, что одно значение больше другого, но неизвестно, *на сколько* больше.

Если переменные ранжированы, то их ранги являются номерами при расположении значений переменной в порядке возрастания какого-то признака. Так, из сравнения рангов $x_3 = 3$ и $x_6 = 6$ следует, что между элементами x_3 и x_6 есть еще два элемента с рангами $x_4 = 4$ и $x_5 = 5$.

Если несколько элементов неразличимы по данному признаку, то им всем присваивается средний ранг из их номеров по порядку. Такие группы переменных называются связками.

Спирмен вывел формулу для оценки тесноты связи между ранжированными переменными, причем при выводе не использовались никакие сомнительные арифметические операции. Формула эта достаточно простая при отсутствии связок, но усложняется при их наличии.

Кендал доказал, что коэффициент ранговой корреляции Спирмена численно равен коэффициенту парной корреляции Пирсона, если ранги считать числовыми значениями переменных.

Еще раз отметим, что вовсе не утверждается, что с рангами всегда можно поступать, как с обычными числами, но коэффициент ранговой корреляции можно рассчитывать обычным образом вручную или по готовым программам на компьютере.

Вывод формулы для коэффициента ранговой корреляции Спирмена

Пусть p_k, q_k – ранги двух показателей \mathcal{A} и \mathcal{B} . Рассмотрим случай отсутствия связок (групп одинаковых рангов).

Наблюдения всегда можно отсортировать в порядке возрастания одной из переменных: $p_k = k = 1, 2, 3, \dots, n$.

Ранги q_k – те же числа, но в другом порядке.

Мерой тесноты связи между показателями \mathcal{A} и \mathcal{B} может быть сумма квадратов разностей рангов:

$$S = \sum_{k=1}^n (p_k - q_k)^2.$$

Если ранги двух показателей совпадают $p_k = q_k$, то $S = 0$, и это соответствует наиболее тесной положительной связи.

Если порядок следования q_k противоположен порядку следования p_k , то $S = S_{\max}$, что соответствует наиболее тесной отрицательной связи. Необходимо найти величину S_{\max} . Для этого случая имеем $p_k + q_k = n + 1$, $p_k = k$, $q_k = n + 1 - k$, $p_k - q_k = 2k - (n + 1)$. Отсюда следует:

$$S_{\max} = \sum (2k - (n + 1))^2 = 4 \cdot \sum k^2 - 4 \cdot (n + 1) \cdot \sum k + (n + 1)^2 \cdot n.$$

Поскольку известны формулы для сумм и сумм квадратов последовательных целых чисел:

$$\sum_{k=1}^n k = 1 + 2 + 3 + \dots + n = \frac{n(n+1)}{2}, \quad \sum_{k=1}^n k^2 = 1^2 + 2^2 + 3^2 + \dots + n^2 = \frac{n(n+1)(2n+1)}{6},$$

то окончательно получаем:

$$S_{\max} = 4 \frac{n(n+1)(2n+1)}{6} - 4(n+1) \frac{n(n+1)}{2} + (n+1)^2 n = \frac{(n-1)n(n+1)}{3}.$$

Вместо меры S вводим меру связи Спирмена:

$$\rho = 1 - 2 \frac{S}{S_{\max}} = 1 - 6 \frac{\sum (p_k - q_k)^2}{(n-1)n(n+1)},$$

которая равна $\rho = 1$ для $S = 0$ (для наиболее тесной положительной связи) и $\rho = -1$ для $S = S_{\max}$ (для наиболее тесной отрицательной связи).

Полученная формула существенно усложняется при наличии связок – групп неразличимых объектов, для которых принимаются одинаковые значения рангов, средних для каждой группы.

Пусть t – количество неразличимых объектов в связке для показателя \mathcal{A} , а τ – количество неразличимых объектов в связке для показателя \mathcal{B} .

Вычисляем поправки: $A = \frac{\sum t(t^2-1)}{n(n^2-1)}$, $B = \frac{\sum \tau(\tau^2-1)}{n(n^2-1)}$ и скорректированный ко-

эффициент ранговой корреляции: $\rho_s = \frac{\rho - \frac{A+B}{2}}{\sqrt{(1-A)} \cdot \sqrt{(1-B)}}$.

Пример. Определим тесноту связи между уровнем механизации работ \mathcal{X} и производительностью труда \mathcal{Y} по 10-ти промышленным предприятиям.

На рис. 15.9 данные ранжированные, в рангах показателя \mathcal{Y} имеется одна связка из двух объектов (два предприятия с одинаковой производительностью труда).

k	p_k	q_k	$p_k - q_k$	$(p_k - q_k)^2$	$(p_k)^2$	$(q_k)^2$	$p_k q_k$
1	1	4	-3	9	1	16	4
2	2	1	1	1	4	1	2
3	3	2	1	1	9	4	6
4	4	3	1	1	16	9	12
5	5	7	-2	4	25	49	35
6	6	5	1	1	36	25	30
7	7	6	1	1	49	36	42
8	8	8,5	-0,5	0,25	64	72,25	68
9	9	8,5	0,5	0,25	81	72,25	76,5
10	10	10	0	0	100	100	100
Суммы	55	55	0	18,5	385	384,5	375,5

Рис. 15.9. Ранжированные данные уровня механизации и производительности труда

Вычисляем коэффициент ранговой корреляции Спирмена без поправки на связку: $\rho = 1 - 6 \frac{\sum (p_k - q_k)^2}{(n-1)n(n+1)} = 1 - 6 \frac{18,5}{9 \cdot 10 \cdot 11} = 0,88788$.

Вычисляем поправку $B = \frac{\sum (\tau-1)\tau(\tau+1)}{(n-1)n(n+1)} = \frac{1 \cdot 2 \cdot 3}{9 \cdot 10 \cdot 11} = \frac{1}{165} = 0,00606$ и скорректированный коэффициент ранговой корреляции:

$$\rho_s = \frac{\rho - \frac{A+B}{2}}{\sqrt{(1-A)} \cdot \sqrt{(1-B)}} = \frac{0,88788 - 0,00303}{\sqrt{1-0} \cdot \sqrt{1-0,00606}} = 0,88754.$$

Для сравнения вычисляем обычный коэффициент парной корреляции Пирсона. Все необходимые суммы подсчитаны в вышеприведенной таблице (см. рис. 15.9).

$$s_p^2 = \overline{p^2} - \bar{p}^2 = 38,5 - 5,5^2 = 8,25; \quad s_q^2 = \overline{q^2} - \bar{q}^2 = 38,45 - 5,5^2 = 8,2;$$

$$s_{pq} = \overline{pq} - \bar{p}\bar{q} = 37,55 - 5,5^2 = 7,3; \quad r_{pq} = \frac{s_{pq}}{s_p s_q} = \frac{7,3}{\sqrt{8,25 \cdot 8,2}} = 0,88754.$$

Полученные значения ρ_s и r_{pq} совпали со всеми десятичными знаками.

Покажем, что коэффициент ранговой корреляции Спирмена совпадает с обычным коэффициентом парной корреляции Пирсона, вычисленным по рангам $\rho = r_{pq}$. Так как $p_k = k$, а q_k – те же числа, но в другом порядке, то будут равны средние $\bar{p} = \bar{q} = \frac{n+1}{2}$ и дисперсии $s_p^2 = s_q^2 = \overline{p^2} - \bar{p}^2 = \overline{q^2} - \bar{q}^2$,

$$s_p^2 = s_q^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}.$$

Преобразуем выражение $S = \sum (p_k - q_k)^2 = \sum ((p_k - \bar{p}) - (q_k - \bar{q}))^2$:

$$S = n(s_p^2 + s_q^2 - 2s_{pq}) = 2n(s_p^2 - s_{pq}) = 2ns_p^2(1 - r_{pq}).$$

$$\text{Отсюда: } \rho = 1 - 6 \frac{\sum (p_k - q_k)^2}{n(n^2-1)} = 1 - 6 \frac{2ns_p^2(1-r_{pq})}{12ns_p^2} = 1 - (1 - r_{pq}) = r_{pq}.$$

Итак, формально коэффициент ранговой корреляции Спирмена равняется обычному коэффициенту парной корреляции Пирсона, вычисленному по рангам p_k, q_k .

Вопросы для самопроверки

1. Какая разница между понятиями значимость корреляционной связи и значимость регрессионной модели?
2. Что такое коэффициент детерминации, каковы его свойства?
3. Как вычисляется корреляционное отношение? Каковы его свойства?
4. Как проверяется адекватность регрессионной модели?
5. Чем отличаются расчеты параметров модели по исходным и по сгруппированным данным?
6. Что такое коэффициент контингенции?
7. Как вычисляются коэффициенты контингенции Крамера и Кендала?
8. Как проверить значимость коэффициентов контингенции?
9. Что такое коэффициент ранговой корреляции Спирмена? Как его можно вычислить?

16. Линейный регрессионный анализ в стандартизованных переменных

Традиционно все формулы многомерного линейного регрессионного анализа записывают в стандартизованных переменных:

$$Y = \frac{y - \bar{y}}{s_y}, \quad X_i = \frac{x_i - \bar{x}_i}{s_{x_i}}.$$

В этих переменных многие формулы принимают простейший вид, поэтому сложные вопросы анализа чаще всего обсуждаются именно в стандартизованных переменных.

Стандартизация позволяет выявить некоторые сомнительные значения данных, например выбросы, которые могут появиться в результате ошибок при переписывании и наборе данных. Кроме описок, опечаток, ошибок измерения, выбросы могут быть следствием принадлежности сомнительных данных до другой совокупности (например, когда в выборку включают данные о продукции другого предприятия за другой временной период, когда часть наблюдений измерена другим прибором с другой шкалой калибровки и т. д.). Конечно, такие данные следует удалить из выборки и изучать отдельно. Возможность выявления выбросов основана на правиле «3-х сигм», которое утверждает, что крайне редко встречаются случайные ошибки, превышающие по модулю утроенное стандартное отклонение. Обычно все значения стандартизованных переменных Y , X_i не выходят за пределы интервала $(-3, 3)$, а если встречаются большие отклонения, то такие данные следует выделять и проверять. Чаще всего границы интервала вариации стандартизованных переменных оказываются близкими к $(-2, 2)$.

Сразу же отметим, что, несмотря на более простой вид формул регрессионного анализа в стандартизованных переменных, никакого сокращения объема вычислительной работы не будет, так как добавляются операции нормирования переменных, более сложного составления системы нормальных уравнений и обратного перехода к исходным переменным после завершения вычислений.

Итак, последовательно преобразуем уравнение регрессии

$$y_p = b_0 + b_1x_1 + b_2x_2 + \dots + b_mx_m + e$$

к центрованной и стандартизованной формам:

$$y - \bar{y} = b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) + \dots + b_m(x_m - \bar{x}_m) + e;$$
$$\frac{y - \bar{y}}{s_y} = \beta_1 \frac{(x_1 - \bar{x}_1)}{s_{x_1}} + \beta_2 \frac{(x_2 - \bar{x}_2)}{s_{x_2}} + \dots + \beta_m \frac{(x_m - \bar{x}_m)}{s_{x_m}} + \frac{e}{s_y};$$

$$Y = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m + \varepsilon,$$

где обозначено $Y = \frac{y - \bar{y}}{s_y}$, $X_i = \frac{x_i - \bar{x}_i}{s_{x_i}}$, $\beta_i = b_i \frac{s_{x_i}}{s_y}$, $\varepsilon = \frac{e}{s_y}$.

Внимание! Обычно коэффициенты регрессии b_j и остатки модели e_i рассматривают как оценки соответствующих генеральных значений β_j , ε_i . Однако теперь обозначения β_j , ε_i используются как выборочные оценки (только в стандартизованных переменных).

На стадии центрирования уже было использовано одно из уравнений нормальной системы $\bar{e} = 0$ (или $\bar{\varepsilon} = 0$), поэтому в окончательной записи уравнения регрессии в стандартизованных переменных отсутствует свободный член $\beta_0 = 0$.

Составляем остальные уравнения нормальной системы ($\overline{\varepsilon X_i} = 0$)

$$\overline{YX_i} = \sum_{j=1}^m \beta_j \overline{X_i X_j} \quad (i=1, 2, \dots, m),$$

которую приводим к виду:

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 r_{x_1 x_2} + \dots + \beta_m r_{x_1 x_m} \\ r_{yx_2} = \beta_1 r_{x_2 x_1} + \beta_2 + \dots + \beta_m r_{x_2 x_m} \\ \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \\ r_{yx_m} = \beta_1 r_{x_m x_1} + \beta_2 r_{x_m x_2} + \dots + \beta_m \end{cases},$$

так как для стандартизованных переменных $\overline{X_i X_j} = r_{x_i x_j}$, $\overline{X_i X_i} = r_{x_i x_i} = 1$.

Формулу для расчета остаточной дисперсии получаем, преобразовывая выражение:

$$\overline{Y Y} = \sum_{j=1}^m \beta_j \overline{X_j Y} + \underbrace{\overline{Y \varepsilon}}_{s_\varepsilon^2},$$

где $\overline{Y Y} = r_{yy} = 1$, $\overline{X_j Y} = r_{yx_j}$, $s_\varepsilon^2 = \overline{\varepsilon \varepsilon} = \overline{Y \varepsilon}$:

$$s_\varepsilon^2 = 1 - \beta_1 r_{yx_1} - \beta_2 r_{yx_2} - \dots - \beta_m r_{yx_m}.$$

Отсюда получаем очень простую и легко запоминаемую формулу для расчета коэффициента детерминации:

$$R^2 = 1 - \frac{s_e^2}{s_y^2} = 1 - s_\varepsilon^2 = \beta_1 r_{yx_1} + \beta_2 r_{yx_2} + \dots + \beta_m r_{yx_m}.$$

После решения системы нормальных уравнений и вычисления коэффициента детерминации делаем обратный переход к исходным переменным, пересчитывая коэффициенты регрессии по формулам:

$$b_i = \beta_i \frac{s_y}{s_{x_i}} \quad (i = 1, 2, \dots, m);$$

$$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - \dots - b_m \bar{x}_m.$$

Наконец, получаем выражение для расчета несмещенной оценки остаточной дисперсии:

$$\hat{\sigma}_e^2 = MSe = \frac{n}{dfe} s_e^2 = \frac{n}{dfe} s_y^2 (1 - R^2),$$

где dfe – число степеней свободы остатка модели $dfe = n - m - 1$.

Значимость модели в целом проверяем с помощью критерия Фишера $F = \frac{R^2}{1-R^2} \cdot \frac{n-1-m}{m}$, который надо сравнивать с табличными значениями $F_\alpha(m, dfe)$.

Если обозначить через c_{ij} элементы матрицы, обратной к матрице коэффициентов корреляции $(c_{ij}) = (r_{x_i x_j})^{-1}$, то есть обратной к матрице системы нормальных уравнений в стандартизованной форме, то можно получить такие формулы для дисперсий и ковариаций стандартизованных β -коэффициентов:

$$\hat{\sigma}_{\beta_i \beta_j} = \frac{c_{ij}}{n} \frac{\hat{\sigma}_e^2}{s_y^2} = c_{ij} \frac{1 - R^2}{dfe},$$

$$\hat{\sigma}_{\beta_i}^2 = \frac{c_{ii}}{n} \frac{\hat{\sigma}_e^2}{s_y^2} = c_{ii} \frac{1 - R^2}{dfe}.$$

Эти формулы дают возможность оценить значимость отдельных членов регрессионной модели по критерию Стьюдента:

$$t_{b_i} = t_{\beta_i} = \frac{\beta_i}{\hat{\sigma}_{\beta_i}} = \frac{\beta_i}{\sqrt{c_{ii} \frac{1 - R^2}{dfe}}},$$

построить доверительные интервалы на коэффициенты регрессии («инструменты экономического воздействия» – по выражению К. Доугерти):

$$\beta_i - t_{0,05}(dfe) \cdot \sqrt{c_{ii} \frac{1 - R^2}{dfe}} \leq M(\beta_i) \leq \beta_i + t_{0,05}(dfe) \cdot \sqrt{c_{ii} \frac{1 - R^2}{dfe}},$$

и вычислить дисперсии расчетных значений:

$$\hat{\sigma}_{y_p}^2 = \frac{\hat{\sigma}_e^2}{n} \left[1 + \sum_{i=1}^m \sum_{j=1}^m c_{ij} \frac{(x_i - \bar{x}_i)(x_j - \bar{x}_j)}{s_{x_i} s_{x_j}} \right].$$

Теперь для любого набора значений аргументов можно вычислить y_p вместе с границами его 95-процентного доверительного интервала $y_p \pm \Delta y_p$, где

$$\Delta y_p = t_{0,05}(dfe) \cdot \hat{\sigma}_{y_p} = t_{0,05}(dfe) \cdot s_y \sqrt{\frac{1-R^2}{dfe}} \cdot \sqrt{1 + \sum_i \sum_j c_{ij} \frac{(x_i - \bar{x}_i)(x_j - \bar{x}_j)}{s_{x_i} s_{x_j}}}.$$

Наличие на графиках 95-процентной доверительной полосы $y_p \pm \Delta y_p$ позволяет установить границы применимости регрессионной модели.

Рассмотрим частный случай однофакторной ($m = 1$) линейной модели:

$$y = b_0 + b_1 \cdot x.$$

Уравнение регрессии в стандартизованных переменных имеет вид:

$$Y = \beta_1 \cdot X + \varepsilon,$$

где обозначено $Y = \frac{y - \bar{y}}{s_y}$, $X = \frac{x - \bar{x}}{s_x}$, $\varepsilon = \frac{e}{s_y}$.

Система нормальных уравнений для этого частного случая сводится к одному равенству $\beta_1 = r_{xy}$. Коэффициент детерминации $R^2 = \beta_1 \cdot r_{xy} = (r_{xy})^2$ здесь равен квадрату коэффициента парной корреляции.

Формулы обратного перехода к исходным переменным:

$$b_1 = \beta_1 \cdot \frac{s_y}{s_x} = r_{xy} \cdot \frac{s_y}{s_x}; \quad b_0 = \bar{y} - b_1 \bar{x}.$$

Выражение для дисперсии остатка модели принимает вид:

$$\hat{\sigma}_e^2 = MSe = \frac{n}{dfe} s_e^2 = \frac{n}{n-2} s_y^2 (1 - r_{xy}^2),$$

где $dfe = (n - 2)$ – число степеней свободы остатка модели при $m = 1$.

Корреляционная матрица состоит из единственного элемента $r_{xx} = 1$, обратная матрица также содержит один элемент $c_{11} = 1$, откуда получаем формулу для расчета дисперсии β -коэффициента в виде:

$$\hat{\sigma}_{\beta_1}^2 = \frac{c_{11}}{n} \frac{\hat{\sigma}_e^2}{s_y^2} = \frac{1 - r_{xy}^2}{n - 2}.$$

Значимость коэффициента регрессии оцениваем по критерию Стьюдента:

$$t_{b_1} = t_{\beta_1} = \frac{\beta_1}{\hat{\sigma}_{\beta_1}} = \frac{r_{xy}}{\sqrt{\frac{1 - r_{xy}^2}{n - 2}}},$$

а значимость модели в целом по критерию Фишера:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-1-m}{m} = \frac{r_{xy}^2}{1-r_{xy}^2} \cdot \frac{n-2}{1}.$$

Нетрудно убедиться, что для одномерного случая эти два критерия совпадают, так как получилось, что $F = t_{b_1}^2$ (значимость коэффициента регрессии автоматически означает значимость модели).

Все вышеприведенные формулы мы уже выводили ранее. Новыми являются интервальная оценка коэффициента регрессии

$$\left(r_{xy} - t_{0,05}(n-2) \cdot \sqrt{\frac{1-r_{xy}^2}{n-2}} \right) \cdot \frac{s_y}{s_x} \leq M(b_1) \leq \left(r_{xy} + t_{0,05}(n-2) \cdot \sqrt{\frac{1-r_{xy}^2}{n-2}} \right) \cdot \frac{s_y}{s_x}$$

и формула для расчета дисперсий расчетных значений

$$\hat{\sigma}_{y_p}^2 = \frac{\hat{\sigma}_e^2}{n} \left[1 + c_{11} \frac{(x - \bar{x})^2}{s_x^2} \right] = s_y^2 \cdot \frac{1-r_{xy}^2}{n-2} \cdot \left[1 + \frac{(x - \bar{x})^2}{s_x^2} \right].$$

Оказывается, что наиболее надежные результаты расчета (с наименьшей случайной ошибкой) будут вблизи центра рассеивания наблюдаемых эмпирических точек (когда $x \approx \bar{x}$). По мере удаления от центра увеличивается случайная ошибка расчетных значений, это ставит пределы применимости регрессионной модели.

На основе центральной предельной теоремы можно утверждать, что при достаточном объеме выборки любые суммарные характеристики, в частности \bar{y} , b_1 , $y_p(x)$, будут распределены асимптотически нормально, для этих характеристик известны несмещенные оценки дисперсий, поэтому для них возможно построить 95-процентные доверительные интервалы. Так, для расчетных значений однофакторной линейной модели $y_p(x) = b_0 + b_1 \cdot x$ доверительная ошибка Δy_p вычисляется по формуле:

$$\Delta y_p(x) = t_{0,05}(n-2) \cdot \hat{\sigma}_{y_p(x)} = t_{0,05} \cdot s_y \cdot \sqrt{\frac{1-r_{xy}^2}{n-2}} \cdot \sqrt{1 + \frac{(x-\bar{x})^2}{s_x^2}},$$

где $t_{0,05}$ для $n \geq 30$ равняется 2.

Доверительный интервал $y_p(x) \pm \Delta y_p(x)$ с гарантией 95 % покрывает неизвестное нам математическое ожидание $M(y|x)$. Границы этих интервалов для каждого расчетного значения образуют доверительную полосу вокруг линии регрессии (полосу неопределенности). Любые кривые, графики которых целиком размещаются в полосе неопределенности, представляют собой множество

равноправных конкурирующих моделей – опытных данных не достаточно, чтобы сделать обоснованный выбор между ними.

Рассмотрим упрощенный графический способ построения границ доверительной полосы для одномерной регрессии. Выражение для $\Delta y_p(x)$ с некоторой заменой обозначений является уравнением сопряженной гиперболы

$Y = b \cdot \sqrt{1 + \frac{X^2}{a^2}}$, или $-\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 1$, где $X = x - \bar{x}$, $Y = \Delta y_p$ – новые переменные;

$a = s_x$, $b = t_{0,05} \cdot s_y \cdot \sqrt{\frac{1-r_{xy}^2}{n-2}}$ –

полуоси гиперболы. График сопряженной гиперболы изображен на рис. 16.1.

Отмечаем следующие особенности этого графика: ширина гиперболической полосы на интервале $[-a, a]$ приблизительно одинакова и равняется $\pm b$; далее границы полосы заметно расширяются, приближаясь к линейным асимптотам $Y = \pm \frac{b}{a} X$ – продолжениям диагоналей прямоугольника со сторонами $(\pm a, \pm b)$.

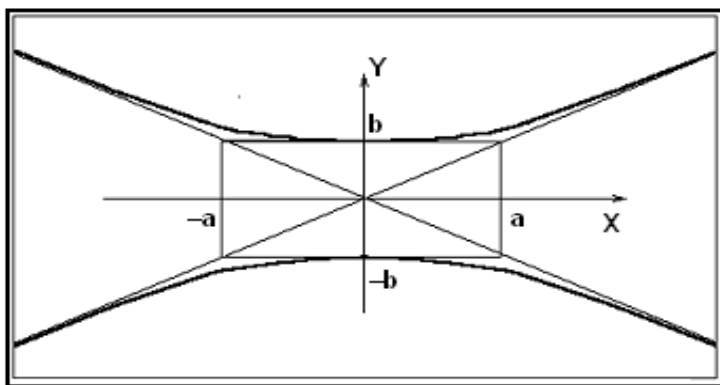


Рис. 16.1. График сопряженной гиперболы

В реальных переменных (x, y) самое узкое место полосы сдвинуто вправо на \bar{x} (с учетом знака) и полоса вытянута вдоль линии регрессии. На интервале $(\bar{x} - s_x, \bar{x} + s_x)$ величина доверительной ошибки $\pm \Delta y_p$ практически постоянна

и равняется $\pm t_{0,05} \cdot s_y \cdot \sqrt{\frac{1-r_{xy}^2}{n-2}} = \pm t_{0,05} \cdot \hat{\sigma}_{\bar{y}}$, где $\hat{\sigma}_{\bar{y}}$ – ошибка среднего.

Наносим эти границы на график $y_p(x) = b_0 + b_1 \cdot x$.

Строим параллелограмм со сторонами $(\bar{x} \pm s_x, b_0 + b_1 x \pm t_{0,05} \cdot \hat{\sigma}_{\bar{y}})$. В этом параллелограмме проводим диагонали и продолжаем их за его границы. Продолжения диагоналей и есть границы 95-процентной доверительной полосы для $|x - \bar{x}| > s_x$.

Саму сглаживающую гиперболу можно не наносить (если график строится вручную).

Пример. Пусть $n = 60$; $\bar{x} = 7,517$, $\bar{y} = 27,700$, $s_x = 1,544$; $s_y = 4,348$; $r_{xy} = 0,669$; $y_p = b_0 + b_1 x = 1,538 + 1,884x$.

Вычисляем:

$$t_{0,05} \cdot \hat{\sigma}_{\bar{y}} \approx 2 \cdot 4,348 \cdot \sqrt{\frac{1 - 0,669^2}{60 - 2}} = 0,8485.$$

На рис. 16.2 сплошной линией изображен график линии регрессии, звездочкой – центр (\bar{x}, \bar{y}) ; от центра вверх и вниз отложено 0,8485 и на интервале $\bar{x} \pm s_x$ построен параллелограмм; две стороны параллелограмма и продолжения его диагоналей представляют границы доверительной полосы на линию регрессии.

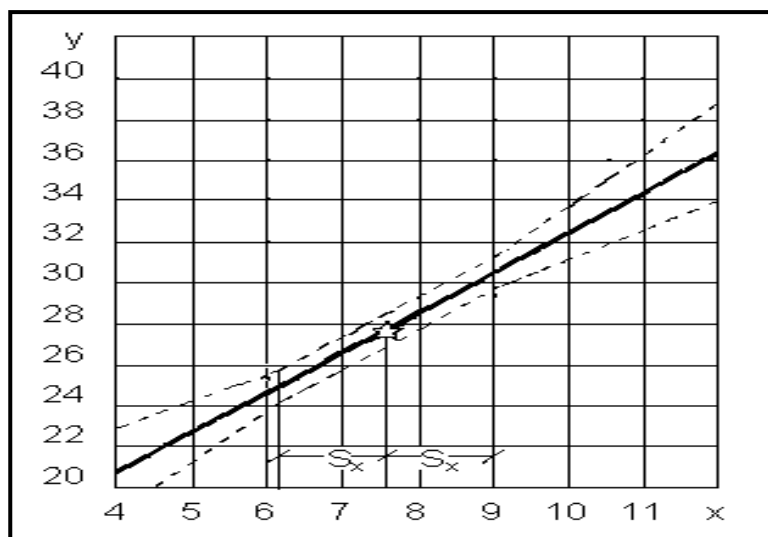


Рис. 16.2. Доверительная полоса на линию регрессии

Кроме доверительной полосы на расчетные значения, можно еще построить доверительную полосу на разброс данных вокруг линии регрессии (на прогнозные значения результативной переменной).

Тут необходимо учесть, что дисперсия прогнозных значений складывается из случайной дисперсии данных и дисперсии расчетных значений $s_{qq} = s_{pp} + MSE$, где обозначено $s_{pp} = \hat{\sigma}_{y_p}^2(x)$.

Способы составления многофакторных моделей

Многофакторная модель может быть значимой в целом (по критерию Фишера), но в то же время состоять из незначимых членов (по критерию Стьюдента). Значимая в целом модель хорошо воспроизводит опытные данные, и уравнение регрессии можно использовать как интерполяционную формулу. Но необходимо получить значимые интервальные оценки коэффициентов регрессии (инструментов экономического воздействия), тогда регрессионная модель приобретает содержательную экономическую интерпретацию.

В методе наименьших квадратов не предусмотрены какие-либо требования к оценкам коэффициентов регрессии, мы добиваемся улучшения только ее качества интерполирования.

Исследования показали, что наиболее существенной ошибкой спецификации многофакторной модели (вида уравнения регрессии) является пропуск важного члена – в этом случае все оценки регрессии будут смещенными, иногда настолько, что нельзя поручиться даже за их знаки. Будем считать, что специалист знает свое дело и ясно представляет себе, какие факторы являются существенными в той или иной проблеме.

Другие последствия будут, если в модель включить несколько незначимых членов – в этом случае снижается значимость всех остальных членов, коэффициенты регрессии оцениваются с большими случайными ошибками (большой дисперсией).

Согласно «принципу экономии» Л. Клейна, незначимые члены не просто бесполезны, они снижают качество модели и поэтому должны быть выбракованы. Однако выбраковки надо производить последовательно, начиная с самых незначимых по критерию Стюдента членов. Сначала каждая очередная выбраковка приводит к улучшению качества модели (повышается значимость оставшихся членов), но начиная с некоторого этапа процесс выбраковок надо прекратить, иначе будет выбракован пока еще незначимый, но важный член модели. Предложено несколько критериев качества модели, из которых наиболее простым является величина несмещенной остаточной дисперсии – для наилучшей модели эта величина принимает наименьшее значение:

$$\hat{\sigma}_e^2 = MSE = \frac{[e^2]}{dfe} \rightarrow \min .$$

Способ последовательных исключений («сверху – вниз») с оценкой значимости по критерию Стюдента оставшихся в модели членов (кандидатов на очередную выбраковку) легко реализуется на компьютере.

При расчетах вручную предпочитают способ последовательных подключений («снизу – вверх») с оценкой значимости еще не подключенных членов (кандидатов на подключение) с помощью так называемых коэффициентов частной корреляции.

Эти два способа приводят к разным моделям с разным числом членов и разным их составом. Теоретически наилучшим способом составления многофакторной модели был бы способ «всех регрессий», когда рассматриваются все возможные комбинации объясняющих переменных. Но таких вариантов $m!$, начиная с отсутствия в модели объясняющих переменных, далее m однофакторных моделей, $\frac{m(m-1)}{2}$ двухфакторных и так далее до включения в

модель всех m аргументов. Число всех регрессий быстро возрастает с увеличением числа членов модели.

Поэтому предложена методика подключения-исключения, которая реализована в специальных статистических пакетах программ на компьютере. Согласно этой методике, на очередном этапе делается попытка подключить в модель самую перспективную переменную, которая объясняет максимум остаточной дисперсии. Эта переменная выбирается по максимуму коэффициента частной корреляции.

Если качество модели при очередном подключении не улучшается (не уменьшается несмещенная оценка остаточной дисперсии), то последнее подключение отменяется и процесс составления модели завершается. При подключении в модель очередного члена изменяются вклады ранее подключенных членов и некоторые из них перестают быть значимыми. Надо попробовать их выбраковать, начиная с самого незначимого по критерию Стьюдента, пока с очередной выбраковкой не ухудшится качество модели. В наилучшую модель ничего больше не подключается и ничто не выбраковывается. В научной литературе этот метод иногда называется методом пошаговой регрессии.

Коэффициенты частной корреляции

Неконтролируемая изменчивость переменных, которые не учтены в модели, может полностью исказить изучаемые корреляционные связи.

На рис. 16.3 изображены две типичные ситуации, когда проявляется этот неприятный эффект.

Рис. 16.3а демонстрирует эффект появления ложной корреляции между показателями x , y . При любом фиксированном значении третьего показателя z облако рассеивания эмпирических точек вытянуто вдоль координатной оси x , следовательно, никакой корреляционной связи между x и y нет. Но с изменением неконтролируемого показателя z облако рассеивания данных сдвигается вдоль некоторой наклонной линии (в координатах x , y), в результате чего по всей совокупности данных проявляется корреляционная зависимость, поскольку большим значениям одной из переменных (x) в среднем соответствуют большие значения другой переменной (y).

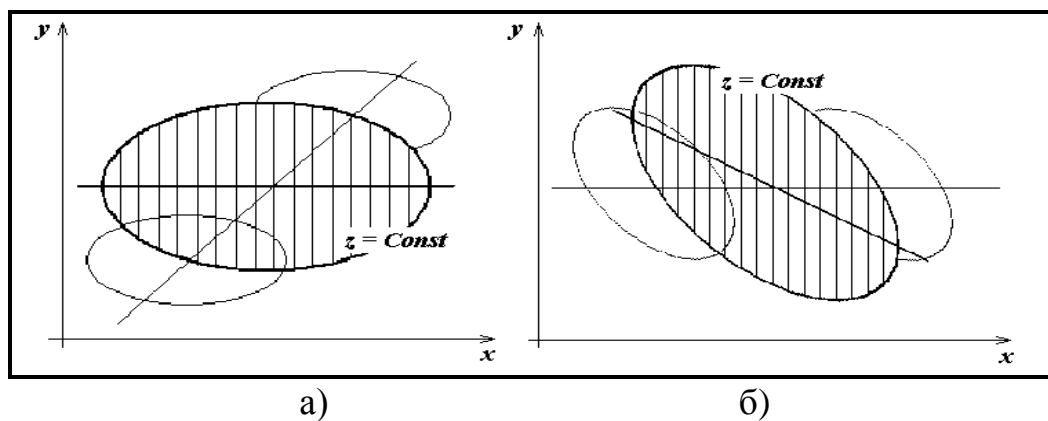


Рис. 16.3. Влияние неконтролируемой изменчивости показателя z на корреляционную зависимость между x и y

На рис. 16.3б показано, как неконтролируемая изменчивость показателя z может скрыть существующую корреляционную зависимость между x и y . Общее (суммарное) облако рассеивания точек тут оказалось вытянутым вдоль оси x , что означает отсутствие корреляционной связи по всей совокупности данных.

Чаще всего подобные искажающие эффекты появляются, когда у исследователя недостаточно наблюдений, и поэтому он дополняет изучаемую выборку данными из других родственных совокупностей (данные за разные годы, продукция разных предприятий и т. п.); иными словами, когда выборка данных неоднородная.

Если для каждого наблюдения известны значения $z = x_k$, то при изучении корреляции между $x = x_i$ и $y = x_j$ появляется возможность предварительно скорректировать все значения переменных x_i и x_j на средний уровень показателя x_k , для чего запишем уравнения регрессии между $x_i - x_k$ и $x_j - x_k$:

$$\frac{(x_i - \bar{x}_i)}{s_{x_i}} = r_{x_i x_k} \frac{(x_k - \bar{x}_k)}{s_{x_k}} + \frac{x_i \cdot x_k}{s_{x_i}} ;$$

$$\frac{(x_j - \bar{x}_j)}{s_{x_j}} = r_{x_j x_k} \frac{(x_k - \bar{x}_k)}{s_{x_k}} + \frac{x_j \cdot x_k}{s_{x_j}} .$$

Обратите внимание на новые обозначения остатков моделей – они тут обозначены через $x_i \cdot x_k$ и $x_j \cdot x_k$, чтобы показать, что они не зависят от переменной x_k , изменчивость x_k учтена в модели, переменная x_k зафиксирована на среднем уровне.

Определяем коэффициент частной корреляции как коэффициент корреляции между остатками моделей, скорректированных на средний уровень x_k :

$$r_{x_i x_j \cdot x_k} = r_{x_i \cdot x_k, x_j \cdot x_k} .$$

После некоторых преобразований получаем формулу:

$$r_{x_i x_j \cdot x_k} = \frac{r_{x_i x_j} - r_{x_i x_k} r_{x_j x_k}}{\sqrt{1 - r_{x_i x_k}^2} \sqrt{1 - r_{x_j x_k}^2}}.$$

По аналогии можно записать:

$$r_{x_i x_j \cdot x_k x_m} = \frac{r_{x_i x_j \cdot x_k} - r_{x_i x_m \cdot x_k} r_{x_j x_m \cdot x_k}}{\sqrt{1 - r_{x_i x_m \cdot x_k}^2} \sqrt{1 - r_{x_j x_m \cdot x_k}^2}}.$$

Здесь при вычислении коэффициента частной корреляции между x_i и x_j фиксируются сразу два показателя x_k и x_m .

Анализируя числитель и оба подкоренных выражения в формуле коэффициента частной корреляции, замечаем, что они получаются как промежуточные результаты при решении системы нормальных уравнений методом Гаусса – Жордана.

Предположим, собираемся составить методом последовательного подключения трехфакторную модель: $Y = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3$. На каждом шагу надо подключать в модель наиболее значимый член и проверять целесообразность дальнейших подключений. На первом шагу значимость пока еще не подключенных членов оценивается обычными коэффициентами корреляции, поэтому в модель первой подключается переменная с максимальным (по модулю) значением коэффициента r_{yx_i} .

Пусть для примера это будет переменная X_3 . Тогда после первого этапа подключения надо будет вычислить коэффициенты частной корреляции $r_{yx_1 \cdot x_3}$, $r_{yx_2 \cdot x_3}$ и найти среди них наибольший (по модулю) – это определит выбор следующей переменной, которая будет подключаться на следующем этапе. На рис. 16.4 изображен первый этап преобразований Гаусса в табличной форме.

Система нормальных уравнений в стандартизованной форме записана в строках 2 – 4 таблицы на рис. 16.4. Очень полезно добавить первой строкой коэффициенты корреляции с результативной переменной и производить преобразования Гаусса также и с этой строкой. На первом этапе подключается X_3 (разрешающий элемент выделен цветом и рамочкой). В строках 5 – 6 приведена преобразованная система уравнений после исключения X_3 (в столбце «Примечания» указаны выполненные действия).

№	Y	X ₁	X ₂	X ₃	Примечания
1	1	r_{yx_1}	r_{yx_2}	r_{yx_3}	
2	r_{yx_1}	1	$r_{x_1x_2}$	$r_{x_1x_3}$	
3	r_{yx_2}	$r_{x_1x_2}$	1	$r_{x_2x_3}$	
4	r_{yx_3}	$r_{x_1x_3}$	$r_{x_2x_3}$	1	
5	$1 - r_{yx_3}^2$	$r_{yx_1} - r_{yx_3} r_{x_1x_3}$	$r_{yx_2} - r_{yx_3} r_{x_2x_3}$	0	$[1] - [4] \cdot r_{yx_3}$
6	$r_{yx_1} - r_{yx_3} r_{x_1x_3}$	$1 - r_{x_1x_3}^2$	$r_{x_1x_2} - r_{x_1x_3} r_{x_2x_3}$	0	$[2] - [4] \cdot r_{x_1x_3}$
7	$r_{yx_2} - r_{yx_3} r_{x_2x_3}$	$r_{x_1x_2} - r_{x_1x_3} r_{x_2x_3}$	$1 - r_{x_2x_3}^2$	0	$[3] - [4] \cdot r_{x_2x_3}$
8		$r_{yx_1 \cdot x_3}$	$r_{yx_2 \cdot x_3}$		

Рис. 16.4. Преобразование системы нормальных уравнений методом Гаусса

В строке 8 подсчитаны коэффициенты частной корреляции:

$$r_{yx_1 \cdot x_3} = \frac{r_{yx_1} - r_{yx_3} r_{x_1x_3}}{\sqrt{1 - r_{yx_3}^2} \sqrt{1 - r_{x_1x_3}^2}} \quad \text{и} \quad r_{yx_2 \cdot x_3} = \frac{r_{yx_2} - r_{yx_3} r_{x_2x_3}}{\sqrt{1 - r_{yx_3}^2} \sqrt{1 - r_{x_2x_3}^2}}.$$

Отмечаем, что числители для этих формул уже вычислены в строке 5 (преобразованная строка 1), а подкоренные выражения расположены в диагональных клетках таблицы.

Следующая переменная выбирается по максимуму абсолютной величины частных коэффициентов корреляции, и выполняется очередной этап преобразований Гаусса.

В диагональной клетке преобразованной строки 1 автоматически получается число, пропорциональное величине остаточной дисперсии, что дает возможность на каждом шагу проверять целесообразность дальнейшего подключения неизвестных.

Вывод формул для дисперсий коэффициентов регрессии и расчетных значений

Приведем систему исходных предпосылок регрессионного анализа (гипотезы Гаусса – Маркова) в порядке убывания их важности:

1. $y_i = \eta(x_i) + \varepsilon_i$; все случайные ошибки относятся только к y ; $\eta(x_i)$ – известная функция.

2. $M(\varepsilon_i) = 0$; систематических ошибок нет; $\hat{y}_i = M(y | x_i) = \eta(x_i)$.

3. $M(\varepsilon_i^2) = \sigma_\varepsilon^2$; наблюдения равноточные (гомоскедастичность – одинаковый разброс).

4. $M(\varepsilon_i \varepsilon_j) = 0$; наблюдения некоррелированные (независимые).

5. $\varepsilon_i \sim N(0; \sigma_\varepsilon)$; ошибки распределены нормально (самая несущественная предпосылка).

По исходным данным (x_i, y_i) вычислены некоторые числовые характеристики:

$$\bar{x}, s_x^2, \bar{y}, b_0, b_1, y_p = b_0 + b_1 x.$$

Первые две из этих характеристик считаются измеренными точно, поскольку все случайные ошибки относятся только к переменной y .

Остальные же характеристики содержат случайную ошибку, поэтому они сами являются случайными величинами с известным законом распределения (такие характеристики называются *статистиками*).

Рассмотрим первую статистику из вышеприведенного списка:

$$\bar{y} = \frac{1}{n} \sum_i y_i = \frac{1}{n} \sum_i (\eta_i + \varepsilon_i) = \bar{\eta} + \frac{1}{n} \sum_i \varepsilon_i.$$

Поскольку все $M(\varepsilon_i) = 0$ (вторая гипотеза), то $M(\bar{y}) = \bar{\eta}$.

Вычисляем дисперсию $\sigma_{\bar{y}}^2 = M(\bar{y} - \bar{\eta})^2$:

$$\sigma_{\bar{y}}^2 = M\left(\frac{1}{n} \sum_i \varepsilon_i\right)^2 = \frac{1}{n^2} \sum_i \sum_j M(\varepsilon_i \varepsilon_j) = \frac{1}{n^2} \sum_i M(\varepsilon_i^2) = \frac{n \sigma_\varepsilon^2}{n^2} = \frac{\sigma_\varepsilon^2}{n}.$$

Здесь использовано:

$$\left(\sum_i \varepsilon_i\right)^2 = \sum_i \sum_j \varepsilon_i \varepsilon_j;$$

свойства математического ожидания;

$$M(\varepsilon_i \varepsilon_j) = 0 \text{ для } i \neq j \text{ (четвертая гипотеза);}$$

$$M(\varepsilon_i^2) = \sigma_\varepsilon^2 \text{ (третья гипотеза).}$$

Получили известный факт – случайная дисперсия среднего арифметического в n раз меньше случайной дисперсии отдельных наблюдений.

