

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Лектор — Пулькин Игорь Сергеевич

Москва 2015

Литература

1. Вентцель Е.С. Теория вероятностей — М.: Высшая школа, 1999. — 576 с.
2. Гмурман В. Е. Руководство к решению задач по теории вероятностей и математической статистике — М.: Высшая школа , 2004. — 404 с.
3. Гмурман В. Е. Теория вероятностей и математическая статистика. — М.: Высшая школа , 2003. — 479 с.
4. Ефимов А. В., Поспелов А. С. (ред.) Сборник задач по математике. Ч. 4. — М.: Издательство физико-математической литературы, 2003. — 432с.
5. Млодинов Л. (Не)совершенная случайность. М.: Livebook, 2013. — 352с,
6. Чжун К. Л., АйтСахлиа Ф. Элементарный курс теории вероятностей. — М.: Бином, 2014 — 454с.

Лекция 1. Классическое определение вероятности. Непосредственный подсчет шансов.

Теория вероятностей начиналась с задач, посвященных азартным играм. Эти задачи часто сводились к непосредственному подсчету шансов. Для игр использовались монеты, игральные кости, карты и т. д. Например, при бросании монеты она может упасть орлом или решкой. Если игрок ставит на орла, то у него один шанс из двух.

Для подсчета шансов используется классическое определение вероятности:

$$\text{Вероятность} = \frac{\text{число благоприятных исходов}}{\text{общее число равновозможных исходов}}.$$

Пример 1. Каковы шансы (какова вероятность) получить ровно одного орла при бросании трех монет?

Решение. Три монеты могут упасть восемью способами:

ooo

oor

opo

opp ✓

roo

rop ✓

pro ✓

rrr

Благоприятных исходов среди них три — они отмечены галочками. Поэтому вероятность равна $3/8$.

Следует обратить внимание на то, что исходы бросания должны быть именно равновозможными. Для каждой монеты таких исходов два, и в сочетании с каждым из них могут случиться два исхода на второй и два исхода на третьей монете. Всего получится $2^3 = 8$ исходов. Аналогично для n монет получится 2^n равновозможных исходов.

В дальнейшем мы встретимся с различными, в том числе

весьма изощренными, способами определения числа шансов, но есть много задач, в которых шансы поддаются непосредственному подсчету.

Пример 2. Каковы вероятности получить при бросании трех игральных костей в сумме 11 очков и 12 очков? Какая из этих вероятностей больше?

Решение. Благоприятных исходов много, для подсчета их количества пальцев рук уже не хватит, поэтому следует как-то упорядочить подсчет. Сумма 11 получится, например, как $1 + 4 + 6$. Если мы будем записывать различные подобные комбинации, то стоит учитывать сразу все возможные перестановки, то есть случаи, когда на какой-то кости — 1 очко, на какой-то другой — 4, а на третьей — 6. Таких перестановок, как нетрудно убедиться, будет шесть: 146, 164, 416, 461, 614, 641. Поэтому мы будем записывать все возможные наборы, дающие нужную нам сумму, а рядом будем писать, сколькими способами такие наборы получаются. Чтобы не запутаться, все наборы будем записывать в возрастающем порядке. Для 11 и 12 очков получим:

11	12
1 4 6 6	1 5 6 6
1 5 5 3	2 4 6 6
2 3 6 6	2 5 5 3
2 4 5 6	3 3 6 3
3 3 5 3	3 4 5 6
3 4 5 3	4 4 4 1

В результате получаем, что вероятности таковы:

$$P(11) = \frac{27}{216}; \quad P(12) = \frac{25}{216}.$$

Пример 3. А какова вероятность получить при бросании трех костей 13 очков?

Легко подсчитать, эта вероятность равна $21/216 = 7/72$.

В предыдущем примере мы столкнулись с необходимо-

стью дать ответ на такой вопрос: сколькими способами можно расположить n различных предметов в определенном порядке. Иными словами: сколько перестановок возможно на множество из n предметов. При 3 предметах таких способов оказалось шесть. В общем случае число таких способов дается формулой

$$n! = 1 \cdot 2 \cdot \dots \cdot n.$$

Число $n!$ читается “эн — факториал”. Например, факториал числа 6 равен, как нетрудно убедиться, 720. В дальнейшем факториал будет использоваться во многих формулах, при этом будет принято соглашение, что $0! = 1$ (пустое множество можно упорядочить единственным способом).

Пример 4. Четырёхтомное сочинение расположено на книжной полке в случайном порядке. Чему равна вероятность того, что тома стоят в должном порядке справа налево или слева направо?

Ответ — 1/12.

Наука о числе возможностей выбора называется комбинаторикой. Мы познакомимся с элементами этой науки.

Основное правило комбинаторики. Если выбор надо произвести дважды, причем при первом выборе есть m возможностей, а при втором — n возможностей, то общее число возможностей для двойного выбора равно mn . Аналогичная формула есть и для k -кратного выбора.

Покажем, например, как с помощью этого правила получается число перестановок. При расположении n предметов в определенном порядке мы расставляем эти предметы: какой-то на первое место, какой-то другой — на второе, и так далее. Предмет для первого места мы можем выбрать n способами, а на второе место останется только $n - 1$ кандидат, ведь один предмет мы уже использовали. На третье место будет уже $n - 2$ кандидата, и так далее. В результате

получим уже знакомую нам формулу.

Число сочетаний. Связь с биномиальными коэффициентами. Треугольник Паскаля.

Пример 5. Есть пять различных предметов, например, 1, 2, 3, 4, 5. Сколькими способами можно из них выбрать два?

Решение. Таких способов 10:

12 23 34 45

13 24 35

14 25

15

Надо обратить внимание на то, что порядок выбора несуществен, то есть выбор 1 и 2 — то же самое, что и выбор 2 и 1.

Число способов, которым из n предметов можно выбрать k (без учета порядка следования предметов), называется числом сочетаний из n по k , и обозначается C_n^k . Выведем формулу для этого числа.