Теперь рассмотрим коэффициент регрессии b_1 :

$$\begin{aligned} b_1 &= \frac{s_{xy}}{s_x^2} = \frac{1}{ns_x^2} \sum_i y_i (x_i - \bar{x}) = \frac{1}{ns_x^2} \sum_i \eta_i (x_i - \bar{x}) + \frac{1}{ns_x^2} \sum_i \varepsilon_i (x_i - \bar{x}) = \\ &= \beta_1 + \frac{1}{ns_x^2} \sum_i \varepsilon_i (x_i - \bar{x}). \end{aligned}$$

Здесь $\beta_1 = M(b_1)$ получается по той же формуле, что и b_1 с заменой y_i на η_i .

Вычисляем дисперсию $\sigma_{b_1}^2 = M(b_1 - \beta_1)^2$:

$$\begin{aligned} \sigma_{b_1}^2 &= M(b_1 - \beta_1)^2 = \\ &= M\left(\frac{1}{ns_x^2} \sum_i \varepsilon_i (x_i - \bar{x})\right)^2 = \frac{1}{n^2 s_x^4} \sum_i \sum_j (x_i - \bar{x})(x_j - \bar{x}) \cdot M(\varepsilon_i \varepsilon_j) = \\ &= \frac{1}{n^2 s_x^4} \sum_i (x_i - \bar{x})^2 M(\varepsilon_i^2) = \frac{\sigma_\varepsilon^2}{ns_x^4} \cdot \frac{1}{n} \sum_i (x_i - \bar{x})^2 = \frac{\sigma_\varepsilon^2}{ns_x^2} = \frac{\sigma_{\bar{y}}^2}{s_x^2}. \end{aligned}$$

Покажем, что ковариация $\sigma_{\bar{y}b_1} = M(\bar{y} - \bar{\eta})(b_1 - \beta_1)$ равна нулю:

$$\begin{aligned} \sigma_{\bar{y}b_1} &= M(\bar{y} - \bar{\eta})(b_1 - \beta_1) = \\ &= M\left(\frac{1}{n} \sum_i \varepsilon_i\right) \left(\frac{1}{ns_x^2} \sum_j (x_j - \bar{x}) \varepsilon_j\right) = \frac{1}{ns_x^2} \sum_i \sum_j (x_j - \bar{x}) M(\varepsilon_i \varepsilon_j) = \\ &= \frac{1}{ns_x^2} \sum_i (x_i - \bar{x}) M(\varepsilon_i^2) = \frac{\sigma_\varepsilon^2}{ns_x^2} \sum_i (x_i - \bar{x}) = 0, \\ &\text{так как } \sum_i (x_i - \bar{x}) = 0. \end{aligned}$$

Расчетное значение y_p для каждого x оказывается комбинацией двух случайных величин \bar{y} и b_1 :

$$y_p(x) = \bar{y} + b_1(x - \bar{x}).$$

Вычисляем его дисперсию по известному свойству дисперсии суммы:

$$\begin{aligned}\sigma_{y_p(x)}^2 &= \sigma_{\bar{y}}^2 + 2(x - \bar{x})\sigma_{\bar{y}b_1} + (x - \bar{x})^2 \sigma_{b_1}^2 = \sigma_{\bar{y}}^2 + \frac{\sigma_{\bar{y}}^2}{s_x^2} (x - \bar{x})^2 \\ &= \sigma_{\bar{y}}^2 \left\{ 1 + \frac{(x - \bar{x})^2}{s_x^2} \right\} = \frac{\sigma_{\varepsilon}^2}{n} \left\{ 1 + \frac{(x - \bar{x})^2}{s_x^2} \right\}.\end{aligned}$$

Осталось заменить дисперсию случайной ошибки на несмещенную дисперсию остатка модели:

$$\hat{\sigma}_{y_p(x)}^2 = \frac{\hat{\sigma}_e^2}{n} \left\{ 1 + \frac{(x - \bar{x})^2}{s_x^2} \right\}.$$

Вопросы для самопроверки

1. Что такое стандартизованные переменные? Каковы их свойства?
2. Что такое β -коэффициенты? Как они связаны с коэффициентами регрессии? Чему равно β_0 ?
3. Какой вид имеет система нормальных уравнений в стандартизованных переменных?
4. Как в стандартизованных переменных записывается выражение для коэффициента детерминации?
5. Как оценивается значимость модели в целом и значимость ее отдельных членов?
6. Как записать интервальные оценки для коэффициентов регрессии?
7. Как строится доверительная полоса на расчетные значения? Что она показывает?
8. Как составляется многофакторная регрессионная модель?
9. Что такое коэффициенты частной корреляции?
10. Каков общепринятый критерий качества регрессионной модели?

17. Случайные функции

Случайной функцией называют функцию неслучайного аргумента t , которая при каждом фиксированном значении аргумента является случайной величиной. Например, если U – случайная величина, то $X(t) = U \cdot \sin(\pi t)$ – случайная функция. Действительно, при каждом значении t эта функция является случайной величиной: при $t_1 = 1/3$ получаем случайную величину $X_1 = 0,5 \cdot U$; при $t_1 = 1/2$ – случайную величину $X_2 = U$; при $t_3 = 1$ – случайную величину $X_3 = -U$ и т. д.

Сечением случайной функции называют *случайную величину*, соответствующую фиксированному значению аргумента t . Выше для случайной функции $X(t) = U \cdot \sin(\pi t)$ получили сечения X_1, X_2, X_3 , соответствующие $t = 1/3, 1/2, 1$.

Реализацией (траекторией) случайной функции называют неслучайную функцию аргумента t , равной которой может оказаться случайная функция в результате испытаний.

Например, если для случайной функции $X(t) = U \cdot \sin(\pi t)$ при первом испытании случайная величина U приняла значение $u_1 = 3$, при втором – $u_1 = -2$, то реализациями $X(t)$ являются, соответственно, $x_1(t) = 3 \cdot \sin(\pi t)$ и $x_2(t) = -2 \cdot \sin(\pi t)$.

Случайную функцию можно рассматривать или как совокупность случайных величин $\{X(t)\}$, зависящих от параметра t , или как совокупность ее возможных реализаций. На рис. 17.1 изображен пример графика случайной функции $X(t)$, более толстой линией выделен график одной ее реализации $x_8(t)$, маркерами выделены два сечения X_4, X_{10} при $t = 4$ и $t = 10$.

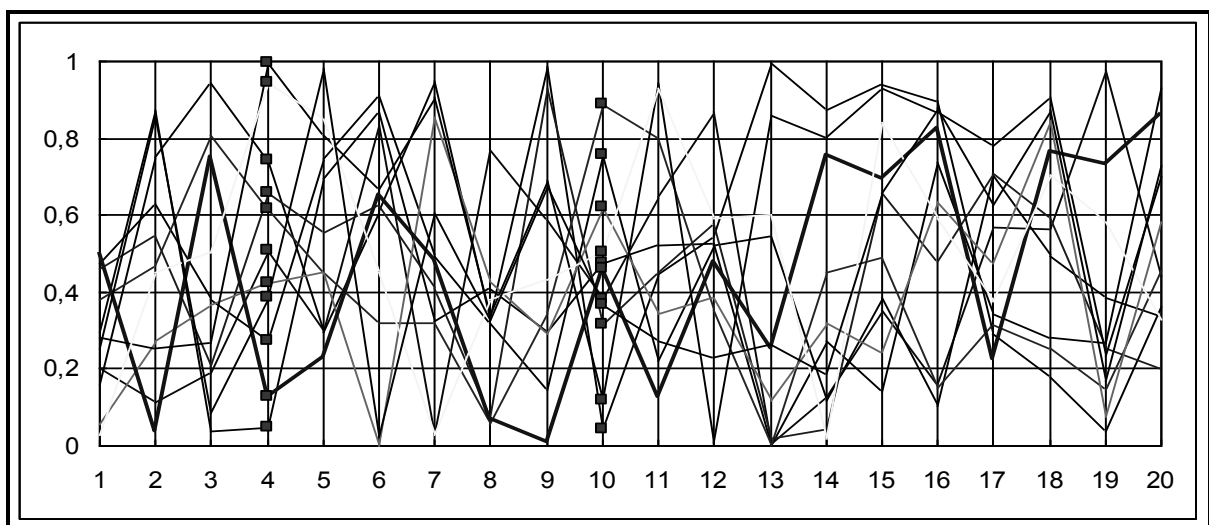


Рис. 17.1. График случайной функции по параметру t

Случайным (стохастическим) процессом называют случайную функцию аргумента t , который истолковывается как время. Если аргумент такой функции изменяется дискретно, реализацию случайного процесса называют временным рядом.

Аргументом случайной функции может быть не только время. Например, диаметр ткацкой нити под воздействием случайных факторов непрерывно изменяется по длине нити.

Характеристики случайных функций

Как уже было сказано выше, при фиксированных значениях параметра t_1, t_2, \dots, t_m случайная функция представляет собой систему случайных величин $X(t_1), X(t_2), \dots, X(t_m)$, и поэтому ее распределение описывается многомерным законом распределения. Однако многомерные распределения практически не изучены; более того, даже если бы они были известны, пользоваться такими громоздкими выражениями было бы крайне неудобно.

В *корреляционной теории случайных функций* (имеется еще *спектральная теория случайных функций*) ограничиваются изучением моментов первого и второго порядков: математических ожиданий, дисперсий, ковариаций. Этого иногда достаточно для решения многих практических задач.

В отличие от числовых характеристик случайных величин, представляющих собой определенные *числа*, характеристики случайных функций в общем случае представляют собой не числа, а *функции*.

Математическим ожиданием случайной функции $X(t)$ называют неслучайную функцию $m_x(t)$, которая при каждом фиксированном значении аргумента t равна математическому ожиданию соответствующего сечения:

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

Геометрически математическое ожидание случайной функции можно трактовать как «среднюю кривую», около которой расположены реализации данной случайной функции; при фиксированном значении аргумента математическое ожидание есть «среднее значение» сечения («средняя ордината»), вокруг которого расположены его возможные значения (ординаты).

На рис. 17.2 изображены 11 реализаций случайной функции. Для каждого сечения $t = 1, 2, \dots, 20$ вычислены средние (оценки математических ожиданий); график средних на рис. 17.2 выделен более толстой линией. Любые закономерные изменения случайной функции по времени (аргументу t) называются трендом. Здесь явно имеется линейный тренд, график которого на рис. 17.2 изобра-

жен пунктиром, а уравнение тренда, полученное методом наименьших квадратов, приведено в заголовке рисунка.

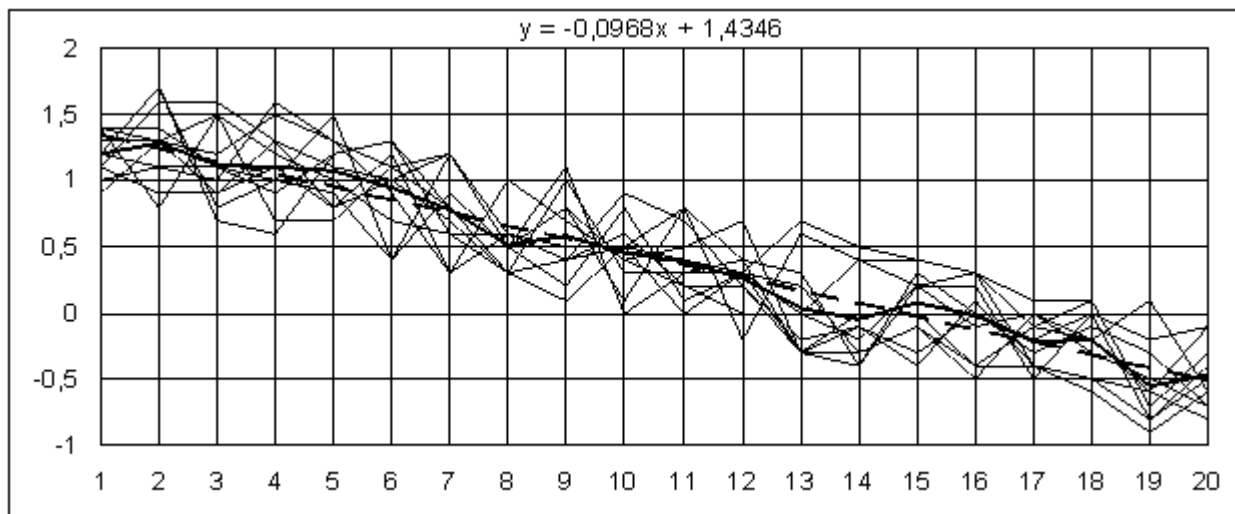


Рис. 17.2. Линейный тренд

Используя свойства математического ожидания случайной величины, несложно получить аналогичные свойства математического ожидания случайной функции:

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

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

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

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

3. Математическое ожидание суммы случайных функций равно сумме их математических ожиданий:

$$M[X(t) + Y(t)] = m_x(t) + m_y(t).$$

Следствие 1. Для того чтобы найти математическое ожидание суммы случайной и неслучайной функций, достаточно к математическому ожиданию случайной функции добавить эту неслучайную функцию:

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

Следствие 2. Математическое ожидание суммы случайной функции $X(t)$ и случайной величины Y равно сумме их математических ожиданий:

$$M[X(t) + Y] = m_x(t) + m_y.$$

Дисперсией случайной функции $X(t)$ называют неслучайную функцию $D_x(t)$, которая при каждом фиксированном значении аргумента t равна дисперсии соответствующего сечения:

$$D_x(t) = D[X(t)] = M[X(t) - m_x(t)]^2.$$

Средним квадратичным отклонением случайной функции называют корень квадратный из ее дисперсии:

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

Дисперсия и среднее квадратичное отклонение характеризуют степень рассеяния возможных реализаций вокруг математического ожидания случайной функции («средней кривой»).

При фиксированном значении аргумента дисперсия и среднее квадратичное отклонение характеризуют степень рассеяния возможных значений (ординат) сечения вокруг его математического ожидания (центра группировки).

Используя свойства дисперсии случайной величины, несложно получить аналогичные свойства дисперсии случайной функции:

1. Дисперсия неслучайной функции $\varphi(t)$ равна нулю:

$$D[\varphi(t)] = 0.$$

2. Дисперсия произведения случайной функции $X(t)$ на неслучайный множитель $\varphi(t)$ равна произведению квадрата неслучайного множителя на дисперсию случайной величины:

$$D[\varphi(t) \cdot X(t)] = \varphi^2(t) \cdot D_x(t).$$

3. Дисперсия суммы случайной функции $X(t)$ и неслучайной функции $\varphi(t)$ равна дисперсии случайной функции:

$$D[X(t) + \varphi(t)] = D_x(t).$$

Корреляционной функцией случайной функции $X(t)$ называют неслучайную функцию $K_x(t_1, t_2)$ двух независимых аргументов t_1, t_2 , значение которой при каждой паре фиксированных аргументов равно ковариации (смешанному центральному моменту) сечений, соответствующих этим значениям аргументов:

$$K_x(t_1, t_2) = M\{[X(t_1) - m_x(t_1)] \cdot [X(t_2) - m_x(t_2)]\}.$$

При равных между собой значениях аргументов $t_1 = t_2 = t$ корреляционная функция случайной функции равна дисперсии этой функции:

$$K_x(t, t) = M[X(t) - m_x(t)]^2 = D_x(t).$$

Используя свойства математического ожидания случайной величины, несложно получить свойства корреляционной функции случайной функции:

1. При перестановке аргументов корреляционная функция не изменяется (свойство симметрии):

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

2. Добавление к случайной функции $X(t)$ неслучайного слагаемого $\varphi(t)$ не изменяет ее корреляционной функции:

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

где $Y(t) = X(t) + \varphi(t)$.

3. При умножении случайной функции $X(t)$ на неслучайный сомножитель $\varphi(t)$ ее корреляционная функция умножается на произведение $\varphi(t_1)\varphi(t_2)$:

$$K_y(t_1, t_2) = K_x(t_1, t_2) \cdot \varphi(t_1)\varphi(t_2),$$

где $Y(t) = X(t) \cdot \varphi(t)$.

Нормированной корреляционной функцией случайной функции $X(t)$ называют неслучайную функцию $\rho_x(t_1, t_2)$ двух независимых аргументов t_1, t_2 , значение которой при каждой паре фиксированных аргументов равно коэффициенту корреляции сечений, соответствующих этим фиксированным значениям аргументов:

$$\rho_x(t_1, t_2) = \frac{K_x(t_1, t_2)}{\sigma_x(t_1) \cdot \sigma_x(t_2)} = \frac{K_x(t_1, t_2)}{\sqrt{K_x(t_1, t_1)} \cdot \sqrt{K_x(t_2, t_2)}}.$$

Абсолютная величина нормированной корреляционной функции не превышает единицы:

$$|\rho_x(t_1, t_2)| \leq 1.$$

Взаимной корреляционной функцией двух случайных функций $X(t)$, $Y(t)$ называют неслучайную функцию $R_{xy}(t_1, t_2)$ двух независимых аргументов t_1, t_2 , значение которой при каждой паре фиксированных аргументов равно ковариации (смешанному центральному моменту) сечений обеих функций, соответствующих этим фиксированным значениям аргументов:

$$R_{xy}(t_1, t_2) = M\{[X(t_1) - m_x(t_1)] \cdot [Y(t_2) - m_y(t_2)]\}.$$

Две случайные функции называют *некоррелированными*, если их взаимная корреляционная функция тождественно равна нулю.

Несложно доказать свойства взаимной корреляционной функции:

1. При одновременной перестановке индексов и аргументов взаимная корреляционная функция не изменяется:

$$R_{xy}(t_1, t_2) = R_{yx}(t_2, t_1).$$

2. Прибавление к случайным функциям $X(t)$, $Y(t)$ неслучайных слагаемых, соответственно $\varphi(t)$ и $\psi(t)$, не изменяет их взаимной корреляционной функции:

$$R_{uv}(t_1, t_2) = R_{xy}(t_1, t_2),$$

где $U(t) = X(t) + \varphi(t)$, $V(t) = Y(t) + \psi(t)$.

3. При умножении случайных функций $X(t)$, $Y(t)$ на неслучайные множители, соответственно $\varphi(t)$ и $\psi(t)$, взаимная корреляционная функция умножается на произведение $\varphi(t) \cdot \psi(t)$:

$$R_{uv}(t_1, t_2) = R_{xy}(t_1, t_2) \cdot \varphi(t) \psi(t),$$

где $U(t) = X(t) \cdot \varphi(t)$, $V(t) = Y(t) \cdot \psi(t)$.

4. Несложно получить следующее обобщение правила вычисления дисперсии суммы случайных величин для вычисления корреляционной функции суммы двух случайных функций.

Корреляционная функция суммы двух случайных функций равна сумме корреляционных функций слагаемых и взаимной корреляционной функцией, которая прибавляется дважды (с разным порядком следования аргументов):

$$K_z(t_1, t_2) = K_x(t_1, t_2) + K_y(t_1, t_2) + R_{xy}(t_1, t_2) + R_{xy}(t_2, t_1),$$

где $Z(t) = X(t) + Y(t)$.

Следствие 1. Корреляционная функция суммы двух некоррелируемых случайных функций равна сумме корреляционных функций слагаемых:

$$K_z(t_1, t_2) = K_x(t_1, t_2) + K_y(t_1, t_2),$$

где $R_{xy}(t_1, t_2) \equiv 0$.

Следствие 2. Корреляционная функция суммы случайной функции $X(t)$ и не коррелируемой с ней случайной величины Y равна сумме корреляционной функции случайной функции и дисперсии случайной величины:

$$K_z(t_1, t_2) = K_x(t_1, t_2) + D_y,$$

где $Z(t) = X(t) + Y$.

Стационарные случайные функции

На практике очень часто встречаются случайные процессы, протекающие во времени приблизительно однородно и имеющие вид непрерывных случайных колебаний вокруг некоторого среднего значения, причем ни средняя амплитуда, ни характер этих колебаний не обнаруживают существенных изменений с течением времени. Такие случайные процессы называются *стационарными*. На рис. 17.1 изображен график такого явно стационарного процесса.

Случайная функция $X(t)$ называется стационарной, если все ее вероятностные характеристики не зависят от t . Поскольку изменение стационарной случайной функции должно протекать однородно во времени, естественно потребовать, чтобы для стационарной случайной функции математическое ожидание было постоянным: $m_x(t) = m_x = \text{Const}$.

Однако это требование не является существенным, так как от случайной функции $X(t)$ всегда можно перейти к центрированной функции $Y(t) = X(t) - m_x(t)$, для которой математическое ожидание тождественно равно нулю: $m_y(t) \equiv 0 = \text{Const}$.

Таким образом, если случайный процесс нестационарен только за счет переменного математического ожидания, это не мешает изучать процесс как стационарный.

На рис. 17.2 изображен график стационарного процесса с добавленной неслучайной функцией $\varphi(t) = 1 - 0,1 \cdot t$ (линейный тренд).

Второе условие, которому, очевидно, должна удовлетворять стационарная случайная функция, – это условие постоянства дисперсии:

$$D_x(t) = D_x = \text{Const}.$$

Установим, какому условию должна удовлетворять корреляционная функция стационарной случайной функции.

Рассмотрим ковариацию $K_x(t, t + \tau)$ двух сечений случайной функции $X(t)$, разделенных интервалом времени τ .

Очевидно, если случайный процесс действительно стационарен, то эта ковариация не должна зависеть от того, где именно на оси времени взяли участок τ , а зависеть только от длины этого участка:

$$K_x(t, t + \tau) = k_x(\tau).$$

Следовательно, корреляционная функция стационарного случайного процесса есть функция не двух, а всего одного аргумента. Это обстоятельство в ряде случаев существенно облегчает операции над стационарными случайными функциями.

Заметим, что при $\tau = 0$ $K_x(t, t) = D_x(t) = k_x(0) = \text{Const}$, то есть требование $K_x(t, t + \tau) = k_x(\tau)$ как частный случай включает $D_x(t) = D_x = \text{Const}$.

Именно это свойство принято за определение стационарности процесса: случайная функция является стационарной, если ее корреляционная функция $K_x(t_1, t_2) = k_x(\tau)$ зависит только от разности аргументов $\tau = t_2 - t_1$.

Из общего условия симметрии корреляционной функции $K_x(t_1, t_2) = K_x(t_2, t_1)$ следует четность корреляционной функции стационарного процесса $k_x(\tau) = k_x(-\tau)$, поэтому рассматривают только неотрицательные значения аргумента $\tau \geq 0$.

На практике вместо корреляционной функции $k_x(\tau)$ часто пользуются нормированной корреляционной функцией:

$$\rho_x(\tau) = \frac{k_x(\tau)}{k_x(0)} = \frac{k_x(\tau)}{D_x},$$

где $D_x = k_x(0)$ – постоянная дисперсия стационарного процесса.

Функция $\rho_x(\tau)$ есть коэффициент корреляции между сечениями случайной функции, разделенными интервалом τ по времени. Очевидно, что $\rho_x(0) = 1$.

Приведем пример расчета нормированной корреляционной функции для процесса, изображенного на рис. 17.3.

t	1	3	5	7	9	11	13	15	17	19
$x_1(t)$	1,1	0,9	0,8	0,3	0,7	0,5	-0,3	-0,1	-0,1	0,1
$x_2(t)$	1,4	1,1	1,2	0,8	0,2	0,1	0,2	-0,1	0	-0,5
$x_3(t)$	1,1	1,6	0,8	1,2	1,1	0,3	0,7	0,4	-0,4	-0,9
$x_4(t)$	1,2	0,7	1,2	0,6	0,4	0,4	-0,3	-0,4	-0,4	-0,6
$x_5(t)$	1,3	1,5	1	0,6	0,4	0,7	-0,3	0	-0,4	-0,8
$x_6(t)$	1,4	0,9	1,1	0,7	1	0,3	-0,3	0,2	0	-0,6
$x_7(t)$	1	1,1	0,9	1,2	0,4	0,2	-0,2	-0,3	-0,2	-0,8
$x_8(t)$	1,4	1,5	0,7	0,8	0,1	0	0	0,2	-0,5	-0,2
$x_9(t)$	1,2	1	1,3	1,2	0,8	0,8	0,6	0,4	0,1	-0,7
$x_{10}(t)$	0,9	1,2	1,3	0,3	0,5	0,8	0,3	0,3	-0,3	-0,3
$x_{11}(t)$	1,2	0,8	1,5	0,9	0,8	0,2	0	0,2	-0,1	-0,7
Ср	1,200	1,118	1,073	0,782	0,582	0,391	0,036	0,073	-0,209	-0,545
D_x	0,025	0,083	0,058	0,098	0,094	0,070	0,124	0,067	0,037	0,082
S_x	0,160	0,289	0,242	0,313	0,307	0,264	0,352	0,260	0,193	0,287

Рис. 17.3. Значения 11-ти реализаций случайной функции

На рис. 17.3 приведены значения 11-ти реализаций случайной функции $X(t)$ через $\Delta t = 2$ при $t = 1, 3, \dots, 19$.

В последних строках таблицы для каждого t подсчитаны средние, оценки дисперсий и средних квадратичных отклонений.

Средние систематически уменьшаются, но дисперсии и средние квадратичные отклонения можно считать постоянными:

$$D_x \approx 0,074; S_x \approx 0,272.$$

Вычислим коэффициенты корреляции между столбцами таблицы на рис. 17.3.

В виду симметрии $\rho_x(t_1, t_2) = \rho_x(t_2, t_1)$ на рис. 17.4 приведена только верхняя треугольная корреляционная матрица.

$t_1 \backslash t_2$	1	3	5	7	5	6	7	8	9	11
1	1	0,039	-0,071	0,073	-0,204	-0,453	-0,226	0,000	0,118	-0,020
3		1	-0,580	0,205	-0,140	-0,034	0,387	0,407	-0,569	-0,166
5			1	-0,043	0,140	0,309	0,140	0,119	0,483	-0,254
7				1	0,243	-0,310	0,493	0,218	0,163	-0,707
9					1	0,244	0,317	0,518	0,350	-0,308
11						1	0,179	0,248	0,123	-0,005
13							1	0,706	0,112	-0,262
15								1	0,122	-0,090
17									1	0,009
19										1
Сум- мы	10	0,921	0,552	1,114	0,465	-0,093	-0,044	-0,822	-0,047	-0,020
$\rho_x(\tau)$	1	0,102	0,069	0,159	0,077	-0,019	-0,011	-0,274	-0,024	-0,020

Рис. 17.4. Верхняя треугольная корреляционная матрица между сечениями

Для всех чисел на линиях, параллельных главной диагонали, одинакова разность $(t_2 - t_1) = \tau = 0, 2, 4, \dots, 18$, поэтому для стационарной функции $X(t)$ числа на этих линиях должны быть одинаковыми.

Однако в данном примере имеются явные нарушения этого условия стационарности, что можно объяснить малым числом учтенных реализаций (всего 11) и малым числом наблюдений.

Поэтому вполне целесообразной будет приближенная замена случайной функции $X(t)$ стационарной (с добавленным линейным трендом).

В последних строках таблицы на рис. 17.4 подсчитаны суммы чисел на каждой линии, параллельной главной диагонали, и средние, которые являются оценками значений нормированной корреляционной функции стационарного процесса $\rho_x(\tau)$.

График этой функции приведен на рис. 17.5.

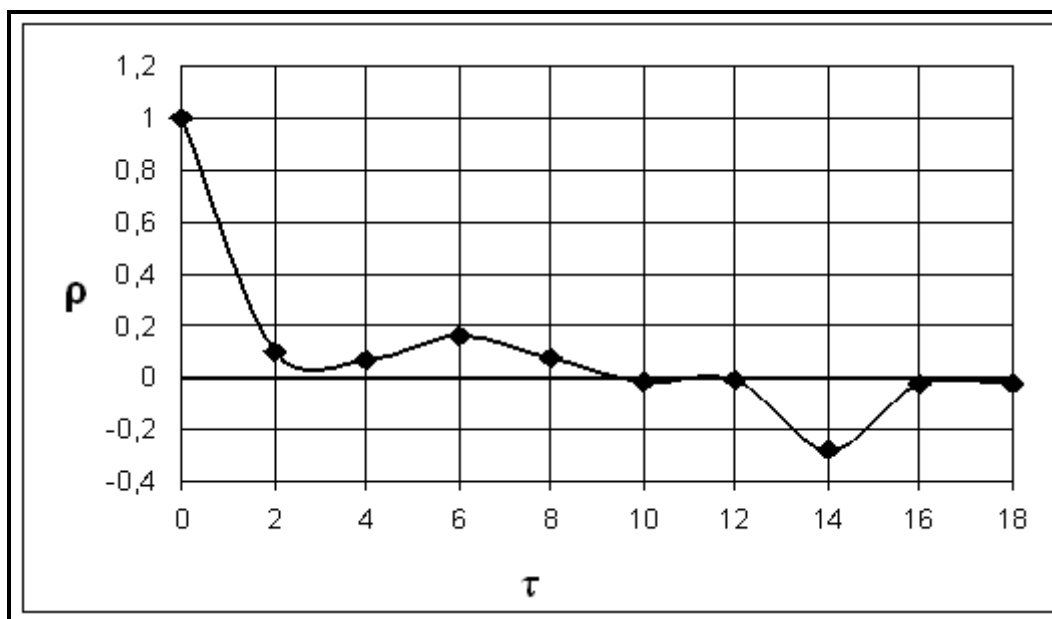


Рис. 17.5. График нормированной корреляционной функции

Последние три точки на этом графике малонадежны, так как их ординаты получены усреднением по слишком малому числу данных.

Корреляционная функция – это «лицо» изучаемой случайной функции; специалист по виду ее графика может многое сказать о характере и дальнейшем развитии процесса.

Эргодичные стационарные процессы

Описанный выше метод обработки данных стационарных процессов является довольно сложным и громоздким и к тому же состоит из двух этапов: приближенного определения характеристик случайной функции и приближенного усреднения этих характеристик. Возникает вопрос: так ли необходимо располагать несколькими реализациями процесса?

Поскольку случайный процесс является стационарным и протекает однородно во времени, естественно предположить, что одна-единственная реализация достаточной продолжительности может служить материалом для определения характеристик случайной величины.

На рис. 17.6 более толстой линией выделена одна из возможных реализаций процесса, откуда видно, что она обладает примерно теми же характеристиками: средним значением, вокруг которого происходят колебания, и средним размахом этих колебаний. Очевидно, при достаточно большом интервале времени эта одна реализация может дать нам адекватное представление о свойствах случайной функции в целом. Усредняя значения этой реализации вдоль

оси абсцисс (оси времени), должны получить приближенные значения математического ожидания и дисперсии стационарной случайной функции.

Кстати, именно для этих данных были рассчитаны оценки математического ожидания $m_x \approx 0,468$ и среднего квадратичного отклонения $\sigma_x \approx 0,272$, а по одной реализации (где данных в 11 раз меньше) были получены близкие значения $\bar{x} = 0,452$ и $s_x = 0,294$.

Эргодическое свойство состоит в том, что каждая отдельная реализация случайной величины является как бы «полномочным представителем» всей совокупности возможных реализаций; одна реализация достаточной продолжительности может заменить при обработке множество реализаций той же продолжительности.

Конечно, далеко не каждая стационарная случайная функция обладает эргодичностью. Например, на рис. 17.6 приведены графики реализаций неэргодичной случайной функции (с добавленным линейным трендом).

Сравнивая графики на рис. 17.2 и рис. 17.6, замечаем, для обоих этих процессов одинаковы математические ожидания и одинаковые дисперсии, однако для процесса на рис. 17.6 ни одна реализация не может быть принята для представления процесса в целом. Здесь, кроме добавленного линейного тренда (что не существенно для анализа), добавлено также случайное слагаемое, которое приводит к случайным сдвигам отдельных реализаций.

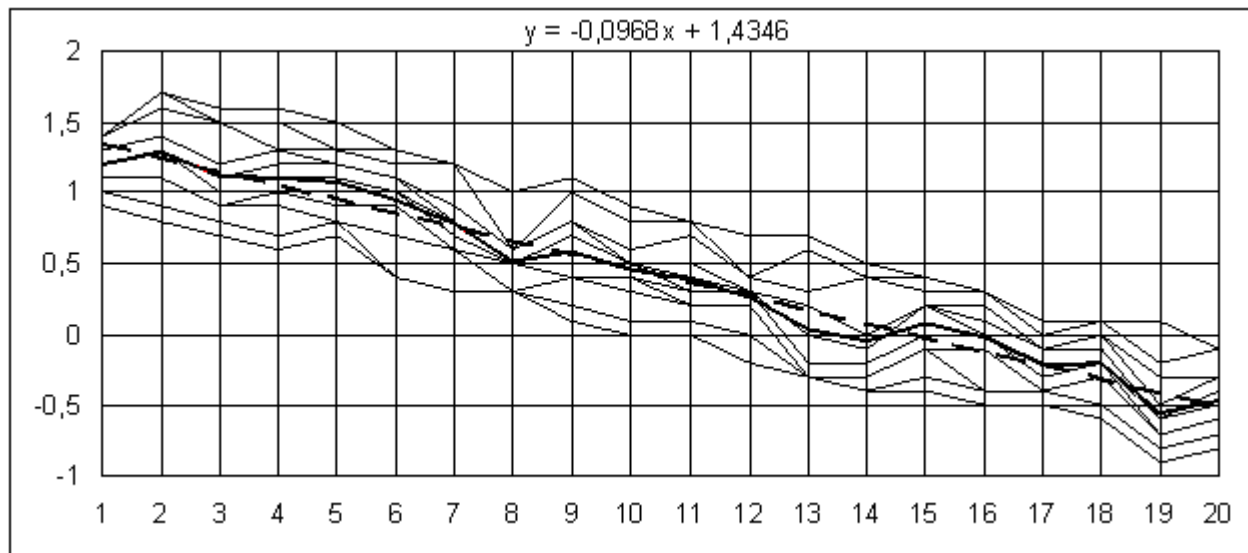


Рис. 17.6. Пример неэргодичного процесса

Рассмотрим случайную функцию:

$$Z(t) = X(t) + Y + \varphi(t),$$

где $X(t)$ – эргодичная стационарная случайная функция с характеристиками m_x , $k_x(\tau)$; Y – случайная величина с характеристиками m_y , D_y ; $\varphi(t)$ – неслучайная функция. Согласно общим правилам сложения случайных функций, определяем характеристики случайной функции $Z(t)$:

$$\begin{aligned} m_z &= m_x + m_y + \varphi(t); \\ k_z(\tau) &= k_x(\tau) + D_y. \end{aligned}$$

Так как корреляционная функция $k_z(\tau)$ зависит только от τ , случайная функция $Z(t)$, по определению, считается стационарной (с добавленным неслучайным трендом). Однако она явно не является эргодичной, поскольку каждая ее реализация отличается от других в зависимости от того, какое значение приняла случайная величина Y .

Об эргодичности или неэргодичности случайного процесса часто судят по виду его корреляционной функции. Считается, что для стационарного процесса корреляционная функция $k_x(\tau)$ должна стремиться к нулю при $\tau \rightarrow \infty$ (корреляционная связь между сечениями случайной функции неограниченно убывает по мере увеличения расстояния между ними). Но для случайной функции $Z(t)$ корреляционная функция $k_z(\tau)$ не стремится к нулю при $\tau \rightarrow \infty$, а приближается к постоянному значению D_y . Поэтому нарушение условия $\lim_{\tau \rightarrow \infty} k_z(\tau) = 0$ считается признаком отсутствия эргодичности процесса.

Если стационарная случайная функция $X(t)$ обладает эргодическим свойством, то для нее среднее и дисперсия *по времени* приближенно равны среднему и дисперсии *по сечениям*, а оценку корреляционной функции $k_x(\tau)$ можно получить как ковариацию между участками одной реализации, разделенными интервалом времени τ :

$$\begin{aligned} k_x(\tau) &\approx \frac{1}{N - \tau} \sum_{t=1}^{N-\tau} x_t x_{t+\tau} - \bar{x}^2, \\ \rho_x(\tau) &= k_x(\tau) / k_x(0) = R(x_t, x_{t+\tau}). \end{aligned}$$

Здесь x_t – последовательные значения временного ряда (одной реализации случайного процесса).

Оценку нормированной корреляционной функции, полученную по данным одной реализации, называют *автокорреляционной функцией*.

Для реализации $x_8(t)$, график которой выделен на рис. 17.1 более толстой линией, были вычислены значения автокорреляционной функции. На рис. 17.7 график полученной автокорреляционной функции изображен тонкой линией.

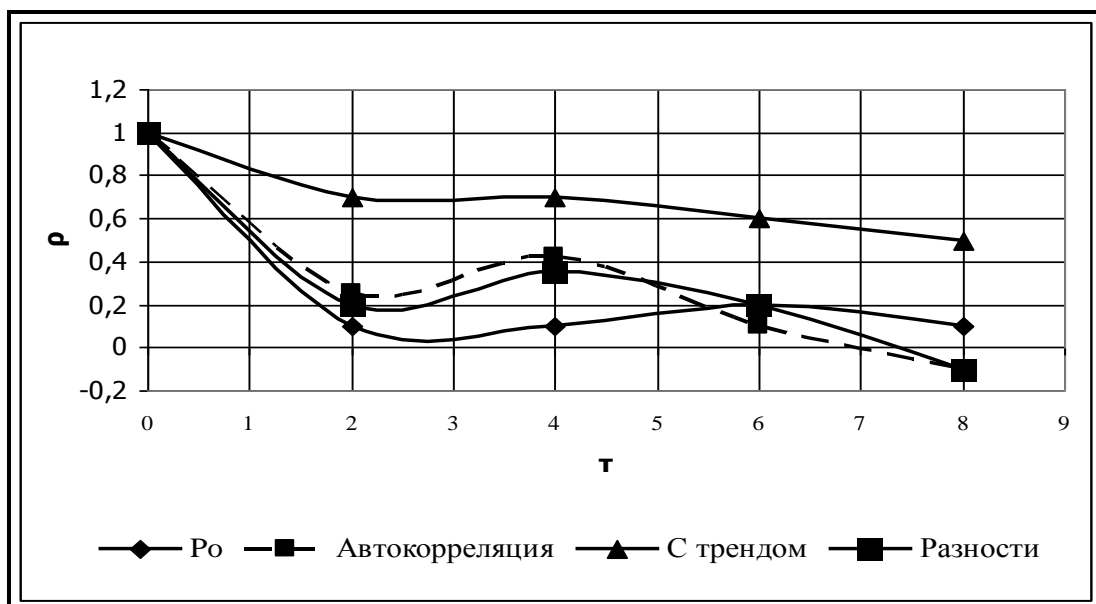


Рис. 17.7. Графики нормированной корреляционной и автокорреляционной функций

Для сравнения приведен построенный ранее (см. рис. 17.5) график нормированной корреляционной функции этого же процесса.

Соответствие не слишком хорошее, но учтем, что данных по одной реализации было слишком мало. Для равноправного сравнения длину временного ряда следовало бы увеличить в 11 раз.

Кроме этого, на рис. 17.7 пунктиром построен график автокорреляционной функции для той же реализации, но с добавленным линейным трендом.

Здесь можно эмпирически заметить интереснейший факт: при наличии трендов (любого порядка) автокорреляционная функция убывает медленно, практически линейно.

Обычно в этих случаях рекомендуется переходить к последовательным разностям $\Delta x_t = x_t - x_{t-1}$ столько раз, пока автокорреляционная функция не будет быстро стремиться к нулю.

В этом примере после перехода к первым разностям график автокорреляционной функции (точечная линия) становится очень похожим на график автокорреляционной функции без добавленного тренда.

Таким приемом определяется порядок тренда (линейный, квадратичный и т. д.). Еще раз сформулируем обнаруженное правило в виде: если автокорреляционная функция быстро спадает до нуля, то никакого тренда временной ряд не имеет.

На рис. 17.8 приведен график биржевых цен акций IBM за 120 рабочих дней, где явно виден квадратичный тренд. На этом же рисунке построена соответствующая линия квадратичной регрессии вместе с границами ее 95-процентной доверительной полосы. Согласно выделенному тренду, ожидается дальнейшее прогрессивное возрастание курса акций IBM. Данных $N = 120$ вполне достаточно для такого прогноза.