Сначала, однако, выведем формулу для числа способов, которым из n предметов можно выбрать k с учетом порядка следования предметов. Это нетрудно: на первом шаге мы можем выбрать любой из n предметов, на втором — любой из оставшихся $n - 1$, и так далее. На последнем, k -м, шаге мы можем выбирать из оставшихся на этот момент $n - k + 1$ предметов. В соответствии с основным правилом комбинаторики искомое число способов равно

$$A_n^k = n \cdot (n - 1) \cdot \dots \cdot (n - k + 1).$$

Полученное число A_n^k часто называют числом размещений из n по k . Формулу для него можно записать по-другому: домножив и разделив на $(n - k)!$, получим в числителе, как легко убедиться, произведение всех целых чисел от 1 до n ,

то есть $n!$. Поэтому

$$A_n^k = \frac{n!}{(n-k)!}.$$

Важный частный случай, когда $k = n$, нам уже знаком: упорядочить n предметов можно $A_n^n = n!$ способами.

Пример 6. Найти вероятность того, что в группе из n человек хотя бы у двух совпадают даты рождения (может быть, в разные годы). Для простоты расчетов високосными годами следует пренебречь.

Это — задача для самостоятельного решения. Хотя можно провести эксперимент. Какова, по мнению читателя, вероятность совпадения хотя бы одной пары дней рождения для 30 человек? для 40 человек? Сколько человек надо собрать, чтобы эта вероятность стала больше $1/2$?

Вернемся к числу сочетаний. Поскольку k предметов можно упорядочить $k!$ способами, каждый неупорядоченный выбор порождается $k!$ упорядоченными. Следовательно

$$C_n^k = \frac{A_n^k}{k!},$$

и окончательно

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

Числа сочетаний называют еще биномиальными коэффициентами, поскольку они появляются в разложении бинома

$$(x+y)^n = \sum_{k=0}^n C_n^k x^k y^{n-k}.$$

Нетрудно понять, почему так получается. Если левую часть представить как произведение

$$\underbrace{(x+y) \cdot (x+y) \cdots \cdots (x+y)}_{n \text{ раз}},$$

то произведение $x^k y^{n-k}$ получается, если из k таких скобок выбрать множитель x , а из остальных $n - k$ — множитель

y . Какой коэффициент будет при $x^k y^{n-k}$? Именно такой, сколькими способами можно этот выбор сделать, то есть как раз C_n^k .

Для подсчета чисел сочетаний, кроме формулы (1), можно использовать треугольник Паскаля:

$$\begin{array}{ccccccc} & & & 1 & & & \\ & & & 1 & 1 & & \\ & & & 1 & 2 & 1 & \\ & & & 1 & 3 & 3 & 1 \\ & & & 1 & 4 & 6 & 4 & 1 \\ & & & 1 & 5 & 10 & 10 & 5 & 1 \\ & & & 1 & 6 & 15 & 20 & 15 & 6 & 1 \\ & & & 1 & 7 & 21 & 35 & 35 & 21 & 7 & 1 \end{array}$$

Этот треугольник строится по простому правилу: стороны слева и справа состоят из одних единиц, а каждое число в середине равно сумме двух чисел, стоящих над ним.

Нетрудно сообразить, что каждая строка треугольника Паскаля представляет собой коэффициенты разложения бинома $(x+y)^n$. Например, в третьей строке стоят числа 1, 3, 3, 1, что соответствует хорошо известной формуле

$$(x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3.$$

Эта строка — третья, а не четвертая, как могло бы показаться, потому что номера строк начинаются с нуля: строка из одной единицы — нулевая строка.

Пример 7. Треугольник Паскаля тоже может быть использован не только для иллюстрации, но и для строгих математических доказательств. В качестве примера докажем формулу

$$\sum_{i=k}^N C_i^k = C_{N+1}^{k+1}.$$

Доказательство. Сумма в левой части — это отрезок диаго-

нали треугольника Паскаля. Эти числа выделены жирным шрифтом на рисунке. Правая часть также выделена жирным шрифтом и обведена в рамочку.

			1								
			1	1	1						
			1	2	1						
			1	3	3	1					
			1	4	6	4	1				
			1	5	10	10	5	1			
			1	6	15	20	15	6	1		
1	7	21	35	35	21	7	1				

Формула теперь легко доказывается индукцией по N . Действительно, при $N = k$ обе части равны 1. Пусть формула доказана для некоторого отрезка диагонали, как на рисунке. Добавляя к диагонали следующее число (на рисунке это число 15), мы получим, что сумма стала равной сумме добавленного числа и числа, взятого в рамочку. В соответствии с правилом построения треугольника Паскаля — это число, стоящее правее и ниже добавленного (на рисунке это число 35). Что и требовалось доказать.

Укажем несколько формул для чисел сочетаний. Все они легко доказываются исходя из формулы бинома, но мы приведем комбинаторные доказательства.

$$1. \ C_n^k = C_n^{n-k}.$$

Число способов отобрать k элементов из n возможных, равно, разумеется, числу способов оставить остальные $n - k$ неотобранными.

$$2. \ C_n^k = C_{n-1}^{k-1} + C_{n-1}^k \ (\text{правило построения треугольника Паскаля}).$$

Для доказательства пометим один из n предметов. Сколькими способами можно выбрать k из n предметов? Если в число выбранных входит помеченный предмет, то таких

способов C_{n-1}^{k-1} , а если не входит, то C_{n-1}^k .

$$3. \sum_{k=0}^n C_n^k = 2^n.$$

Поскольку C_n^k — не что иное, как число подмножеств из k элементов у множества из n элементов, доказываемая формула утверждает, что число всех подмножеств множества из n элементов равно 2^n . Действительно, выбирая некоторое подмножество, мы для каждого из n элементов должны сделать выбор из двух вариантов: войдет он в подмножество или нет. Поскольку выбор делается n раз, в соответствии с основным правилом комбинаторики получим, что всего подмножеств (включая пустое) будет 2^n .

Пример 8. Задача о выборке. В урне находятся n шаров, из них k белых, а остальные $n - k$ — красные. Из урны извлекают m шаров. Какова вероятность того, что будет извлечено r белых шаров и $m - r$ красных?

Решение. Выбрать k шаров из n возможных можно C_n^k способами, причем все эти способы равновозможны. Теперь подсчитаем число благоприятных исходов. Выбрать r белых шаров из k можно C_r^k способами, а выбрать $m - r$ красных шаров из $n - k$ можно C_{m-r}^{n-k} способами. В соответствии с основным правилом комбинаторики число благоприятных исходов равно $C_r^k \cdot C_{m-r}^{n-k}$, а искомая вероятность равна

$$p = \frac{C_r^k \cdot C_{m-r}^{n-k}}{C_n^k}.$$

Модели случайного выбора

С двумя моделями случайного выбора мы уже познакомились. Это

а). Упорядоченный выбор без возвращения. Число способов выбрать таким образом k предметов из n , как мы установили, равно A_n^k .

б). Неупорядоченный выбор без возвращения. Число способов выбрать таким образом k предметов из n , как мы

установили, равно C_n^k .

в). Упорядоченный выбор с возвращением. Это новая модель случайного выбора. Можно представлять себе, что в урне лежит n различных предметов, и мы k раз извлекаем случайным образом предмет из урны, отмечаем, какой это был предмет, после чего возвращаем его в урну. Поскольку каждый раз мы выбираем один из n предметов, в соответствии с основным правилом комбинаторики число исходов равно n^k .

г). Неупорядоченный выбор с возвращением. Здесь мы не обращаем внимания на то, в каком порядке выбирались предметы, а следим только за тем, сколько раз был выбран каждый предмет. Вывод формулы для числа способов такого выбора является образцом применения комбинаторного мышления.

Для каждого способа выбора составим строку из n клеточек, где крестиком будем отмечать, что соответствующий предмет был выбран, а клеточки будем разделять вертикальными черточками. Например, если мы выбираем 4 раза из 5 предметов, то строка

$xx| |x| |x$

означает, что первый предмет был выбран 2 раза, 3-й и 5-й — по одному разу, а 2-й и 4-й не были выбраны ни разу. Мы не потеряем в информативности, если сотрем все пробелы в этой строке:

$xx||x||x$

Обратно, любая строка из 4 крестиков и 4 вертикальных черточек будет определять некоторый вариант выбора. Например, строка

$||| |xxxx$

соответствует тому варианту выбора, когда все 4 раза был выбран 5-й предмет.

Таким образом, число вариантов неупорядоченного выбо-

ра с возвращением k раз из n возможных предметов находится во взаимно однозначном соответствии с числом строк из k крестиков и $n - 1$ палочек. Ясно, что таких строк ровно столько, сколькими способами можно из $n + k - 1$ предметов выбрать k (крестиков, а остальные будут палочками), то есть C_{n+k-1}^k .

Пример 9. Пусть мы выбираем дважды из 4 предметов. Сколько будет способов такого выбора при различных моделях выбора?

Решение.

а). Упорядоченный выбор без возвращения. $A_4^2 = 12$.

- (1 2) (1 3) (1 4)
- (2 1) (2 3) (2 4)
- (3 1) (3 2) (3 4)
- (4 1) (4 2) (4 3)

б). Неупорядоченный выбор без возвращения. $C_4^2 = 6$.

- (1 2) (1 3) (1 4)
- (2 3) (2 4)
- (3 4)

в). Упорядоченный выбор с возвращением. $4^2 = 16$.

- (1 1) (1 2) (1 3) (1 4)
- (2 1) (2 2) (2 3) (2 4)
- (3 1) (3 2) (3 3) (3 4)
- (4 1) (4 2) (4 3) (4 4)

г). Неупорядоченный выбор с возвращением. $C_5^2 = 10$.

- (1 1) (1 2) (1 3) (1 4)
- (2 2) (2 3) (2 4)
- (3 3) (3 4)
- (4 4)

В задачах на нахождение вероятностей следует четко представлять себе, каковы все равновозможные исходы опыта. Позже об этом еще будет сказано при обсуждении пространства элементарных исходов, а сейчас рассмотрим при-

мер.

Пример 10. Из n предметов выбирают k с возвращением. Найти вероятность того, что все выбранные предметы различны. При взгляде на предыдущий пример ясно, что есть два возможных решения:

1) если рассматривать упорядоченный выбор, то

$$p = \frac{A_n^k}{n^k},$$

2) если выбор неупорядоченный, то

$$p = \frac{C_n^k}{C_{n+k-1}^k}.$$

Эти формулы, однако, дают разные ответы. Какая из них верна?

Решение.

При реальном выборе по одному предмету выбирают k раз, то есть выбор упорядоченный. Например, если выбор делается дважды, то выбрать 1 и 2 предметы можно двумя способами: сначала 1, затем 2, или наоборот. А дважды выбрать первый предмет можно только одним способом. То есть вероятность выбрать 1 и 2 вдвое больше, чем выбрать два раза 1. При неупорядоченном же выборе эти вероятности равны. Мы уже встречались с этим в примере 2. Таким образом, верна первая формула.

Этот пример может создать представление, что неупорядоченный выбор в реальной жизни не встречается. Его действительно не так просто реализовать в урновой схеме, но в других случаях такой выбор вполне имеет право на существование. Например, в квантовой механике все электроны считаются одинаковыми, неразличимыми. И тогда задача о заполнении электронных уровней приводит именно к неупорядоченному выбору: важно только то, сколько электронов на каких уровнях.

Задачи к лекции 1.

Задача 1. Сколько равновозможных исходов будет при бросании 2 игральных костей? 3 костей? n костей? Какова вероятность выпадения 7 очков при бросании 2 костей? 8 очков?

Задача 2. Кубик сначала окрасили, а затем распилили на тысячу одинаковых кубиков. Какова вероятность того, что наудачу выбранный кубик будет иметь как минимум две окрашенные грани?