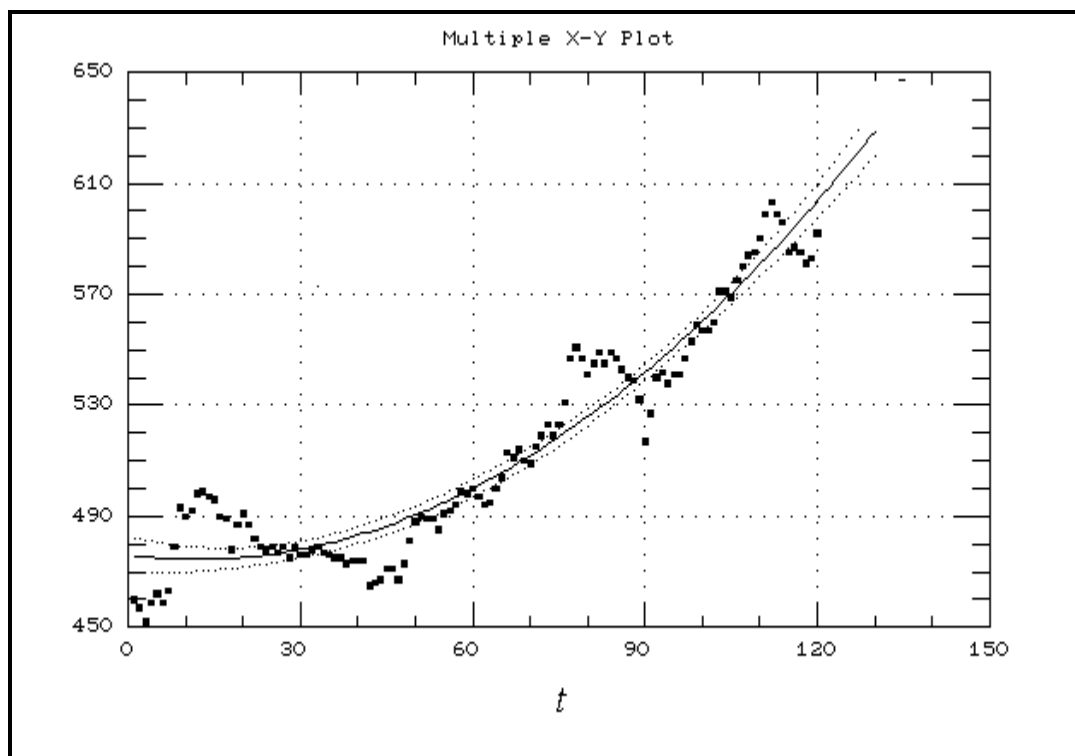
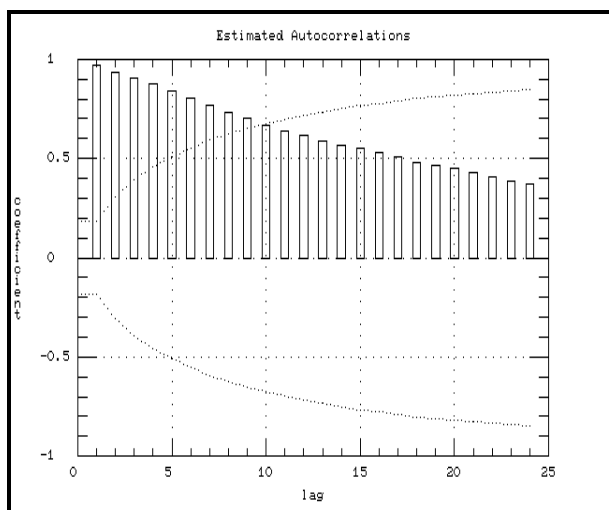


Рис. 17.8. Динамика биржевых цен акций IBM за 120 рабочих дней
с 17.05.61

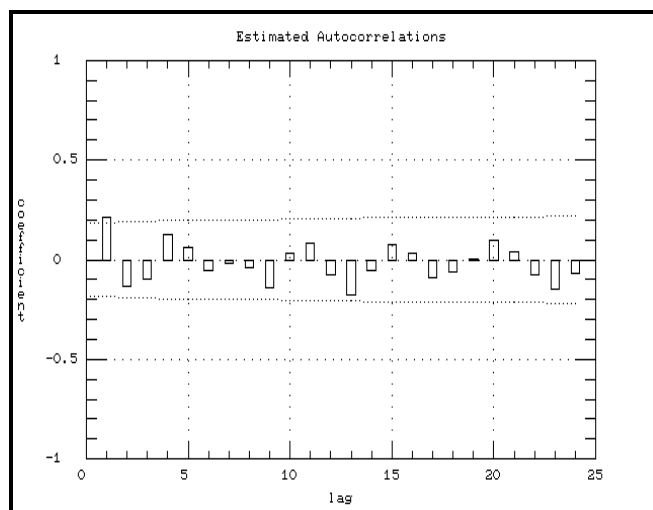
Оказывается, что в действительности здесь нет ни квадратичного, ни даже линейного тренда, поведение курса акций является случайным блужданием без фиксированного среднего уровня, поэтому наилучшим прогнозом будет последнее значение наблюдений.

Для анализа на рис. 17.9 приведены графики автокорреляционной функции курса акций и их первых разностей.

Автокорреляции курса акций (рис. 17.9а) спадают медленно, практически по линейному закону, следовательно, какой-то тренд действительно имеет место. Но поведение автокорреляций для первых разностей (рис. 17.9б) имеет совершенно другой характер: они все практически близки к нулю. Следовательно, линейного тренда (и, тем более, квадратичного тренда) нет. Явно наблюдаемый тренд имеет нулевой порядок – случайные сдвиги среднего уровня.



а)



б)

Рис. 17.9. Автокорреляционные функции курса акций и их первых разностей

Сейчас нам известны сведения о дальнейшем поведении курса акций IBM, график которых изображен на рис. 17.10.

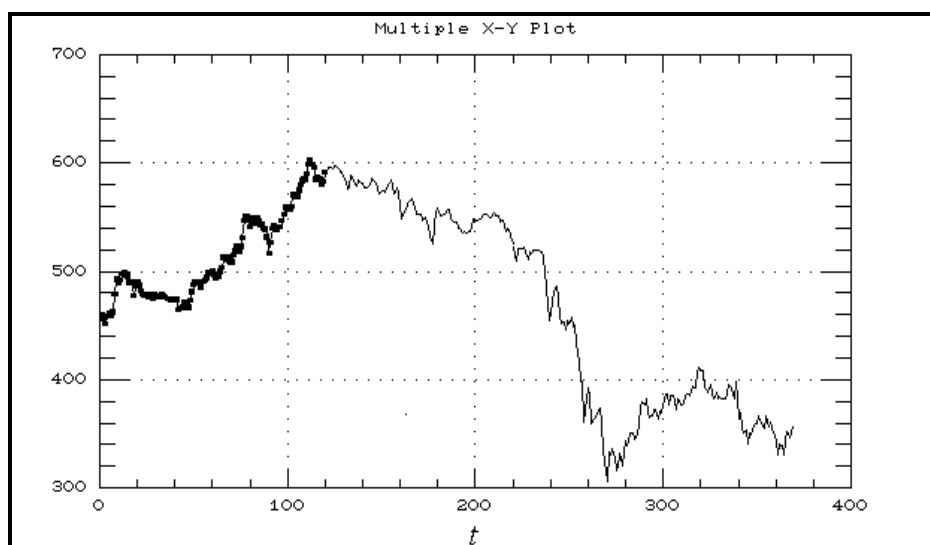


Рис. 17.10. Курс акций IBM з 17.05.61 до 02.11.62

Оказывается, обнадеживающий прогрессивный рост курса акций на протяжении достаточно длительного периода (120 рабочих дней, то есть 4 месяца) был всего лишь небольшим эпизодом действительного развития событий. Примечательно, что математический инструмент автокорреляционных функций по возрастающему участку временного ряда четко выявил действительную природу процесса и не позволил сделать неправильных выводов. К сожалению,

до сих пор бытует порочная практика визуального выделения трендов – на этом основании делать те или иные прогнозы.

Кроме упомянутого правила для выявления трендов, сформулированы правила выявления скрытых периодичностей и возможных последствий. Все это составляет предмет изучения раздела «Анализ временных рядов».

Вопросы для самопроверки

1. Что такое случайная функция, ее сечения и реализации?
2. Что такое случайный процесс и временной ряд?
3. Сформулируйте предмет изучения корреляционной теории случайных функций. Какие есть еще теории случайных функций?
4. Что такое математическое ожидание случайной функции? Каковы ее свойства?
5. Что такое дисперсия случайной функции? Каковы ее свойства?
6. Что такое корреляционная функция случайной функции? Каковы ее свойства?
7. Что такое взаимная корреляционная функция двух случайных функций? Каковы ее свойства? Какие случайные функции называются некоррелированными?
8. Сформулируйте правило вычисления корреляционной функции суммы двух случайных функций.
9. Что такое нормированная корреляционная функция случайной функции? Каковы ее свойства?
10. Приведите определение стационарности случайной функции.
11. Что такое тренд, какие бывают тренды?
12. Изложите методику оценки корреляционной функции стационарной случайной функции.
13. Что такое эргодичность стационарного случайного процесса?
14. Сформулируйте необходимое условие эргодичности.
15. Что такое автокорреляционная функция?
16. Сформулируйте правило выявления трендов по виду автокорреляционной функции. Применимо ли это правило для выявления периодичных трендов?
17. В каком разделе теории случайных процессов изучается структура временных рядов, на основе чего делаются прогнозы о дальнейшем поведении временного ряда?

Лабораторная работа 1

Изучение распределения Бернулли средствами Excel

Говорить о математической статистике и не вычислять – это все равно как говорить о музыке, но не исполнять и не слушать ее. К тому же вычисления любой сложности уже не представляют проблемы. Прошли времена громоздких ручных расчетов, теперь практически на каждом персональном компьютере установлена универсальная вычислительная система MS Excel или ее аналоги типа OpenOffice.org Calc. Знание этого вычислительного инструмента обязательно для студентов всех специальностей всех форм обучения. Существуют также мощные специализированные пакеты прикладных программ, такие, как StatGraphics, Statistica, MatLab, MathCad. Однако для целей обучения специализированные пакеты мало подходят, так как нет ничего поучительного в том, когда компьютер автоматически генерирует готовое решение, а пользователь иногда даже не знает, какой алгоритм был использован и на каких предположениях он основан.

В первых двух лабораторных работах в среде MS Excel исследуется общая проблема о повторении однородных независимых испытаний, изучаются особенности распределения Бернулли – Пуассона – Лапласа в зависимости от параметров и устанавливаются границы применимости асимптотических формул.

Задача теории вероятностей о повторении однородных независимых испытаний была решена Бернулли для небольшого числа испытаний $n < 30$ в виде

$$P_n(m) = C_n^m p^m (1-p)^{n-m} = \frac{n!}{m!(n-m)!} p^m q^{n-m},$$

где n – число испытаний;

p – вероятность успеха (появления некоторого события) в каждом испытании;

$q = (1 - p)$ – вероятность противоположного результата (непоявления успеха);

m – случайная величина – общее число успехов;

$P_n(m)$ – вероятность появления m успехов при n испытаниях.

При большом числе испытаний ($n \geq 30$) используются предельные формулы Пуассона и Лапласа. Компьютер позволяет представить особенности распределений Бернулли – Пуассона – Лапласа в зависимости от параметров распределения в наглядной графической форме.

На рабочем листе Excel предлагается построить графики распределения Бернулли для различных значений параметра p ($0 < p < 1$) и при разных значе-

ниях другого параметра n ($n = 10, 20, 30, 50$). Эти графики позволяют заметить характерные особенности распределения.

Кроме этого, полезно убедиться, что характеристики распределения Бернулли правильно отображаются известными формулами:

$$M(m) = np, D(m) = npq.$$

Полезно также убедиться в справедливости правила «3-х сигм»:

$$M(m) - 3\sigma_m < m < M(m) + 3\sigma_m,$$

где $\sigma_m = \sqrt{npq}$; значения m , которые выходят за границы указанного интервала, маловероятны.

Ниже на рис. Л1.1 приведен первый фрагмент выполнения задания на рабочем листе Excel.

	A	B	C	D	E	F	G	H	I	J
1	Распределение Бернулли									
2	$P_n(m) = n! / (m!(n-m)!) \cdot p^m \cdot q^{(n-m)}$									
3	$P_n(m) = P_n(m-1) \cdot (n-m+1) / m \cdot p/q$									
4	$P_n(0) = q^n$									
5	n = 10		n = 10		n = 10		n = 10		n = 10	
6	p = 0,1		p = 0,3		p = 0,5		p = 0,7		p = 0,9	
7	q = 0,9		q = 0,7		q = 0,5		q = 0,3		q = 0,1	
8	M = 1		M = 3		M = 5		M = 7		M = 9	
9	D = 0,9		D = 2,1		D = 2,5		D = 2,1		D = 0,9	
10	M-3σ_m = -1,8461		M-3σ_m = -1,3474		M-3σ_m = 0,2566		M-3σ_m = 2,6526		M-3σ_m = 6,1540	
11	M+3σ_m = 3,8461		M+3σ_m = 7,3474		M+3σ_m = 9,7434		M+3σ_m = 11,3474		M+3σ_m = 11,8461	
12										
13	m	p=0,1	m	p=0,3	m	p=0,5	m	p=0,7	m	p=0,9
14	0	0,34868	0	0,02825	0	0,00098	0	5,9E-06	0	1E-10
15	1	0,38742	1	0,12106	1	0,00977	1	0,00014	1	9E-09
16	2	0,19371	2	0,23347	2	0,04395	2	0,00145	2	3,6E-07
17	3	0,05740	3	0,26683	3	0,11719	3	0,00900	3	8,8E-06
18	4	0,01116	4	0,20012	4	0,20508	4	0,03676	4	0,00014
19	5	0,00149	5	0,10292	5	0,24609	5	0,10292	5	0,00149
20	6	0,00014	6	0,03676	6	0,20508	6	0,20012	6	0,01116
21	7	8,8E-06	7	0,00900	7	0,11719	7	0,26683	7	0,05740
22	8	3,7E-07	8	0,00145	8	0,04395	8	0,23347	8	0,19371
23	9	9E-09	9	0,00014	9	0,00977	9	0,12106	9	0,38742
24	10	1E-10	10	5,9E-06	10	0,00098	10	0,02825	10	0,34868

Рис. Л1.1. Изучение зависимости распределения Бернулли от параметра p

Рассмотрим внимательно первый блок (столбцы А, В таблицы Excel).

В строках 5 и 6 задаем значения параметров $n = 10, p = 0,1$.

В следующих строках вычисляем $q = 1 - p, M = \Sigma(m \cdot P_n(m)), D = \Sigma(m^2 \cdot P_n(m)) - M^2, M - 3S_m, M + 3S_m$, где $S_m = \sqrt{D}$.

Последние 4 формулы можно набрать позже, когда будет заполнен диапазон В14:В24, который содержит значения $P_n(m)$. Отметим полезный прием: в столбце А записываем текст и смещаем его *вправо*, а в столбце В вычисляем числовое значение и смещаем его *влево*. Получается понятный комментарий к выполненным операциям.

Рабочий лист Excel, кроме всего прочего, является отчетным документом, поэтому не стоит экономить на комментариях и заголовках.

Из информации в строках 8 – 11 первого блока видно, что, действительно, $M = np = 10 \cdot 0,1 = 1$; $D = npq = 10 \cdot 0,1 \cdot 0,9 = 0,9$; и что все вероятные значения m не превосходят $M + 3S_m = 3,85 \approx 4$.

Значения $P_n(m)$ удобно рассчитывать по рекуррентной формуле, которая приведена в строке 3 рабочего листа Excel:

$$P_n(m) = P_n(m-1) \cdot (n-m+1) / m \cdot p / q.$$

Начальное значение $P_n(0) = q^n$ вычислено в ячейке В14.

При наборе рекуррентной формулы в ячейке В15 следует зафиксировать (знаками \$) номера строк постоянных значений n, p, q . Далее формула копируется ниже до ячейки В24.

Заполнив первый блок, копируем его несколько раз вправо и в новых блоках заменяем значения параметра p на $p = 0,3; p = 0,5; p = 0,7; p = 0,9$. Все автоматически пересчитывается.

В блоках серым фоном выделены значения m , которые признаны значимыми по правилу «3-х сигм».

Теперь строим графики. Выделяем значения m вместе с заголовком в ячейке А13, далее при нажатой клавише **Ctrl** выделяем мышкой значения $P_n(m)$ для $p = 0,1; 0,3; 0,5; 0,7; 0,9$.

Выделять диапазоны надо вместе с заголовками в строке 13, тогда эти заголовки автоматически будут отображены в легенде (пояснениях к каждой линии на графике). Вызываем *Мастер диаграмм*, выбираем тип диаграммы – точечная, легенда – внизу, линии сетки – основные, заголовок: «Распределение Бернулли при разных p ($n=10$)».

В результате получим график (рис. Л1.2), который почти не требует дополнительного форматирования.

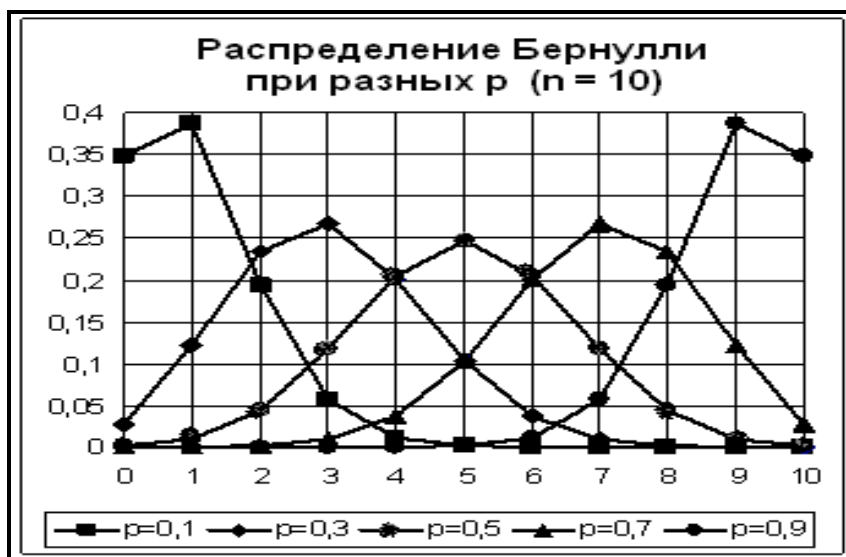


Рис. Л1.2. Зависимость распределения Бернулли от параметра p

Из этого графика видно, как изменяется асимметрия распределения при увеличении параметра p : при $p = 0,5$ – распределение симметричное, при $p < 0,5$ – график скошен влево (положительная асимметрия), а при $p > 0,5$ – скошен вправо (отрицательная асимметрия).

Как уже указывалось выше, заголовки из строки 13 автоматически переносятся в легенду диаграммы. Но тогда желательно, чтобы они автоматически корректировались при замене параметра p . Поэтому в качестве заголовка в ячейке В13 набрана формула = "p=" & ТЕКСТ(В6;"0,0"). Функция ТЕКСТ (Число; Формат) переводит в символьную форму значения p из ячейки В9; в тексте заголовка это число будет округлено до одного знака после десятичной запятой.

Остальные заголовки в строке 13 корректируются автоматически при копировании.

Значения t , удовлетворяющие условию $M - 3Sm \leq t \leq M + 3Sm$, выделены на рабочем листе (см. рис. Л1.1) в столбцах **m** с помощью условного форматирования. Для этого в первом блоке предварительно был выделен весь диапазон изменения t (ячейки А14:А24) и в меню *Формат* выбран пункт *Условное форматирование*.

На панели *Условного форматирования* (рис. Л1.3) задано Условие 1: «значение между =В\$10 и =В\$11»; далее, нажав кнопку **Формат**, на вкладке **Вид** выбран серый фон для диапазона, указанного в Условии 1. Обратим внимание на то, что в ссылках на ячейки В\$10 и В\$11 зафиксированы только номера строк, чтобы при копировании первого блока вправо условные форматы автоматически корректировались.

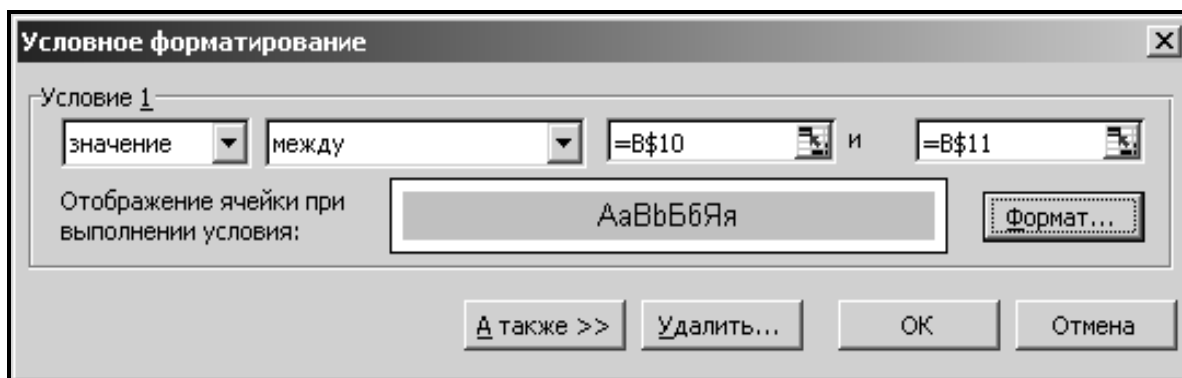


Рис. Л1.3. Панель условного форматирования

Теперь переходим к изучению зависимости распределения Бернулли от другого параметра n (рис. Л1.4).

	K	L	M	N	O	P	Q	R	S	T
5	n =10		n =20		n =30		n =40		n =50	
6	p =0,1		p =0,1		p =0,1		p =0,1		p =0,1	
7	q =0,9		q =0,9		q =0,9		q =0,9		q =0,9	
8	M =1		M =2		M =3		M =4		M =5	
9	D =0,9		D =1,8		D =2,7		D =3,6		D =4,5	
10	M-3Sm -1,8461		M-3Sm -2,0249		M-3Sm -1,9295		M-3Sm -1,6921		M-3Sm -1,3640	
11	M+3Sm 3,8461		M+3Sm 6,0249		M+3Sm 7,9295		M+3Sm 9,6921		M+3Sm 11,3640	
13	m	n=10	m	n=20	m	n=30	m	n=40	m	n=50
14	0	0,3487	0	0,1216	0	0,0424	0	0,0148	0	0,0052
15	1	0,3874	1	0,2702	1	0,1413	1	0,0657	1	0,0286
16	2	0,1937	2	0,2852	2	0,2277	2	0,1423	2	0,0779
17	3	0,0574	3	0,1901	3	0,2361	3	0,2003	3	0,1386
18	4	0,0112	4	0,0898	4	0,1771	4	0,2059	4	0,1809
19	5	0,0015	5	0,0319	5	0,1023	5	0,1647	5	0,1849
20	6	0,0001	6	0,0089	6	0,0474	6	0,1068	6	0,1541
21	7	0,0000	7	0,0020	7	0,0180	7	0,0576	7	0,1076
22	8	0,0000	8	0,0003	8	0,0058	8	0,0264	8	0,0643
23	9	0,0000	9	0,0001	9	0,0016	9	0,0104	9	0,0333
24	10	0,0000	10	0,0000	10	0,0004	10	0,0036	10	0,0152
25	11	0	11	0,0000	11	0,0001	11	0,0011	11	0,0061

Рис. Л1.4. Изучение зависимости распределения Бернулли от параметра n

Скопируем все 5 готовых блоков вправо, начиная со столбца K, и заменим в новых блоках значения параметров: $n = 10, 20, 30, 40, 50$ и $p = 0,1$ (для всех новых блоков).

Новые таблицы следует продолжить вниз с запасом до строки 64 (теперь блоки будут иметь разную длину).

Соответствующим образом надо изменить верхнюю границу суммирования при вычислении M и D (только для первого блока новой таблицы, далее откорректированные формулы копируются вправо).

Для того чтобы на графиках автоматически изменялась легенда, следует заголовки в строке 13 заменить на формулы $=\text{"n=" \& ТЕКСТ(L5;"0")}$.

Теперь все готово для построения нового графика (рис. Л1.5).



Рис. Л1.5. Зависимость распределения Бернулли от параметра n

Из рис. Л1.5 видно, как с увеличением параметра n распределение Бернулли приближается к какой-то стандартной форме – к распределению Лапласа или к так называемому нормальному закону распределения Гаусса.

Считается, что при $n \geq 30$ распределение уже практически нормально.

Лабораторная работа 2

Асимптотические формулы Пуассона и Лапласа

Изучение распределения Пуассона средствами Excel

Распределение Пуассона $P(m) = e^{-a} \frac{a^m}{m!}$ зависит от одного параметра a .

Фрагмент выполнения задания на рабочем листе приведен на рис. Л2.1, который оформлен аналогично рабочему листу с расчетами по формулам Бернулли.

	A	B	C	D	E	F	G	H	I	J
1	Распределение Пуассона									
2	$P(m)=e^{(-a)}*a^m/m!$									
3	$P(m)=P(m-1)*a/m$									
4	$P(0)=e^{(-a)}$									
5										
6	a =1		a =2		a =3		a =4		a =5	
7										
8	M =0,9810		M =1,9669		M =2,9643		M =3,9675		M =4,9727	
9	D =0,9383		D =1,8877		D =2,8757		D =3,8848		D =4,90223	
10	M-3Sm -2		M-3Sm -2,2426		M-3Sm -2,1962		M-3Sm -2		M-3Sm -1,7082	
11	M+3Sm 4		M+3Sm 6,2426		M+3Sm 8,1962		M+3Sm 10		M+3Sm 11,7082	
12										
13	m	a=1	m	a=2	m	a=3	m	a=4	m	a=5
14	0	0,367879	0	0,135335	0	0,049787	0	0,018316	0	0,006738
15	1	0,367879	1	0,270671	1	0,149361	1	0,073263	1	0,033690
16	2	0,183940	2	0,270671	2	0,224042	2	0,146525	2	0,084224
17	3	0,061313	3	0,180447	3	0,224042	3	0,195367	3	0,140374
18	4	0,015328	4	0,090224	4	0,168031	4	0,195367	4	0,175467
19	5	0,003066	5	0,036089	5	0,100819	5	0,156293	5	0,175467
20	6	0,000511	6	0,012030	6	0,050409	6	0,104196	6	0,146223
21	7	7,3E-05	7	0,003437	7	0,021604	7	0,059540	7	0,104445
22	8	9,12E-06	8	0,000859	8	0,008102	8	0,029770	8	0,065278
23	9	1,01E-06	9	0,000191	9	0,002701	9	0,013231	9	0,036266
24	10	1,01E-07	10	3,82E-05	10	0,000810	10	0,005292	10	0,018133
25	11	9,22E-09	11	6,94E-06	11	0,000221	11	0,001925	11	0,008242

Рис. Л2.1. Изучение зависимости распределения Пуассона от параметра a

Предлагается на рабочем листе Excel построить графики распределения Пуассона для разных значений параметра $a = 1, 2, 3, 4, 5$; вычислить для этих значений параметра характеристики M , D и убедиться, что математическое ожидание и дисперсия правильно воспроизводятся известными формулами $M(m) = a$, $D(m) = a$.

В распределении Пуассона m теоретически не ограничено ($m \rightarrow \infty$). Однако правило «3-х сигм» остается в силе: все вероятные значения m попадают в интервал $a - 3\sigma_m < m < a + 3\sigma_m$, где $\sigma_m = \sqrt{a}$; значения m , которые выходят за пределы указанного интервала, маловероятны.

Рабочий лист на рис. Л2.1 оформлен аналогично рабочему листу с расчетами по формулам Бернулли, только характеристики M и D на рабочем листе вычислены по диапазону $m \leq a + 3\sqrt{a}$ (правило «3-х сигм»), поэтому они слегка меньше теоретических значений $M = D = a$. Последовательные значения вероятностей удобно получать по рекуррентной формуле $P(m) = P(m-1) \cdot \frac{a}{m}$, где $P(0) = e^{-a}$ (эти формулы приведены в строке 3 рабочего листа). Значимые по правилу «3-х сигм» значения m выделены серым фоном. Подготовив первый блок (столбцы А, В), копируем его несколько раз вправо и в копиях заменяем значения параметра на $a = 2, 3, 4, 5$. Практически все пересчитывается автоматически (за исключением характеристик M и D , где необходимо вручную уточнять верхнюю границу диапазонов $m \leq a + 3\sqrt{a}$).

Все подготовлено для построения графиков распределения Пуассона для разных значений параметра a (рис. Л2.2). Поскольку в легенду переносятся заголовки столбцов из строки 13, заголовок в ячейке В13 был задан формулой `="a"&ТЕКСТ(В6;"0")`; при копировании вправо этот заголовок корректируется автоматически. На рис. Л2.2 видно, как с увеличением параметра a распределение Пуассона приближается к симметричному распределению Лапласа (нормальному закону Гаусса) и что при $a \geq 5$ распределение Пуассона уже можно считать нормальным.



Рис. Л2.2. Зависимость распределения Пуассона от параметра a

Распределение Пуассона можно считать асимптотическим приближением для распределения Бернулли, когда $n \gg 1$, $p \ll 1$, $a = np$. Здесь выражение $n \gg 1$ означает « n очень большое», а выражение $p \ll 1$ – « p очень маленькое». Считается, что предельным распределением можно пользоваться при $n \geq 30$, $p \leq 0,1$ и $a = np \leq 5$. На компьютере можно проверить и уточнить эти утверждения.

Изучение распределения Лапласа средствами Excel

Распределение Лапласа $P_n(m) = \frac{1}{\sqrt{2\pi} \cdot \sigma_m} e^{-\frac{(m-a)^2}{2\sigma_m^2}}$ зависит от двух параметров n и p ; характеристики распределения выражаются через эти параметры: $a = np$, $\sigma_m = \sqrt{npq}$. При расчетах вручную используют таблицы дифференциальной $\varphi(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}$ и интегральной $\Phi(t) = \int_0^t \varphi(s) ds = \frac{1}{\sqrt{2\pi}} \cdot \int_0^t e^{-\frac{s^2}{2}} ds$ функций Лапласа. Тогда $P_n(m) = \frac{\varphi(t_m)}{\sigma_m}$, где $t_m = \frac{m-a}{\sigma_m}$. Обычно эта формула называется локальной теоремой Лапласа. Интегральная функция Лапласа используется для вычисления вероятности попадания случайной величины $\mathcal{M} = \{m\}$ в полуоткрытый интервал: $P(m_1 \leq m \leq m_2) \approx \Phi(t_{m_2}) - \Phi(t_{m_1}) = F(m_2) - F(m_1)$. Обычно эта формула называется интегральной теоремой Лапласа. Заметим, что интегральная функция *распределения* Лапласа $F(m) = P(\mathcal{M} \leq m)$ отличается от интегральной функции Лапласа постоянным слагаемым: $F(m) = \Phi(t_m) + 0,5$. На компьютере $P_n(m)$ и $F(m)$ вычисляются функцией НОРМРАСП($m; a; \sigma_m; k$), которая при $k = 0$ возвращает значения вероятности $P_n(m)$, а при $k = 1$ возвращает значения интегральной функции распределения Лапласа $F(m)$. Считается, что асимптотическими формулами Лапласа можно пользоваться при $n \geq 30$, $a = np \geq 5$ и $nq \geq 5$. Предлагается прямыми расчетами проверить точность локальной и интегральной теорем Лапласа для $n = 10, 20, 30, 40, 50$ и $p = 0,3$.

Интегральную теорему Лапласа можно уточнить, если расширить интервал $[m_1; m_2]$ влево и вправо на 0,5: $P(m_1 \leq m \leq m_2) = F(m_2 + 0,5) - F(m_1 - 0,5)$.

На рис. Л2.3 изображен фрагмент рабочего листа Excel с выполненным заданием. Слева в столбцах А, В размещен блок расчетов по формуле Бернулли (сейчас принято $n = 10$). В соседнем столбце С приведены значения кумуляты (накопленных вероятностей) $F(m) = \Sigma(P_n(k))$, вычисленных по рекуррентной

формуле $F(m) = F(m - 1) + P_n(m)$, где $F(0)=P_n(0)=q^n$. С помощью кумуляты одним вычитанием определяются вероятности попадания случайной величины m в заданные интервалы $P(m_1 \leq m \leq m_2) = F(m_2) - F(m_1 - 1)$. Так, вероятность $P(2 \leq m \leq 5)$ можно вычислить как разность $P(2 \leq m \leq 5) = F(5) - F(1) = 0,9526 - 0,1493 = 0,8433$; это же значение можно получить непосредственно, суммируя вероятности $P_n(m)$ для $m = 2, 3, 4, 5$ (цифры на сером фоне в столбце А). Отметим существенную особенность интегральной теоремы для дискретных случайных величин $P(m_1 \leq m \leq m_2) = F(m_2) - F(m_1 - 1)$ в отличие от аналогичной интегральной теоремы для непрерывных случайных величин $P(x_1 \leq x \leq x_2) = F(x_2) - F(x_1)$.

Для сравнения с точными расчетами по формуле Бернулли в графе Laplas (столбец D) вычислены вероятности по локальной формуле Лапласа.

	A	B	C	D	E	F	G	H	I	J
1	Распределение Лапласа									
2	Pn(m)=НОРМРАСП(m;a;Sm;0)				a=np	Sm=КОРЕНЬ(npq)				
3	P(m1≤m≤m2)=НОРМРАСП(m2;a;Sm;1)–НОРМРАСП(m1;a;Sm;1)									
4	P(m1≤m≤m2)=НОРМРАСП(m2+0.5;a;Sm;1)–НОРМРАСП(m1-0,5;a;Sm;1)									
5										
6	n =10		Φ(2)-Φ(-1) 0,8186		P([m1;m2])	n=10	n=20	n=50	n=100	n=200
7	p =0,3				Bernully	0,8033	0,8758	0,8359	0,8159	0,8137
8	q =0,7		M-Sm=1,5509		Laplas	0,6712	0,8100	0,7907	0,7837	0,7907
9	a = M =3		M+2Sm 5,8983		Correct	0,8074	0,8747	0,8375	0,8179	0,8152
10	Dm =2,1		m1 =2		Error%	-16,5%	-7,51%	-5,41%	-3,92%	-2,83%
11	Sm =1,4491		m2 =5		ErrCorr	0,51%	-0,12%	0,20%	0,24%	0,18%
12	Б е р н у л л и									
13	m	n=10	F(m)	Laplas	P(a)	n=10	n=20	n=30	n=40	n=50
14	0	0,0282	0,0282	0,0323	Bernully	0,26683	0,19164	0,15729	0,13657	0,12235
15	1	0,1211	0,1493	0,1062	Laplas	0,27530	0,19466	0,15894	0,13765	0,12312
16	2	0,2335	0,3828	0,2170	Correct	0,26993	0,19275	0,15790	0,13697	0,12263
17	3	0,2668	0,6496	0,2753	Error%	3,17%	1,58%	1,05%	0,79%	0,63%
18	4	0,2001	0,8497	0,2170	ErrCorr	1,16%	0,58%	0,39%	0,29%	0,23%
19	5	0,1029	0,9527	0,1062						
20	6	0,0368	0,9894	0,0323						

Рис. Л2.3. Проверка точности асимптотических формул Лапласа

По результатам расчетов построены сравнительные графики распределения Бернулли и распределения Лапласа с теми же самыми характеристиками.

Изменяя n в ячейке В6, можно наблюдать, насколько хорошо распределение Бернулли описывается предельным распределением Лапласа.

На рис. Л2.4 приведены два сравнительных графика для $n = 10$ и $n = 20$, откуда видно, что при $n \geq 20$ распределение Бернулли для $p = 0,3$ практически совпало с предельным распределением Лапласа.

Таким образом, убедились в том, что локальная формула Лапласа достаточно точная.

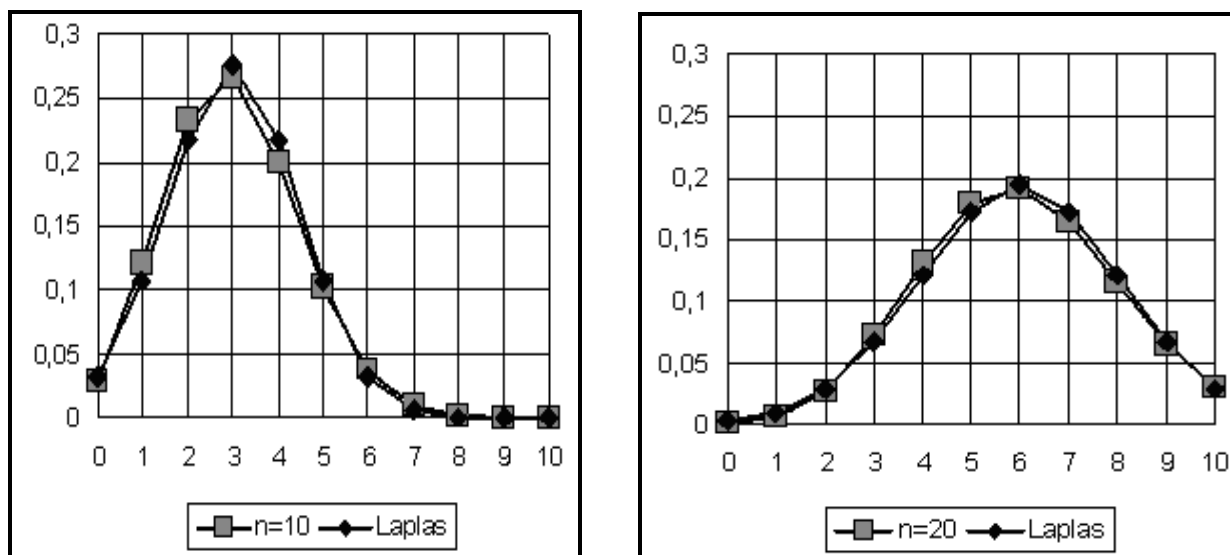


Рис. Л2.4. Сравнительные графики распределений Бернулли и Лапласа

Переходим к проверке точности расчетов по интегральной формуле Лапласа. Для примера вычислим вероятность попадания случайной величины m в интервал $[M - \sigma_m; M + 2\sigma_m]$.

Согласно стандартной записи интегральной теоремы Лапласа, эта вероятность подсчитывается как разность $\Phi(2) - \Phi(-1) = 0,8186$.

Однако интегральная теорема Лапласа выведена в предположении непрерывности распределения случайной величины, а распределение Лапласа дискретное. Поэтому значение, вычисленное по интегральной теореме Лапласа (или близкое к нему), может быть получено только для очень большого $n > 200$. Проверим данное утверждение. Выше над блоком расчетов по формуле Бернулли для $n = 10$ вычислены границы интервала: $M - S_m = 1,551$ и $M + 2S_m = 5,898$; эти границы округлены до целых $m_1 = 2$, $m_2 = 5$ как для целочисленной величины $[1,551 \leq m \leq 5,898] \sim [2 \leq m \leq 5]$ (при изменении n все эти числа будут автоматически пересчитаны).

В ячейке F7 для заданного $n = 10$ подсчитывается точное значение вероятности $P(2 \leq m \leq 5) = F(5) - F(1) = 0,8033$.

В таблице A14:D100 (таблицу можно продлить автозаполнением) поиск нужных значений интегральной функции $F(m_2)$ и $F(m_1 - 1)$ производится с помощью функции ВПР (Искомое_значение; Таблица; Номер_столбца), где Искомое_значение = m_2 или $(m_1 - 1)$, Таблица = A14:D100, Номер_столбца = 3.

В ячейке F8 вычисляется та же вероятность $P(2 \leq m \leq 5)$ по стандартной интегральной теореме Лапласа $P(m_1 \leq m \leq m_2) = \text{НОРМРАСП}(m_2; a; S_m; 1) - \text{НОРМРАСП}(m_1; a; S_m; 1)$. Для $n = 10$ получилось $P(2 \leq m \leq 5) = 0,6712$. В сравнении с точным значением (0,8033) ошибка составила -16,46 %.

В ячейке F9 для расчета $P(2 \leq m \leq 5)$ использована уточненная формула: $P(m_1 \leq m \leq m_2) = \text{НОРМРАСП}(m_2 + 0,5; a; S_m; 1) - \text{НОРМРАСП}(m_1 - 0,5; a; S_m; 1)$. Для $n = 10$ получилось $P(2 \leq m \leq 5) = 0,8074$ с ошибкой всего 0,51 %.

При изменении n в ячейке B6 все цифры в столбце F автоматически пересчитываются. Для того чтобы сохранить результаты предыдущих расчетов, они были скопированы Специальной вставкой – Только значения вправо (столбцы G, H, I, J).

Рассматривая заполненную таблицу (рис. Л2.5), убеждаемся, что стандартная интегральная теорема Лапласа может использоваться только для очень больших значений $n > 200$, даже когда аргументы m_1, m_2 правильно округлены до целых значений. Но обычно это не делается.

Например, в третьей форме интегральной теоремы Лапласа $P\left(\left|\frac{m}{n} - p\right| \leq \varepsilon\right) = 2\Phi\left(\varepsilon \sqrt{\frac{n}{pq}}\right)$ не предпринимается никаких попыток учесть целочисленность m и n , что наверняка приводит к дополнительным ошибкам. Для данного примера по стандартной асимптотической формуле Лапласа получено значение $P([m_1, m_2]) = \Phi(2) - \Phi(-1) = 0,8186$, которое еще не достигнуто даже для $n = 200$.

P([m1;m2])	n=10	n=20	n=50	n=100	n=200
Bernully	0,8033	0,8758	0,8359	0,8159	0,8137
Laplas	0,6712	0,8100	0,7907	0,7837	0,7907
Correct	0,8074	0,8747	0,8375	0,8179	0,8152
Error %	-16,46%	-7,51%	-5,41%	-3,92%	-2,83%
ErrCorr %	0,51%	-0,12%	0,20%	0,24%	0,18%

Рис. Л2.5. Сравнение точности асимптотических формул Лапласа

Теперь можно оценить эффективность скорректированной интегральной формулы Лапласа $P(m_1 \leq m \leq m_2) = F(m_2 + 0,5) - F(m_1 - 0,5)$, которая уже для

$n = 10$ привела к практически точному значению 0,8074 (по сравнению с точным значением 0,8033 ошибка составляет всего 0,51 %).

Эффективность скорректированной формулы была также проверена на вычислении вероятности моды (самого вероятного значения $m = a = np$). Эти расчеты выполнены в таблице диапазона E13:J18 (рис. Л2.6).

P(a)	n=10	n=20	n=30	n=40	n=50
Bernully	0,26683	0,19164	0,15729	0,13657	0,12235
Laplas	0,27530	0,19466	0,15894	0,13765	0,12312
Correct	0,26993	0,19275	0,15790	0,13697	0,12263
Error%	3,17%	1,58%	1,05%	0,79%	0,63%
ErrCorr%	1,16%	0,58%	0,39%	0,29%	0,23%

Рис. Л2.6. Сравнительные расчеты вероятности наивероятнейшего значения

Обычная интегральная формула Лапласа тут дает значения 0, поэтому для расчетов использовалась локальная формула Лапласа.

Погрешности локальной формулы невелики и быстро убывают с увеличением n (при $n = 20$ погрешность составила всего 1,6 %).

Скорректированная интегральная формула оказалась еще более точной; она дает практически точные значения уже при $n = 10$.

Вопросы для самопроверки

1. Сформулируйте задачу Бернулли.
2. Приведите формулы для расчета характеристик распределения Бернулли.
3. Опишите особенности распределения Бернулли при разных значениях параметров.
4. Сформулируйте закон редких событий Пуассона.
5. В каких случаях распределение Бернулли допустимо аппроксимировать предельным распределением Пуассона?
6. Сформулируйте локальную и интегральную теоремы Лапласа.
7. Сформулируйте три основных формы интегральной теоремы Лапласа.

Лабораторная работа 3

Обработка данных наблюдений

Следующие три лабораторных работы (3 – 5) посвящены изучению проблем математической статистики на примере обработки реальных данных (собранные Воловельской С. Н.).

Цель работы:

1. По данным наблюдений (по данным выборки) требуется представить вид распределения случайной величины. Для этого необходимо сгруппировать данные на ряд интервалов и построить графики гистограммы и кумуляты.

2. По виду эмпирического распределения следует выбрать наиболее подходящий теоретический закон распределения и проверить соответствие эмпирического и теоретического распределений с помощью критериев согласия. В качестве теоретических распределений обязательно следует рассмотреть нормальный и равномерный законы. Для этого предварительно требуется вычислить выборочные характеристики (среднее, дисперсию, стандартное отклонение, коэффициент вариации), рассчитать и построить графики дифференциальной и интегральной функций теоретического распределения, найти ожидаемые частоты попадания случайной величины в заданные интервалы, наконец, проверить согласие эмпирических и теоретических частот по критерию Пирсона. Кроме критерия согласия Пирсона, предлагается ознакомиться еще с каким-либо критерием, например с критерием согласия Колмогорова – Смирнова.

3. Полезно ознакомиться с графиком нормальной вероятностной кривой для графического подбора наиболее подходящего функционального преобразования случайной величины с целью приблизить ее распределение к нормальному или к другому предполагаемому теоретическому распределению.

4. По выборочным данным предлагается сделать заключения о совокупности в целом – найти интервальные оценки на математическое ожидание (центр совокупности, генеральное среднее) и на генеральную дисперсию. Требуется определить объем выборки, необходимый для оценки математического ожидания с заданной точностью и надежностью.

5. Предлагается найти медиану и квартили распределения и построить блочную диаграмму Тьюки («ящик и усы»). С помощью блочной диаграммы Тьюки уверенно выделяются все выбросы – слишком далеко отклоняющиеся наблюдения. Полезно сравнить эффективность этого приема с правилом «3-х сигм». Найденные выбросы следует удалить из данных и тем самым скорректировать результаты анализа.

6. Помимо всего прочего, одной из целей данных лабораторных работ является освоение эффективных приемов работы в среде Excel – универсального вычислительного инструмента, который необходимо знать специалисту любого профиля.

Описательная статистика

Представление данных

Исходные данные допустимо представлять компактно в виде таблицы из нескольких столбцов (не обязательно одинаковой длины). Это удобно для последующего отчета. Например, ниже на рис. ЛЗ.1 приведены данные одного из вариантов индивидуального задания.

Исходные данные X						
0,66	1,5	1,52	0,98	0,39	0,88	2,45
1,23	1,9	0,93	1,21	0,99	0,64	1,13
2,05	2,5	1,11	1,74	0,98	1,82	0,68
1,57	1	0,9	1,54	1,4	1,42	0,78
0,78	1,1	1,29	1,23	1,33	2,55	1,35
1,23	1,2	1,43	1,13	0,7	1,69	2
1,06	0,94	1,11	0,68	0,81	1,39	1,23
0,93	0,87	1,04	0,78	1,54	0,7	1,02
1,54	2,2	0,78	1,27	1,58	1,7	1,7
0,74	0,67	1,21	1,12	1,09	1,5	1,38
0,68	0,88	1,03	1,39	0,91	0,99	

Рис. ЛЗ.1. Пример представления данных наблюдений

Функциями СЧЁТ(X), МАКС(X), МИН(X), где X – диапазон данных, находим количество наблюдений (n), максимальное (X_{max}) и минимальное (X_{min}) значения. Желательно все числовые расчеты сопровождать пояснениями. Предлагается следующая форма записи.

n =	76
X_{max} =	2,55
X_{min} =	0,39

Рис. ЛЗ.2. Образец

Пояснения записывать в ячейках слева и эти *тексты сдвигать вправо*; формулы набирать в ячейках справа и полученные числа *сдвигать влево* (рис. ЛЗ.2). Тогда сразу получается понятный отчет – немаловажная часть любого исследования.

Очень удобно в таблице данных показывать максимальное и минимальное значения. Для этого выделяем таблицу данных и в меню *Формат* находим

пункт *Условное форматирование*. На панели условного форматирования (рис. Л3.3) заполняем поля ввода. **Условие 1**: значение равно X_{max} (указываем мышкой на число); **Формат** – шрифт полужирный, красный. А также **Условие 2**: значение равно X_{min} (адрес числа); **Формат** – шрифт полужирный, синий.

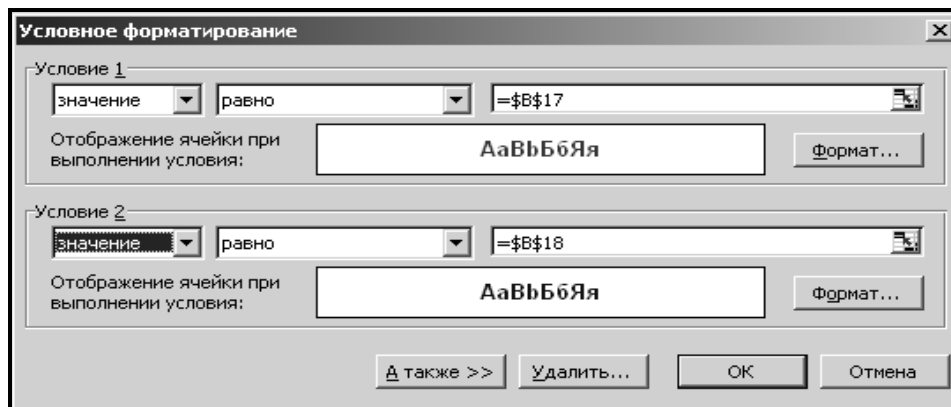


Рис. Л3.3. Панель условного форматирования

Если сейчас в таблице данных (см. рис. Л3.1) удалить наименьшее значение **0,39**, тут же изменится число наблюдений (n) и в таблице будет выделено новое минимальное значение X_{min} .

Группировка данных

Обработка данных представляет собой маленькое научное исследование, группировка – первый шаг к созданию модели, цель которой заключается в отображении основных особенностей изучаемого явления, пренебрегая мелкими второстепенными и случайными различиями.

Как правило, первоначальную группировку приходится в дальнейшем уточнять путем выбора другой ширины интервалов группировки и другого начала отсчета. Надо предусмотреть заранее, чтобы группировка корректировалась автоматически и не требовалось выполнять все расчеты заново. В идеале заполненный лист Excel должен представлять собой универсальный шаблон, позволяющий производить обработку любых других данных, а по внешней форме иметь вид готового отчета, не требующего никаких дополнительных пояснений. Конечно, автоматизировать все не имеет смысла, так как это может потребовать слишком много неоправданных усилий, но вмешательство человека в работу программы можно свести к минимуму.

При группировке данных на интервалы равной ширины следует учесть некоторые требования ГОСТа (ДСТУ). Во-первых, шаг группировки (ширина интервалов) с точностью до множителя $10^{\pm m}$ должен равняться одному из трех чисел: 1, 2, 5. Во-вторых, границы групп должны быть кратны шагу или полу-

шагу. Ориентировочно начальное число интервалов можно принять равным корню квадратному из объема выборки $k \approx \sqrt{n}$, но не меньше 10.

Итак, ориентировочно определяем шаг группировки на 10 интервалов: $h = (X_{max} - X_{min})/10 = 0,216$. Принимаем шаг **$h = 0,2$** и начало 1-го интервала **$s_0 = 0,3$** (рекомендуем выделять полужирным красным шрифтом все числа, которые требуется вводить вручную). Начало отсчета $s_0 = 0,3$ меньше $X_{min} = 0,39$ и кратно полшагу $h / 2 = 0,1$.

Для принятого шага и начала отсчета находим число групп по формуле $k = \text{ОКРУГЛВВЕРХ}((X_{max} - s_0)/h; 0) = 12$ и конец последнего интервала $s_k = s_0 + k \cdot h = 2,7$ (в Excel имеется несколько функций для округления чисел: ЦЕЛОЕ, ОКРУГЛ, ОКРУГЛВВЕРХ, ОКРУГЛВНИЗ; можно было бы обойтись одной первой функцией, аналоги которой присутствуют в любом другом программном обеспечении).

№	Х _{лв}	Х _{пр}	х
0	0,1	0,3	0,2
1	0,3	0,5	0,4
2	0,5	0,7	0,6
3	0,7	0,9	0,8
4	0,9	1,1	1
5	1,1	1,3	1,2
6	1,3	1,5	1,4
7	1,5	1,7	1,6
8	1,7	1,9	1,8
9	1,9	2,1	2
10	2,1	2,3	2,2
11	2,3	2,5	2,4
12	2,5	2,7	2,6
13	2,7	2,9	2,8

Рис. ЛЗ.4. Границы интервалов

Генерируем таблицу границ интервалов (рис. ЛЗ.4). Правая граница интервала № 0 равна s_0 . Границы интервала № 1 определены по формулам: левая граница $X_{лев}$ равна правой границе предыдущего интервала; правая граница отличается от нее на шаг $X_{пр} = X_{лев} + h$ (ссылка на h должна быть абсолютной, чтобы она не изменялась при копировании формулы). Для остальных интервалов формулы копируются автозаполнением. К интервалам № 1 – 12 добавлены пустые интервалы № 0 и 13 – они понадобятся для построения полигона. В последнем столбце таблицы вычислены центры интервалов $x = (X_{лев} + X_{пр})/2$.

В Excel имеется специальная функция ЧАСТОТА, которая производит автоматическую группировку на интервалы произвольной ширины, лишь бы были заданы правые границы интервалов. К сожалению, есть одна тонкость, не документированная в описании, в результате чего группировка с помощью этой функции может отличаться от группировок, полученных другими способами. Например, в Excel имеется надстройка (встроенная подпрограмма) «Анализ данных» и группировки с помощью этой надстройки могут отличаться от группировок с функцией ЧАСТОТА.

Предлагаем два надежных способа автоматической группировки данных, один из которых будет рассмотрен в лабораторных работах 6 – 8. Оба способа основаны на применении функции Excel СЧЁТЕСЛИ (Диапазон; Критерий). «Диапазон» – это массив данных X . В справке по этой функции указано, что, если «Критерий» содержит знаки сравнений ($<$, $<=$, $>=$, $>$), он должен быть *текстовой константой*. Иными словами, входящие в критерий числа должны быть преобразованы в текст, для чего в Excel имеется функция ТЕКСТ (Число; Формат). Формат "0,0###" показывает, что Число будет округлено до трех десятичных знаков, незначимые нули в конце дробной части будут отброшены. Фрагменты текста, заключенные в двойные кавычки, объединяются операторами &.

С помощью формулы СЧЁТЕСЛИ ($X; "<=" & \text{ТЕКСТ}(X_{np}; "0,0###")$) можно найти количество наблюдений, не превышающих заданные правые границы интервалов X_{np} . В таблице на рис. ЛЗ.5 эти числа обозначены как Σm . Разность двух соседних значений накопленных частот (текущее минус предыдущее) равна m – количеству наблюдений в данном интервале.

№	Х _{лв}	Х _{пр}	х	Σm	m	Частота	f	F
0	0,1	0,3	0,2	0	0	0	0	0
1	0,3	0,5	0,4	1	1	1	0,065789	0,013158
2	0,5	0,7	0,6	9	8	8	0,526316	0,118421
3	0,7	0,9	0,8	19	10	9	0,657895	0,25
4	0,9	1,1	1	34	15	15	0,986842	0,447368
5	1,1	1,3	1,2	48	14	15	0,921053	0,631579
6	1,3	1,5	1,4	58	10	8	0,657895	0,763158
7	1,5	1,7	1,6	67	9	9	0,592105	0,881579
8	1,7	1,9	1,8	70	3	4	0,197368	0,921053
9	1,9	2,1	2	72	2	3	0,131579	0,947368
10	2,1	2,3	2,2	73	1	1	0,065789	0,960526
11	2,3	2,5	2,4	75	2	2	0,131579	0,986842
12	2,5	2,7	2,6	76	1	1	0,065789	1
13	2,7	2,9	2,8	76	0	0	0	1
Суммы				76		76		

Рис. ЛЗ.5. Группировка исходных данных

Теперь посмотрим, как работает функция ЧАСТОТА. Ее синтаксис: ЧАСТОТА (Массив_данных; Массив_интервалов). Эта функция возвращает не одно число, а целый диапазон, и поэтому надо знать особенности работы с такими «функциями диапазонов». Прежде всего надо выделить интервал

будущих результатов работы функции (столбец Частота). Далее, *не снимая выделения*, вызвать функцию ЧАСТОТА и заполнить ее поля ввода: Массив_данных – диапазон X , Массив_интервалов – *правые* границы X_{np} . Наконец (*внимание!*), надо нажать одновременно три клавиши **Ctrl+Shift+Enter**. Как видно из приведенной таблицы (см. рис. Л3.4), получились расхождения с предыдущим способом группировки. Все дело в том (это не указано в справке по функции ЧАСТОТА), что правые границы интервалов должны быть *точными*. Числа, которые мы видим на экране, могут отличаться от своего внутреннего представления на компьютере на очень малую величину (на так называемый «машинный ноль»), но именно это приводит к ошибкам группировки. Давайте изменим слегка формулу для вычисления правых границ интервалов на $X_{np} = \text{ОКРУГЛ}(X_{лв} + h; 6)$, то есть округлим правые границы до 6-го знака после запятой. Внешне ничего не изменится, но теперь функция ЧАСТОТА будет давать правильные результаты. В данном способе округление значений правых границ было произведено при задании формата текстового представления чисел «0,0##».

Гистограмма и кумулята

В столбце **f** на рис. Л3.5 для каждого интервала вычислены значения эмпирической плотности вероятности $f_i = m_i / n / h$. Ссылки на n и h должны быть абсолютными, чтобы они не изменялись при копировании формулы в другие интервалы.

График эмпирической плотности вероятности ступенчатый и называется гистограммой (в Excel же гистограммой называются любые столбчатые диаграммы). В столбце **F** на рис. Л3.5 для каждого *правого края* интервалов вычислены значения кумуляты $F = \sum m / n$ – эмпирической оценки интегральной функции распределения. Для непрерывной случайной величины график кумуляты кусочно-линейный.

Оба графика приведены ниже на рис. Л3.6.

Гистограмма получилась вполне удовлетворительной, хотя крайние интервалы придется впоследствии укрупнять. На графике кумуляты показано, как находить медиану и квартили (об этом будет сказано дальше). При построении гистограммы возникнут некоторые проблемы, так как потребуется совмещать гистограмму с графиками дифференциальных функций теоретических законов (нормального и равномерного), а это графики разных типов.

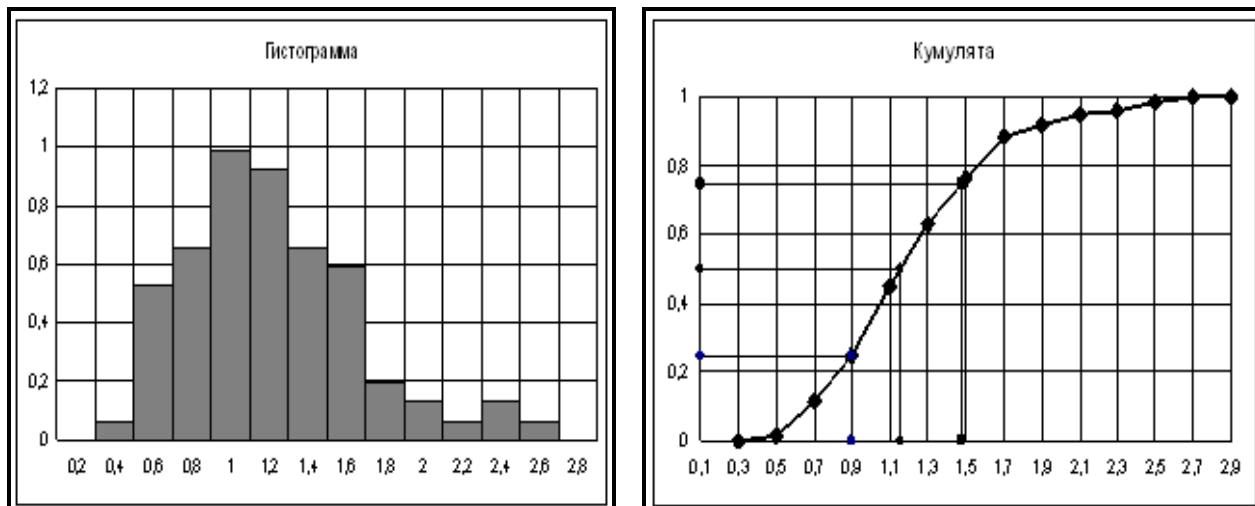


Рис. Л3.6. Графики гистограммы и кумуляты

Расчет характеристик

Для расчета характеристик по исходным данным в Excel имеются функции СРЗНАЧ(X), ДИСПР(X), ДИСП(X) и др. Функция ДИСП(X) возвращает несмещенную оценку дисперсии, а ДИСПР(X) – обычную (в справке по этой функции сказано, что она возвращает генеральную дисперсию – это, конечно, не так). Для сравнения предлагается также вычислить характеристики по сгруппированным данным и определить погрешности группировки (рис. Л3.7).

Расчеты по сгруппированным данным выполняются по формулам:

$$X_{cp} = \text{СУММПРОИЗВ}(x;m) / n, \quad (XX)_{cp} = \text{СУММПРОИЗВ}(x;x;m) / n,$$

$$Dx = (XX)_{cp} - (X_{cp})^2, \quad S_x = \text{КОРЕНЬ}(D_x), \quad V_x = S_x / X_{cp} \cdot 100 \, \%. \quad .$$

Здесь x – центры интервалов, m – частоты.

По исходным данным	По сгруппированным данным	Погрешности группировки
$X_{cp} = 1,2286$	$X_{cp} = 1,2158$	-1,04 %
$Dx = 0,2040$	$Dx = 0,2071$	1,53 %
$Sx = 0,4517$	$Sx = 0,4551$	0,76 %
$Vx = 36,76\%$	$Vx = 37,43 \, \%$	1,82 %

Рис. Л3.7. Сравнительный расчет характеристик

Как видно из приведенной на рис. Л3.7 таблицы, погрешности группировок при $k = 12$ оказались порядка $(1 \div 2) \, \%$.

Лабораторная работа 4

Проверка статистических гипотез

Критерии согласия

По виду гистограммы делаются предположения о виде распределения и формулируется гипотеза, какой из известных теоретических законов наиболее соответствует наблюдаемому распределению. Параметры предполагаемого теоретического закона определяются методом моментов путем приравнивания теоретических характеристик (математического ожидания и дисперсии) к их эмпирическим оценкам (среднему и оценке дисперсии). Для выбранного теоретического закона надо построить графики его функций распределения и сравнить их с соответствующими эмпирическими графиками – гистограммой и кумулятой. Кроме чисто визуального сравнения, используются числовые критерии согласия, из которых наиболее распространенным является критерий Пирсона «Хи-квадрат». В познавательных целях, независимо от вида распределения, студенту в обязательном порядке предлагается проверить согласие с нормальным и равномерным законами распределения.

Расчет функций распределения нормального закона

Приравниваем теоретические характеристики к их эмпирическим значениям: $a = X_{cp} = 1,216$; $\sigma_x = s_x = 0,455$, и вычисляем значения дифференциальной и интегральной функций нормального закона распределения (рис. Л4.1).

№	Х _{лв}	Х _{пр}	х	м	fn	Fn	mn	mnf
0	0,1	0,3	0,2	0	0,072613	0,022096	1,7	1,1
1	0,3	0,5	0,4	1	0,175819	0,057882	2,7	2,7
2	0,5	0,7	0,6	8	0,350949	0,128535	5,4	5,3
3	0,7	0,9	0,8	10	0,577496	0,243877	8,8	8,8
4	0,9	1,1	1	15	0,783394	0,399584	11,8	11,9
5	1,1	1,3	1,2	14	0,87607	0,5734	13,2	13,3
6	1,3	1,5	1,4	10	0,807651	0,733849	12,2	12,3
7	1,5	1,7	1,6	9	0,613811	0,856326	9,3	9,3
8	1,7	1,9	1,8	3	0,384568	0,933634	5,9	5,8
9	1,9	2,1	2	2	0,198627	0,973985	3,1	3,0
10	2,1	2,3	2,2	1	0,084573	0,991398	1,3	1,3
11	2,3	2,5	2,4	2	0,029686	0,997612	0,5	0,5
12	2,5	2,7	2,6	1	0,00859	0,999445	0,1	0,1
13	2,7	2,9	2,8	0	0,002049	0,999893	0,0	0,0
Суммы				76			76,0	75,5

Рис. Л4.1. Расчет функций распределения нормального закона

Дифференциальная функция нормального закона f_n рассчитывается по *центрам* интервалов (x) функцией $f_n = \text{НОРМРАСП}(x; X_{cp}; S_x; 0)$, интегральная функция F_n рассчитывается по *правым границам* интервалов (X_{np}) той же функцией $F_n = \text{НОРМРАСП}(X_{np}; X_{cp}; S_x; 1)$ – разница только в значении последнего параметра функции. Ссылки на X_{cp} и S_x должны быть абсолютными, чтобы они не изменялись при копировании формул.

Знание интегральной функции теоретического распределения позволяет рассчитать ожидаемые частоты $m_n = \Delta F \cdot n$, где ΔF – разность текущего и предыдущего значений интегральной функции F_n (ссылка на n должна быть абсолютной). Для того чтобы сумма ожидаемых частот точно равнялась общему количеству наблюдений $\sum m_n = n$, крайние дополнительные интервалы № 0 и 13 условно расширяем; ожидаемую частоту в начальном интервале № 0 вычисляем по формуле $(F_0 - 0) \cdot n \approx 1,7$, а в последнем (дополнительном) интервале № 13 – по формуле $(1 - F_{12}) \cdot n \approx 0,0$.

В недавнем прошлом, когда все расчеты выполнялись вручную, была необходимость вычислять ожидаемые частоты по значениям дифференциальной функции (поскольку таблицы интегральных функций некоторых законов распределения были только в специальных справочниках). В учебных целях в последней колонке вышеприведенной таблицы найдены ожидаемые частоты по приближенной формуле $m_{nf} \approx f_n \cdot h \cdot n$. На удивление, получилось неплохое соответствие с расчетами по интегральной функции.

Критерий согласия Пирсона

Проверяем нормальность распределения по критерию Пирсона χ^2 , сравнивая два ряда частот: наблюдаемых (m) и ожидаемых по теоретическому закону (m_n). Последовательность расчетов показана на рис. Л4.2.

№	m	mn	Хи2	Скоррект.
0 – 2	9	9,8	0,06	0,07
3	10	8,8	0,17	0,20
4	15	11,8	0,85	1,00
5	14	13,2	0,05	0,06
6	10	12,2	0,39	0,47
7	9	9,3	0,01	0,01
8 – 13	9	10,9	0,34	0,39
Суммы	76	76,0	1,87	2,20

Рис. Л4.2. Расчеты по критерию Пирсона

Малонасыщенные интервалы надо укрупнить так, чтобы в каждый укрупненный интервал попало не менее 5-ти наблюдений ($m, m_n \geq 5$).

Пришлось укрупнить (объединить) крайние интервалы (0, 1, 2) и (8, 9, 10, 11, 12, 13). В столбце **Хи2** подсчитаны стандартные члены формулы Пирсона $(m - m_n)^2 / m_n$, а в со-

седнем столбце (**Скоррект**) – с поправкой $\chi^2 / (1 - m_n / n)$. Стандартное значение статистики Пирсона оказалось равно $\chi^2 = 1,87$; скорректированное значение – немного больше: $\chi^2 = 2,20$. Число укрупненных интервалов (число сравниваемых пар частот) равно 7; для двухпараметрического нормального закона вычисляем $\chi^2_{СС} = 7 - 3 = 4$. Функцией ХИ2ОБР (Вероятность; $\chi^2_{СС}$) находим критические значения: ХИ2ОБР (0,99;4) = 0,30; ХИ2ОБР (0,95;4) = 0,71; ХИ2ОБР (0,05;4) = 9,49; ХИ2ОБР (0,01;4) = 13,28.

Оба вычисленных значения χ^2 попали в область принятия нулевой гипотезы $0,71 < \chi^2 < 9,49$. Гипотеза о нормальности распределения не может быть отвергнута.

Критерий согласия Колмогорова – Смирнова

Находим максимальную разницу (по абсолютной величине) между кумулятой (F) и интегральной функцией теоретического закона (F_n). Наибольшее расхождение оказалось для интервала № 5: $d = |0,632 - 0,574| = 0,058$. Вычисляем статистику Колмогорова – Смирнова $KS = d \cdot \text{КОРЕНЬ}(n) = 0,51$, которую сравниваем с критическими значениями $KS_{0,05} = 1,36$ и $KS_{0,01} = 1,63$. Так как вычисленное значение KS оказалось меньше $KS_{0,05}$, принимается нуль-гипотеза об отсутствии значимых различий между рядами частот. Как уже неоднократно указывалось, простой критерий Колмогорова – Смирнова слишком «либеральный»; его выводы заслуживают внимания, когда критерий отвергает нуль-гипотезу.

Равномерный закон распределения

Параметрами равномерного закона распределения являются границы единственной ступеньки a и b ; высота ступеньки равна $f_p = 1 / (b - a)$; характеристики закона выражаются через параметры:

$$M(x) = (a + b) / 2, D_x = (b - a)^2 / 12.$$

Для определения наилучших значений параметров теоретического закона приравниваем его характеристики к соответствующим эмпирическим оценкам $X_{cp} = 1,216$ и $S_x = 0,455$.

Получаем систему уравнений для определения параметров:

$$a = X_{cp} - S_x \cdot \text{КОРЕНЬ}(3) = 0,428;$$

$$b = X_{cp} + S_x \cdot \text{КОРЕНЬ}(3) = 2,004.$$

Высота единственной ступеньки равна $f_p = 1 / (b - a) = 0,634$.

	x		fp	Fp
s0 =	0,1		0	0
a =	0,428		0	0
a =	0,428	fp =	0,634	0
b =	2,004	fp =	0,634	1
b =	2,004		0	1
sk =	2,9		0	1

Рис. Л4.3. **Равномерный закон**

Графики функций распределения равномерного закона состоят из последовательных отрезков прямых. На рис. Л4.3 составлена таблица координат начала и конца каждого отрезка, после чего строятся графики распределения. Оба графика имеют один тип («точечный», то есть кусочно-линейный).

Графики функций распределения

Построим на одной диаграмме графики всех дифференциальных функций, а на другой – всех интегральных функций (рис. Л4.4).

Если все сделано без ошибок, графики интегральных функций должны переплетаться (при ошибках получаются явные сдвиги между графиками разных распределений – эмпирического, нормального, равномерного).

На диаграмме дифференциальных функций вместо гистограммы построен полигон – ломаная, которая соединяет середины ступенек гистограммы (именно для полного построения полигона при группировке данных были добавлены крайние пустые интервалы № 0 и 13).



Рис. Л4.4. **Стандартные графики функций распределения**

Полигон эквивалентным образом представляет гистограмму, его площадь равна единице. График полигона имеет тот же тип, что и остальные графики, поэтому при построении диаграммы с графиками дифференциальных функций никаких затруднений не возникает.

Однако традиционно пытаются совместить на одной диаграмме графики разных типов, что в принципе сделать можно и не одним способом. На рис. Л4.5 приведены два нестандартных графика дифференциальных функций распределения, построенных разными способами, которые будут описаны в следующей лабораторной работе.

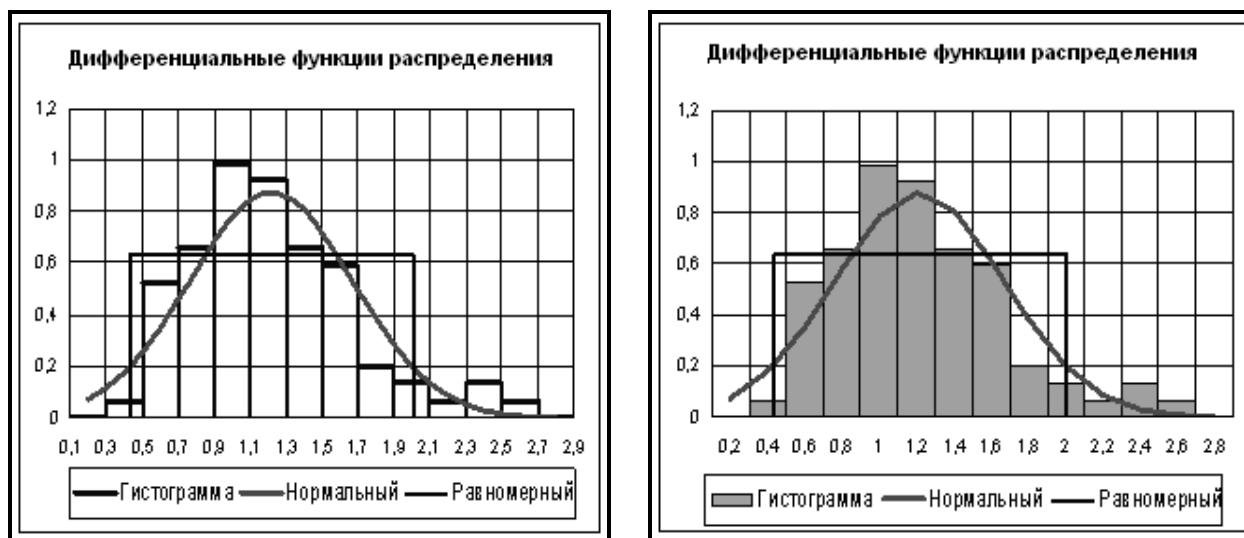


Рис. Л4.5. Нестандартные комбинированные графики

Интервальная оценка математического ожидания

При заданном уровне доверия $P = 1 - \alpha$ границы доверительного интервала, который с вероятностью P покрывает неизвестное генеральное среднее (математическое ожидание), вычисляются по формуле $\bar{X}_{cp} \pm HCP$, где HCP – наименьшая существенная разность, которая равна $HCP = t_{1-P} \cdot S_x / \text{КОРЕНЬ}(n)$; табличное значение статистики Стьюдента t_{1-P} вычисляется функцией $t_{1-P} = \text{СТЮДРАСПОБР}(1 - P; n - 1)$.

Доверительные границы на $M(x)$
$\bar{X}_{cp} - HCP < M(x) < \bar{X}_{cp} + HCP$
P = 0,95
$t_{1-P} = 1,99$
$HCP = 0,103$
$1,113 < M(x) < 1,319$
Погрешность = 8,49 %

Рис. Л4.6. Расчет интервальной оценки центра совокупности

На рис. Л4.6 изображен блок расчетов доверительных границ на неизвестное математическое ожидание, где для принятого уровня доверия $P = 0,95$ ($\alpha = 0,05$) получен 95-процентный интервал $1,113 < M(x) < 1,319$ с относительной погрешностью $HCP / \bar{X}_{cp} = 0,085 = 8,5 \%$. Если задать меньший уровень доверия (гарантию) $P = 0,9$, доверительная погрешность снизится

до 7,1 %. Если задать больший уровень доверия $P = 0,99$, доверительная погрешность увеличится до 11,3 %.

Определение потребного объема выборки

Когда заданы и уровень доверия (P), и предельная относительная погрешность (q), требуется определить объем выборки (n), при котором результаты будут получены с заданной точностью и надежностью. Потребный объем выборки зависит от коэффициента вариации совокупности. Предполагаем, что он близок к выборочному коэффициенту вариации $V_x = 0,37$ или немного больше. Объем выборки n находим как решение неравенства $n > (t_{1-P} \cdot V_x / q)^2$, где табличное значение статистики Стьюдента также зависит от n . Поэтому данное неравенство решаем последовательными приближениями (итерациями).

Потребный объем выборки	
P = 95 %	q = 5 %
V_x = 40 %	t_{1-P} = 1,97
t _{1-P} = СТЬЮДРАСПОБР(1-P; n-1)	
Задаем: n = 250	
Получаем: n > (t_{1-P} * V_x / q)² = 248,3	

Рис. Л4.7. Блок расчета потребного объема выборки

На рис. Л4.7 изображен блок расчетов потребного объема выборки, где для заданного уровня доверия $P = 0,95$ (95 %), предельной погрешности $q = 0,05$ (5 %) и коэффициента вариации $V_x = 0,40$ (40 %), вычислен потребный объем выборки $n = 250$.

Если снизить уровень доверия до $P = 0,9$ (90 %), потребный объем выборки снизится до $n = 175$.

Интервальная оценка на генеральную дисперсию

Интервальная оценка Пирсона на генеральную дисперсию σ^2 выведена в предположении нормальности распределения X :

$$\frac{SSX}{\chi^2_{\alpha}} < \sigma^2 < \frac{SSX}{\chi^2_{1-\alpha}}.$$

Здесь $SSX = n \cdot D_x = n \cdot (S_x)^2$;

P – уровень доверия;

$\alpha = (1 - P) / 2$;

$\chi^2_{\alpha} = \text{ХИ2ОБР}(\alpha; n - 1)$;

$\chi^2_{1-\alpha} = \text{ХИ2ОБР}(1 - \alpha; n - 1)$.

Доверительные границы на дисперсию	
P =95 %	SSX =15,7411
$\alpha=(1-P)/2=2,5 \%$	$1 - \alpha =97,5 \%$
$SSX / \chi^2_{\alpha} < \sigma^2 < SSX / \chi^2_{1-\alpha}$	
$\chi^2_{\alpha}=100,839$	
$\chi^2_{1-\alpha}=52,942$	
$0,1561 < \sigma^2 < 0,2973$	
$0,3951 < \sigma^2 < 0,5453$	

Рис. Л4.8. Блок расчета интервальной оценки дисперсии

На рис. Л4.8 изображен блок расчетов интервальных границ на дисперсию совокупности, где для $P = 0,95$ (95 %) получены границы:

$$0,395 < \sigma < 0,545.$$

Для меньшего уровня $P = 0,90$ (90 %) получены границы: $0,404 < \sigma < 0,530$.

Нормальная вероятностная кривая

Для графической проверки соответствия предполагаемого теоретического закона и подбора подходящего функционального преобразования случайной величины с целью приблизить ее распределение к теоретическому предлагается построить следующий график (этот график называется нормальной вероятностной кривой, если проверяется согласие распределения с нормальным законом).

Рассматриваются ординаты кумуляты (эмпирической интегральной функции), и по этим значениям находят квантили предполагаемого теоретического распределения (то есть по теоретической интегральной функции находят значения аргумента, для которых теоретическая интегральная функция равна соответствующим значениям кумуляты).

Для нормального закона в Excel имеется две функции: НОРМСТОБР(Вероятность) и НОРМОБР(Вероятность; Среднее; Станд_Откл).

С помощью одной из этих функций (например, НОРМСТОБР) для всех значений кумуляты, кроме крайних значений 0 и 1, вычисляются значения $t = \text{НОРМСТОБР}(F)$ и строится график $(X_{np} - t)$. В случае соответствия распределений все точки нормальной вероятностной кривой должны группироваться вокруг прямой $t = (X_{np} - X_{cp}) / S_x$.

На рис. Л4.9а по данным рассматриваемого примера построен график нормальной вероятностной кривой, откуда видно, что точки (X_{np}, t) систематически уклоняются от указанной прямой, что говорит о несоответствии наблюдаемого распределения нормальному закону.

На сравнительных графиках полигона (гистограммы) и кривой нормального закона (см. рис. Л4.4 и Л4.5) также визуальны явные различия между этими распределениями.

Однако критерий согласия Пирсона не выявил расхождений между наблюдаемыми и теоретическими частотами значимыми, поэтому был сделан вывод о допустимости использования в дальнейшем нормального закона как аппроксимации действительного распределения.

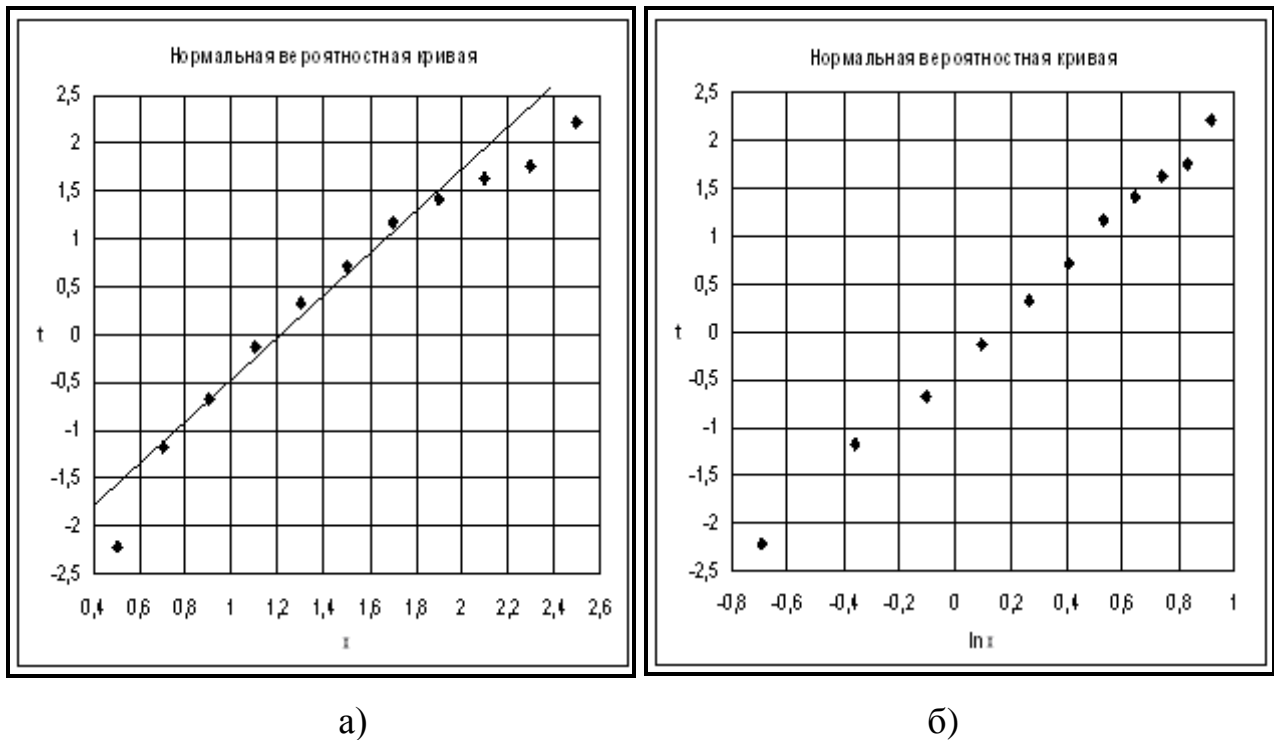


Рис. Л4.9. Нормальная вероятностная кривая для x и $\ln x$

Вид кривой, вокруг которой группируются точки (X_{np}, t) , подсказывает возможное функциональное преобразование случайной величины.

В данном примере эта кривая похожа на график логарифма, поэтому все значения X_{np} были прологарифмированы и построен новый график (см. рис. Л4.9б) в полулогарифмическом масштабе $(\ln X_{np}, t)$. После этого преобразования точки нормальной вероятностной кривой стали тесно группироваться вокруг некоторой прямой, что свидетельствует в пользу предположения о лог-нормальном законе распределения.

Лабораторная работа 5

Нестандартная графика

Определение квартилей и выявление выбросов

На графике кумуляты показан графический способ определения квартилей (рис. Л5.1): на оси ординат задаем значения $F = 0,25; 0,5; 0,75$ и на оси абсцисс находим соответствующие значения квартилей $X_{0,75}$ (нижняя квартиль), $X_{0,5} = Me$ (средняя квартиль, или медиана), $X_{0,25}$ (верхняя квартиль). Квартили делят весь размах варьирования на четыре равнонасыщенных интервала по 25 % наблюдений в каждом.

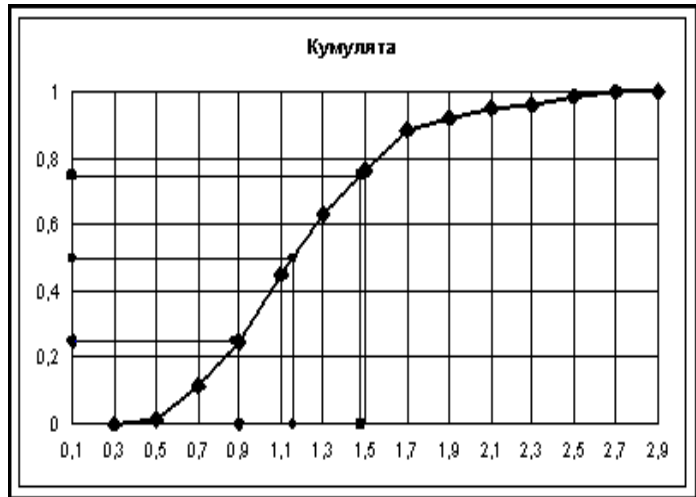


Рис. Л5.1. Графический способ определения квартилей

По определению квантиль X_α является решением уравнения $F(X_\alpha) = 1 - \alpha$. Для данного примера нижняя квартиль находится сразу, так как для $X_{np} = 0,9$ значение кумуляты оказалось точно равно $F = 0,25$; отсюда $X_{0,75} = 0,9$. Для вычисления остальных квартилей потребуется линейное интерполирование. Пусть F_1 – ближайшее меньшее, а F_2 – ближайшее большее заданного значения $F = 1 - \alpha$; X_1 и X_2 – правые границы интервалов, соответствующие F_1 и F_2 . Для заданного F вычисляем квантиль $X_{1-F} = X_1 + (X_2 - X_1) / (F_2 - F_1) \cdot (F - F_1)$. Так, для вычисления медианы из таблицы значений кумуляты (см. рис. Л3.4) выписываем: при $X_1 = 1,1$ значение $F_1 = 0,447 < 0,5$, а для следующего $X_2 = 1,3$ $F_2 = 0,632 > 0,5$.

Отсюда $X_{0,5} = 1,1 + (1,3 - 1,1) / (0,632 - 0,447) \cdot (0,5 - 0,447) = 1,157$ – это и есть медиана. По этой же формуле вычисляем верхнюю квартиль: $X_{0,25} = 1,3 + (1,5 - 1,3) / (0,763 - 0,632) \cdot (0,75 - 0,632) = 1,480$.

Полезно создать блок линейного интерполирования (рис. Л5.2) и продублировать его 3 раза для вычисления нижней, средней и верхней квартилей.

$F_1 = 0,250$	$X_1 = 0,9$
$F_2 = 0,250$	$X_2 = 0,9$
$F = 0,25$	$X = 0,900$

$F_1 = 0,447$	$X_1 = 1,1$
$F_2 = 0,632$	$X_2 = 1,3$
$F = 0,50$	$X = 1,157$

$F_1 = 0,632$	$X_1 = 1,3$
$F_2 = 0,763$	$X_2 = 1,5$
$F = 0,75$	$X = 1,480$

Рис. Л5.2. Вычисление квартилей

В Excel имеется функция КВАРТИЛЬ(Массив; Часть), которая для данного примера дает близкие значения квартилей: $X_{0,75} = 0,908$; $X_{0,5} = 1,13$; $X_{0,25} = 1,50$.