Задача 3. Вам предлагают такую игру: вы кидаете две монеты, а ваш противник — одну. При этом вы выигрываете, если у вас выпадет больше орлов, чем у вашего противника, в противном случае — проигрываете. Какова вероятность вашего выигрыша?

Задача 4. Раньше в поездах, на вокзалах иногда предлагали сыграть в такую азартную игру. Игрок зажимает в кулаке носовой платок так, что наружу торчат только 4 угла. Вы выбираете два угла и тяните за них. Если вы ухватились за противоположные углы платка — вы выиграли, а если за смежные — проиграли. Какова вероятность вашего выигрыша?

Задача 5. На шахматную доску случайным образом ставят белого короля и черную ладью. Какова вероятность, что белый король окажется под шахом?

Задача 6. А какова вероятность получить при бросании трех костей 13 очков?

Задача 7. Секретарша раскладывает четыре письма по четырём конвертам с адресами случайным образом. 1) Какова вероятность, что ни одно письмо не попадёт по назначению? Какова вероятность, что по назначению попадёт 2) ровно одно письмо? 3) ровно два? 4) ровно три? 5) все четыре?

Задача 8. Найти вероятность того, что в группе из n

человек хотя бы у двух совпадают даты рождения (может быть, в разные годы). Для простоты расчетов високосными годами следует пренебречь. Чему должно быть равно n , чтобы вероятность совпадения стала больше $1/2$?

Задача 9. Доказать приведенные формулы, исходя из формулы бинома.

1. $C_n^k = C_n^{n-k}$.
2. $C_n^k = C_{n-1}^{k-1} + C_{n-1}^k$ (правило построения треугольника Паскаля).
3. $\sum_{k=0}^n C_n^k = 2^n$.

Задача 10. Написать и доказать “зеркально-симметричную” формулу к примеру 7.

Задача 11. Доказать формулу $\sum_{k=0}^n (-1)^k C_n^k = 0$. Эта формула легко доказывается на основании формулы бинома. Легкое комбинаторное доказательство получается при нечетных n . Можно ли придумать комбинаторное доказательство для четных n ?

Задача 12. Случайным образом выбирают 4 раза цифру от 0 до 9. Какова вероятность того, что все 4 выбранные цифры окажутся разными?

Задача 13. Студент знает 30 из 40 вопросов экзаменационных билетов. В билете 3 вопроса. Какова вероятность того, что студент знает все вопросы своего билета?

Задача 14. Трамвай состоит из переднего и прицепного вагонов. В трамвай случайным образом садятся 4 человека. Найти вероятности: а) все четверо окажутся в одном вагоне; б) и в переднем вагоне, и в прицепном окажутся по два человека.

Задача 15. Саша, Маша и еще 6 человек рассаживаются на стоящих в ряд 8 стульях. а) Какова вероятность того, что Саша и Маша окажутся сидящими рядом? б) Как изменится эта вероятность, если стульев будет не 8, а 12?

Задача 16. Саша, Маша и еще 6 человек рассаживаются за круглый стол. Какова вероятность того, что Саша и Маша окажутся сидящими рядом?

Задача 17. В урне 5 белых и 3 красных шара. Из урны случайным образом выбирают 2 шара. Какова вероятность того, что оба они окажутся белыми?

Задача 18. В урне 4 белых и 5 красных шаров. Из урны случайным образом выбирают 4 шара. Какова вероятность того, что среди них будут 2 белых и 2 красных?

Задача 19. У одного из игроков в преферанс на руках 4 бубны. Какова вероятность, что еще ровно одна бубна в прикупе?

Задача 20. У игрока в покер на руках 3 туза. Какова вероятность, что после обмена двух других карт у него будет комбинация “4 туза”? По правилам покера тузом можно объявить джокера.

Задача 21. k предметов раскладывают по n коробкам ($k < n$). Все предметы считаются эквивалентными, то есть два размещения предметов по коробкам различны, если различно число предметов в коробках. Найти число таких размещений.

Задача 22. Как и в предыдущей задаче, k предметов раскладывают по n коробкам ($k < n$), но теперь в каждую коробку можно поместить не более 2 предметов. Найти число таких размещений.

Лекция 2. Геометрическая вероятность. Аксиоматическое определение вероятности.

Классическое определение вероятности может применяться только в том случае, когда число возможных исходов конечно. В реальной жизни это не всегда так. В качестве первого примера рассмотрим классическую задачу о встрече.

Пример 11. Два студента условились встретиться в определенном месте между 12 и 13 часами дня. Пришедший первым ждет второго в течение 15 минут, после чего уходит. Найти вероятность того, что встреча состоится, если каждый студент наудачу выбирает момент своего прихода в промежутке от 12 до 13 часов.

Обозначим время прихода первого студента через x , а второго — через y . Будем считать, что $0 \leq x \leq 1$, $0 \leq y \leq 1$, то есть время прихода измеряется в часах. Например, если первый студент пришел в 12-45, то $x = 3/4$.

Здесь все возможные исходы занимают единичный квадрат на плоскости, их бесконечно много, и нам необходимо ввести новую формулу для определения вероятности. Будем называть эту вероятность геометрической вероятностью. Обозначим ее буквой P .

$$P = \frac{\text{площадь множества благоприятных исходов}}{\text{площадь множества всех равновозможных исходов}}.$$

Студенты встретятся, если выполнено неравенство $|y - x| < 1/4$. Раскрывая модуль, получим двойное неравенство

$$x - \frac{1}{4} < y < x + \frac{1}{4}.$$

Соответствующие множества изображены на рисунке. При этом множество благоприятных исходов не заштриховано — оставлено белым. А заштрихованные области соответству-

ют тем исходам, когда студенты не встретятся.