Вычисляем межквартильный размах $\Delta = X_{0,25} - X_{0,75} = 1,48 - 0,90 = 0,58$. Все данные, выходящие за пределы интервала $(X_{0,75} - 1,5 \cdot \Delta; X_{0,25} + 1,5 \cdot \Delta) = (0,90 - 0,87; 1,48 + 0,87) = (0,03; 2,35)$, считаются выбросами. Так как $X_{min} = 0,39 > 0,03$, выбросов влево нет. Но $X_{max} = 2,55 > 2,35$ можно считать выбросом вправо. Функцией НАИБОЛЬШИЙ (Массив; К) находим остальные выбросы вправо:

НАИБОЛЬШИЙ(X ; 2) = 2,5 > 2,35; НАИБОЛЬШИЙ(X ; 3) = 2,45 > 2,35; но НАИБОЛЬШИЙ(X ; 4) = 2,2 – уже меньше 2,35 (рис. Л5.3).

Итак, три наибольших значения похожи на выбросы вправо. Если их удалить, все автоматически пересчитывается и распределение приближается к нормальному. Интересно, что правило «3-х сигм» не обнаружило ни одного выброса. По всей видимости, выбросы сдвинули положение среднего X_{cp} .



Рис. Л5.3. Гистограмма после удаления трех выбросов

Построение блочной диаграммы Тьюкки

Английский статистик Джон Тьюкки предложил вместо гистограмм строить диаграммы «ящик и усы» (рис. Л5.4).

Границами «ящика» являются нижняя и верхняя квартили, они ограничивают диапазон «лучшей половины» наблюдений.



Рис. Л5.4. Блочная диаграмма Тьюки

«Усы» простираются до X_{min} и до X_{max} , но не более полутора межквартильного размаха от границ «ящика». Данные, которые выходят за пределы «усов», отмечаются отдельными точками как выбросы. Крестиком обозначено X_{cp} .

Эту диаграмму можно построить следующим образом. Определяем квантили, предельную длину «усов» вправо и влево, с помощью функций НАИБОЛЬШИЙ и НАИМЕНЬШИЙ находим все выбросы. Далее составляем таблицу последовательности узлов блочной диаграммы так, чтобы ее можно было нарисовать отрезками прямых (для наглядности на рис. Л5.5 немного изменены абсциссы некоторых точек):



Рис. Л5.5. Способ построения блочной диаграммы Тьюки

Первый способ построения комбинированных диаграмм

В Excel удобно совмещать в одной диаграмме графики одного типа, например кусочно-линейные (*Точечные*). При этом столбчатый график гистограммы следует заменить на эквивалентный кусочно-линейный график полигона. Однако иногда, по традиции, желают в одной диаграмме совместить графики разных типов – привычную гистограмму с наложенными на нее графиками теоретических законов распределения. Сделать это возможно, хотя работа нестандартная и требует определенных трудозатрат. Ниже на рис. Л5.6 приведена диаграмма с графиками дифференциальных функций распределения и полигоном вместо гистограммы. Мы желаем преобразовать полигон в гистограмму,

то есть совместить на одной диаграмме графики разных типов. Щелкнем правой кнопкой мышки по полю диаграммы и в контекстном меню выберем *Тип диаграммы* / **Нестандартные** / *График* | *Гистограмма* (рис. Л5.7).



Рис. Л5.6. Кусочно-линейные графики дифференциальных функций

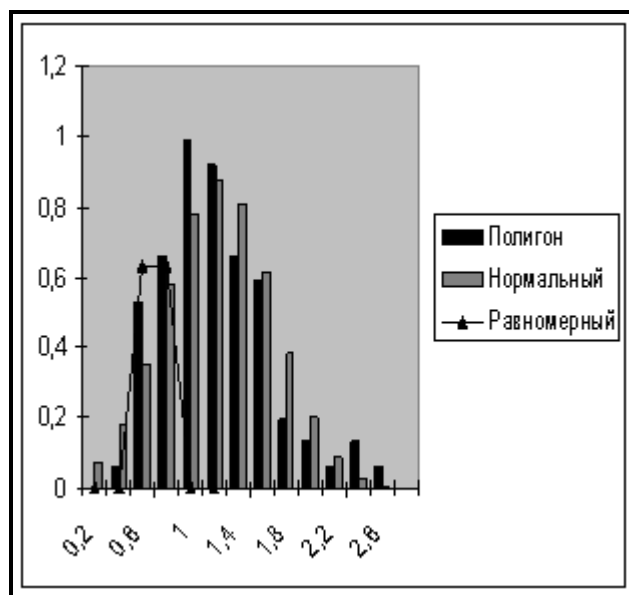


Рис. Л5.7. Изменение типа диаграммы
Нестандартные / *График* | *Гистограмма*

В столбчатую форму преобразовались сразу два графика (полигона и нормального закона), график равномерного закона теперь необычный и сдвинут влево. Форматирование потеряно.

Прежде всего щелкнем правой кнопкой мышки по столбчатому графику нормального закона и в контекстном меню выбираем *Тип диаграммы* / **Стандартные** / *Точечная*. Ту же операцию сделаем с графиком равномерного закона. Вид графиков нормального и равномерного законов изменился к лучшему, но они почему-то сдвинуты влево (рис. Л5.8). Восстанавливаем форматирование: *Легенда* – внизу, *Фон диаграммы* – белый, *Рамка* – черная, *Линии сетки* – по обеим осям, в легенде слово «Полигон» заменяем на «Гистограмма». Щелкнем правой кнопкой мышки по столбику гистограммы, в контекстном меню выбираем пункт *Формат рядов данных*. Во вкладке *Параметры* устанавливаем *ширину зазора*, равную нулю; во вкладке *Вид* изменяем цвет гистограммы.



Рис. Л5.8. Диаграмма после форматирования



Рис. Л5.9. Окончательный вид диаграммы

Теперь разберемся, почему графики нормального и равномерного законов сдвинуты влево. Дело в том, что для гистограммы по оси абсцисс откладываются номера интервалов, начиная с единицы. Метки на оси, которые мы видим на комбинированной диаграмме, – всего лишь подписи, никак не связанные с графиками нормального и равномерного законов.

Для точечных графиков (нормального и равномерного) надо задать другие Значения X преобразованием реальных значений аргументов по формуле $X = (x - x_0) / h + 1$. Для нормального закона новые значения X будут совпадать с номерами интервалов, увеличенными на единицу (у нас нумерация интервалов начинается с № 0, а для комбинированной диаграммы требуется нумерацию начинать с единицы). Для равномерного закона таблицу координат узлов (x, f_p) на рис. Л4.3 придется дополнить справа новыми значениями абсцисс X в другом масштабе (рис. Л5.10). Заменяем на диаграмме рис. Л5.8 Значения X для теоретических законов на новые значения в другом масштабе и получаем желаемую комбинированную диаграмму (рис. Л5.9). Работа закончена.

Равномерный закон		
X	x	f_p
0,500	0,1	0
2,138	0,428	0
2,138	0,428	0,634
10,020	2,004	0,634
10,020	2,004	0
14,500	2,9	0

Рис. Л5.10. Новые абсциссы X для графика равномерного закона

Второй способ построения комбинированных диаграмм

№	X	f
0	0,1	0
	0,3	0
	0,3	0
1	0,3	0,065789
	0,5	0,065789
	0,5	0
2	0,5	0,526316
	0,7	0,526316
	0,7	0
3	0,7	0,657895
	0,9	0,657895
	0,9	0
4	0,9	0,986842
	1,1	0,986842
	1,1	0
5	1,1	0,921053
	1,3	0,921053
	1,3	0
6	1,3	0,657895
	1,5	0,657895
	1,5	0
7	1,5	0,592105
	1,7	0,592105
	1,7	0
8	1,7	0,197368
	1,9	0,197368
	1,9	0
9	1,9	0,131579
	2,1	0,131579
	2,1	0
10	2,1	0,065789
	2,3	0,065789
	2,3	0
11	2,3	0,131579
	2,5	0,131579
	2,5	0
12	2,5	0,065789
	2,7	0,065789
	2,7	0
13	2,7	0
	2,9	0
	2,9	0

Рис. Л5.11. Координаты узлов гистограммы

Кроме большой трудоемкости, первый способ представления распределений имеет еще один недостаток: все интервалы должны быть одинаковой ширины. Иногда это ограничение является существенным. Кроме этого, довольно трудно добавить на такую комбинированную диаграмму еще хотя бы один график. Поэтому предлагается вычертить гистограмму так, как мы вычерчивали одну ступеньку для равномерного закона. При этом все графики будут одного типа, что снимает массу проблем. Для вычерчивания каждой ступеньки гистограммы требуется построить три отрезка прямых (вверх, вправо, вниз). Надо спроектировать это для одного интервала, для всех остальных формулы будут скопированы.

На рис. Л5.11 в столбце № через две строки набираем номера интервалов (достаточно набрать два номера и воспользоваться автозаполнением). В столбце X с помощью функции ВПР(№; Таблица; Номер столбца; 0) находим левую границу Х_{лв} (Номер столбца – 2) и два раза правую границу Х_{пр} (Номер столбца – 3). Ссылка на Таблицу (рис. Л3.4) должна быть абсолютной, чтобы она не менялась при копировании. В столбце f с помощью той же функции ВПР находим два раза значение f (Номер столбца – 8). Третье число в блоке из трех строк – нуль. Копируем заполненный блок для всех остальных №.

Далее строим ступенчатый график гистограммы как обычный точечный график, к которому легко добавляются графики теоретических законов распределения (рис. Л5.12). В готовую комбинированную диаграмму добавили еще график полигона, соединяющего середины ступенек гистограммы.



Рис. Л5.12. Готовая комбинированная диаграмма

Вопросы для самопроверки

1. Перечислите основные характеристики генеральной совокупности и их выборочные оценки (центра группировки, меры изменчивости, функций распределения).
2. Запишите сравнительные формулы для вычисления характеристик и их выборочных оценок (математического ожидания, дисперсии, ковариации).
3. Поясните, что означает состоятельность, несмещенность и эффективность статистических оценок. Приведите формулу для вычисления несмещенной оценки дисперсии.
4. Перечислите свойства математического ожидания и выборочного среднего, приведите примеры их использования (сформулируйте нулевое свойство математического ожидания, упростите формулы для вычисления дисперсии и ковариации, обоснуйте возможность применения условных переменных).
5. Перечислите свойства дисперсии. Запишите формулы для вычисления дисперсии суммы и дисперсии разности случайных величин (зависимых и независимых). Сформулируйте правило «3-х сигм». Поясните, что такое коэффициент вариации и в каких ситуациях он используется.
6. Сформулируйте утверждение центральной предельной теоремы. Перечислите основные особенности нормального распределения. Поясните, что такое ошибка среднего.
7. Дайте определение понятию о распределении Стьюдента. Покажите, как определяются границы 95-процентного доверительного интервала на генеральное среднее (математическое ожидание); как определяется необходимый объем выборки для оценки центра группировки совокупности с заданной точностью и надежностью.

8. Поясните, что такое гистограмма, как она строится, чем отличается от дифференциальной функции распределения, чему равна ее полная площадь и площадь на заданном интервале.

9. Поясните, что такое кумулята, как она строится, чем отличается от интегральной функции распределения. Дайте определение функции распределения (интегральной функции распределения) и сформулируйте суть общей интегральной теоремы.

10. Дайте определение понятию о квантилях распределения, покажите, как определяются медиана и квартили. Опишите блочную диаграмму Тьюкки.

11. Покажите, как по данным выборки найти параметры теоретического закона распределения и как по выбранному закону рассчитать ожидаемые частоты попадания случайной величины в заданные интервалы (с помощью интегральной и дифференциальной функций распределения).

12. Дайте определение понятию о распределении Пирсона, о критерии согласия Пирсона, опишите последовательность расчетов по критерию Пирсона.

13. Дайте понятие о числе степеней свободы и о числе связей. Покажите, какие связи есть при сравнении эмпирических и ожидаемых частот (в критерии Пирсона).

14. Обоснуйте условия, необходимые для корректного применения критерия Пирсона (учесть особенности распределения Бернулли – Пуассона – Лапласа).

15. Покажите, как определяются границы 90-процентного доверительного интервала на генеральную дисперсию с помощью распределения Пирсона.

16. Дайте определение понятию о критерии согласия Колмогорова – Смирнова.

17. Дайте определение понятию о нормальной вероятностной кривой, покажите, как она строится и применяется.

18. Опишите особенности нормального закона распределения, его параметры и характеристики, дифференциальную и интегральную функции, структуру таблиц, область применения.

19. Опишите особенности равномерного закона распределения, его параметры и характеристики, дифференциальную и интегральную функции, область применения.

20. Опишите особенности показательного закона распределения, его параметры и характеристики, дифференциальную и интегральную функции, область применения.

Лабораторная работа 6

Анализ корреляционных связей

Следующие три лабораторные работы 6 – 8 посвящены проблемам регрессионного и дисперсионного анализов.

Цель работы:

1. По данным наблюдений двух показателей предлагается определить тип корреляционной связи и выявить возможные выбросы – аномальные наблюдения, явно не относящиеся к данной совокупности (чаще всего выбросы появляются как ошибки в записи чисел). Для этого, прежде всего, надо построить график разброса точек в осях (X, Y) . В докомпьютерную эпоху строили корреляционное поле, группируя данные на ряд интервалов по каждой переменной. Корреляционное поле полезно также для изучения некоторых других проблем анализа связей.

2. Независимо от расположения экспериментальных точек предлагается найти МНК-оценки параметров *линейной* зависимости, построить ее график, оценить тесноту линейной связи и ее значимость. Для сравнения параметры линейной зависимости (коэффициенты регрессии) следует найти по исходным и по сгруппированным данным.

3. Так как у нас имеется корреляционное поле, можно построить эмпирическую линию регрессии – кусочно-линейный график с узлами (X_i, U_i) , где $U_i = \bar{y}_{x_i}$ – средние интервальные (средние значения результативной переменной в каждой группе по X_i). С помощью дисперсионного анализа проверяется значимость существующей корреляционной связи. Вычисляется более объективная оценка тесноты существующей связи (корреляционное отношение вместо коэффициента корреляции). Для ординат узлов эмпирической линии регрессии полезно вычислить интервальные оценки (доверительные интервалы).

4. Проверяем адекватность линейной модели, сравнивая корреляционное отношение с коэффициентом корреляции (по готовой формуле или заполняя соответствующую таблицу дисперсионного анализа).

5. Корреляционное поле можно рассматривать как таблицу сопряженности категорий двух качественных показателей. Появляется возможность оценить значимость существующей связи по критерию Пирсона, а тесноту связи – по коэффициентам контингенции. Полезно сравнить результаты оценок тесноты связи и ее значимости по разным методикам.

Представление данных

Исходные данные обычно записываются в виде таблицы из двух столбцов X , Y ; можно также добавить столбец номеров наблюдений № (рис. Л6.2).

Такое расположение удобно для расчетов, но не удобно для отчета, если длинный столбец данных не помещается на одной странице.

Имеется возможность компактно расположить данные в несколько столбцов, но тогда появляются несмежные диапазоны значений переменных.

В формулах адреса несмежных диапазонов надо указывать мышкой при нажатой клавише **Ctrl**.

Можно любому диапазону присвоить краткое имя и использовать его в формулах.

На компьютере имеется возможность посмотреть на график разброса точек x , y (рис. Л6.1).

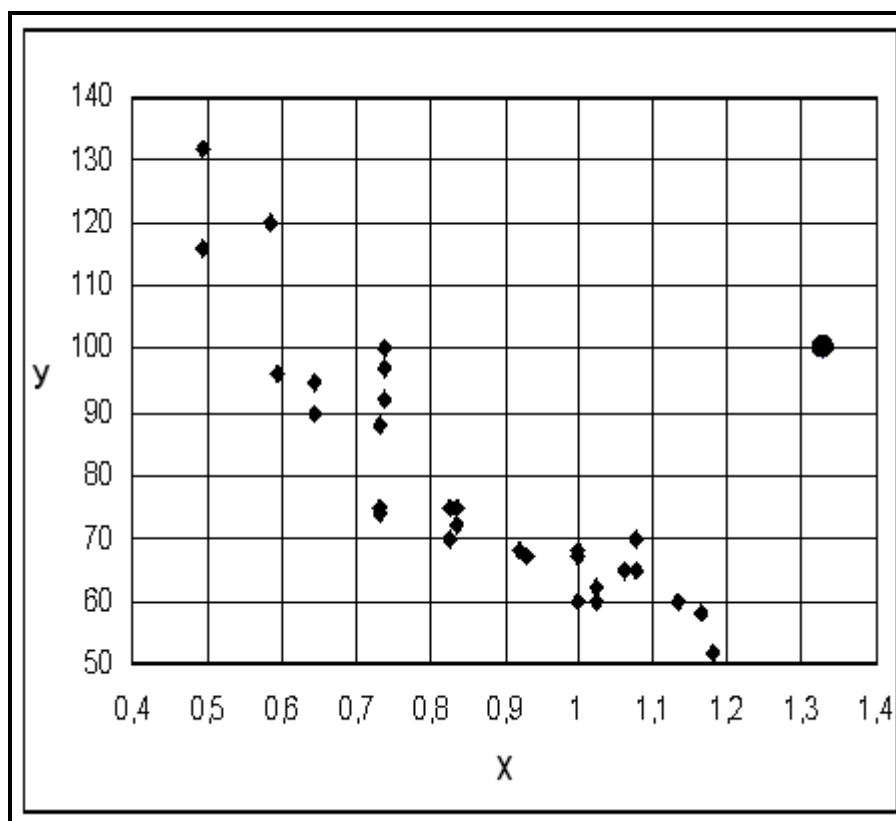


Рис. Л6.1. График разброса точек

Из этого графика визуально определяем наличие выброса (точка № 22 – выделена курсивом на графике и в таблице на рис. Л6.2). Выбросы, естественно, надо удалять.

№	X	Y	№X	№Y	Шифр
1	1,182	52	8	1	801
2	1,076	70	7	2	702
3	0,999	60	6	1	601
4	0,646	95	3	5	305
5	0,740	97	4	5	405
6	0,646	95	3	5	305
7	0,740	97	4	5	405
8	0,920	68	6	2	602
9	1,063	65	7	2	702
10	1,076	65	7	2	702
11	1,024	60	7	1	701
12	1,063	65	7	2	702
13	0,646	90	3	4	304
14	0,733	75	4	3	403
15	0,74	92	4	5	405
16	0,826	75	5	3	503
17	1,063	65	7	2	702
18	1,076	65	7	2	702
19	0,999	67	6	2	602
20	0,931	67	6	2	602
21	0,495	132	1	9	109
22	1,329	100	10	5	1005
23	0,835	72	5	3	503
24	1,166	58	8	1	801
25	1,135	60	8	1	801
26	0,931	67	6	2	602
27	1,076	65	7	2	702
28	1,024	62	7	2	702
29	0,999	68	6	2	602
30	0,835	75	5	3	503
31	0,826	75	5	3	503
32	0,740	100	4	5	405
33	0,733	74	4	3	403
34	0,495	116	1	7	107
35	0,585	120	2	7	207
36	0,999	68	6	2	602
37	0,999	68	6	2	602
38	0,733	88	4	4	404
39	0,835	75	5	3	503
40	0,826	70	5	2	502
41	0,999	67	6	2	602
42	0,999	67	6	2	602
43	0,999	67	6	2	602
44	0,594	96	2	5	205

Рис. Л6.2. Исходные данные
и номера классов (интервалов)

Двойная группировка данных

При отсутствии компьютера данные следует сгруппировать и отобразить их на так называемом корреляционном поле. Нам все равно понадобятся сгруппированные данные для изучения некоторых проблем анализа связей, поэтому переходим к рассмотрению метода двойной группировки данных.

Здесь будет описан другой надежный способ группировки данных, отличный от того, который был использован в лабораторной работе 3.

Предварительно заполняем блок на рис. Л6.3, где для каждой переменной выбираем шаг группировки и начало первого интервала. Находим для каждой переменной максимальные и минимальные значения и вычисляем ориентировочно величину шага группировки на 10 интервалов (обозначим ориентировочные значения hh_x и hh_y). Принимаем значения шага (h_x и h_y), ближайшие к ориентировочным, но допускаемым по ГОСТу (ДСТУ). Принимаем также значения начала первого интервала (они должны быть меньше минимальных значений и кратны принятому шагу или полушагу). Принятые значения выделены в блоке на рис. Л6.3 красным цветом.

Определяем количество интервалов группировки (p, q) по каждой переменной и значения правых границ последних интервалов:

$$p = \text{ОКРУГЛВВЕРХ}((X_{\max} - x_0) / h_x; 0);$$

$$q = \text{ОКРУГЛВВЕРХ}((Y_{\max} - y_0) / h_y; 0);$$

$$x_p = x_0 + p \cdot h_x, \quad y_q = y_0 + q \cdot h_y.$$

Теперь для каждого наблюдения на рис. Лб.2 вычисляем номера интервалов:

$$\text{№}X = \text{ОКРУГЛВВЕРХ}((x - x_0) / h_x; 0);$$

$$\text{№}Y = \text{ОКРУГЛВВЕРХ}((y - y_0) / h_y; 0).$$

В последней колонке таблицы (см. рис. Лб.2) для каждого наблюдения вычислен Шифр = $100 \cdot \text{№}X + \text{№}Y$, объединяющий в краткой записи номера интервалов по X и по Y .

Функцией СЧЁТЕСЛИ подсчитываем количество одинаковых шифров и записываем найденные частоты m_{ij} в таблицу размером $p \times q$ (рис. Лб.4).

В верхней строке этой таблицы записываем номера групп по переменной X (индексы i). В левом столбце таблицы записываем номера групп по переменной Y (индексы j), причем для визуального сходства с графиком разброса индексы j записываем в порядке убывания (так как ось Y направлена вверх). Для каждого индекса можно сгенерировать значения центров интервалов. Частоты вычисляются формулой $m_{ij} = \text{СЧЁТЕСЛИ}(\text{Шифры}; 100 \cdot \text{Индекс}_i + \text{Индекс}_j)$, которая набирается для одной ячейки таблицы и копируется в остальные.

n =44	n =44
Xmax =1,329	Ymax =132
Xmin =0,495	Ymin =52
hhx =0,083	hhy =8
hx =0,1	hy =10
x0 =0,4	y0 =50
p =10	q =9
xp =1,4	yq =140

Рис. Лб.3. Шаг и границы интервалов группировки

	$i \Rightarrow$	1	2	3	4	5	6	7	8	9	10	
j	$\begin{smallmatrix} Y \\ X \end{smallmatrix}$	0,45	0,55	0,65	0,75	0,85	0,95	1,05	1,15	1,25	1,35	l
9	135	1	0	0	0	0	0	0	0	0	0	1
8	125	0	0	0	0	0	0	0	0	0	0	0
7	115	1	1	0	0	0	0	0	0	0	0	2
6	105	0	0	0	0	0	0	0	0	0	0	0
5	95	0	1	2	4	0	0	0	0	0	1	8
4	85	0	0	1	1	0	0	0	0	0	0	2
3	75	0	0	0	2	5	0	0	0	0	0	7
2	65	0	0	0	0	1	10	8	0	0	0	19
1	55	0	0	0	0	0	1	1	3	0	0	5
	k	2	2	3	7	6	11	9	3	0	1	44

Рис. Лб.4. Данные, сгруппированные на 9×10 интервалов по Y и X

При копировании не должны изменяться ссылки на диапазон Шифров, ссылки на строку Индекс_i и ссылки на столбец Индекс_j (при наборе в формуле адресов этих операндов клавишей **F4** делаются абсолютные ссылки или на весь адрес, или только на адрес строки, или только на адрес столбца).

Корреляционное поле – это центральная часть таблицы частот m_{ij} . С помощью условного форматирования можно нулевые частоты вывести серыми символами, чтобы они не отвлекали внимания. На корреляционном поле ясно виден выброс, который мы уже обнаружили ранее на графике разброса (см. рис. Л6.1).

Выбросы должны быть удалены, после чего автоматически перевычисляются число наблюдений, минимальные и максимальные значения, число интервалов и правые границы последних интервалов. Иногда после этой операции приходится задавать другие значения шага группировки или другое начало отсчета. На рис. Л6.5 изменено начало отсчета на $x_0 = 0,45$.

n =43	n =43
Xmax =1,182	Ymax =132
Xmin =0,495	Ymin =52
hhx =0,069	hhy =8
hx =0,1	hy =10
x0 =0,45	y0 =50
p =8	q =9
xp =1,25	yq =140

Рис. Л6.5. Шаг и границы интервалов группировки после удаления выброса

В данном примере из таблицы на рис. Л6.4 удаляются два последних столбца (с индексами $i = 9, 10$), размер таблицы сокращается до 9×8 . Новое корреляционное поле изображено на рис. Л6.6.

	$i \Rightarrow$	1	2	3	4	5	6	7	8			
j	y^x	0,5	0,6	0,7	0,8	0,9	1	1,1	1,2	l	ΣmX	V
9	135	1	0	0	0	0	0	0	0	1	0,5	0,500
8	125	0	0	0	0	0	0	0	0			
7	115	1	1	0	0	0	0	0	0	2	1,1	0,550
6	105	0	0	0	0	0	0	0	0			
5	95	0	3	4	0	0	0	0	0	7	4,6	0,657
4	85	0	1	1	0	0	0	0	0	2	1,3	0,650
3	75	0	0	2	5	0	0	0	0	7	5,4	0,771
2	65	0	0	0	1	3	8	7	0	19	19,2	1,011
1	55	0	0	0	0	0	2	1	2	5	5,5	1,100
	k	2	5	7	6	3	10	8	2	43		
	ΣmY	250	485	615	440	195	630	510	110			
	U	125,00	97,00	87,86	73,33	65,00	63,00	63,75	55,00			

Рис. Л6.6. Корреляционное поле после удаления выброса

Подсчитываем суммы частот по столбцам и строкам таблицы:

$$k_i = \sum_j m_{ij}; \quad l_j = \sum_i m_{ij}.$$

Делается это так: выделяем диапазон частот *вместе* с дополнительной строкой снизу и дополнительным столбцом справа; далее на панели инструментов нажимаем мышкой кнопку $\boxed{\Sigma}$.

При этом в дополнительные ряды записываются *формулы* суммирования, а не числа.

В последних двух строках таблицы вычислены средние интервальные значения $U_i = \bar{Y}_{x_i} = \sum_j m_{ij} Y_j / k_i$ для каждого интервала по X .

Аналогично в последних двух столбцах таблицы вычислены средние интервальные значения $V_j = \bar{X}_{y_j} = \sum_i m_{ij} X_i / l_j$ для каждого интервала по Y .

Суммы $\sum mY$ и $\sum mX$ еще понадобятся в дальнейшем.

Расчет параметров линейной модели

По графику разброса и виду корреляционного поля делаем заключение, что эмпирические точки явно уклоняются от линейной зависимости и в принципе требуется дополнительное исследование по определению типа нелинейной зависимости.

Однако изучение линейной модели является обязательным, поэтому рассчитываем ее параметры как по исходным, так и по сгруппированным данным.

Последовательность расчетов и необходимые вычислительные формулы как по исходным, так и по сгруппированным данным сведены в таблицу на рис. Л6.7.

В основном различия в методиках расчета заключаются в способах вычисления средних: X_{cp} , Y_{cp} , $(XX)_{cp}$, $(YY)_{cp}$, $(XY)_{cp}$.

При расчете на компьютере по исходным данным можно использовать также некоторые альтернативные формулы, например, в Excel имеются готовые формулы для вычисления коэффициента корреляции R_{xy} , коэффициента регрессии b_1 и свободного члена b_0 линейной модели $y_p = b_0 + b_1 x$.

Расчеты по исходным данным x, y – диапазоны исходных данных	Расчеты по сгруппированным данным X, Y – диапазоны центров интервалов; k, l – диапазоны частот
$X_{cp} = \text{СУММ}(x) / n$ $X_{cp} = \text{CPЗНАЧ}(x)$	$X_{cp} = \text{СУММПРОИЗВ}(X; k) / n$
$Y_{cp} = \text{СУММ}(y) / n$ $Y_{cp} = \text{CPЗНАЧ}(y)$	$Y_{cp} = \text{СУММПРОИЗВ}(Y; l) / n$
$(XX)_{cp} = \text{СУММКВ}(x) / n$	$(XX)_{cp} = \text{СУММПРОИЗВ}(X; X; k) / n$
$(YY)_{cp} = \text{СУММКВ}(y) / n$	$(YY)_{cp} = \text{СУММПРОИЗВ}(Y; Y; l) / n$
$(XY)_{cp} = \text{СУММПРОИЗВ}(x; y) / n$	$(XY)_{cp} = \text{СУММПРОИЗВ}(X; \Sigma mY) / n$ $(XY)_{cp} = \text{СУММПРОИЗВ}(Y; \Sigma mX) / n$
$Dx = (XX)_{cp} - (X_{cp})^2$ $Dx = \text{ДИСПР}(x)$	$Dx = (XX)_{cp} - (X_{cp})^2$
$Dy = (YY)_{cp} - (Y_{cp})^2$ $Dy = \text{ДИСПР}(y)$	$Dy = (YY)_{cp} - (Y_{cp})^2$
$S_{xy} = (XY)_{cp} - X_{cp}Y_{cp}$ $S_{xy} = \text{КОВАР}(x; y)$	$S_{xy} = (XY)_{cp} - X_{cp}Y_{cp}$
$Sx = \text{КОРЕНЬ}(Dx)$	$Sx = \text{КОРЕНЬ}(Dx)$
$Sy = \text{КОРЕНЬ}(Dy)$	$Sy = \text{КОРЕНЬ}(Dy)$
$R_{xy} = S_{xy} / Sx / Sy$ $R_{xy} = \text{КОРРЕЛ}(x, y)$	$R_{xy} = S_{xy} / Sx / Sy$
$b1 = R_{xy} * Sy / Sx$	$b1 = R_{xy} * Sy / Sx$
$b0 = Y_{cp} - b1 * X_{cp}$	$b0 = Y_{cp} - b1 * X_{cp}$

Рис. Л6.7. Формулы для расчета параметров линейной модели

Относительные погрешности группировки на $p = 8$ интервалов в этом примере (рис. Л6.8) оказались довольно большими (больше 5 %), тем не менее, судя по графикам (рис. Л6.9), имеется очень хорошее соответствие между наблюдениями и расчетными значениями.

Исходные данные	Сгруппированные	Погрешн.
$X_{cp} = 0,8848$	$X_{cp} = 0,8744$	-1,17%
$Y_{cp} = 76,628$	$Y_{cp} = 75,233$	-1,82%
$(XX)_{cp} = 0,8164$	$(XX)_{cp} = 0,8037$	-1,55%
$(YY)_{cp} = 6180,2$	$(YY)_{cp} = 5978,5$	-3,26%
$(XY)_{cp} = 64,899$	$(XY)_{cp} = 62,721$	-3,36%
$Dx = 0,0335$	$Dx = 0,0391$	16,77%
$Dy = 308,37$	$Dy = 318,55$	3,30%
$S_{xy} = -2,903$	$S_{xy} = -3,064$	5,55%
$Sx = 0,1830$	$Sx = 0,1978$	8,06%
$Sy = 17,561$	$Sy = 17,848$	1,64%
$R_{xy} = -0,903$	$R_{xy} = -0,868$	-3,90%
$b1 = -86,66$	$b1 = -78,33$	-9,61%
$b0 = 153,31$	$b0 = 143,73$	-6,25%
$a1 = -0,0094$	$a1 = -0,0096$	2,18%
$a0 = 1,6061$	$a0 = 1,5980$	-0,51%

Рис. Л6.8. Сравнительные расчеты по исходным и сгруппированным данным

Кроме параметров (коэффициентов регрессии) линейной модели $y_p = b_0 + b_1 \cdot x$, рассчитаны также параметры сопряженной модели, когда y считается объясняющей, а x – результативной переменной: $a_1 = R_{xy} \cdot S_x / S_y$, $a_0 = \bar{X} - b_1 \cdot \bar{Y}$, $x_p = a_0 + a_1 \cdot y$.

Графики обеих линий регрессии приведены ниже на рис. Л6.9.

На этих же графиках построены эмпирические линии регрессии (X_i, U_i) и (V_j, Y_j) .

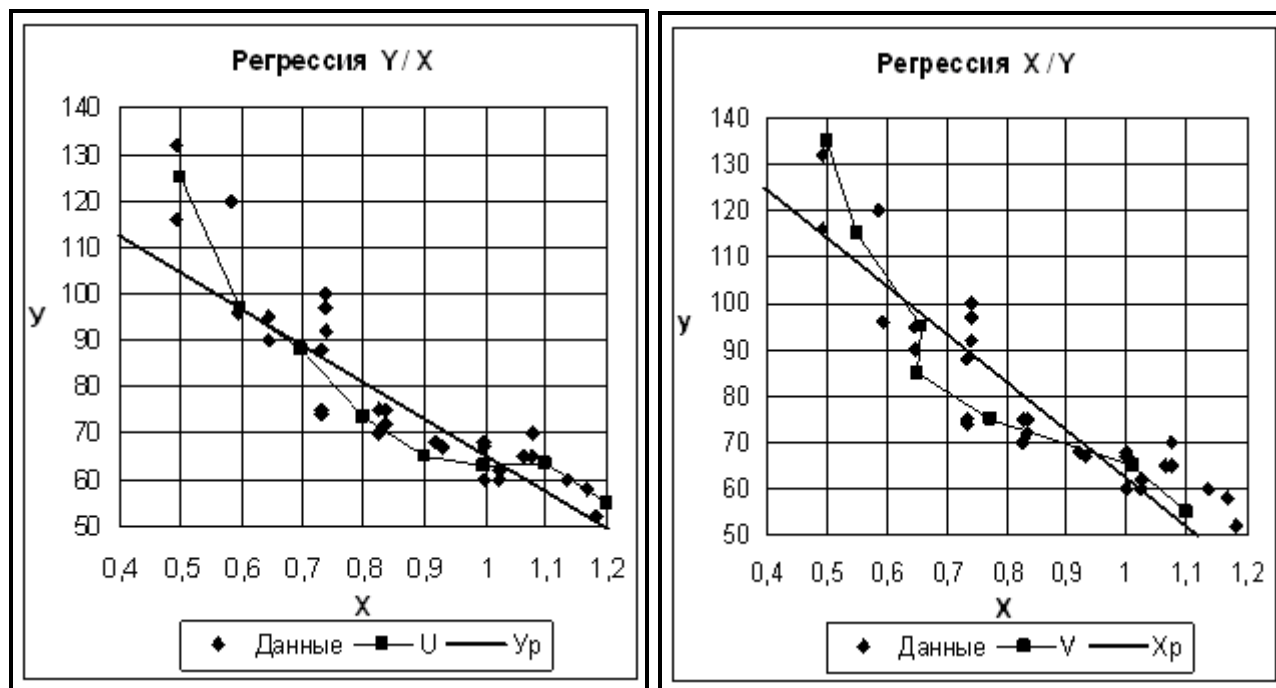


Рис. Л6.9. Графики взаимно сопряженных линий регрессии

При построении эмпирической линии регрессии X/Y для сопряженной модели было небольшое затруднение: диапазоны значений V (а значит, и соответствующие диапазоны Y) не являются смежными, поэтому их надо задавать при нажатой клавише **Ctrl** (иначе линия регрессии будет с пропусками).

Кстати, иногда, наоборот, требуется, чтобы некоторые соединительные линии не изображались.

Вопрос на сообразительность – как это можно сделать с наименьшей трудоемкостью?

Эмпирические линии регрессии явно указывают на нелинейный характер зависимости.

Лабораторная работа 7

Проблемы тесноты, значимости и адекватности

Различают тесноту и значимость существующей корреляционной связи и тесноту, значимость и адекватность корреляционной модели. Анализ существующей корреляционной зависимости можно провести лишь в том случае, когда имеются параллельные наблюдения (повторения) с одинаковыми значениями аргумента в каждой группе или же когда исходные данные предварительно сгруппированы на несколько интервалов варьирования аргумента.

Основным математическим инструментом в этом случае является дисперсионный анализ.

Дисперсионный анализ.

Оценка тесноты и значимости корреляционной связи

После двойной группировки мы представили данные в виде p групп с примерно одинаковыми значениями аргумента X_i в каждой группе; количество наблюдений в каждой группе обозначено через k_i ($\sum k_i = n$); среднее значение результативной переменной Y_{ij} в каждой группе обозначено через U_i .

Модель дисперсионного анализа заключается в разложении «полного сигнала» Y_{ij} на «полезный сигнал» U_i плюс помеху ε_{ij} : $Y_{ij} = U_i + \varepsilon_{ij}$.

Точно такое же разложение имеет полная сумма квадратов $SSY = SSU + SS\varepsilon$, или обычная оценка дисперсии $D_Y = D_U + D_\varepsilon$.

Отношение $\eta^2_{y/x} = D_u / D_y$ называется индексом детерминации, оно показывает, какая часть полной изменчивости Y_{ij} объясняется различиями между группами с разными значениями аргумента X_i или какая часть изменчивости Y_{ij} определяется наличием корреляционной связи Y/X .

Индекс детерминации есть мера тесноты корреляционной связи – если индекс детерминации равен нулю, корреляционной связи нет; если индекс детерминации приближается к единице, корреляционная связь приближается к наиболее тесной связи – функциональной.

В результате двойной группировки данные также можно представить в виде q групп с примерно одинаковыми значениями аргумента Y_j в каждой группе; количество наблюдений в каждой группе обозначено через l_j ($\sum l_j = n$); среднее значение результативной переменной X_{ij} в каждой группе обозначено через V_j .

Рассуждая аналогично вышеприведенному, вводим меру тесноты сопряженной корреляционной зависимости в виде $\eta_{x/y}^2 = D_v / D_x$.

Если один из индексов детерминации существенно превосходит другой, то это может являться доводом в пользу того или иного направления причинно-следственной связи.

Определим тесноту корреляционных связей для данного примера.

Вычисляем дисперсии средних интервальных:

$$\begin{aligned}(UU)_{cp} &= \text{СУММПРОИЗВ}(U; U; k) / n = 5942,356; & U_{cp} &= Y_{cp} = 75,233; \\ (VV)_{cp} &= \text{СУММПРОИЗВ}(V; V; l) / n = 0,79862; & V_{cp} &= X_{cp} = 0,87442; \\ D_u &= (UU)_{cp} - (U_{cp})^2 = 282,418; & D_v &= (VV)_{cp} - (V_{cp})^2 = 0,034012.\end{aligned}$$

Вычисляем индексы детерминации:

$$\eta_{y/x}^2 = D_u / D_y = 0,8866; \quad \eta_{x/y}^2 = D_v / D_x = 0,8696; \quad r_{xy}^2 = 0,7534.$$

Таким образом, 88,7 % общей изменчивости Y объясняется наличием корреляционной связи Y/X (то есть различиями между группами наблюдений с разными значениями X_i); 87,0 % общей изменчивости X объясняется наличием корреляционной связи X/Y (то есть различиями между группами наблюдений с разными значениями Y_j); линейной моделью объясняется 75,3 % общей изменчивости (одинаково для взаимно сопряженных моделей).

Значимость корреляционной связи устанавливается или после заполнения таблицы дисперсионного анализа 1 (рис. Л7.1), или по готовой формуле.

Изменчивость	Суммы квадратов	ЧСС	Средние квадраты	Дисперс. отношение	Табл. значения
Между группами U	$SS_u = 12144$	$df_u = 7$	$MS_u = 1735$	$F_\eta = 39,08$	$F_{0,05} = 2,29$
Внутри групп ε	$SS_\varepsilon = 1554$	$df_\varepsilon = 35$	$MS_\varepsilon = 44,39$		$F_{0,01} = 3,20$
Общая Y	$SS_y = 13698$	$df_y = 42$			$\text{Alpha} = 0,00$

Рис. Л7.1. Таблица дисперсионного анализа для проверки значимости корреляционной связи

Суммы квадратов пропорциональны дисперсиям: $SS_y = n \cdot D_y$, $SS_u = n \cdot D_u$. Все остальные графы заполняются стандартным образом.

Так как вычисленное дисперсионное отношение $F_\eta = 39,08$ превышает табличное $F_{0,01} = 3,20$, нуль-гипотеза о случайности различий между группами с

разными значениями X_i *отвергается*; имеется значимая корреляционная связь Y/X между X и Y .

Дисперсионное отношение можно было вычислить по готовой формуле:

$$F_{\eta} = \frac{\eta^2}{1-\eta^2} \cdot \frac{n-p}{p-1} = \frac{0,887}{1-0,887} \cdot \frac{43-8}{8-1} = 39,08.$$

Вместо сравнения с табличными значениями $F_{0,05}$ и $F_{0,01}$ (которые в Excel вычисляются функцией ФРАСПОБР), можно функцией ФРАСП найти вероятность (Alpha) того, что вычисленное значение F является случайным отклонением от единицы.

Нуль-гипотеза принимается, если эта вероятность окажется больше 0,05, и отвергается, если $\text{Alpha} < 0,01$.

Доверительные интервалы на центры групп

После того как с помощью дисперсионного анализа найдено, что между группами имеются значимые различия, желательно выяснить, между какими именно группами имеются значимые различия.

Считаем, что группы отличаются только средними интервальными – средними значениями результативной переменной; дисперсия же – мера изменчивости данных в каждой группе – одинакова и равна $MS_{\varepsilon} = 44,39$.

Можно, конечно, с помощью критерия Стьюдента оценить значимость разницы между каждой парой групп, но таких сравнений будет $C_p^2 = \frac{p(p-1)}{2} = \frac{8 \cdot 7}{2} = 28$.

Гораздо проще построить интервальные оценки на центры каждой группы (доверительные интервалы на математическое ожидание результативной переменной в каждой группе).

Ширина этих доверительных интервалов равна:

$$\pm HCP_{\alpha} = \pm t_{\alpha}(df_{\varepsilon}) \cdot \sqrt{MS_{\varepsilon} / k_i}.$$

Принимаем уровень доверия 95 % и по таблицам Стьюдента находим $t_{0,05}(35) = 2,03$ (в Excel табличные значения вычисляются функцией СТБДРАСПОБР).

Ниже в таблице на рис. Л7.2 подсчитаны нижние ($U - HCP$) и верхние ($U + HCP$) границы 95-процентных доверительных интервалов на узлы эмпирической линии регрессии.

X	X	k	U	НСР	U – НСР	U + НСР
0,5	0,5	2	125,00	9,56	115,44	134,56
0,6	0,6	5	97,00	6,05	90,95	103,05
0,7	0,7	7	87,86	5,11	82,75	92,97
0,8	0,8	6	73,33	5,52	67,81	78,86
0,9	0,9	3	65,00	7,81	57,19	72,81
1	1	10	63,00	4,28	58,72	67,28
1,1	1,1	8	63,75	4,78	58,97	68,53
1,2	1,2	2	55,00	9,56	45,46	64,56

Рис. Л7.2. Расчет интервальных оценок на центры в каждой группе

С какой-то целью абсциссы узлов на рис. Л7.2 повторены два раза. (Зачем? Напоминаем, что нам надо изобразить отдельные вертикальные отрезки с маркерами на краях. Не должно быть соединительных линий между этими отрезками).