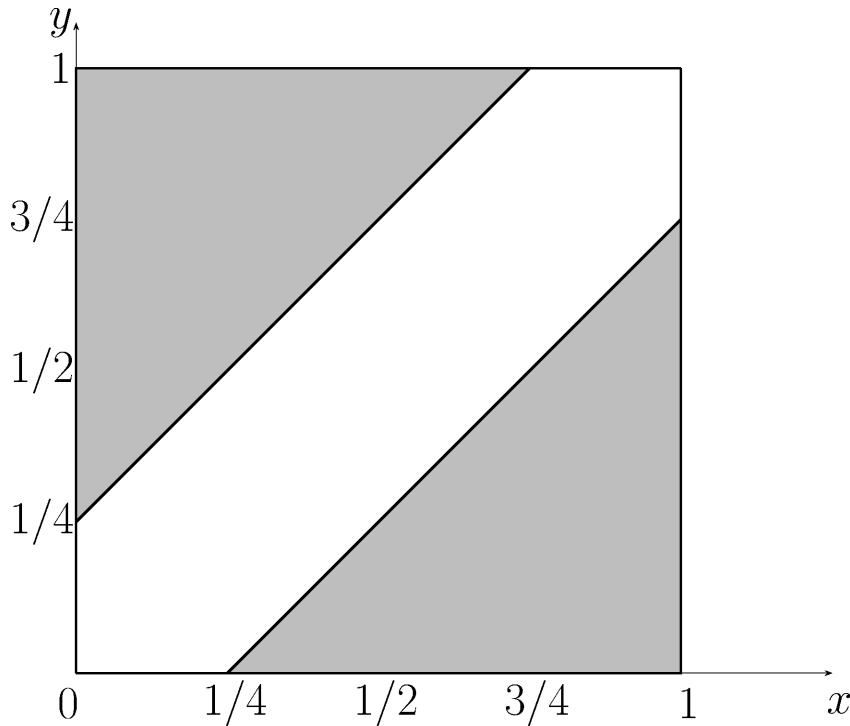


Рисунок 1. Задача о встрече.

В нашем примере площадь множества благоприятных исходов равна $7/16$, а площадь множества всех возможных исходов равна 1. Поэтому вероятность равна $7/16$.

Теперь становится понятно, что же такое вероятность. Это — мера, аналог площади.

Рассмотрим еще один пример.

Пример 12. Цилиндрический предмет с диаметром основания r и высотой h бросают на ровную поверхность. Найти вероятность того, что он упадет на боковую сторону.

Решение. Случайное направление падения — это луч, начинаящийся в геометрическом центре цилиндра. Можно считать, что цилиндр упадет на боковую поверхность в том и только в том случае, если этот луч пересекает боковую поверхность, а не основание цилиндра.

Пусть около цилиндра описана сфера. Тогда радиус этой

сферы равен

$$R_{\text{сф}} = \sqrt{\frac{h^2}{4} + r^2}.$$

Если луч проходит через одно из оснований цилиндра, то он проходит через один из сферических сегментов, опирающихся на одно из оснований. Если же луч проходит через боковую сторону, то он проходит через сферический пояс — оставшуюся часть сферы.

Площадь поверхности всей сферы равна

$$S_{\text{общ}} = 4\pi R_{\text{сф}}^2 = 4\pi \left(\frac{h^2}{4} + r^2 \right).$$

Осталось вычислить площадь сферического пояса. Он образован вращением части окружности

$$y = \sqrt{\frac{h^2}{4} + r^2 - x^2}$$

при x от $-h/2$ до $h/2$ вокруг оси $0x$. Поэтому, в соответствии с формулой площади поверхности вращения

$$S_{\text{бок}} = 2\pi \int_{-h/2}^{h/2} y \sqrt{1 + y'^2} dx = 2\pi h \sqrt{\frac{h^2}{4} + r^2}.$$

Следовательно, искомая вероятность равна

$$P = \frac{S_{\text{бок}}}{S_{\text{общ}}} = \frac{h}{\sqrt{h^2 + 4r^2}}.$$

В только что рассмотренном примере множеством всех равновозможных исходов были точки сферы, а множеством благоприятных исходов — точки шарового пояса.

Переходим к формализации введенных понятий.

Все возможные исходы вероятностного испытания составляют пространство элементарных исходов Ω .

Событие — подмножество пространства элементарных исходов. Для бесконечных множеств Ω событием может быть

не всякое подмножество, а только такое, которое имеет площадь.

Для событий определены операции взятия дополнения $\overline{A} = \Omega \setminus A$, объединения $A \cup B$ и пересечения $A \cap B$.

Множество событий вместе с определенными на нем операциями образует *алгебру событий*.

Приведем основные формулы алгебры событий.

Двойное отрицание: $(\overline{\overline{A}}) = A$.

Коммутативность: $A \cap B = B \cap A; A \cup B = B \cup A$.

Ассоциативность:

$$(A \cap B) \cap C = A \cap (B \cap C);$$

$$(A \cup B) \cup C = A \cup (B \cup C).$$

Дистрибутивность:

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C);$$

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C).$$

Формулы Моргана: $(A \cap B) = \overline{A} \cup \overline{B}; (\overline{A} \cup \overline{B}) = \overline{A \cap B}$.