График средних интервальных вместе с границами 95-процентных интервалов в виде вертикальных отрезков с маркерами построен на рис. Л7.3.

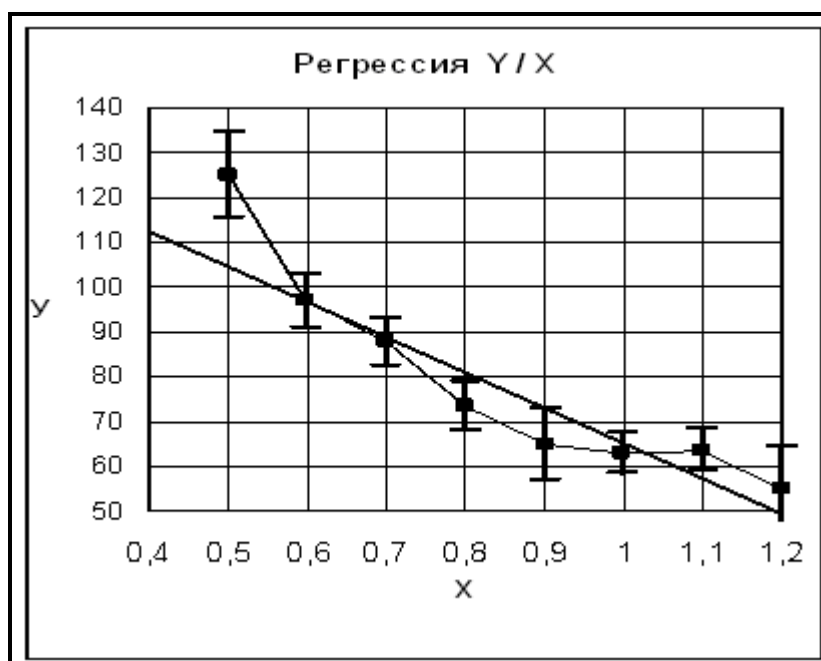


Рис. Л7.3. Сравнение линейной и эмпирической линий регрессии

Из рис. Л7.3 видно, что последние 5 групп неразличимы по уровню варьирования результативной переменной (их доверительные интервалы перекрываются).

Сравнивая эмпирическую и линейную линии регрессии, замечаем значимые отклонения наблюдаемых значений от расчетных по линейной модели – график теоретической регрессии не пересекает некоторые доверительные

интервалы на центры групп. Это свидетельствует о нелинейном характере существующей корреляционной связи.

Оценка тесноты и значимости линейной модели

Теснота линейной связи оценивается с помощью коэффициента детерминации (квадрата коэффициента корреляции).

Коэффициент детерминации $r_{xy}^2 = 0,7534$ показывает, какая часть полной изменчивости определяется регрессионной моделью (75,3 %): значимость регрессионной модели устанавливается или по готовой формуле, или же после заполнения таблицы дисперсионного анализа 2 (рис. Л7.4), в которой сумма квадратов расчетных значений вычисляется с помощью коэффициента детерминации:

$$SSp = (R_{xy})^2 \cdot SSy.$$

Остальные графы таблицы заполняются обычным образом.

Изменчивость	Суммы квадратов	ЧСС	Средние квадраты	Дисперсионное отношение	Табличные значения
Регрессия Y_p	$SSp = 10320$	$dfp = 1$	$MSp = 10320$	$Fr = 125,26$	$F_{0,05} = 4,08$
Остаток e	$SSe = 3378$	$dfe = 41$	$MSe = 82,39$		$F_{0,01} = 7,30$
Общая Y	$SSy = 13698$	$dfy = 42$			$Alpha = 0,00$

Рис. Л7.4. Дисперсионный анализ для проверки значимости линейной модели

Так как вычисленное дисперсионное отношение $F_r = 125,26$ превышает табличное $F_{0,01} = 7,30$, нуль-гипотеза о незначимости линейной модели отвергается.

Дисперсионное отношение можно было вычислить по готовой формуле:

$$F_r = \frac{r_{xy}^2}{1-r_{xy}^2} \cdot \frac{n-2}{1} = \frac{0,753}{1-0,753} \cdot \frac{43-2}{1} = 125,26.$$

Проверка адекватности (линейности) модели

Остаток модели (e) складывается из случайной ошибки (ε) и систематической ошибки неверного выбора формы связи (ошибки спецификации модели).

Обозначим эту ошибку через A (неадекватность).

Если нам известна мера изменчивости случайной ошибки, можно проверить значимость ошибки неадекватности.

Из таблицы дисперсионного анализа 1 (см. рис. Л7.1) выписываем все сведения об изменчивости данных внутри групп – это и есть случайная изменчивость.

Из таблицы дисперсионного анализа 2 (см. рис. Л7.4) выписываем сведения об остатке модели.

Разности между суммами квадратов и числами степеней свободы характеризуют систематическую ошибку спецификации модели.

Заполняем таблицу дисперсионного анализа 3 (рис. Л7.5).

Изменчивость	Суммы квадратов	ЧСС	Средние квадраты	Дисперсионное отношение	Табличные значения
Неадекватность A	$SS_A = 1824$	$dfp = 6$	$MSp = 304$	$F_A = 6,84$	$F_{0,05} = 2,37$
Случайность ε	$SS_\varepsilon = 1554$	$df_\varepsilon = 35$	$MS_\varepsilon = 44,4$		$F_{0,01} = 3,37$
Остаток e	$SS_e = 3378$	$dfe = 41$	$MSe = 82,4$		Alpha = 0,00

Рис. Л7.5. Дисперсионный анализ для проверки адекватности модели

Так как вычисленное дисперсионное отношение $F_A = 6,84$ превышает табличное $F_{0,01} = 3,37$, нуль-гипотеза о незначимости систематической ошибки отвергается.

Линейная модель неадекватная, надо искать более подходящую форму связи.

Дисперсионное отношение можно было вычислить по готовой формуле:

$$F_A = \frac{\eta_{y/x}^2 - r_{xy}^2}{1 - \eta_{y/x}^2} \cdot \frac{n-p}{p-2} = \frac{0,887 - 0,753}{1 - 0,887} \cdot \frac{43-8}{8-2} = 6,84.$$

Лабораторная работа 8

Специальные вопросы регрессионного анализа

Выбор нелинейной формы связи

Линейная модель для описания данных оказалась неадекватной, и требуется найти более подходящую нелинейную модель. Обычно принимают двухпараметрическую форму связи $F(x, y) = a + b \cdot \Phi(x, y)$, которая является линейной после функциональных преобразований переменных. Чаще всего проверяют или обратные преобразования переменных, или же их логарифмирование. Таким образом, предлагается рассмотреть шесть двухпараметрических моделей:

- $y = a + b/x$ – с обратным преобразованием переменной x ;
- $1/y = a + b \cdot x$ – с обратным преобразованием переменной y ;
- $1/y = a + b/x$ – с обратным преобразованием обеих переменных;
- $y = a + b \cdot \ln x$ – с логарифмированием переменной x ;
- $\ln y = a + b \cdot x$ – с логарифмированием переменной y ;
- $\ln y = a + b \cdot \ln x$ – с логарифмированием обеих переменных.

Наилучшей среди двухпараметрических моделей будет та, для которой получится наибольший коэффициент корреляции. Графиком линейной зависимости является прямая линия; отсюда следует такой визуальный критерий выбора наилучшей нелинейной формы связи: в функциональных масштабах эмпирические точки должны группироваться вокруг некоторой прямой.

Сделаем копию рабочего листа Excel и в копии заменим одну из переменных на функционально преобразованную. Сначала попробуем заменить аргументы x в таблице исходных данных (см. рис. Л6.2) на обратные величины: $z = 1/x$.

Практически все расчеты по исходным данным корректируются автоматически, и потребуется в основном только заменить разметку осей на графиках (но в расчетах по сгруппированным данным придется просмотреть и самостоятельно откорректировать всю последовательность вычислений).

$Z_{cp} = 1,197$
$Y_{cp} = 75,233$
$S_z = 0,292$
$S_y = 17,848$
$R_{zy} = 0,939$
$b = 57,414$
$a = 6,536$

Рис. Л8.1. Параметры нелинейной модели

На рис. Л8.1 приведены результаты перерасчета по сгруппированным данным после обратного преобразования объясняющей переменной $z = 1/x$. Получено уравнение регрессии $y_p = 6,536 + 57,414/x$ с коэффициентом корреляции $R_{zy} = 0,939$ (для линейной модели $R_{xy} = -0,868$).

На графиках (рис. Л82а, б) в функциональных масштабах ($1/x$, y) видно, что эмпирические точки тесно группируются вокруг прямой (линии регрессии), а теоретическая линия регрессии пересекает доверительные интервалы для всех узлов эмпирической линии регрессии. Удалась первая же попытка найти подходящую нелинейную форму связи.

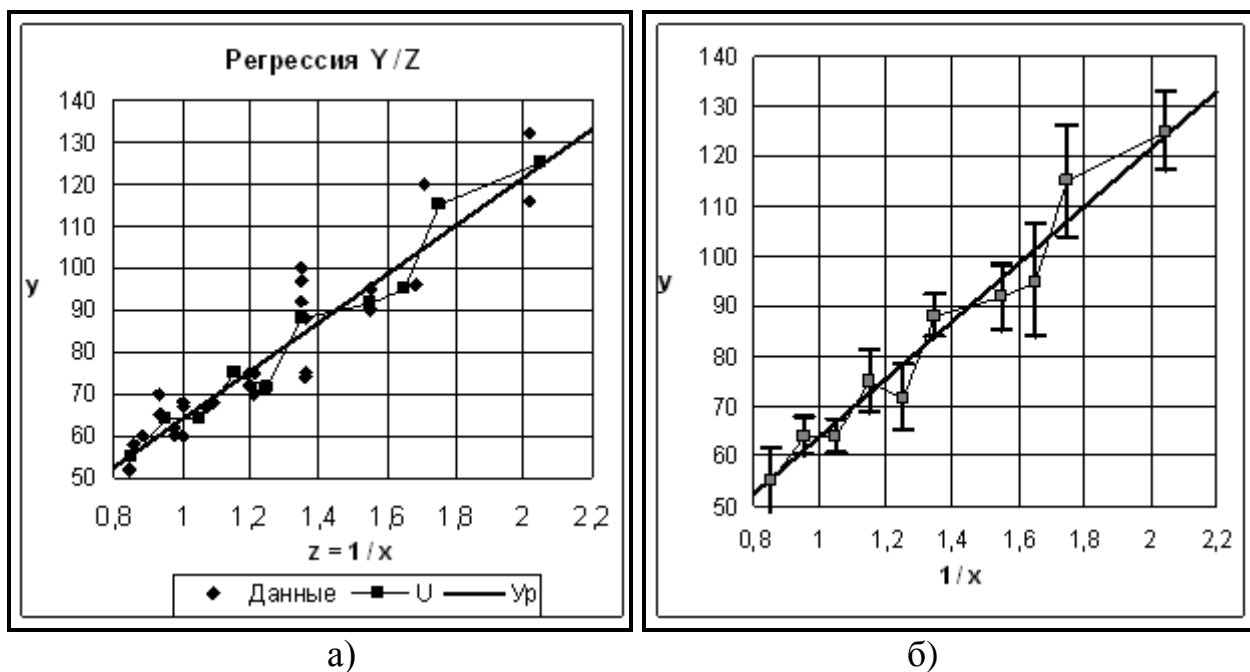


Рис. Л8.2. Графики нелинейной зависимости в функциональных масштабах

Для построения эмпирической линии регрессии U и доверительных интервалов на ее узлы предлагается два способа. Первый способ: после замены в копии рабочего листа значений x на их обратные значения $z = 1/x$ следует вручную внести все соответствующие коррективы. Например, здесь был выбран новый шаг группировки $h_z = 0,1$ и начало первого интервала $z_0 = 0,8$. В результате получилось новое число интервалов $p = 13$, из которых три интервала оказались пустыми (для них $k_i = 0$). Графики $U \pm HCP$ по z и по x пришлось построить заново. Именно такая работа была проделана, чтобы получить рис. Л8.2б. Второй способ: надо сделать еще одну копию рабочего листа и в этой новой копии только на корреляционном поле (см. рис. Л6.6) заменить на обратные значения центры интервалов по оси абсцисс. Тогда на графиках понадобится всего лишь изменить разметку по оси абсцисс. Таким способом график на рис. Л7.3 был преобразован к виду рис. Л8.3а. Для сравнения рядом (рис. Л8.3б) расположен аналогичный график, построенный ранее (см. рис. Л8.2б) первым, более трудоемким способом.

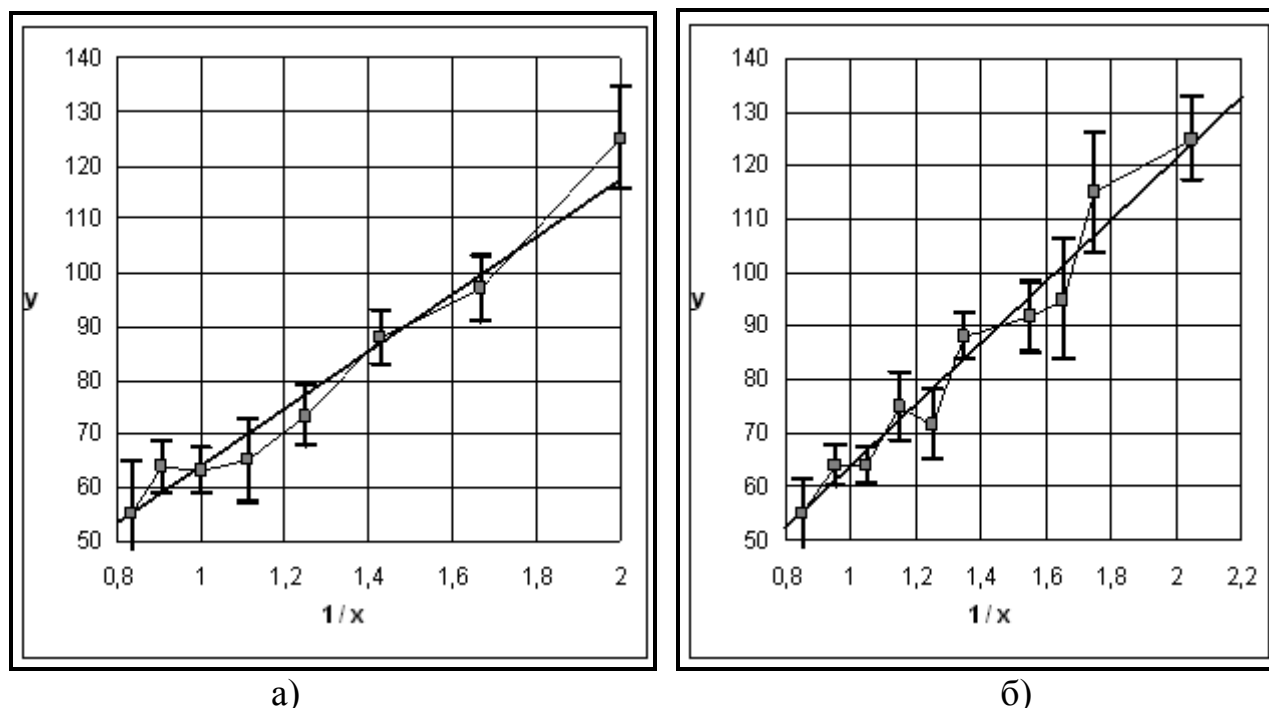


Рис. Л8.3. Графики эмпирических линий регрессии, построенные разными способами

Мы получили три нелинейных уравнения регрессии.

По исходным данным:

$$Y_p = 9,595 + 56,384 / X, R_{xy} = 0,938.$$

По сгруппированным по $z = 1/x$ данным:

$$Y_p = 6,536 + 57,414 / X, R_{xy} = 0,939.$$

По сгруппированным по x данным с заменой центров интервалов на $z = 1/x$:

$$Y_p = 10,773 + 53,194 / X, R_{xy} = 0,922.$$

Две последних регрессионных модели эквивалентные, так как их графики пересекают доверительные интервалы всех узлов эмпирических линий регрессии на рис. Л8.3а, б.

На рис. Л8.4 найденная нелинейная зависимость изображена в привычных исходных масштабах (x, y).

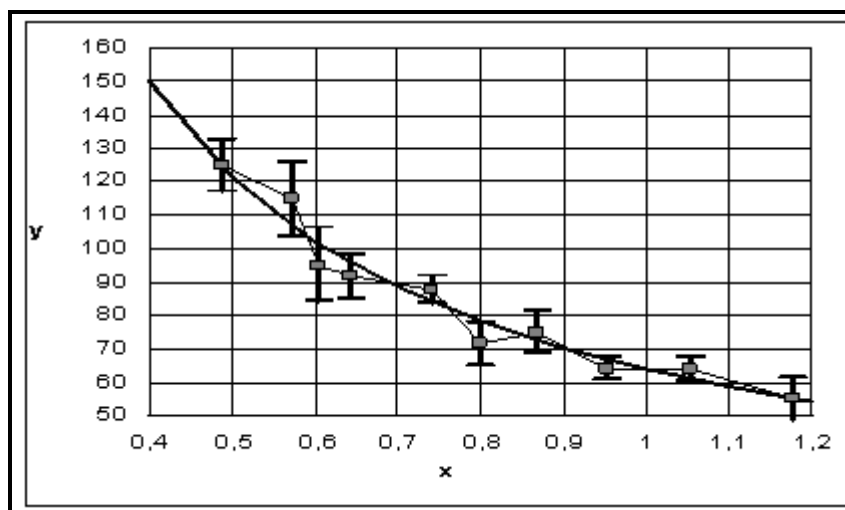


Рис. Л8.4. Графики нелинейной зависимости в исходных масштабах

Доверительные интервалы на расчетные значения

Ширина 95-процентного доверительного интервала на расчетное значение для *линейной* модели равна $\Delta y_p(x) = t_{0,05} \cdot s_y \cdot \sqrt{\frac{1-r_{xy}^2}{n-2}} \cdot \sqrt{1 + \frac{(x-\bar{x})^2}{s_x^2}}$.

n =43
Zcp =1,1889
Sz =0,2921
Sy =17,561
Rzy =0,9380
b =56,384
a =9,5952

Рис. Л8.5. Расчет по преобразованным данным

У нас линейной оказалась модель $Y_p = a + b \cdot z$, после функционального преобразования $z = 1/x$. Выписываем все необходимые сведения (рис. Л8.5).

Находим $\Delta_{\min} = t_{0,05} \cdot s_y \cdot \sqrt{\frac{1-r_{xy}^2}{n-2}}$, где $t_{0,05} = 2,02$;
 $\Delta_{\min} = 2,020 \cdot 17,561 \cdot \sqrt{\frac{1-0,938^2}{43-2}} = 1,920$.

Далее для $z = 0,8 - 1,2$ с шагом $\Delta z = 0,2$ вычисляем:
 $x = 1/z$, $HCP = 1,92 \cdot \sqrt{1 + \frac{(z-1,189)^2}{0,292^2}}$, $Y_p(z) = a + b \cdot z$, $Y_p(z) -$

HCP , $Y_p(z) + HCP$ (рис. Л8.6).

x	1,250	1,000	0,833	0,714	0,625	0,556	0,500	0,455
z = 1/x	0,8	1	1,2	1,4	1,6	1,8	2	2,2
HCP	3,197	2,287	1,922	2,369	3,315	4,452	5,667	6,918
Yp	54,702	65,979	77,256	88,533	99,810	111,087	122,363	133,640
Yp – HCP	51,506	63,693	75,334	86,164	96,495	106,634	116,696	126,722
Yp + HCP	57,899	68,266	79,178	90,902	103,125	115,539	128,030	140,558

Рис. Л8.6. Расчет границ 95-процентной доверительной полосы на линию регрессии

Границы доверительной полосы $Y_p(Z) - HCP$, $Y_p(Z) + HCP$ наносим на график линейной зависимости пунктиром (рис. Л8.7а).

Наносим эти же границы (ординаты) для соответствующих значений x на график нелинейной зависимости (рис. Л8.7б).

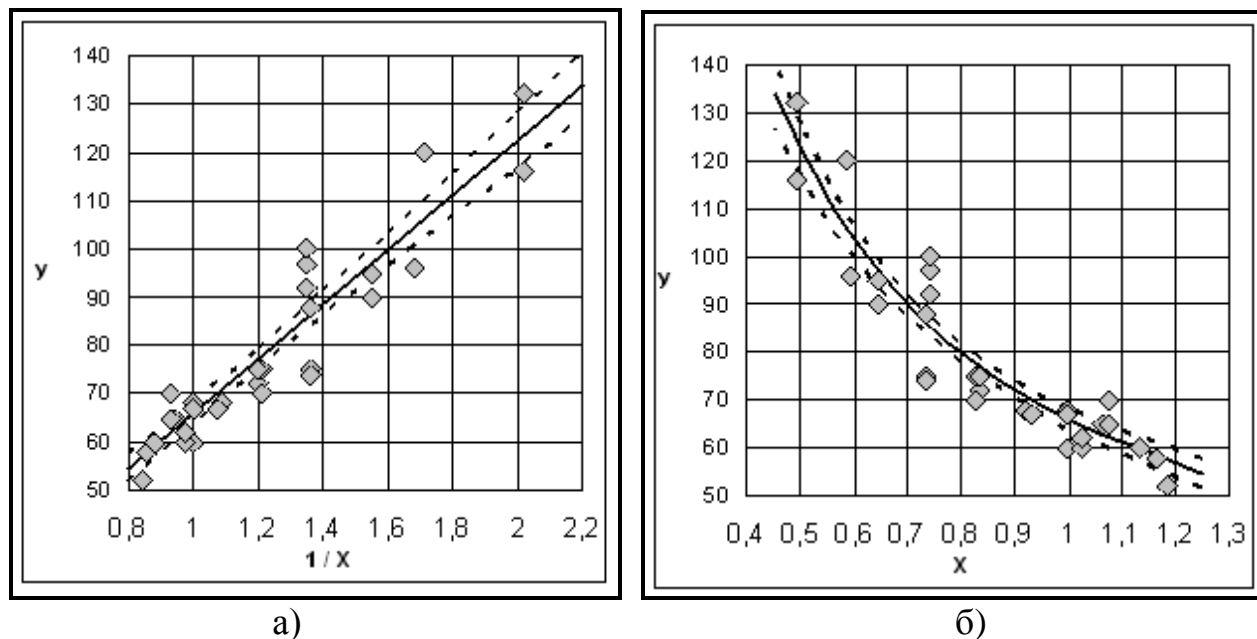


Рис. Л8.7. Графики линий регрессии с 95-процентной доверительной полосой

Наличие на графике регрессии 95-процентной доверительной полосы определяет пределы применимости модели.

Таблицы сопряженности и коэффициенты контингенции

Корреляционное поле очень похоже на таблицу сопряженности категорий двух качественных показателей (назовем их A и B , чтобы не было ассоциаций с количественными переменными X и Y).

Мы уже один раз понизили шкалу измерения переменных до дискретной и получили таблицу частот m_{ij} совместного появления разных комбинаций дискретных значений (X_i , Y_j).

Далее мы вообще абстрагировались от числовых значений переменной X (то есть понизили шкалу этой переменной до шкалы имен) и получили более объективную меру тесноты связи – корреляционное отношение вместо коэффициента корреляции ($r_{xy} = -0,868$; $\eta_{y/x} = 0,942$).

Сделаем еще один шаг: абстрагируемся также от числовых значений переменной Y и будем считать, что задано $p = 8$ категорий качественного показателя

теля A и $q = 7$ категорий качественного показателя B (категории B_2 и B_4 ни разу не появились, поэтому их сократили).

Таблица частот на рис. Л8.8 теперь называется "таблицей сопряженности" качественных показателей A и B .

Нуль-гипотеза заключается в утверждении о независимости показателей A и B . Но тогда, согласно теореме умножения вероятностей, ожидаемые частоты совместного появления категорий (A_i, B_j)

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	l
B_1	1	0	0	0	0	0	0	0	1
B_3	1	1	0	0	0	0	0	0	2
B_5	0	3	4	0	0	0	0	0	7
B_6	0	1	1	0	0	0	0	0	2
B_7	0	0	2	5	0	0	0	0	7
B_8	0	0	0	1	3	8	7	0	19
B_9	0	0	0	0	0	2	1	2	5
k	2	5	7	6	3	10	8	2	43

Рис. Л8.8. Таблица сопряженности A и B

должны быть равны $\tilde{m}_{ij} = \frac{k_i l_j}{n}$. Составляем статистику Пирсона для сравнения двух рядов частот – наблюдаемых и ожидаемых при справедливости нуль-гипотезы:

$$\chi^2 = \sum \sum \frac{(m_{ij} - \tilde{m}_{ij})^2}{\tilde{m}_{ij}};$$

$$\chi^2 = n \cdot \left(\sum \sum \frac{m_{ij}^2}{k_i l_j} - 1 \right).$$

На рис. Л8.9 вычислены отношения $\frac{m_{ij}^2}{k_i l_j}$, их сумма равна 3,613, откуда получаем $\chi^2 = 43 \cdot (3,613 - 1) = 112,36$.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
B_1	0,500	0	0	0	0	0	0	0
B_3	0,250	0,100	0	0	0	0	0	0
B_5	0	0,257	0,327	0	0	0	0	0
B_6	0	0,100	0,071	0	0	0	0	0
B_7	0	0	0,082	0,595	0	0	0	0
B_8	0	0	0	0,009	0,158	0,337	0,322	0
B_9	0	0	0	0	0	0,080	0,025	0,400

Рис. Л8.9. Вычисление статистики Пирсона

Эту статистику надо сравнивать с табличными значениями критерия Пирсона при $ЧСС = (p - 1)(q - 1)$.

Для данного примера $ЧСС = (8 - 1)(7 - 1) = 42$.

Функцией ХИ2ОБР находим $\chi_{0,05}^2 = 58,12$; $\chi_{0,01}^2 = 66,21$.

Так как вычисленное значение χ^2 больше табличного $\chi_{0,01}^2$, нуль-гипотеза отвергается и делается заключение о существовании значимой связи между показателями A, B .

Теснота этой связи оценивается с помощью коэффициентов контингенции Крамера (C) и Кендала (K, KK):

$$C = \sqrt{\frac{\chi^2}{n \cdot (d-1)}} = \sqrt{\frac{112,36}{43 \cdot (7-1)}} = 0,660,$$

где $d = \min\{p, q\} = 7$;

$$K = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{112,36}{112,36 + 43}} = 0,850;$$

$$KK = K \cdot \sqrt{\frac{d}{d-1}} = 0,919.$$

Заметим, что коэффициент контингенции Кендала $KK = 0,919$, не использующий никаких сведений о числовой природе переменных, оказался очень близким к наиболее объективной мере тесноты связи – корреляционному отношению $\eta_{y/x} = 0,942$.

Вопросы для самопроверки

1. Дайте определение функциональной, статистической и корреляционной зависимостей. Продемонстрируйте различия между сопряженными корреляционными моделями. Приведите пример статистической, но не корреляционной зависимости.

2. Сформулируйте идею принципа Лежандра (МНК), разъясните смысл системы нормальных уравнений, составьте систему нормальных уравнений для линейной и квадратичной моделей с одной объясняющей переменной.

3. Сформулируйте основные предпосылки дисперсионного анализа. Докажите, что средние по группам являются наилучшими МНК-оценками центров каждой группы. Разложите общую сумму квадратов на межгрупповую и внутригрупповую составляющие.

4. Опишите методику сравнения двух выборок по критерию Стьюдента. Сформулируйте основные предпосылки (гипотезы) этого метода. Покажите, что этот анализ является частным случаем дисперсионного анализа, когда количество сравниваемых групп равно двум.

5. Покажите, как строится эмпирическая линия регрессии, как оценивается теснота корреляционной связи. Поясните, что такое индекс детерминации и корреляционное отношение, чем они отличаются от коэффициента детерминации и коэффициента корреляции соответственно.

6. Изложите последовательность расчетов для оценки значимости корреляционной связи. Опишите таблицу дисперсионного анализа, разъясните смысл ее отдельных граф (столбцов) — сумм квадратов, чисел степеней свободы, средних квадратов. Поясните, какой смысл имеет дисперсионное отношение Фишера, что такое уровень значимости и как им пользоваться.

7. Изложите последовательность расчетов для оценки значимости регрессионной модели. Опишите таблицу дисперсионного анализа, разъясните смысл ее отдельных граф. Выразите для этой проблемы дисперсионное отношение через коэффициент детерминации.

8. Опишите методику оценки значимости коэффициента регрессии и коэффициента парной корреляции по критерию Стьюдента. Докажите, что эта методика является частным случаем дисперсионного анализа для оценки значимости линейной одномерной модели.

9. Изложите последовательность расчетов для оценки адекватности модели. Опишите таблицу дисперсионного анализа, разъясните смысл ее отдельных граф. Покажите, в чем разница между оценкой дисперсии остатка модели и дисперсией случайной ошибки.

10. Выведите формулы для расчета параметров парной линейной регрессии. Дайте определение коэффициента парной корреляции, перечислите его свойства. Поясните, что такое коэффициент детерминации, чем он отличается от индекса детерминации.

11. Перечислите основные предпосылки регрессионного анализа. Сформулируйте идею принципа максимального правдоподобия и покажите, что по этому принципу наилучшими оценками параметров модели будут МНК-оценки.

12. Сформулируйте идею расчета дисперсий коэффициентов регрессии и дисперсий расчетных значений. Опишите графический способ построения 95-процентной доверительной полосы на линию регрессии.

13. Поясните способ выбора формы связи. Рассмотрите стандартные преобразования переменных (логарифмирование и переход к обратным величинам).

Теория вероятностей в вопросах и ответах

1. Что такое вероятность?

Вероятность – это число, которое показывает, как часто происходит событие (А) при испытаниях. Это число изменяется в пределах $0 \leq p_A \leq 1$. Если $p_A = 0$, событие А является невозможным, невероятным, оно никогда не происходит, сколько бы не повторять испытания. Если, наоборот, $p_A = 1$, то событие А обязательно произойдет в каждом испытании, иными словами, такое событие не является случайным, его называют детерминированным, достоверным.

2. Какие известны способы определения вероятности?

Известны 4 способа определения вероятности – в смысле 4 способа вычисления вероятности – статистический, геометрический, классический, экспертный.

При *статистическом* (или *стохастическом*) способе производят n испытаний и фиксируют, сколько раз при этом появилось событие А; число появления события называется частотой m . Относительная частота $\frac{m}{n}$ (частость) изменяется от 0 до 1 и показывает, как часто появлялось событие при n испытаниях.

Ожидается, что с увеличением числа испытаний это отношение приближается к некому пределу – вероятности события А:

$$p_A = \lim_{n \rightarrow \infty} \frac{m(n)}{n}.$$

При *геометрическом* способе область всех возможных исходов Ω и область исходов, при которых появляется событие А, пытаются изобразить в виде геометрических фигур. Все точки этих фигур считаются равновероятными.

Тогда вероятность события А можно вычислить как отношение площадей указанных фигур:

$$p_A = \frac{S_A}{S_{\Omega}}.$$

Классический способ применим, если исходы испытания представлены набором элементарных исходов. *Элементарные исходы* равновероятные, несовместные и составляют полную группу.

Тогда вероятность события A равна отношению числа элементарных исходов, при которых появляется событие A (m), к общему числу элементарных исходов (n):

$$P_A = \frac{m_A}{n}.$$

Когда невозможен ни один из вышеупомянутых способов, применяют способ *экспертных оценок*. Группа экспертов обсуждает вероятности неких начальных простых событий, а вероятности более сложных последствий уже рассчитываются на основе известных теорем теории вероятностей.

3. Какие бывают события? Приведите их краткую классификацию.

Во-первых, события подразделяются на *совместные* и *несовместные*. Несовместные события не могут появиться одновременно в одном испытании. Во-вторых, совместные события подразделяются на *зависимые* и *независимые*. Вероятности независимых событий не зависят от того, появилось или не появилось перед этим другое событие. Наконец, независимые события подразделяются на *однородные* и *неоднородные*. Вероятности однородных событий не зависят также от номера испытания (они постоянны).

С вероятностной точки зрения, события бывают *невозможными* ($p_A = 0$), *детерминированными* (достоверными, $p_A = 1$) и *случайными* ($0 < p_A < 1$).

4. Что такое полная группа событий?

События составляют полную группу, если при испытании одно из них обязательно произойдет. Если события составляют полную группу *несовместных* событий, то сумма их вероятностей равна единице. *Противоположные* события составляют полную группу несовместных событий, поэтому всегда $p_A + p_{\bar{A}} = 1$.

5. Приведите краткую классификацию испытаний.

Существует две принципиально разные схемы испытаний – повторения испытаний заданное число раз (n) и повторения испытаний до появления первого успеха (до первого появления события A). Обе схемы испытаний предложены Бернулли. Для краткости в отечественной научной литературе часто используют словосочетания: «задача Бернулли», или «задача о повторении однородных независимых испытаний». Эти формулировки надо понимать так: производятся

испытания заданное число раз (n), в каждом из испытаний событие A может появиться с вероятностью (p), которая не зависит ни от номера испытания, ни от того, сколько раз появилось событие A до этого испытания; требуется найти вероятность появления m успехов $P_n(m)$. Если же применяется схема испытаний до первого успеха, никакие сомнительные сокращенные словосочетания не используются, задача формулируется подробно и корректно.

6. Сформулируйте аксиому и теорему сложения вероятностей.

Аксиома сложения. Вероятность появления одного из *несовместных* событий равна сумме вероятностей указанных событий: $p_{A+B} = p_A + p_B$, если $AB = \emptyset$.

Теорема сложения. Вероятность появления одного из двух событий равна сумме вероятностей указанных событий минус вероятность их совместного появления:

$$p_{A+B} = p_A + p_B - p_{AB}.$$

Вероятность появления одного из трех событий равна сумме вероятностей указанных событий, минус вероятности совместного появления каждой пары событий плюс вероятность совместного появления всех трех событий:

$$p_{A+B+C} = (p_A + p_B + p_C) - (p_{AB} + p_{AC} + p_{BC}) + p_{ABC}.$$

Формулировка теоремы усложняется с увеличением числа событий.

7. Сформулируйте теорему умножения вероятностей.

Теорема умножения для независимых событий. Вероятность совместного появления нескольких *независимых* событий равна произведению их вероятностей:

$$p(AB) = p(A) \cdot p(B).$$

Теорема умножения в общем виде. Вероятность совместного появления двух событий равна произведению вероятности одного из них на условную вероятность другого:

$$p(AB) = p(A) \cdot p(B|A) \text{ или } p(AB) = p(B) \cdot p(A|B).$$

8. Сформулируйте теорему о полной вероятности

В рассматриваемой задаче событие A появляется совместно с одним из событий H_i , которые составляют полную группу несовместных событий и называются гипотезами. Даны вероятности гипотез $p(H_i)$ и условные вероятности появления события A в присутствии каждой гипотезы $p(A|H_i)$.

Требуется найти вероятность события A :

$$p(A) = p(H_1)p(A|H_1) + p(H_2)p(A|H_2) + p(H_3)p(A|H_3).$$

9. Сформулируйте теорему Байеса.

Теорема (формула) Байеса оценивает относительные вклады каждого члена в формуле полной вероятности:

$$\frac{p_{H_1} p_{A|H_1}}{p_A} = p_{H_1|A}, \quad \frac{p_{H_2} p_{A|H_2}}{p_A} = p_{H_2|A}, \quad \frac{p_{H_3} p_{A|H_3}}{p_A} = p_{H_3|A}.$$

10. Поясните, что такое дискретная случайная величина, как задается закон ее распределения.

Понятие «случайная величина» является исходным неопределяемым понятием, которое (как и понятие «случайное событие») демонстрируется только на примерах. Если все возможные значения случайной величины можно заранее перечислить, то такую случайную величину называют *дискретной*. У *непрерывной* случайной величины значения заполняют сплошь некоторый интервал. Соответствие между отдельными значениями случайной величины и вероятностью их появления называют *законом распределения* данной случайной величины. Закон распределения дискретной случайной величины может быть задан в табличной форме в виде ряда распределения, где каждому возможному значению случайной величины поставлена в соответствие вероятность его появления. Сумма этих вероятностей равна единице, так как все возможные значения случайной величины составляют полную группу несовместных событий.

11. Что такое математическое ожидание, как оно вычисляется?

Математическое ожидание – это среднее значение случайной величины; центр, вокруг которого группируются отдельные значения случайной величины; среднее взвешенное, где «весами» являются вероятности появления отдельных значений случайной величины; центр тяжести полигона распределения (для дискретной величины) или центр тяжести фигуры, ограниченной графиком плотности вероятности (для непрерывной величины).

Для дискретной случайной величины математическое ожидание вычисляется как сумма произведений значений случайной величины на их вероятности:

$$a = M(x) = \sum_{i=1}^k x_i p_i, \text{ где } \sum_{i=1}^k p_i = 1.$$

Для непрерывной случайной величины дискретная сумма заменяется на интеграл:

$$a = M(x) = \int_{-\infty}^{\infty} x f(x) dx, \text{ где } \int_{-\infty}^{\infty} f(x) dx = 1.$$

Здесь $f(x)$ – дифференциальная функция распределения (плотность вероятности).

12. Что такое дисперсия, как она вычисляется?

Дисперсия – мера разброса, мера изменчивости, мера рассеяния значений случайной величины вокруг центра – математического ожидания. Вычисляется по формуле:

$$D(x) = M(x - a)^2 = M(x^2) - M^2(x).$$

Дисперсия постоянной равна нулю.

13. Что такое среднее квадратичное отклонение?

Корень квадратный из дисперсии называется *средним квадратичным отклонением*, или *стандартным отклонением* (но не стандартной ошибкой!). Обозначение:

$$\sigma_x = \sqrt{D(x)}.$$

Размерность дисперсии равна квадрату размерности случайной величины. Это не удобно. Хотелось бы в качестве меры изменчивости ввести средний размер отклонений, но характеристика $M(x - a) \equiv 0$ оказалась равной нулю из-за разных знаков отклонений (так называемое *нулевое*, или *центральное*, *свойство* математического ожидания). Именно поэтому дисперсия введена как среднее квадратов отклонений. Стандартное отклонение σ_x действительно есть среднее отклонение, но не среднее арифметическое, а среднее квадратичное, что и отражено в его названии. На практике несколько длинное название *среднее квадратичное отклонение* постепенно заменили на более энергичное *стандартное отклонение* и, к сожалению, на стандартную ошибку, а это уже не правильно. В зарубежной литературе четко разделяют два понятия: Standard Deviation

(это σ_x) и Standard Error ($\frac{\sigma_x}{\sqrt{n}}$ – совсем другая характеристика). В отечественной научной литературе характеристика $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$ называется ошибкой среднего.

14. Перечислите свойства математического ожидания.

Свойства математического ожидания совпадают со свойствами обычного среднего:

1. Математическое ожидания постоянной равно этой постоянной	1. Среднее постоянной равно этой постоянной
2. Постоянный множитель можно вынести за знак математического ожидания	2. Постоянный множитель можно вынести за знак среднего
3. Математическое ожидание суммы случайных величин равно сумме их математических ожиданий	3. Среднее значение суммы величин равно сумме их средних
4. Математическое ожидание случайных отклонений равно нулю: $M(x - a) \equiv 0$	4. Сумма отклонений от среднего равна нулю: $\sum (x_i - \bar{x}) = 0$

15. Перечислите свойства дисперсии.

1. Дисперсия постоянной равна нулю.
2. При вынесении постоянного множителя за знак дисперсии множитель возводится в квадрат.
3. Дисперсия суммы *независимых* случайных величин равна сумме их дисперсий.
4. Дисперсия суммы двух случайных величин равна сумме их дисперсий плюс удвоенная ковариация: $D(x + y) = D(x) + D(y) + 2\sigma_{xy}$. Ковариация – смешанный центральный момент $\sigma_{xy} = M(x - a)(y - b)$, где $a = M(x)$, $b = M(y)$. Для независимых случайных величин ковариация равна нулю.

16. Перечислите характеристики случайных величин. Что такое коэффициент асимметрии и коэффициент эксцесса?

Для описания положения, рассеяния и формы распределения используют так называемые *моменты распределения*: $m_1 = M(x)$, $m_2 = M(x^2)$, $m_3 = M(x^3)$, $m_4 = M(x^4)$. Заметим, что момент 1-го порядка есть основная характеристика положения – математическое ожидание $m_1 = M(x) = a$. Кроме этого, введены центральные моменты распределения: $\mu_1 = M(x - a)$, $\mu_2 = M(x - a)^2$, $\mu_3 = M(x - a)^3$, $\mu_4 = M(x - a)^4$. Первый центральный момент всегда равен нулю (центральное

свойство математического ожидания); второй центральный момент есть основная характеристика разброса – дисперсия $\mu_2 = M(x - a)^2 = (\sigma_x)^2$. Моменты 2-го и 4-го порядков имеют размерности куба и четвертой степени от размера исходной величины, поэтому применяют еще безразмерные нормированные или стандартизированные моменты распределения: $\rho_3 = \frac{\mu_3}{\sigma_x^3} = A$, $\rho_4 = \frac{\mu_4}{\sigma_x^4}$. Нормированный момент 3-го порядка называется коэффициентом асимметрии; если $A = 0$ – распределение симметричное, при $A > 0$ – скошено влево, при $A < 0$ – скошено вправо. Нормированный момент 4-го порядка называется коэффициентом плосковершинности, или игольчатости, или же коэффициентом эксцесса. В отечественной литературе принято вычитать тройку из 4-го нормированного момента $E = \rho_4 - 3$, тогда $E = 0$ соответствует стандартная «нормальная» форма распределения, $E < 0$ – плосковершинная, $E > 0$ – игольчатая форма распределения. Наличие игольчатости трактуется как результат смеси двух случайных величин с одинаковыми центрами, но с разными дисперсиями (например, смесь продукции мастера и ученика).

17. Сформулируйте правило «3-х сигм».

Случайные отклонения, большие 3-х σ_x , маловероятны. На практике такие отклонения обычно считаются выбросами и появление их объясняют наличием грубых ошибок в записи чисел. Правило «3-х сигм» основано на неравенстве Чебышева:

$$P(|x - a| \leq t\sigma_x) > 1 - \frac{1}{t^2},$$

откуда при $t = 3$ следует, что $P(|x - a| \leq 3\sigma_x) > 1 - \frac{1}{9} > 0,89$. Иными словами, с уровнем доверия, не меньшим 90 %, можно утверждать, что случайные отклонения $|x - a|$ не превышают $3\sigma_x$.

18. Сформулируйте задачу о повторении однородных независимых испытаний, приведите ее другие названия.

Производится n испытаний, в каждом из которых событие A может появиться с вероятностью p , которая не зависит ни от номера испытания, ни от того, сколько раз появилось событие A до этого испытания. Тогда вероятность появления m успехов $P_n(m)$ определяется по формуле Бернулли:

$$P_n(m) = C_n^m p^m (1-p)^{n-m} = \frac{n!}{m!(n-m)!} p^m q^{n-m}.$$

Другие названия этой проблемы – задача Бернулли, распределение Бернулли, биномиальный закон распределения. При увеличении числа испытаний распределение Бернулли приближается к некой стандартной форме – распределению Лапласа или к нормальному закону Гаусса. Характеристики распределения Бернулли: $M(m) = np$, $D(m) = npq$.

19. Приведите асимптотическую формулу Пуассона и укажите область применения этого распределения. Что такое рекуррентная формула?

Для $n > 30$ производить расчеты по формуле Бернулли становится затруднительным из-за слишком больших величин факториалов, поэтому используют асимптотические формулы Пуассона и Лапласа, которые становятся все более и более точными с увеличением n (именно в тех случаях, когда расчеты по исходной формуле Бернулли практически невозможны).

Формула Пуассона применяется для больших $n > 30$ и малых $p < 0,05$, таких, что $np < 5$ (поэтому распределение Пуассона применяется для изучения распределения числа редких событий). Если обозначить $a = np$, то формула Пуассона приобретает вид:

$$P(m) = e^{-a} \cdot \frac{a^m}{m!}.$$

Характеристики распределения Пуассона: $M(m) = a$, $D(m) = a$.

Вычисления вероятностей $P(m)$ распределения Пуассона удобно производить по рекуррентной (возвратной) формуле:

$$P(m) = P(m-1) \cdot \frac{a}{m}, \text{ где } P(0) = e^{-a}.$$

При увеличении параметра a распределение Пуассона приближается к некой стандартной форме – распределению Лапласа (или к нормальному распределению Гаусса).

20. Сформулируйте локальную и интегральную теоремы Лапласа, сравните распределение Лапласа с нормальным законом Гаусса.

Для $n > 30$, $np \geq 5$, $nq \geq 5$ распределение Бернулли практически точно аппроксимируется асимптотической формулой Лапласа:

$$P_n(m) = \frac{1}{\sqrt{2\pi} \cdot \sigma_m} \cdot e^{-\frac{(m-a)^2}{2\sigma_m^2}} = \frac{\varphi(t_m)}{\sigma_m},$$

где $t_m = \frac{m-a}{\sigma_m}$, $a = M(m) = np$, $\sigma_m = \sqrt{npq}$, $\varphi(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}$.

Эта аппроксимация называется *локальной теоремой Лапласа*.

Вероятность попадания случайной величины m в интервал $[m_1, m_2]$ можно записать как разность значений интегральной функции Лапласа на краях этого интервала:

$$P(m_1 \leq m \leq m_2) = \Phi(t_{m_2}) - \Phi(t_{m_1}),$$

где $\Phi(t) = \int_0^t \varphi(s) ds = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-\frac{s^2}{2}} ds$.

Это утверждение называется *интегральной теоремой Лапласа*.

Для нормального закона распределения Гаусса имеем похожие выражения:

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \cdot e^{-\frac{(x-a)^2}{2\sigma_x^2}} = \frac{\varphi(t_x)}{\sigma_x},$$

где $t_x = \frac{x-a}{\sigma_x}$, $a = M(x)$, $\varphi(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}$;

$$P(x_1 \leq m \leq x_2) = \Phi(t_{x_2}) - \Phi(t_{x_1}),$$

где $\Phi(t) = \int_0^t \varphi(s) ds = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-\frac{s^2}{2}} ds$.

Сходство очень большое, но есть и различия. Распределение Лапласа дискретное и $P_n(m) = \frac{\varphi(t_m)}{\sigma_m}$ – вероятность; нормальное распределение непрерывное и $f(x) = \frac{\varphi(t_x)}{\sigma_x}$ – плотность вероятности. В распределении Лапласа характеристики a и σ_m связаны между собой, причем $0 < \sigma_m < a$; в нормальном распределении a может быть каким угодно, хоть нулем, хоть отрицательным.

21. Перечислите особенности дифференциальной и интегральной функций Лапласа. Является ли интегральная функция Лапласа интегральной функцией распределения Лапласа?

Дифференциальная функция Лапласа $\varphi(t)$ четная $\varphi(-t) = \varphi(t)$, поэтому таблицы этой функции составлены только для неотрицательных значений аргумента $t \geq 0$.

Функция $\varphi(t)$ строго положительная и имеет горизонтальную асимптоту – ось абсцисс, поэтому таблицы составлены только для $0 \leq t \leq 4$, для больших значений аргумента функция практически равна нулю. Единственный экстремум – максимум $\varphi_{\max} = \varphi(0)$ при $t = 0$.

Площадь под дифференциальной кривой равна единице, площадь фигуры на интервале $[-3; 3]$ равна 0,997; на интервале $[-2; 2]$ – 0,954.

Интегральная функция Лапласа $\Phi(t)$ нечетная $\Phi(-t) = -\Phi(t)$, поэтому таблицы этой функции составлены только для неотрицательных значений аргумента $t \geq 0$.

Функция $\Phi(t)$ строго возрастает и имеет горизонтальную асимптоту – $\Phi_{\max} = 0,5$, поэтому таблицы составлены только для $0 \leq t \leq 4$, для больших значений аргумента функция практически равна 0,5.

По определению интегральная функция *распределения* есть вероятность того, что случайная величина примет значение не меньше заданного: $F(m) = P(\mathcal{X} \leq m)$.

Согласно интегральной теореме Лапласа имеем $F(m) = \Phi(t_m) + 0,5$; то есть интегральная функция распределения Лапласа $F(m)$ отличается от интегральной функции Лапласа $\Phi(t_m)$ постоянным слагаемым 0,5.

22. Сформулируйте три формы интегральной теоремы Лапласа.

В качестве первой формы теоремы примем общую формулу для вычисления вероятности попадания случайной величины в заданные интервалы:

$$P(m_1 \leq m \leq m_2) = \Phi(t_{m_2}) - \Phi(t_{m_1}).$$

Вторая форма предназначена для вычисления вероятности попадания случайной величины в интервалы, симметричные относительно центра:

$$P(|m - a| \leq t \cdot \sigma_m) = 2\Phi(t);$$

$$P(|m - np| \leq t \sqrt{npq}) = 2\Phi(t).$$

Третью форму интегральной теоремы получим, разделив на n обе части неравенства $|m - np| \leq t\sqrt{npq}$:

$$P\left(\left|\frac{m}{n} - p\right| \leq t\sqrt{\frac{pq}{n}}\right) = 2\Phi(t);$$

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

23. Перечислите задачи, которые решаются с помощью третьей формы интегральной теоремы Лапласа.

1. Заданы параметры распределения n, p и погрешность утверждения:

$$\left|\frac{m}{n} - p\right| \leq \varepsilon.$$

Найти уровень доверия P этого утверждения.

2. Заданы параметры распределения n, p и уровень доверия P утверждения:

$$\left|\frac{m}{n} - p\right| \leq \varepsilon.$$

Найти погрешность ε .

3. Заданы параметр p , уровень доверия P и погрешность ε утверждения:

$$\left|\frac{m}{n} - p\right| \leq \varepsilon.$$

При каких значениях n это утверждение выполняется с заданным уровнем доверия?

24. Как по результатам статистических испытаний делаются заключения о величине параметров распределения?

Это еще одна задача на применение третьей формы интегральной теоремы Лапласа. Заданы m, n, P . Какие значения параметра p согласуются с утверждением:

$$P\left(\left|\frac{m}{n} - p\right| \leq t\sqrt{\frac{pq}{n}}\right) = 2\Phi(t) = P ?$$

Из соотношения $2\Phi(t) = P$ находят t и получают квадратичное неравенство:

$$\left| \frac{m}{n} - p \right|^2 \leq t^2 \frac{p \cdot (1-p)}{n},$$

откуда определяются границы $p_1 \leq p \leq p_2$ допустимых значений параметра p .

25. Что такое функция распределения, каковы ее свойства и график, каковы ее особенности для дискретной случайной величины?

Функция распределения (или *интегральная функция распределения*) – вероятность того, что случайная величина \mathcal{X} примет значение, не меньшее заданного x :

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

Для дискретной случайной величины эта функция называется кумулятой и представляет собой функцию накопленных вероятностей $F(x_j) = p_1 + p_2 + \dots + p_j$. Между соседними значениями дискретной величины $x_i \leq x < x_{i+1}$ кумулята сохраняет постоянное значение $F(x_i)$. График кумуляты ступенчатый.

Функция распределения непрерывной случайной величины непрерывная.

Функция распределения неубывающая, большим значениям случайной величины соответствуют не меньшие значения функции распределения: $(x_1 < x_2) \rightarrow (F(x_1) \leq F(x_2))$.

Функция распределения изменяется от 0 до 1: $F(-\infty) = 0, F(\infty) = 1$.

Интегральная теорема. Вероятность попадания случайной величины в полуоткрытый интервал $x_1 < x \leq x_2$ равна разности значений интегральной функции распределения на краях этого интервала: $P(x_1 < x \leq x_2) = F(x_2) - F(x_1)$.

Для непрерывной случайной величины можно также записать $P(x_1 \leq x \leq x_2) = F(x_2) - F(x_1)$.

26. Что такое функция плотности вероятности, каковы ее свойства и график?

Плотность вероятности – это отношение вероятности попадания случайной величины в малый интервал к длине этого интервала:

$$f(x) = \lim_{\Delta x \rightarrow 0} \frac{P(x \leq X \leq x + \Delta x)}{\Delta x} = F'(x).$$

Функция плотности вероятности $f(x)$ оказалась равной производной от функции распределения $F(x)$, поэтому функцию $f(x)$ называют также *дифферен-*

циальной функцией распределения. Наоборот, функция распределения $F(x)$ выражается как интеграл от функции плотности вероятности

$$F(x) = \int_{-\infty}^x f(s) ds ,$$

поэтому функцию $F(x)$ называют также *интегральной функцией распределения*.

Функция плотности вероятности неотрицательная $f(x) \geq 0$.

Площадь под дифференциальной кривой равна единице. Площадь под частью дифференциальной кривой на интервале (x_1, x_2) равна вероятности попадания случайной величины в этот интервал $P(x_1 \leq x \leq x_2)$.

27. Как вычисляются основные характеристики непрерывной случайной величины?

Поскольку почти все основные характеристики случайной величины выражаются через оператор математического ожидания, достаточно вывести формулу для расчета математического ожидания непрерывной случайной величины:

$$M(x) = \sum_x x \cdot p(x) = \lim_{\Delta x \rightarrow 0} \sum x \cdot P(x \leq x \leq x + \Delta x) = \lim_{\Delta x \rightarrow 0} \sum x \cdot f(x) \Delta x = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Остальные моменты распределения рассчитываются аналогично, например:

$$M(x^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x) dx .$$

28. Что такое квантили? Перечислите их разновидности.

Для непрерывной случайной величины в качестве дополнительных характеристик используются так называемые *квантили*, к которым относятся медиана, квартили, децили и процентиля. Квантили делят фигуру под дифференциальной кривой на равновеликие части, или же они делят интервал изменения (варьирования) случайной величины на равновероятные части.

Медиана делит интервал варьирования на две части с вероятностью 50 % попадания случайной величины в каждую часть.

Квартили делят интервал варьирования на четыре части с вероятностью 25 % попадания случайной величины в каждую часть.

Децили делят интервал варьирования на десять частей с вероятностью 10 % попадания случайной величины в каждую часть.

Проценти делят интервал варьирования на сто частей с вероятностью 1 % попадания случайной величины в каждую часть.

Обозначения квантилей x_α ,: $P(x > x_\alpha) = \alpha$ (где α – вероятность того, что случайная величина примет значение, большее квантиля x_α ; площадь фигуры плотности вероятности *справа* от квантиля). Выражение для вероятности противоположного события $P(x \leq x_\alpha) = F(x_\alpha) = 1 - \alpha$ приводит к вычислительной формуле $F(x_\alpha) = 1 - \alpha$.

29. Что такое нормальный закон распределения Гаусса, каковы его характерные особенности?

Плотность вероятности и функция распределения нормального закона выражаются через дифференциальную и интегральную функции Лапласа:

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \cdot e^{-\frac{(x-a)^2}{2\sigma_x^2}} = \frac{\varphi(t_x)}{\sigma_x}, \text{ где } t_x = \frac{x-a}{\sigma_x}, \varphi(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}};$$

$$F(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \int_{-\infty}^x e^{-\frac{(s-a)^2}{2\sigma_x^2}} ds = \Phi(t_x) + 0,5, \text{ где } \Phi(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-\frac{s^2}{2}} ds.$$

Параметры нормального закона a и σ_x совпадают с его характеристиками: $a = M(x)$ – математическое ожидание, σ_x – стандартное отклонение. Коэффициенты асимметрии и эксцесса для нормального закона равны нулю.

Особенности нормального закона – одномодальность (одновершинность), симметрия и правило «2-х сигм»: с вероятностью 95 % отклонения от центра не превышают $2\sigma_x$ (только в 5-ти случаях из 100 отклонения $|x - a|$ могут превысить $2\sigma_x$).

30. Сформулируйте три формы интегральной теоремы нормального закона.

1. Основная форма – вероятность попадания нормально распределенной величины x в заданный интервал с границами $[x_1, x_2]$:

$$P(x_1 \leq x \leq x_2) = F(x_2) - F(x_1) = \Phi(t_{x_2}) - \Phi(t_{x_1}).$$

2. Вторая форма предназначена для вычисления вероятности попадания случайной величины в интервал с симметричными границами:

$$P(|x - a| \leq t\sigma_x) = 2\Phi(t).$$

3. Утверждение, которое называем третьей формой интегральной теоремы нормального закона, фактически является записью второй формы интегральной теоремы для случайной величины \mathcal{X}_{cp} :

$$P\left(|\bar{x} - a| \leq t \frac{\sigma_x}{\sqrt{n}}\right) = 2\Phi(t).$$

Согласно *центральной предельной теореме*, \mathcal{X}_{cp} распределено асимптотически нормально (независимо от распределения отдельных слагаемых) с характеристиками:

$$X_{cp} \sim N\left(a; \frac{\sigma_x}{\sqrt{n}}\right).$$

31. Какие задачи решаются с помощью 3-й формы интегральной теоремы нормального закона?

1. Известны параметры нормального распределения a и σ_x . Дополнительно заданы n и уровень доверия P утверждения:

$$|\bar{x} - a| \leq \varepsilon.$$

Найти погрешность ε .

2. Известны параметры нормального распределения a и σ_x . Дополнительно заданы n и погрешность ε в утверждении:

$$|\bar{x} - a| \leq \varepsilon.$$

Найти уровень доверия P этого утверждения.

3. Известны параметры нормального распределения a и σ_x . Дополнительно заданы погрешность ε и уровень доверия P утверждения:

$$|\bar{x} - a| \leq \varepsilon.$$

Найти, при каких значениях n это утверждение выполняется с заданным уровнем доверия P .

32. Сформулируйте закон равномерного распределения, опишите область применения, приведите выражения для функций распределения и его характеристик.

По закону равномерной плотности распределены ошибки округления чисел, время ожидания транспорта, который движется через равные интервалы времени, место остановки тела под воздействием сухого трения.

Плотность вероятности (дифференциальная функция распределения) постоянна на интервале $[a, b]$ и равна нулю за его пределами:

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

Интегральная функция распределения линейна на интервале $[a, b]$:

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

Характеристики распределения выражаются через его параметры a, b :

$$M(x) = \frac{a+b}{2}; \quad \sigma_x = \frac{b-a}{\sqrt{12}}.$$

33. Сформулируйте показательный (экспоненциальный) закон распределения, опишите область применения, приведите выражения для функций распределения и его характеристик.

По этому закону распределено время работы оборудования до первого отказа. Дифференциальная функция (плотность вероятности) выражается формулами:

$$f(t) = \begin{cases} 0, & t < 0 \\ \lambda e^{-\lambda t}, & t \geq 0 \end{cases}.$$

Интегральная функция распределения задается формулами:

$$F(t) = \begin{cases} 0, & t < 0 \\ 1 - e^{-\lambda t}, & t \geq 0 \end{cases}.$$

Характеристики распределения выражаются через единственный параметр λ :

$$M(t) = \frac{1}{\lambda}; \quad \sigma_t = \frac{1}{\lambda}.$$

Коэффициент вариации показательного закона равен 100 %.

Математическая статистика в вопросах и ответах

1. Перечислите основные характеристики генеральной совокупности и их выборочные оценки (центра группировки, меры изменчивости, функций распределения).

Генеральные характеристики	Выборочные оценки
p – вероятность	m / n – относительная частота
$M(x)$ – математическое ожидание, центр совокупности, генеральное среднее	$\bar{x} = X_{cp}$ – центр выборки, выборочное среднее (просто среднее)
$D(x) = \sigma_x^2$ – дисперсия, мера изменчивости	$s_x^2, \hat{\sigma}_x^2$ – оценки дисперсии; $\hat{\sigma}_x^2$ – несмещенная оценка дисперсии
$f(x)$ – дифференциальная функция распределения, плотность вероятности	Гистограмма – ступенчатый график плотности вероятности $f_i = \frac{m_i}{nh_i}$, где h_i – ширина интервала
$F(x)$ – функция распределения (интегральная функция распределения); $F(x) = P(X \leq x)$	Кумулята – график накопленных относительных частот $F_i = \frac{1}{n} \sum_{j=1}^i m_j$

2. Запишите сравнительные формулы для вычисления характеристик и их выборочных оценок (математического ожидания, дисперсии, ковариации).

Генеральные характеристики	Выборочные оценки
1	2
$M(x) = \sum x_i p_i$ – для дискретной случайной величины; $M(x) = \int_{-\infty}^{\infty} x \cdot f(x) dx$ – для непрерывной случайной величины	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ – для несгруппированных данных; $\bar{x} = \frac{1}{n} \sum_{j=1}^k m_j x_j$ – для сгруппированных данных (m_j – частоты, x_j – центры интервалов)
$\sigma_x^2 = M(x - M(x))^2$; $\sigma_x^2 = M(x^2) - M^2(x)$	$s_x^2 = \overline{(x - \bar{x})^2}$; $s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$; $s_x^2 = \overline{x^2} - (\bar{x})^2$; $s_x^2 = \frac{1}{n} \sum_{j=1}^k m_j (x_j - \bar{x})^2$

1	2
$\sigma_{xy} = M[(x - M(x)) \cdot (y - M(y))];$ $\sigma_{xy} = M(xy) - M(x) \cdot M(y).$	$s_{xy} = \overline{(x - \bar{x})(y - \bar{y})};$ $s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y});$ $s_{xy} = \overline{xy} - \bar{x} \cdot \bar{y};$ $s_{xy} = \frac{1}{n} \sum_{j=1}^n m_j (x_j - \bar{x})(y_j - \bar{y})$

3. Поясните, что означает состоятельность, несмещенность и эффективность статистических оценок. Приведите формулу для вычисления несмещенной оценки дисперсии.

Оценка b параметра β называется состоятельной, если при увеличении объема выборки ($n \rightarrow \infty$) оценка стремится к своему генеральному значению $b \rightarrow \beta$. Оценка называется несмещенной, если $M(b) = \beta$, то есть отдельные значения оценок по разным выборкам группируются вокруг генерального значения β ; несмещенная оценка не имеет систематической ошибки. Оценка называется эффективной, если по сравнению с другими оценками она имеет наименьшую дисперсию (эффективная оценка имеет наименьшую погрешность). Оценка дисперсии $s_x^2 = \overline{x^2} - (\bar{x})^2$ систематически занижена. Несмещенная оценка (исправленная на ЧСС – число степеней свободы) несколько больше: $\hat{\sigma}_x^2 = \frac{n}{\text{ЧСС}} \cdot s_x^2$, где ЧСС равно разности между числом случайных величин и числом наложенных на них линейных связей. Обычно $\text{ЧСС} = n - 1$.

4. Перечислите свойства математического ожидания и выборочного среднего, приведите примеры их использования (сформулируйте нулевое свойство математического ожидания, упростите формулы для вычисления дисперсии и ковариации, обоснуйте возможность применения условных переменных).

Свойства:

$$M(a) = a; M(kx) = k \cdot M(x); M(x+y) = M(x) + M(y); M(xy) = M(x) \cdot M(y).$$

$$\overline{a} = a; \quad \overline{kx} = k\bar{x}; \quad \overline{(x+y)} = \bar{x} + \bar{y} \quad \overline{xy} = \bar{x} \cdot \bar{y}$$

для независимых x, y .

Следствия: если $a = M(x)$, то $M(x - a) = 0$ – нулевое свойство; $\overline{(x - \bar{x})} = 0$.

$$\sigma_x^2 = M(x - a)^2 = M(x^2) - M^2(x); \quad s_x^2 = \overline{(x - \bar{x})^2} = \overline{x^2} - (\bar{x})^2.$$

$$\sigma_{xy} = M[(x - M(x)) \cdot (y - M(y))] = M(xy) - M(x) \cdot M(y); s_{xy} = \overline{xy} - \bar{x} \cdot \bar{y}.$$

Для независимых x, y ковариация σ_{xy} (s_{xy}) равна нулю.

Условная переменная: $X = \frac{x-c}{h}$; обратно: $x = c + hX$; тогда $\bar{x} = c + h\bar{X}$.

5. Перечислите свойства дисперсии. Запишите формулы для вычисления дисперсии суммы и дисперсии разности случайных величин (зависимых и независимых). Сформулируйте правило «3-х сигм». Поясните, что такое коэффициент вариации и в каких ситуациях он используется.

Свойства: $D(a) = 0$; $D(kx) = k^2 \cdot D(x)$; $D(x+y) = D(x) + D(y) + 2 \cdot Cov(x, y)$.

$$s^2(a) = 0; \quad s^2(kx) = k^2 s_x^2; \quad s^2(x+y) = s_x^2 + s_y^2 + 2 \cdot s_{xy}.$$

Следствия: $D(\lambda x + \mu y) = \lambda^2 \cdot D(x) + \mu^2 \cdot D(y) + 2 \cdot \lambda \mu \cdot Cov(x, y)$.

Для независимых x, y : $D(\lambda x + \mu y) = \lambda^2 \cdot D(x) + \mu^2 \cdot D(y)$; $D(x \pm y) = D(x) + D(y)$.

Условная переменная: $X = \frac{x-c}{h}$; обратно: $x = c + hX$; тогда $s_x = h \cdot s_X$.

Правило «3-х сигм»: с уровнем доверия, не меньшим 90 %, можно утверждать, что случайные отклонения от центра не превышают 3-х стандартных (среднеквадратичных) отклонений: $P(|x - a| \leq 3\sigma_x) > 0,89$, где $a = M(x)$. Это правило имеет место для любых случайных величин, независимо от вида их распределения.

Если коэффициент вариации $v_x = \frac{s_x}{\bar{x}} \cdot 100$ % меньше 2 %, случайной изменчивостью можно пренебречь.

6. Сформулируйте утверждение центральной предельной теоремы. Перечислите основные особенности нормального распределения. Поясните, что такое ошибка среднего.

ЦПТ: чем больше членов в сумме, тем ближе распределение суммы случайных величин к нормальному закону (независимо от вида распределения отдельных слагаемых). Сумма 10-ти и более слагаемых распределена практически нормально. Если закон распределения отдельных слагаемых хотя бы симметричен, то нормально распределена сумма значительно меньшего числа слагаемых (порядка 5-ти).

Нормальный закон одномодальный, симметричный, характеризуется правилом «2-х сигм». Выборочное среднее распределено асимптотически нормально с характеристиками $M(\bar{x}) = M(x)$; $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$.

Величина $\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}_x}{\sqrt{n}}$ называется ошибкой среднего или стандартной ошибкой (не путать со стандартным отклонением!).

7. Дайте понятие о распределении Стьюдента. Покажите, как определяются границы 95-процентного доверительного интервала на генеральное среднее (математическое ожидание); как определяется необходимый объем выборки для оценки центра группировки совокупности с заданной точностью и надежностью.

Если величина z распределена нормально $N(z; a; \sigma_z)$, то величина $t = \frac{z-a}{\hat{\sigma}_z}$ распределена по закону Стьюдента с параметром $ЧСС = n - 1$.

В частности, по Стьюденту распределена статистика $t = \frac{\bar{x}-a}{\hat{\sigma}_x/\sqrt{n}}$.

Распределение Стьюдента очень похоже на нормальное (одномодальное, симметричное), но правило «2-х сигм» заменяется на $|\bar{x} - a| \leq t_{0,05} \cdot \frac{\hat{\sigma}_x}{\sqrt{n}}$, где значение $t_{0,05}$ зависит только от $ЧСС$ и определяется по таблицам Стьюдента; при $n \geq 30$ уже соблюдается правило «2-х сигм», то есть $t_{0,05} \approx 2$.

Уровень доверия этого утверждения $P = (1 - 0,05) = 0,95$.

Отсюда получаем такой 95-процентный доверительный интервал на математическое ожидание (генеральное среднее): $\bar{x} - t_{0,05} \cdot \frac{\hat{\sigma}_x}{\sqrt{n}} \leq a \leq \bar{x} + t_{0,05} \cdot \frac{\hat{\sigma}_x}{\sqrt{n}}$ – с гарантией 95 % этот доверительный интервал со случайными границами покрывает неизвестное значение $a = M(x)$.

Если потребовать, чтобы относительная погрешность в определении $a = M(x)$ была не больше q %, получаем соотношение для определения необходимого объема выборки в виде:

$$n \geq \left[t_{0,05}(n-1) \cdot \frac{v_x}{q} \right]^2.$$

Это неравенство решается последовательными приближениями.

8. Поясните, что такое гистограмма, как она строится, чем отличается от дифференциальной функции распределения, чему равна ее полная площадь и площадь на заданном интервале.

Гистограмма – ступенчатый график плотности вероятности попадания случайной величины в заданные интервалы. Ординаты гистограммы: $f_i = \frac{m_i}{nh_i}$, где h_i – ширина интервала. Гистограмма – эмпирическая оценка дифференциальной функции распределения. Площадь одного столбца гистограммы равна оценке вероятности $p_i = \frac{m_i}{n}$ попадания случайной величины в этот интервал. Площадь всей гистограммы равна 1.

9. Поясните, что такое кумулята, как она строится, чем отличается от интегральной функции распределения. Дайте определение функции распределения (интегральной функции распределения) и сформулируйте суть общей интегральной теоремы.

Кумулята – график относительных частот, накопленных к правым краям интервалов, с ординатами на правых краях $F_i = \frac{1}{n} \sum_{j=1}^i m_j$. Для дискретной случайной величины график кумуляты ступенчатый, для непрерывной величины – кусочно-линейный. Кумулята – эмпирическая оценка функции распределения (интегральной функции распределения) $F(x) = P(X \leq x)$. Интегральная теорема: $P(x_1 < x \leq x_2) = F(x_2) - F(x_1)$.

10. Дайте понятия о квантилях распределения, покажите, как определяются медиана и квартили. Опишите блочную диаграмму Тьюкки.

Квантили – границы равнонасыщенных интервалов. Квартили – границы 4-х равнонасыщенных интервалов по 25 % наблюдений в каждом. Средняя квартиль называется также медианой. По определению $P(X > x_\alpha) = \alpha$ или $P(X \leq x_\alpha) = 1 - \alpha$, откуда следует способ вычисления квантилей с помощью кумуляты: $F(x_\alpha) = 1 - \alpha$, в частности $F(x_{0,75}) = 0,25$, $F(Me) = 0,50$, $F(x_{0,25}) = 0,75$. По предложению Дж. Тьюкки, особенности распределения достаточно информативно описываются диаграммой «ящик и усы». Границами «ящика» являют-

ся нижняя $x_{0,75}$ и верхняя $x_{0,25}$ квартили, средняя линия ящика – медиана; «усы» простираются до x_{min} и x_{max} , но не далее полутора межквартильного размаха от границ «ящика». Точки за пределами «усов» считаются выбросами.

11. Покажите, как по данным выборки найти параметры теоретического закона распределения и как по выбранному закону рассчитать ожидаемые частоты попадания случайной величины в заданные интервалы (с помощью интегральной и дифференциальной функций распределения).

Оценки параметров выбранного закона распределения определяются методом моментов – теоретические моменты распределения приравниваются к их выборочным оценкам; число таких соотношений (связей) равно числу параметров теоретического закона. Нормальный и равномерный законы распределения зависят от двух параметров, для их определения вводятся две связи: $M(x) = \bar{x}$; $\sigma_x = s_x$. Например, для равномерного закона эти соотношения имеют вид: $\frac{a+b}{2} = \bar{x}$; $\frac{b-a}{\sqrt{12}} = s_x$, откуда определяются параметры a и b . Показательный закон распределения зависит только от одного параметра, который определяется из уравнения $M(x) = \bar{x}$. Зная параметры теоретического закона, рассчитываются значения дифференциальной функции распределения \tilde{f}_i (для центров интервалов) и интегральной функции \tilde{F}_j (для правых краев интервалов). Теоретические частоты попадания случайной величины в заданные интервалы в соответствии с выбранным законом распределения определяются с помощью интегральной теоремы $\tilde{m}_j = n \cdot (\tilde{F}_j - \tilde{F}_{j-1}) = n \cdot \Delta \tilde{F}_j$. Если интервалы достаточно узкие, то приближенно $\tilde{m}_i \approx n \cdot h_i \cdot \tilde{f}_i$, где h_i – ширина интервала.

12. Дайте понятие о распределении Пирсона, о критерии согласия Пирсона, опишите последовательность расчетов по критерию Пирсона.

Если x_i распределены нормально с нулевым математическим ожиданием и единичной дисперсией $x_i \sim N(0; 1)$, то сумма квадратов таких величин распределена по закону Пирсона, который зависит от единственного параметра – числа степеней свободы (ЧСС). Область случайной изменчивости этой статистики $(\chi^2_{0,95}; \chi^2_{0,05})$; $M(\chi^2) = ЧСС$.

В критерии согласия Пирсона сравниваются частоты попадания случайной величины в заданные интервалы: эмпирические – m_i и соответствующие предполагаемому закону распределения – \hat{m}_i .

Статистика $\chi^2 = \sum_{i=1}^k \frac{(m_i - \hat{m}_i)^2}{\hat{m}_i} = \sum_{i=1}^k \frac{m_i^2}{\hat{m}_i} - n$ при выполнении некоторых условий распределена по закону Пирсона с $ЧСС = k - 1 - p$, где k – число интервалов, p – число параметров теоретического закона. Если вычисленное значение χ^2 попадает в интервал $(\chi_{0,95}^2; \chi_{0,05}^2)$, нуль-гипотеза об отсутствии значимых различий между двумя рядами частот не может быть отвергнута; считаем, что проверяемый закон соответствует данным. Если $\chi^2 > \chi_{0,01}^2$, нуль-гипотеза отвергается; проверяемый закон отклоняется, так как он неудовлетворительно описывает распределение данных. Если соответствие слишком хорошее $\chi^2 < \chi_{0,99}^2$, возникает сомнение в достоверности данных. Все остальные случаи являются областями неопределенности критерия.

13. Дайте понятие о числе степеней свободы и о числе связей. Покажите, какие связи есть при сравнении эмпирических и ожидаемых частот (в критерии Пирсона).

Число степеней свободы (ЧСС) равно разности между числом случайных величин и числом линейных связей, наложенных на эти величины. В критерии согласия Пирсона рассматриваются разности $(m_i - \tilde{m}_i)$, число которых равно k – числу (укрупненных) интервалов. Суммы наблюдаемых и теоретических частот одинаковы и равны общему числу наблюдений n , откуда имеем первую связь: $\sum_{i=1}^k (m_i - \tilde{m}_i) = 0$. Центр теоретического распределения совмещен с центром выборки $M(x) = \bar{x}$, откуда имеем вторую связь: $\sum_{i=1}^k (m_i - \tilde{m}_i) \cdot x_i = 0$.

Наконец, для двухпараметрических законов уравнивается также мера изменчивости теоретическая и выборочная $\sigma_x^2 = s_x^2$, откуда добавляется еще одна связь: $\sum_{i=1}^k (m_i - \tilde{m}_i) \cdot x_i^2 = 0$. Здесь x_i – центры заданных интервалов, то есть неслучайные коэффициенты.

14. Обоснуйте условия, необходимые для корректного применения критерия Пирсона (учесть особенности распределения Бернулли – Пуассона – Лапласа).

Прежде всего обеспечиваем выполнение условия $\sum_{i=1}^k \tilde{m}_i = n$, для чего расширяем границы крайних интервалов (если требуется, до $\pm\infty$). Число интервалов должно быть достаточно большим, чтобы в каждый интервал попало не более 10 % наблюдений, то есть $\tilde{p}_i \leq 0,1$. Тогда частоты m_i в каждом интервале будут распределены по закону Пуассона с характеристиками $a_i = n\hat{p}_i = \hat{m}_i$; $D_i = n\tilde{p}_i = \tilde{m}_i$. С другой стороны, предполагается, что в каждый интервал попадает не менее 5-ти наблюдений, для чего малонасыщенные интервалы укрупняем (объединяем с соседними). Тогда распределение Пуассона будет близким к распределению Лапласа (нормальному распределению), стандартизованные величины $\frac{m_i - \hat{m}_i}{\sqrt{\hat{m}_i}}$ будут распределены нормально с нулевым математическим ожиданием и единичной дисперсией, следовательно, сумма квадратов этих величин $\sum_{i=1}^k \frac{(m_i - \hat{m}_i)^2}{\hat{m}_i}$ будет иметь распределение Пирсона χ^2 . Несколько противоречивые требования $\tilde{p}_i = \frac{\tilde{m}_i}{n} \leq 0,1$; $\tilde{m}_i \geq 5$ могут быть удовлетворены только для больших выборок $n \approx 200$. Для выборок меньшего размера условие $\tilde{p}_i \leq 0,1$ не выполняется, частоты m_i в каждом интервале распределены по биномиальному закону с характеристиками $a_i = n\tilde{p}_i = \tilde{m}_i$; $D_i = n\tilde{p}_i\tilde{q}_i = \tilde{m}_i(1 - \frac{\tilde{m}_i}{n})$ и в статистике Пирсона следует учесть поправку: $\chi^2 = \sum_{i=1}^k \frac{(m_i - \tilde{m}_i)^2}{\tilde{m}_i(1 - \frac{\tilde{m}_i}{n})}$.

15. Покажите, как определяются границы 90-процентного доверительного интервала на генеральную дисперсию с помощью распределения Пирсона.

Если X распределено нормально, то статистика $\sum \frac{(x_i - \bar{x})^2}{\sigma_x^2} = \frac{n \cdot s_x^2}{\sigma_x^2}$ распределена по закону Пирсона χ^2 с ЧСС = $n - 1$, потому с вероятностью 90 % она заключена в пределах $(\chi_{0,95}^2; \chi_{0,05}^2)$, откуда получаем 90-процентные границы на дисперсию совокупности: $\frac{n}{\chi_{0,05}^2} s_x^2 \leq \sigma_x^2 \leq \frac{n}{\chi_{0,95}^2} s_x^2$.

16. Дайте понятие о критерии согласия Колмогорова – Смирнова.

Простой критерий Колмогорова – Смирнова основан на сравнении кумуляты с интегральной функцией теоретического закона. Найдем максимальное расхождение $D = \max |F_i - \hat{F}_i|$. Если статистика $K = \sqrt{n} \cdot D$ окажется больше 1,63, теоретический закон отвергается, а если 1,36 – принимается. Корректное применение критерия предполагает, что известны параметры теоретического закона, а не определять их по выборочным данным (как в критерии Пирсона).

17. Дайте понятие о нормальной вероятностной кривой, покажите, как она строится и применяется.

Для визуальной проверки соответствия нормальному распределению применяется графический метод, который основан на сравнении кумуляты и интегральной функции нормального закона. На краях укрупненных интервалов s_i (в которые должно попасть не менее 5-ти наблюдений) определяются ординаты кумуляты $F(s_i)$; предполагается, что эти значения порождены нормальным законом распределения $F(s_i) = \Phi(t_i) + 1/2$, откуда находят значения $\Phi(t_i) = F(s_i) - 1/2$; далее по таблице интегральной функции Лапласа находят соответствующие t_i и строят график $s_i - t_i$. Для нормального закона этот график представляет собой прямую $t_i = \frac{s_i - \bar{x}}{s_x}$, или $s_i = \bar{x} + s_x t_i$. Если точки (s_i, t_i) явно не группируются вокруг некоторой прямой, гипотеза о нормальности распределения отклоняется. В таком случае вид графика нормальной вероятностной кривой подсказывает, после какого функционального преобразования переменной распределение будет более близким к нормальному.

18. Опишите особенности нормального закона распределения, его параметры и характеристики, дифференциальную и интегральную функции, структуру таблиц, область применения.

Нормальное распределение Гаусса является наиболее распространенным законом природы и занимает среди других распределений особое положение. Применения его настолько разнообразны, что перечислить их практически невозможно. В частности, выборочное среднее распределено асимптотически нормально (или по близкому к нему распределению Стьюдента для малых выборок).

Дифференциальная функция нормального закона $f(x)$ (плотность вероятности) выражается через дифференциальную функцию Лапласа $\varphi(t_x)$, которая затабулирована:

$$f(x) = \frac{\varphi(t_x)}{\sigma_x} , \quad t_x = \frac{x-a}{\sigma_x} , \quad \varphi(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-t^2/2} .$$

Интегральная функция нормального закона $F(x)$ выражается через другую затабулированную функцию $\Phi(t_x)$ – интегральную функцию Лапласа:

$$F(x) = \Phi(t_x) + \frac{1}{2} , \quad t_x = \frac{x-a}{\sigma_x} , \quad \Phi(t) = \int_0^t \varphi(s) ds = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-s^2/2} ds .$$

Два параметра закона – a , σ_x – являются основными характеристиками распределения: $a = M(x)$, $\sigma_x^2 = D(x)$. Основные особенности нормального закона – одномодальность, симметричность, правило «2-х сигм».

19. Опишите особенности равномерного закона распределения, его параметры и характеристики, дифференциальную и интегральную функции, область применения.

По закону равномерной плотности (равномерный закон) распределены ошибки округления, время ожидания транспорта, который движется по графику строго через равные интервалы времени, и т. д.

Дифференциальная функция этого распределения постоянна на интервале $[a, b]$ и равняется нулю за его пределами. Интегральная функция линейна на интервале $[a, b]$:

$$f(x) = \begin{cases} 0 & (x < a) \\ \frac{1}{b-a} & (a \leq x \leq b) ; \\ 0 & (x > b) \end{cases} ; \quad F(x) = \begin{cases} 0 & (x < a) \\ \frac{x-a}{b-a} & (a \leq x \leq b) . \\ 1 & (x > b) \end{cases}$$

Параметрами закона являются границы интервала a, b . Характеристики выражаются через эти параметры: $M(x) = \frac{a+b}{2}$; $D(x) = \frac{(b-a)^2}{12}$.

20. Опишите особенности показательного закона распределения, его параметры и характеристики, дифференциальную и интегральную функции, область применения.

По показательному закону распределено время работы оборудования до первого отказа, время ожидания вызова на АТС и т. д. Дифференциальная

функция показательного закона для $x \geq 0$: $f(x) = \lambda \cdot e^{-\lambda x}$. Интегральная функция (для $x \geq 0$):

$$F(x) = 1 - e^{-\lambda x}.$$

Единственный параметр закона – λ .

Основная черта показательного закона – равенство основных характеристик:

$$M(x) = \sigma_x = 1/\lambda.$$

Коэффициент вариации $v_x = 100\%$.

5. Опишите особенности логнормального закона распределения, его параметры и характеристики, область применения.

Случайная величина распределена по логарифмически нормальному закону, если ее логарифм $y = \ln x$ распределен нормально.

Основная область применения логнормального закона – социологические и экономические исследования. В частности, этим законом хорошо описываются распределения таких экономических показателей, как доход, заработная плата, потребительский спрос.

Если случайная величина $y = \ln x$ распределена нормально с характеристиками $\mu_0 = M(y)$ и $\sigma_0 = \sqrt{D(y)}$, то этот факт кратко обозначается как $y \sim N(\mu_0; \sigma_0)$.

При этом величина $x > 0$ имеет логнормальное распределение с этими же параметрами, что кратко обозначается как $x \sim \Lambda(\mu_0; \sigma_0)$.

Характеристики нормального распределения y ($\mu_0 = M(y)$, $\sigma_0 = \sqrt{D(y)}$) связаны с характеристиками исходного показателя x ($\mu_x = M(x)$, $\sigma_x = \sqrt{D(x)}$) достаточно простыми соотношениями, так что нет необходимости заново определять характеристики после логарифмирования. Как и гамма-распределение, логнормальное распределение скошено влево.

Регрессионный анализ в вопросах и ответах

1. Дайте определение функциональной, статистической и корреляционной зависимостей. Продемонстрируйте различия между сопряженными корреляционными моделями. Приведите пример статистической, но не корреляционной зависимости.

По определению *функциональной* зависимости, каждому значению аргумента (набору значений аргументов) соответствует единственное значение результативного признака. В *стохастических* (статистических) зависимостях каждому значению аргумента соответствует свой ряд распределения результативного признака. Частным случаем стохастической зависимости является *корреляционная* зависимость, когда следят за изменением только одной характеристики распределения результативного признака – центром группировки Y при каждом значении X (то есть за изменением условного математического ожидания $M(y|x)$ при изменении аргумента x). График корреляционной зависимости называется также линией регрессии, а ее уравнение – уравнением регрессии. Для корреляционных, как и для функциональных зависимостей, имеет место однозначное соответствие между значениями аргумента и откликом (средними значениями результативной переменной \bar{y}_x). Однако между этими видами зависимостей остается принципиальное различие – корреляционные зависимости необратимы относительно замены направления причинно-следственных связей. В наиболее распространенном случае совместного нормального распределения двух случайных величин (X, Y) облако рассеяния точек (X, Y) имеет форму вытянутого эллипса. Линия регрессии \bar{y}_x представляет собой диаметр этого эллипса, сопряженный семейству вертикальных хорд (середины вертикальных хорд). Если же в качестве результативного признака выбрана другая переменная X (y – причина, x – следствие), то линия регрессии представляет собой диаметр эллипса, сопряженный семейству горизонтальных хорд (середины горизонтальных хорд \bar{x}_y). Это совсем разные диаметры (разные корреляционные зависимости не только по аналитической форме записи, но и по существу, так называемые *сопряженные* регрессии). Существуют также стохастические, но не корреляционные зависимости, когда при изменении аргумента X изменяется не центр группировки Y , а другие характеристики распределения отклика, например изменчивость (дисперсия).