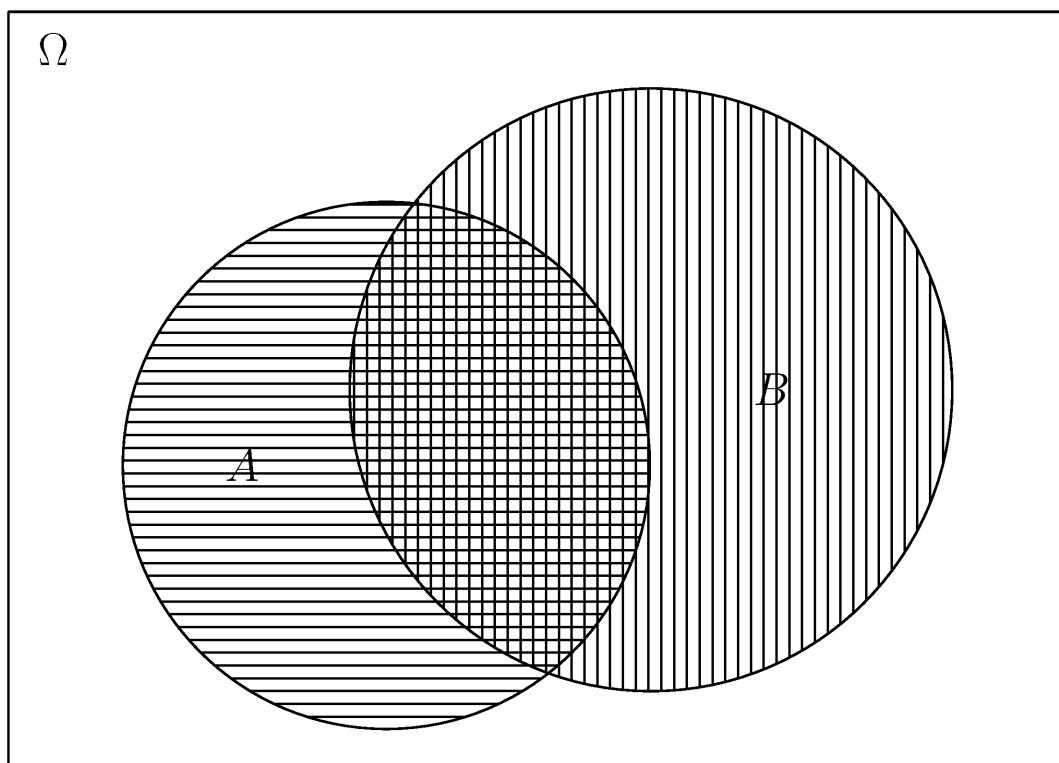


Рисунок 2. Диаграмма Венна.

Для иллюстрации (а также для доказательства) приведенных формул часто используются специальные рисунки,

они называются *диаграммы Венна* (или диаграммы Эйлера — Венна). Рисунок 2, например, иллюстрирует вторую из формул Моргана. Действительно, незакрашенная часть — это дополнение к объединению, то есть левая часть формулы. Глядя на ту же диаграмму, легко убедиться, что эта область действительно совпадает с пересечением дополнений.

Аксиоматическое определение вероятности.

Это аксиоматическое определение было введено советским математиком, академиком А. Н. Колмогоровым в 30-е годы XX века. Приводимая ниже аксиоматика А. Н. Колмогорова лежит в основе современной теории вероятностей.

Как уже упоминалось, вероятность — аналог площади. Однако можно вообразить такое хитро зазубренное множество, у которого нельзя определить площадь. Поэтому для аксиоматического определения вероятности следует ввести ограничения на рассматриваемые подмножества — рассматривать только измеримые множества.

Оказывается, можно корректно определить меру на таком наборе подмножеств, который образует *сигма-алгебру*.

Сигма-алгеброй \mathbb{S} называется такое множество подмножеств Ω , которое удовлетворяет следующим аксиомам:

- 1) для всякого $A \in \mathbb{S}$ его дополнение также принадлежит \mathbb{S} ;
- 2) для всякого конечного или счетного набора A_1, \dots, A_n, \dots множеств из \mathbb{S} их объединение также принадлежит \mathbb{S} .

Вероятность (вероятностная мера) — функция, определенная для каждого из подмножества сигма-алгебры и удовлетворяющая следующим аксиомам:

- 1) $P(A) \geq 0$;
- 2) $P(A \cup B) = P(A) + P(B)$ если $AB = \emptyset$;
- 3) $P(\Omega) = 1$.

Таким образом, вероятность — это мера на пространстве

элементарных исходов.

Такой взгляд на вероятность позволяет сразу написать несколько важных формул, которые далее будут часто использоваться.

1. Формулы вероятности суммы и произведения событий

$$P(A \cup B) = P(A) + P(B) - P(AB);$$

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

2. Вероятность противоположного события.

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

Здесь использована часто встречающаяся сокращенная запись $A \cap B = AB$.

В справедливости этих формул можно убедиться, рассматривая уже знакомую нам диаграмму Венна (рисунок 2). Например, для первой из приведенных формул: площадь объединения двух фигур, A и B , равна сумме их площадей минус площадь того куска, который посчитали дважды, то есть минус площадь их пересечения.

Для дальнейшего введем понятие независимости событий. В теории вероятностей события A и B называются независимыми, если $P(AB) = P(A) \cdot P(B)$. Обоснование такого определения будет дано в следующей лекции.

Пример 13. Найти вероятность того, что при четырех бросаниях игральной кости хотя бы раз выпадет шестерка.

Часто приходится встречаться с таким неправильным решением: при одном бросании шестерка выпадает с вероятностью $1/6$, поэтому при четырех будет $4/6$. Это, разумеется, неверно. Если так рассуждать, то при восьми, скажем, бросаниях вероятность будет больше единицы.

Только что выписанные формулы позволяют получить правильный ответ. Сначала надо рассчитать, какова вероятность того, что шестерка не выпадет ни разу. При каждом бросании вероятность того, что шестерка не выпадет, равна

$5/6$, поэтому при четырех бросаниях она не выпадет с вероятностью $(5/6)^4 = 625/1296$. Поэтому искомая вероятность равна $1 - 625/1296 = 671/1296$.

По сути, выведена часто применяемая формула вероятности хотя бы одного события. Если в каждом из n независимых испытаний некоторое событие может произойти с вероятностью p , то вероятность того, что оно произойдет хотя бы один раз, равна $1 - (1 - p)^n$.

Пример 14. Пусть элементы соединены в цепь так, как показано на рисунке 3. Цепь считается работающей, если от левого к правому концу может идти ток. Допустим, что надежность, то есть вероятность безотказной работы для элемента A равна $0,7$, а для элемента B — $0,8$, причем элементы выходят из строя независимо друг от друга. Какова вероятность безотказной работы цепи?

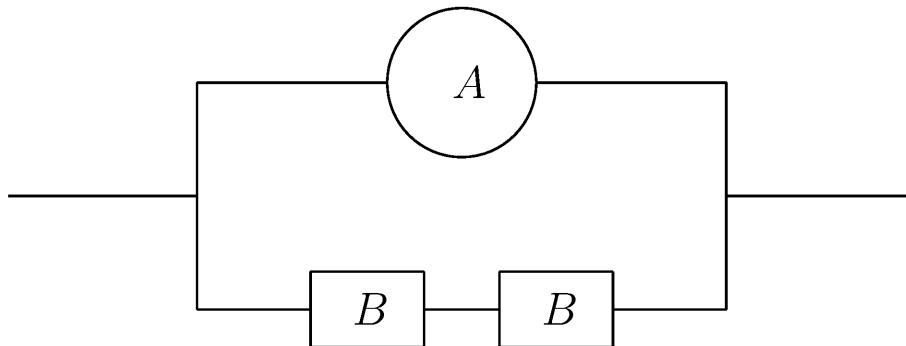


Рисунок 3. Схема цепи.

Какова вероятность выхода из строя нижнего фрагмента цепи? Для его выхода из строя нужно, чтобы отказал хотя бы один из элементов B . Поэтому вероятность безотказной работы этого фрагмента будет равна $0,8 \cdot 0,8 = 0,64$, а вероятность отказа — $0,36$.

Для верхнего фрагмента вероятность отказа равна $0,3$. Чтобы цепь отказалась, необходимо, чтобы отказали и верхний, и нижний фрагменты. Вероятность этого события равна $0,36 \cdot 0,3 = 0,108$. Следовательно, вероятность безотказ-

ной работы цепи равна $1 - 0,108 = 0,892$.

Задачи к лекции 2.

Задача 23. Винни-Пух и Пятачок раздобыли где-то бокал с шампанским конической формы. Винни-Пух, как честный медвежонок, решил выпить ровно половину содержимого. И выпил: до середины высоты бокала. Какую часть содержимого выпил Винни-Пух?

Задача 24. На стержне длины 1 случайным образом выбирают две точки, а затем стержень ломают в этих точках. Найти вероятность того, что из получившихся обломков можно составить треугольник.

Задача 25. На стержне длины 1 случайным образом выбирают точку, а затем стержень ломают в этой точке. Затем больший из получившихся обломков также ломают в случайно выбранной точке. Найти вероятность того, что из получившихся обломков можно составить треугольник.

Задача 26. Случайным образом выбраны три числа на отрезке $[0, 1]$. Какова вероятность того, что их сумма также не превосходит 1?

Задача 27. Случайным образом выбраны два числа p и q на отрезке $[0, 1]$. Найти вероятность того, что квадратное уравнение $x^2 + px + q = 0$ имеет два действительных корня. Найти вероятность того, что это квадратное уравнение имеет ровно один действительный корень.

Задача 28. На окружности случайным образом выбирают три точки. Найти вероятность того, что треугольник с вершинами в этих точках — остроугольный.

Задача 29. (Задача Льюиса Кэрролла). В ожесточенном бою не менее 70% рыцарей потеряли один глаз, не менее 75% — одно ухо, не менее 80% — одну руку, и не менее 85% — одну ногу. Каково минимальное число потерявших одновременно глаз, ухо, руку и ногу?

Задача 30.(*) Написать формулу, выражающую вероятность объединения n событий A_1, \dots, A_n через вероятности этих событий и вероятности их пересечений. Для простоты рассмотреть сначала случай $n = 3$.

Задача 31. Жюри состоит из трех судей. Первый и второй принимают правильное решение с вероятностью p , а третий для принятия решения подбрасывает монету. Окончательное решение принимается большинством голосов. Какова вероятность того, что жюри примет правильное решение?

Задача 32. Жюри состоит из трех судей. Каждый из них принимает правильное решение с вероятностью p . Каким должно быть p , чтобы это жюри принимало правильное решение с вероятностью большей, чем жюри из предыдущей задачи?

Задача 33. Жюри состоит из трех судей. Первый и второй принимают правильное решение с вероятностью p , а третий поступает следующим образом. Если мнения первых двух совпадают, то он к ним присоединяется, а если отличаются — бросает монетку. Какова вероятность того, что это жюри примет правильное решение?

Задача 34. За некоторый промежуток времени амеба может погибнуть с вероятностью $1/4$, выжить с вероятностью $1/4$ и разделиться на две с вероятностью $1/2$. В следующий такой же промежуток с любой амебой независимо от ее “возраста” происходит то же самое. Если первоначально была одна амеба, то сколько амеб и с какими вероятностями будут существовать к концу второго промежутка времени?

Задача 35. Двое игроков поочередно бросают монету. Выигрывает тот из них, у кого первым выпадет орел. Найти вероятность выигрыша первого игрока.

Задача 36. Как и в предыдущей задаче, двое игроков

бросают монету, но после каждого бросания первого игрока второй имеет право на два бросания. Как и ранее, выигрывает тот из них, у кого первым выпадет орел. Найти вероятность выигрыша первого игрока при этих условиях.

Задача 37. (Задача Пачоли о разделе ставки) Двое равносильных игроков играют матч до 6 побед в игру, где нет ничьих, то есть каждая партия может кончиться только победой одного из игроков (во времена Лука Пачоли, в XIV — начале XV веков, скорее всего, это была игра в орлянку). Игроки были вынуждены прервать игру при счете 5:3 в пользу одного из них. Как по справедливости должна быть разделена ставка между ними?

Задача 38. n студентов пришли на семинар по борьбе с забывчивостью. После семинара никто из них не смог вспомнить, в какой шляпе он пришел, и поэтому каждый из них выбрал шляпу с вешалки случайным образом. Найти вероятность p_n того, что никто из них не выбрал свою шляпу. Найти $\lim_{n \rightarrow \infty} p_n$.

Указание. Можно воспользоваться результатом задачи 30.