2. Разъясните смысл термина «диагональная регрессия». Поясните, является ли диагональная регрессия регрессией вообще (согласно определению этого понятия), в каких случаях целесообразно использовать эту модель.

Наличие связи вовсе не означает, что одна из переменных определяет другую. Вполне возможно, что две переменные изменяются синхронно («в такт») потому, что обе они являются следствиями некой общей причины. В этом случае неверно будет приписывать какой-либо из этих переменных роль результативного признака и выбирать соответствующую связь из числа взаимно сопряженных; наилучшим графиком существующей зависимости в этом случае была бы главная ось эллипса рассеивания, вдоль которой он вытянут.

Заметим, что уравнение главной оси облака рассеивания формально не является уравнением регрессии по определению, поскольку точки главной оси не есть средние значения одной переменной при фиксированных значениях другой.

Уравнением этой диагональной регрессии является:

$$\frac{y-\bar{y}}{s_y} = \pm \frac{x-\bar{x}}{s_x},$$

где знак «+» выбирается для возрастающей, а знак «-» для убывающей зависимости.

Для сравнения приведем здесь же уравнение регрессии (y/x): $\frac{y-\bar{y}}{s_y} = r_{xy} \frac{x-\bar{x}}{s_x}$ и уравнение сопряженной регрессии (x/y): $\frac{x-\bar{x}}{s_x} = r_{xy} \frac{y-\bar{y}}{s_y}$.

3. Сформулируйте идею принципа Лежандра (МНК), разъясните смысл системы нормальных уравнений, составьте систему нормальных уравнений для линейной и квадратичной моделей с одной объясняющей переменной.

По методу наименьших квадратов (МНК) параметры модели $y = a_0 + a_1x_1 + a_2x_2 + e$ необходимо подбирать таким образом, чтобы была минимальной сумма квадратов ошибок (e) по всем наблюдениям.

Условия минимума суммы квадратов ошибок $\sum e^2$ приводят к требованию ортогональности (нормальности) вектора ошибок к каждому члену модели: $\sum e = 0$, $\sum ex_1 = 0$, $\sum ex_2 = 0$.

Отсюда получаем такую систему «нормальных» уравнений для определения параметров:

$$\begin{aligned}\Sigma y &= a_0 n + a_1 \Sigma x_1 + a_2 \Sigma x_2 ; \\ \Sigma y x_1 &= a_0 \Sigma x_1 + a_1 \Sigma (x_1)^2 + a_2 \Sigma x_1 x_2 ; \\ \Sigma y x_2 &= a_0 \Sigma x_2 + a_1 \Sigma x_1 x_2 + a_2 \Sigma (x_2)^2 .\end{aligned}$$

Для квадратичной модели $y = a_0 + a_1 x + a_2 x^2 + e$ условия ортогональности ошибки к каждому члену модели $\Sigma e = 0, \Sigma e x = 0, \Sigma e x^2 = 0$ приводят к такой нормальной системе уравнений:

$$\begin{aligned}\Sigma y &= a_0 n + a_1 \Sigma x + a_2 \Sigma x^2 ; \\ \Sigma y x &= a_0 \Sigma x + a_1 \Sigma x^2 + a_2 \Sigma x^3 ; \\ \Sigma y x^2 &= a_0 \Sigma x^2 + a_1 \Sigma x^3 + a_2 \Sigma x^4 .\end{aligned}$$

4. Сформулируйте основные предпосылки дисперсионного анализа. Докажите, что средние по группам являются наилучшими МНК-оценками центров каждой группы. Разложите общую сумму квадратов на межгрупповую и внутригрупповую составляющие.

Имеется p групп наблюдений y_{ij} . Группы описываются значениями некоторого фактора, например разными значениями объясняющей переменной x_i . Количество наблюдений в каждой группе – k_i , общее количество наблюдений – $n = \Sigma k_i$. Необходимо выяснить, имеются ли между группами значимые различия (то есть имеется ли зависимость y от x). Оценку значимости различий между группами в целом производят с помощью дисперсионного анализа Фишера, а между каждой парой групп – по критерию Стьюдента. Модель дисперсионного анализа имеет вид: $y_{ij} = u_i + \varepsilon_{ij}$. Основные предпосылки анализа – группы различаются только средними значениями (\bar{y}_{x_i}), изменчивость данных (дисперсия) по группам одинакова, все наблюдения независимые.

Величины u_i , которые характеризуют каждую группу, определяем методом наименьших квадратов (МНК):

$$\sum_{i=1}^p \sum_{j=1}^{k_i} \varepsilon_{ij}^2 = \sum_{j=1}^{k_1} (y_{1j} - u_1)^2 + \sum_{j=1}^{k_2} (y_{2j} - u_2)^2 + \dots + \sum_{j=1}^{k_p} (y_{pj} - u_p)^2 \rightarrow \min .$$

Приравниваем нулю частные производные суммы квадратов ошибок по u_i и получаем $-2 \sum_{j=1}^{k_i} (y_{ij} - u_i) = 0$, откуда следует: $u_i = \frac{1}{k_i} \sum_{j=1}^{k_i} y_{ij} = \bar{y}_{x_i}$, то есть наилучшими оценками для u_i являются средние групповые \bar{y}_{x_i} . Для каждой

группы теперь выполняется нулевое свойство: $\sum_{j=1}^{k_i} \varepsilon_{ij} = 0$, откуда $\bar{\varepsilon} = 0$,
 $\bar{u} = \bar{y} = y_{cp}$.

Аналогично разложению $y_{ij} = u_i + \varepsilon_{ij}$ разлагается сумма квадратов отклонений $SSY = SSU + SS\varepsilon$, где $SSY = \sum \sum (y_{ij} - y_{cp})^2$ – общая сумма квадратов отклонений; $SSU = \sum \sum (u_i - y_{cp})^2$ – сумма квадратов отклонений между группами; $SS\varepsilon = \sum \sum (\varepsilon_{ij})^2$ – сумма квадратов отклонений внутри групп. Действительно, $SSY = \sum \sum (y_{ij} - y_{cp})^2 = \sum \sum [(y_{ij} - u_i) - (u_i - y_{cp})]^2 = SS\varepsilon + SSU - 2 \sum \sum [(y_{ij} - u_i) \cdot (u_i - y_{cp})]$, где сумма произведений $\sum \sum [(y_{ij} - u_i) \cdot (u_i - y_{cp})] = \sum (u_i - y_{cp}) \sum (y_{ij} - u_i)$ равна нулю, так как в каждой группе $\sum (y_{ij} - u_i) = 0$. Точно также разлагается общее число степеней свободы $dfY = dfU + df\varepsilon$ (df – degree of freedom), где $dfY = n - 1$, $dfU = p - 1$, $df\varepsilon = n - p$. Средние квадраты (несмещенные оценки дисперсий) вычисляются по формулам $MS = SS/df$. Дисперсионное отношение Фишера $F = \frac{MSU}{MS\varepsilon} = \frac{SSU}{SS\varepsilon} \cdot \frac{n-p}{p-1}$ показывает, во сколько раз изменчивость средних групповых u_i превосходит изменчивость помехи ε_{ij} . Если дисперсионное отношение окажется меньше табличного значения $F < F_{0,05}(p-1; n-p)$, нуль-гипотеза об отсутствии значимых различий между группами не может быть отвергнута. Различия между группами считаются значимыми, если $F > F_{0,01}(p-1; n-p)$.

5. Опишите методику сравнения двух выборок по критерию Стьюдента. Сформулируйте основные предпосылки (гипотезы) этого метода. Покажите, что этот анализ является частным случаем дисперсионного анализа, когда количество сравниваемых групп равно двум.

Когда с помощью дисперсионного анализа установлено, что между группами в целом имеются значимые различия, далее следует выяснить, между какими именно группами имеются значимые различия. Различия между каждой парой групп можно проще (и быстрее) проверить с помощью критерия Стьюдента. Предпосылки этого анализа совпадают с предпосылками дисперсионного анализа – группы различаются только значениями средних групповых (u_1 , u_2); случайная изменчивость данных по группам одинакова ($s_1^2 \approx s_2^2$); все наблюдения независимые.

Общую случайную дисперсию (несмещенную оценку) получаем объединением дисперсий по группам: $\hat{\sigma}_{\varepsilon}^2 = \frac{k_1 s_1^2 + k_2 s_2^2}{k_1 + k_2 - 2}$, где $df_{\varepsilon} = (k_1 + k_2 - 2)$ – ЧСС случайной изменчивости (две связи, так как в каждой группе суммы ошибок равны нулю). Случайная дисперсия среднего u_i будет в k_i раз меньше. Рассматриваем разность средних групповых $\Delta = |u_1 - u_2|$. Дисперсия разности независимых величин равна сумме их дисперсий $\hat{\sigma}_{\Delta}^2 = \frac{k_1 s_1^2 + k_2 s_2^2}{k_1 + k_2 - 2} \cdot \left(\frac{1}{k_1} + \frac{1}{k_2} \right)$. Если статистика Стьюдента $t_{\Delta} = \frac{\Delta}{\hat{\sigma}_{\Delta}}$ меньше табличного значения $t_{0,05}(k_1 + k_2 - 2)$, нуль-гипотеза об отсутствии значимых различий между двумя группами не может быть отвергнута. Различия между группами считаются значимыми, если $t_{\Delta} > t_{0,01}(k_1 + k_2 - 2)$. Применение дисперсионного анализа для выявления различий между двумя выборками ($p = 2$) приведет к тому же выводу, так как $F(1; n - 2) = t^2(n - 2)$.

6. Покажите, как строится эмпирическая линия регрессии, как оценивается теснота корреляционной связи. Поясните, что такое индекс детерминации и корреляционное отношение, чем они отличаются от коэффициента детерминации и коэффициента корреляции соответственно.

Данные следует сгруппировать на несколько интервалов по возрастающим значениям объясняющей переменной так, чтобы в каждую группу попало не менее 5-ти наблюдений (для малой выборки – не менее 5 % наблюдений); малонасыщенные группы объединяем с соседними. Обозначим через x_i центры интервалов, y_{ij} – значения отклика (результативной переменной) в группе, k_i – количество наблюдений в группе, $n = \sum k_i$ – общее количество наблюдений. В каждой группе вычисляем среднее значение результативной переменной

$u_i = \bar{y}_{x_i}$, где $\bar{y}_{x_i} = \frac{1}{k_i} \sum_{j=1}^{k_i} y_{ij}$. Строим кусочно-линейный график с узлами $(x_i; u_i)$,

который называется эмпирической линией регрессии. В модели дисперсионного анализа предполагается, что группы различаются только средними значениями отклика: $y_{ij} = u_i + \varepsilon_{ij}$, где ε_{ij} – случайные ошибки, которые не зависят ни от x_i , ни от u_i . Дисперсия суммы независимых случайных величин равна сумме

дисперсий $s_y^2 = s_u^2 + s_{\varepsilon}^2$. Обозначим $\eta_{y/x}^2 = \frac{s_u^2}{s_y^2} = 1 - \frac{s_{\varepsilon}^2}{s_y^2}$ – относительный вклад в

общую дисперсию, который определяется различиями между группами (то есть

влиянием объясняющей переменной x). Эту величину называют индексом детерминации, а корень квадратный из нее $\eta_{y/x}$ – корреляционным отношением.

Из определения индекса детерминации следует: $0 \leq \eta_{y/x}^2 \leq 1$; при $\eta_{y/x}^2 = 1$ все $\varepsilon_{ij} = 0$, то есть каждому значению аргумента x_i соответствует единственное значение отклика y_i , что является характерной особенностью функциональной зависимости; при $\eta_{y/x}^2 = 0$ все $\bar{y}_{x_i} = \text{Const}$, то есть корреляционной связи нет (никакой). Таким образом, индекс детерминации является объективной мерой тесноты корреляционной связи.

В регрессионном анализе принимают иную модель: $y_i = y_{pi} + e_i$, где y_{pi} – расчетные значения по уравнению регрессии, e_i – остатки модели, которые не зависят от аргументов (и от расчетных значений). Поэтому $s_y^2 = s_p^2 + s_e^2$. Отно-

шение $R^2 = \frac{s_p^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$ называется коэффициентом детерминации, а корень

квадратный из этой величины – коэффициентом корреляции (коэффициентом парной корреляции r_{xy} – если зависимость линейная от одного аргумента, или коэффициентом множественной корреляции R – в остальных случаях). Коэффициент детерминации является мерой тесноты корреляционной связи *указанного типа*. Например, если для линейной модели $y_p = b_0 + b_1 x$ оказалось $R^2 \approx 0$, то нельзя утверждать, что нет корреляционной связи вообще; правильный вывод – между x и y нет *линейной* корреляционной зависимости.

7. Изложите последовательность расчетов для оценки значимости корреляционной связи. Опишите таблицу дисперсионного анализа, разъясните смысл ее отдельных граф (столбцов) – сумм квадратов, чисел степеней свободы, средних квадратов. Поясните, какой смысл имеет дисперсионное отношение Фишера, что такое уровень значимости и как им пользоваться.

Данные сгруппированы на p интервалов по возрастающим значениям объясняющей переменной x . В каждой группе вычислены средние значения результирующего признака $u_i = \bar{y}_{x_i}$, подсчитаны значения: $s_y^2 = SSY/n$ – общей дис-

персии, $s_u^2 = SSU/n$ – дисперсии средних групповых, их отношение $\eta_{y/x}^2 = \frac{s_u^2}{s_y^2}$ –

индекс детерминации.

Аналогично разложению «общего сигнала» на «полезный сигнал» и «помеху» $y_{ij} = u_i + \varepsilon_{ij}$ разлагается общая сумма квадратов отклонений (SS – summa of squares) на межгрупповую и внутригрупповую суммы квадратов $SSY = SSU + SS\varepsilon$. Точно так же разлагается общее число степеней свободы $dfY = dfU + df\varepsilon$ (df – degree of freedom).

Расчеты всех компонент сведены в таблицу дисперсионного анализа 1.

Таблица дисперсионного анализа 1 для оценки значимости корреляционной связи

Изменчивость	Суммы квадратов	ЧСС	Средние квадраты	Дисперсионное отношение
Между группами (u)	$SSU = \eta_{y/x}^2 \cdot SSY$	$dfU = p - 1$	$MSU = \frac{\eta_{y/x}^2}{p-1} \cdot SSY$	$F_\eta = \frac{\eta_{y/x}^2}{1-\eta_{y/x}^2} \cdot \frac{n-p}{p-1}$
Внутри групп (ε)	$SS\varepsilon = \left(1 - \eta_{y/x}^2\right) \cdot SSY$	$df\varepsilon = n - p$	$MS\varepsilon = \frac{1-\eta_{y/x}^2}{n-p} \cdot SSY$	
Общая (y)	SSY	$dfY = n - 1$		

Здесь $dfY = (n - 1)$, так как вследствие нулевого свойства сумма отклонений от среднего всегда равна нулю; $dfU = (p - 1)$ – по той же причине; $df\varepsilon = (n - p)$, так как суммы ошибок в каждой группе равны нулю.

Средние квадраты (несмещенные оценки дисперсий) вычисляются по формулам $MS = SS/df$ (MS – mean of squares). Дисперсионное отношение Фишера $F = \frac{MSU}{MS\varepsilon} = \frac{SSU}{SS\varepsilon} \cdot \frac{n-p}{p-1}$ показывает, во сколько раз изменчивость средних групповых u_i превосходит изменчивость помехи ε_{ij} .

Если $F < F_{0,05}(p-1; n-p)$, нуль-гипотеза об отсутствии значимых различий между группами не может быть отвергнута. Различия между группами считаются значимыми, если $F > F_{0,01}(p-1; n-p)$. Вместо таблиц квантилей $F_{0,05}$, $F_{0,01}$ можно использовать таблицы уровня значимости $\alpha = P(F > F_\alpha)$. Если получилось, что $\alpha < 0,01$, то это означает $F > F_{0,01}$ (корреляционная связь значима), а если $\alpha < 0,05$, то $F < F_{0,05}$ (корреляционной связи нет).

8. Изложите последовательность расчетов для оценки значимости регрессионной модели. Опишите таблицу дисперсионного анализа, разъясните смысл ее отдельных граф. Выразите для этой проблемы дисперсионное отношение через коэффициент детерминации.

Для модели $y = y_p + e = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_m x_m + e$ (линейной относительно параметров) получены МНК-оценки $(m + 1)$ параметра (коэффициентов регрессии) и коэффициент детерминации $R^2 = \frac{s_{yp}^2}{s_y^2} = \frac{SSR}{SSY}$. Аналогично разложе-

нию $y = y_p + e$, разлагается сумма квадратов отклонений $SSY = SSR + SSE$ и число степеней свободы $dfY = dfR + dfE$.

Расчеты всех компонент сведены в таблицу дисперсионного анализа 2.

Таблица дисперсионного анализа 2 для оценки значимости регрессионной модели

Изменчивость	Суммы квадратов	ЧСС	Средние квадраты	Дисперсионное отношение
Регрессия (y_p)	$SSR = R^2 \cdot SSY$	$dfR = m$	$MSR = \frac{R^2}{m} \cdot SSY$	$F_R = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}$
Остаток (e)	$SSE = (1 - R^2) \cdot SSY$	$dfE = n - m - 1$	$MSE = \frac{1-R^2}{n-m-1} \cdot SSY$	
Общая (y)	SSY	$dfY = n - 1$		

Здесь $dfE = n - m - 1$, так как для определения $(m + 1)$ параметра модели на остатки e наложено $(m + 1)$ связей (система нормальных уравнений).

MSE – несмещенная оценка остаточной дисперсии: $MSE = \frac{1-R^2}{n-m-1} \cdot SSY = s_y^2 (1 - R^2) \frac{n}{n-m-1}$. Дисперсионное отношение Фишера $F_R = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}$ показывает, во сколько раз изменчивость расчетных значений y_p превосходит изменчивость помехи e .

Если $F < F_{0,05}(m; n-m-1)$, нуль-гипотеза об отсутствии значимой корреляционной связи не может быть отвергнута. Регрессионная модель признается значимой, если $F > F_{0,01}(m; n-m-1)$.

9. Опишите методику оценки значимости коэффициента регрессии и коэффициента парной корреляции по критерию Стьюдента. Покажите, что эта методика является частным случаем дисперсионного анализа для оценки значимости линейной одномерной модели.

Выборочная оценка коэффициента парной корреляции r_{xy} с математическим ожиданием ρ_{xy} и оценкой дисперсии $\hat{\sigma}_r^2 = \frac{1-r_{xy}^2}{n-2}$ для малых $\rho_{xy} \approx 0$ распределена приблизительно нормально, поэтому для малых ρ_{xy} статистика $t_r = \frac{r_{xy} - \rho_{xy}}{\hat{\sigma}_r}$ распределена по закону Стьюдента с ЧСС = $n - 2$. Проверяется нуль-гипотеза о том, что действительное значение параметра равно нулю $\rho_{xy} = 0$. В этом случае абсолютная величина статистики Стьюдента показывает, во сколько раз параметр (коэффициент корреляции) превышает свою оценку стандартного отклонения.

Если статистика $t_r = \frac{|r_{xy}|}{\hat{\sigma}_r} = \frac{|r_{xy}|}{\sqrt{1-r_{xy}^2}} \cdot \sqrt{n-2}$ будет меньше табличного значения $t_{0,05}(n-2)$, нуль-гипотеза не может быть отвергнута.

Корреляционная зависимость признается значимой, если статистика Стьюдента будет больше $t_{0,01}(n-2)$. Сравним статистику Стьюдента с дисперсионным отношением Фишера $F_r = \frac{r_{xy}^2}{1-r_{xy}^2} \cdot \frac{n-2}{1} = t_r^2$ и убедимся в эквивалентности этих двух критериев. Что касается оценок коэффициентов регрессии b_i , то они (на основании центральной предельной теоремы) всегда распределены асимптотически нормально, поэтому с помощью критерия Стьюдента можно не только проверить нуль-гипотезу $\beta_i = M(b_i) = 0$, но также вычислить границы 95-процентного доверительного интервала на генеральные значения коэффициентов регрессии:

$$b_i - t_{0,05} \cdot \hat{\sigma}_{b_i} \leq \beta_i \leq b_i + t_{0,05} \cdot \hat{\sigma}_{b_i}.$$

Для линейной однофакторной модели имеем оценку:

$$\hat{\sigma}_{b_1} = \frac{\hat{\sigma}_{\bar{y}}}{s_x} = \frac{s_y}{s_x} \cdot \sqrt{\frac{1-r_{xy}^2}{n-2}}$$

и значимость коэффициента регрессии b_1 оказывается эквивалентной значимости коэффициента корреляции и значимости модели в целом $t_{b_1} = t_r = \sqrt{F_r}$.

10. Изложите последовательность расчетов для оценки адекватности модели. Опишите таблицу дисперсионного анализа, разъясните смысл ее отдельных граф. Покажите, в чем разница между оценкой дисперсии остатка модели и дисперсией случайной ошибки.

Адекватность (форму связи) принятой модели можно проверить, если имеются дополнительные данные, которые не были использованы для оценки параметров модели (контрольная выборка); или же известна величина дисперсии случайной ошибки, с которой можно сравнить величину дисперсии остатков модели. Дисперсия случайной изменчивости известна, когда данные для каждого значения аргумента x приведены в нескольких повторениях (так называемый активный эксперимент). Можно получить оценку дисперсии случайной изменчивости, если сгруппировать данные на классы по возрастающим значениям аргумента (объясняющей переменной) x . Дисперсию данных внутри групп можно принять за оценку случайной дисперсии. Обозначим через u_i средние

\bar{y}_{x_i} в каждой группе, s_u^2 – их дисперсию, $\eta_{y/x}^2 = \frac{s_u^2}{s_y^2}$ – индекс детерминации, ко-

торый показывает, во сколько раз изменчивость между группами превышает изменчивость внутри групп, то есть во сколько раз изменчивость, связанная с влиянием x , превышает случайную изменчивость. Тогда оценку случайной дис-

персии можно выразить как $s_\varepsilon^2 = \left(1 - \eta_{y/x}^2\right) \cdot s_y^2$. Если рассчитать коэффициент

корреляции по сгруппированным данным, то дисперсию остатка модели $s_e^2 = \left(1 - r_{xy}^2\right) \cdot s_y^2$ можно сравнивать с оценкой случайной дисперсии. Остатки

модели, кроме случайной компоненты, содержат ошибку спецификации модели, ошибку неадекватности, систематическую ошибку из-за выбора неверной

формы связи: $e = \varepsilon + A$, с дисперсией $s_A^2 = s_e^2 - s_\ell^2 = \left(\eta_{y/x}^2 - r_{xy}^2\right) \cdot s_y^2$. Оценку

значимости ошибки неадекватности модели получаем, заполнив таблицу дисперсионного анализа 3, где $dfE = n - 2$, так как для определения 2-х параметров линейной модели на остатки e наложены две связи (система нормальных уравнений); $df\varepsilon = n - p$, так как в каждой группе сумма случайных отклонений

$\varepsilon_{ij} = (y_{ij} - u_i)$ равна нулю (нулевое свойство средних групповых). $MS = \frac{SS}{df}$ – не-

смещенные оценки дисперсий. Дисперсионное отношение Фишера

$F_A = \frac{\eta_{y/x}^2 - r_{xy}^2}{1 - \eta_{y/x}^2} \cdot \frac{n-p}{p-2}$ показывает, во сколько раз изменчивость систематической

ошибки (ошибки неадекватности) превышает случайную изменчивость. Если окажется, что $F > F_{0,01}$, то модель признается неадекватной (недоброкачественной).

Таблица дисперсионного анализа 3 для проверки адекватности регрессионной модели

Изменчивость	Суммы квадратов	ЧСС	Средние квадраты	Дисперсионное отношение
Неадекватность (A)	$SSA = \left(\eta_{y/x}^2 - r_{xy}^2 \right) \cdot SSY$	$p - 2$	$MSA = \frac{\eta_{y/x}^2 - r_{xy}^2}{p-2} \cdot SSY$	$F_A = \frac{\eta_{y/x}^2 - r_{xy}^2}{1 - \eta_{y/x}^2} \cdot \frac{n-p}{p-2}$
Случайность (ε)	$SS\varepsilon = \left(1 - \eta_{y/x}^2 \right) \cdot SSY$	$n - p$	$MS\varepsilon = \frac{1 - \eta_{y/x}^2}{n-p} \cdot SSY$	
Остаток модели (У)	$SSE = \left(1 - r_{xy}^2 \right) \cdot SSY$	$n - 2$		

Принятая модель считается адекватной если $F < F_{0,05}$, в этом случае нуль-гипотеза об отсутствии систематической ошибки не может быть отвергнута.

11. Выведите формулы для расчета параметров парной линейной регрессии. Дайте определение коэффициента парной корреляции, перечислите его свойства. Поясните, что такое коэффициент детерминации, чем он отличается от индекса детерминации.

Для линейной однофакторной модели $y = b_0 + b_1 x + e$ составляем условия ортогональности вектора ошибок к каждому члену модели: $\bar{e} = 0$; $\overline{ex} = 0$. Кроме этого, учтем, что $\overline{ey} = s_e^2$. Получим:

$$\bar{y} = b_0 + b_1 \bar{x}; \quad \overline{xy} = b_0 \bar{x} + b_1 \overline{x^2}; \quad \overline{y^2} = b_0 \bar{y} + b_1 \overline{xy} + s_e^2.$$

С помощью первого уравнения исключаем b_0 из остальных равенств:

$$s_{xy} = b_1 s_x^2; \quad s_y^2 = b_1 s_{xy} + s_e^2.$$

Отсюда получаем:

$$b_1 = \frac{s_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x}; \quad R^2 = 1 - \frac{s_e^2}{s_y^2} = b_1 \frac{s_{xy}}{s_y^2} = r_{xy}^2.$$

Из первого уравнения имеем $b_0 = \bar{y} - b_1 \bar{x}$. Квадрат коэффициента парной корреляции (нормированного смешанного момента) $r_{xy} = \frac{s_{xy}}{s_x s_y}$ оказался равен

коэффициенту детерминации $R^2 = \frac{s_p^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$, который показывает, какая часть

полной изменчивости y определяется моделью (линейной зависимостью от x). Отсюда следует, что $-1 \leq r_{xy} \leq +1$; при $|r_{xy}| = 1$ ошибок нет (все $e_i = 0$), связь точная, функциональная; при $r_{xy} = 0$ все $y_p = \bar{y}$, корреляционной связи указанного типа нет (то есть нет линейной связи). В отличие от коэффициента детер-

минации индекс детерминации $\eta_{y/x}^2$ является более объективной оценкой тесноты корреляционной связи; если индекс детерминации равен нулю, корреляционной связи нет (никакой). Для определения индекса детерминации требуется сгруппировать данные на классы с разными значениями аргумента x_i , вычислить средние в каждой группе $u_i = \bar{y}_{x_i}$, дисперсию этих средних s_u^2 , после

чего найти отношение $\eta_{y/x}^2 = \frac{s_u^2}{s_y^2}$, которое показывает, какая часть полной изменчивости y определяется различиями между классами (то есть разными значениями аргумента x).

12. Перечислите основные предпосылки регрессионного анализа. Сформулируйте идею принципа максимального правдоподобия и покажите, что по этому принципу наилучшими оценками параметров модели будут МНК-оценки.

Предпосылки регрессионного анализа: 1) $y(x_i) = y_p(x_i) + e_i$ (все ошибки относятся только к результативной переменной y , объясняющие переменные x измерены без ошибок); 2) $M(e_i) = 0$ – систематических ошибок нет, выбранная модель адекватная; 3) $M(e_i e_j) = 0$ – ошибки разных наблюдений не коррелиро-

ваны (наблюдения независимые); 4) $M(e_i^2) = \sigma_e^2 = Const$ – наблюдения равноточные; 5) ошибки распределены нормально. Отсюда получаем плотность вероятности для отдельных ошибок $f_i = f(e_i) = \frac{1}{\sqrt{2\pi}\sigma_e} \cdot \exp\left\{-\frac{e_i^2}{2\sigma_e^2}\right\}$ и плотность совместного распределения системы независимых ошибок $f(e_1, e_2, \dots, e_n) = f_1 f_2 \dots f_n = \frac{1}{(\sqrt{2\pi})^n \sigma_e^n} \cdot \exp\left\{-\frac{\sum e_i^2}{2\sigma_e^2}\right\}$. Согласно принципу максимума правдоподобия, параметры модели (и оценку дисперсии σ_e^2) надо выбрать так, чтобы получить максимум функции распределения f (наблюдаемая система ошибок e_i должна быть наиболее вероятной). Из условий максимума плотности совместного распределения следует метод наименьших квадратов (параметры модели необходимо определять из условий минимума $\sum e^2$), а оценка дисперсии случайной ошибки оказывается равной $s_e^2 = \overline{e^2}$.

13. Сформулируйте идею расчета дисперсий коэффициентов регрессии и дисперсий расчетных значений. Опишите графический способ построения 95-процентной доверительной полосы на линию регрессии.

Согласно предпосылкам регрессионного анализа, все объясняющие переменные x считаются измеренными точно, все случайные ошибки относятся только к результативному признаку y . Эти ошибки независимые, равноточные (имеют одинаковую дисперсию для любых наблюдений), систематических ошибок нет (то есть $M(e_i) = 0$). МНК-оценки коэффициентов регрессии являются линейными комбинациями значений результативной переменной y_i с неслучайными коэффициентами, отсюда можно получить выражение ошибки коэффициента регрессии как комбинацию ошибок наблюдений и вычислить ее дисперсию (дисперсия суммы независимых величин равна сумме дисперсий, неслучайные множители возводятся в квадрат).

Так, для линейной однофакторной модели коэффициент регрессии вычисляется по формуле $b_1 = \frac{s_{xy}}{s_x^2} = \frac{1}{ns_x^2} \sum (x_i - \bar{x}) \cdot y_i$, откуда получаем выражение случайной ошибки коэффициента регрессии в виде $\frac{1}{ns_x^2} \sum (x_i - \bar{x}) \cdot e_i$, где x и s_x – не случайные.

Дисперсия этой ошибки равна $\hat{\sigma}_{b_1}^2 = \frac{\hat{\sigma}_e^2}{ns_x^2} = \frac{\hat{\sigma}_{\bar{y}}^2}{s_x^2} = \frac{s_y^2}{s_x^2} \cdot \frac{1-r_{xy}^2}{n-2}$.

Теперь рассмотрим случайную дисперсию расчетных значений $y_p = \bar{y} + b_1(x - \bar{x})$ как линейную комбинацию случайных величин \bar{y} и b_1 с уже известными дисперсиями $\sigma_{\bar{y}}^2$, $\sigma_{b_1}^2$ и нулевой ковариацией $\sigma_{\bar{y}, b_1} = 0$.

В результате получим $\hat{\sigma}_{\bar{y}(x)}^2 = \hat{\sigma}_{\bar{y}}^2 \left[1 + \frac{(x - \bar{x})^2}{s_x^2} \right]$.

Как функция x это равенство является уравнением сопряженной гиперболы $Y^2 = b^2 \left(1 + \frac{X^2}{a^2} \right)$, сдвинутой вправо на величину \bar{x} с полуосями $a = s_x$ и $b = \hat{\sigma}_{\bar{y}} = s_y \sqrt{\frac{1-r_{xy}^2}{n-2}}$.

Достаточно построить только каркас доверительной полосы на линию регрессии.

На интервале $\bar{x} \pm s_x$ ширина доверительной полосы практически постоянна и равна удвоенной ошибке среднего $t_{0,05} \hat{\sigma}_{\bar{y}} \approx 2 \hat{\sigma}_{\bar{y}}$; строим на линии регрессии параллелограмм с такими границами; далее доверительная полоса расширяется, приближаясь к продолжениям диагоналей построенного параллелограмма.

14. Поясните способ выбора формы связи. Продемонстрируйте возможности обобщенной линейной модели, нелинейной по аргументам, но линейной по параметрам. Рассмотрите стандартные преобразования переменных (логарифмирование и переход к обратным величинам).

Для МНК важно, чтобы форма связи была *линейной относительно параметров* (а не относительно x), тогда система нормальных уравнений для определения параметров будет линейной.

Общий вид двухпараметрических моделей, линейных относительно параметров, имеет вид:

$$Y = a + b X,$$

где $Y = F(x, y)$; $X = \Phi(x, y)$.

Если эмпирические точки в преобразованных координатах (X, Y) не группируются вокруг некоторой прямой, то принятая форма связи должна быть отвергнута (надо подбирать другую, более подходящую).

Чаще всего применяется или логарифмирование переменных, или переход к обратным величинам, что дает дополнительно 7 нелинейных моделей, приведенных ниже.

Двухпараметрические зависимости $Y(y) = a + b X(x)$

Преобразования	$X = x$	$X = \ln x$	$X = 1/x$
$Y = y$	Линейная $y = a + b x$	Логарифмическая $y = a + b \ln x$	Гиперболическая 1 $y = a + b / x$
$Y = \ln y$ $g^2 = y^2$	Показательная $\ln y = a + b x$ $y = A e^{b x}$	Степенная $\ln y = a + b \ln x$ $y = A x^b$	S-образная $\ln y = a + b / x$ $y = A e^{b / x}$
$Y = 1/y$ $g^2 = y^4$	Гиперболическая 2 $1/y = a + b x$ $y = \frac{1}{a+bx}$		Гиперболическая 3 $1/y = a + b/x$ $y = \frac{x}{ax+b}$

Если применяется функциональное преобразование результативной переменной, желательно во всех расчетах заменить обычные средние на средние взвешенные с весовой функцией g^2 , то есть вместо обычных средних \bar{X} , \bar{Y} , $\overline{X^2}$, $\overline{Y^2}$, \overline{XY} надо использовать взвешенные средние:

$$X_{cp} = \frac{\sum g^2 X}{\sum g^2}; Y_{cp} = \frac{\sum g^2 Y}{\sum g^2}; (XX)_{cp} = \frac{\sum g^2 X^2}{\sum g^2};$$

$$(YY)_{cp} = \frac{\sum g^2 Y^2}{\sum g^2}; (XY)_{cp} = \frac{\sum g^2 XY}{\sum g^2}.$$

Использованная литература

Вентцель Е. С. Теория вероятностей / Вентцель Е. С. – М. : Наука, 1969. – 576 с.

Смирнов Н. В. Курс теории вероятности и математической статистики для технических приложений / Н. В. Смирнов, И. В. Дунин-Барковский. – М. : Наука, 1969. – 512 с.

Гмурман В. Е. Теория вероятностей и математическая статистика / Гмурман В. Е. – М. : Высшая школа, 2000. – 480 с.

Егоршин А. А. Корреляционно-регрессионный анализ. Курс лекций и лабораторных работ / А. А. Егоршин, Л. М. Малярец. – Х. : Основа, 1998. – 208 с.

Содержание

Введение.....	3
1. Основные понятия теории вероятностей	5
Вопросы для самопроверки.....	14
2. Теоремы о вероятностях	15
Теорема умножения вероятностей	15
Теорема о полной вероятности	21
Теорема (формула) Байеса.....	23
Теорема сложения вероятностей	23
Принцип практической невозможности редких событий.....	24
Вопросы для самопроверки.....	25
3. Случайные величины.....	26
Дискретная случайная величина.....	26
Числовые характеристики случайных величин	27
Свойства математического ожидания.....	31
Свойства дисперсии	33
Вопросы для самопроверки.....	36
4. Распределение Бернулли – Пуассона – Лапласа	37
Распределение Бернулли	37
Распределение Пуассона	42
Вывод формулы для расчета вероятностей распределения Пуассона.....	44
Вопросы для самопроверки.....	46
5. Распределение Лапласа.....	47
Интегральная теорема Лапласа.....	48
Три основных формы интегральной теоремы Лапласа	50
Доказательство локальной теоремы Лапласа	54
Вопросы для самопроверки.....	56
6. Непрерывная случайная величина.....	56
Нормальный закон распределения Гаусса.....	61
Показательный, или экспоненциальный, закон распределения	65
Квантили распределения	66
Некоторые соображения, приводящие к нормальному закону	67
Вопросы для самопроверки.....	68
7. Предельные теоремы теории вероятностей.....	69
Закон больших чисел	69
Центральная предельная теорема	71
Композиция распределений случайных величин.....	72
Функции случайного аргумента	75
Вывод формулы композиции двух непрерывных величин.....	78
Вопросы для самопроверки.....	78
8. Система случайных величин.....	79
Закон распределения дискретной двумерной величины.....	79
Характеристики дискретной двумерной случайной величины	80
Закон распределения непрерывной двумерной величины.....	82

Характеристики непрерывной двумерной величины	84
Двумерный нормальный закон	85
Вопросы для самопроверки	87
9. Проблемы математической статистики	89
Способы составления выборочных подсовокупностей	90
Статистическое оценивание	91
Вопросы для самопроверки	96
10. Свойства статистических оценок	97
Оценка параметров распределения	100
Статистические критерии	103
Вопросы для самопроверки	107
11. Критерии согласия	108
Критерий согласия Пирсона	108
Критерий согласия Колмогорова – Смирнова	113
Интервальные оценки характеристик и параметров	114
Вывод дифференциальной функции распределения Пирсона	115
Вопросы для самопроверки	117
12. Проверка статистических гипотез	118
Распределение Стьюдента	118
Интервальная оценка для математического ожидания	119
Проверка гипотезы о равенстве центров двух совокупностей	121
Сравнение двух дисперсий	125
Вывод дифференциальной функции распределения Стьюдента	127
Вывод дифференциальной функции распределения Фишера	128
Вопросы для самопроверки	129
13. Дисперсионный анализ	130
Сравнение групп	130
Ранговый дисперсионный анализ Краскала – Уоллиса	138
Дополнение к выводу формул Краскала – Уоллиса	142
Вопросы для самопроверки	143
14. Регрессионный анализ	144
Метод наименьших квадратов (МНК)	146
Пример расчета МНК-оценок параметров	148
Оценка тесноты принятой формы связи	150
Однофакторная линейная зависимость	152
Нелинейные двухпараметрические модели	154
Вопросы для самопроверки	155
15. Проблема значимости и адекватности регрессионной модели	156
Оценка значимости регрессионной модели	156
Оценка значимости корреляционной связи	157
Проверка адекватности модели	160
Таблицы сопряженности и коэффициенты контингенции	163
Коэффициент ранговой корреляции Спирмена	166
Вывод формулы для коэффициента ранговой корреляции Спирмена	167
Вопросы для самопроверки	169
16. Линейный регрессионный анализ в стандартизованных переменных	170

Способы составления многофакторных моделей	176
Коэффициенты частной корреляции	178
Вывод формул для дисперсий коэффициентов регрессии и расчетных значений	181
Вопросы для самопроверки	184
17. Случайные функции	185
Характеристики случайных функций	186
Стационарные случайные функции	190
Эргодичные стационарные процессы	194
Вопросы для самопроверки	200
Лабораторная работа 1. Изучение распределения Бернулли средствами Excel	201
Лабораторная работа 2. Асимптотические формулы Пуассона и Лапласа	207
Изучение распределения Пуассона средствами Excel	207
Изучение распределения Лапласа средствами Excel	209
Вопросы для самопроверки	213
Лабораторная работа 3. Обработка данных наблюдений	214
Описательная статистика	215
Лабораторная работа 4. Проверка статистических гипотез	221
Критерии согласия	221
Интервальная оценка математического ожидания.	225
Определение потребного объема выборки.	226
Интервальная оценка на генеральную дисперсию.	226
Нормальная вероятностная кривая.	227
Лабораторная работа 5. Нестандартная графика	229
Определение квартилей и выявление выбросов.	229
Построение блочной диаграммы Тьюки	230
Первый способ построения комбинированных диаграмм	231
Второй способ построения комбинированных диаграмм.	234
Вопросы для самопроверки	235
Лабораторная работа 6. Анализ корреляционных связей	237
Представление данных.	238
Двойная группировка данных	239
Расчет параметров линейной модели.	242
Лабораторная работа 7. Проблемы тесноты, значимости и адекватности	245
Дисперсионный анализ. Оценка тесноты и значимости корреляционной связи.	245
Доверительные интервалы на центры групп.	247
Оценка тесноты и значимости линейной модели	249
Проверка адекватности (линейности) модели	249
Лабораторная работа 8. Специальные вопросы регрессионного анализа	251
Выбор нелинейной формы связи	251
Доверительные интервалы на расчетные значения.	254
Таблицы сопряженности и коэффициенты контингенции	255
Вопросы для самопроверки	257
Теория вероятностей в вопросах и ответах	259
Математическая статистика в вопросах и ответах	275
Регрессионный анализ в вопросах и ответах	286
Использованная литература	301

НАВЧАЛЬНЕ ВИДАННЯ

Малярець Людмила Михайлівна

Єгоршин Олександр Олександрович

ТЕОРІЯ ЙМОВІРНОСТЕЙ ТА МАТЕМАТИЧНА СТАТИСТИКА

**Навчально-практичний посібник
для іноземних студентів
галузі знань 0305 "Економіка та підприємництво"**

(рос. мовою)

Відповідальний за випуск Малярець Л. М.

Відповідальний редактор Сєдова Л. М.

Редактор Семенова І. М.

Коректор Мартовицька-Максимова В. А.

Викладено теоретичний матеріал з навчальної дисципліни, що сформований за лекціями, кожна з яких супроводжується прикладами та запитаннями для самоперевірки. Розроблено лабораторні роботи для закріплення теоретичних знань і формування практичних навичок з теорії ймовірностей та математичної статистики.

Рекомендовано для всіх студентів, які вивчають теорію ймовірностей та математичну статистику.

План 2013 р. Поз. № 118-П.

Підп. до друку Формат 60 × 90 1/16. Папір MultiCopy. Друк Riso.

Ум.-друк. арк. 19,0. Обл.-вид. арк. 23,75. Тираж прим. Зам. №

Видавець і виготівник – видавництво ХНЕУ, 61166, м. Харків, пр. Леніна, 9а

*Свідомо про внесення до Державного реєстру суб'єктів видавничої справи
Дк № 481 від 13.06.2001 р.*