Задача 39. Троє стреляють в мишень. Вероятность того, что попадет A , равна 0,2, B — 0,3, C — 0,4. Найти вероятность того, что мишень будет поражена.

Задача 40. Для разрушения моста достаточно попадания одной авиабомбы. Найти вероятность того, что мост будет разрушен, если на него сбросили 4 авиабомбы, вероятности попадания которых равны 0,3, 0,4, 0,6 и 0,7.

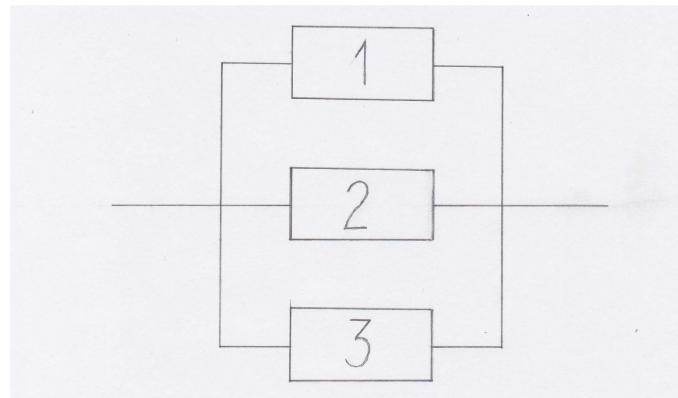
Задача 41. Сколько раз надо подбросить монету, чтобы с вероятностью не менее 0,9 хотя бы раз выпали два орла подряд? три орла подряд?

Задача 42. Сколько раз надо бросить игральную кость, чтобы вероятность того, что хотя бы один раз выпадет шестерка, стала больше 0,6?

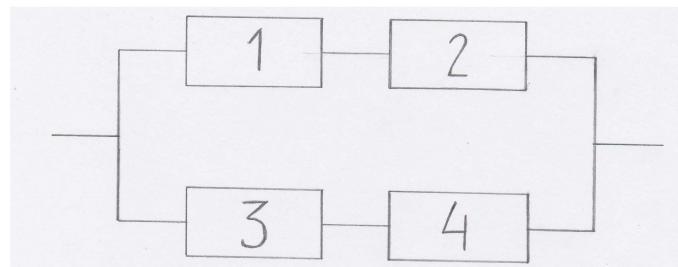
Задача 43. Сколько раз нужно бросить пару игральных костей, чтобы вероятность того, что хотя бы раз выпадет пара шестерок, стала больше 0,5? больше 0,9?

В следующих задачах приведены схемы соединения элементов, образующих цепь. Цепь считается работающей, если от левого к правому концу может идти ток. Допустим, что надежность, то есть вероятность безотказной работы для элемента с номером i равна p_i , причем элементы выходят из строя независимо друг от друга. Какова вероятность безотказной работы цепи?

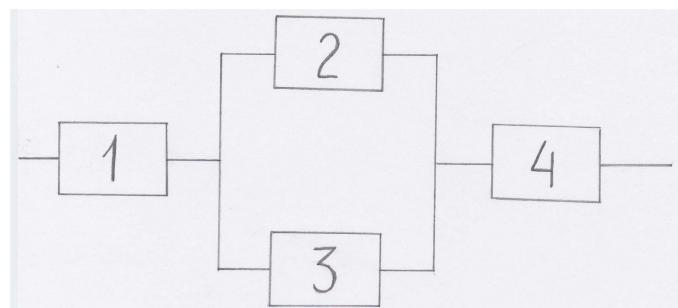
Задача 44.



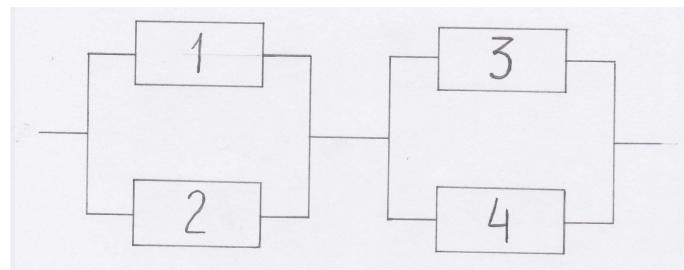
Задача 45.



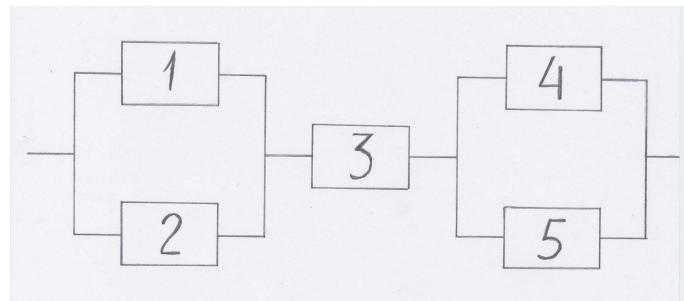
Задача 46.



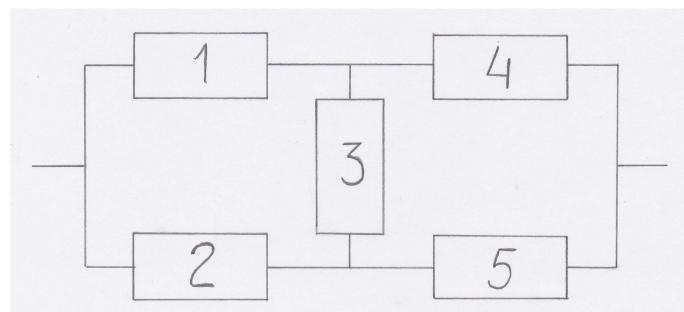
Задача 47.



Задача 48.



Задача 49.



Лекция 3. Формула полной вероятности. Формула Байеса.

Условной вероятностью события A при условии, что событие B произошло, называется

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

Поскольку известно, что событие B произошло, оно становится всем пространством элементарных исходов. А множеством благоприятных исходов становится пересечение AB .

Пример 15. Вероятность попасть в самолет равна 0,4, а вероятность его сбить равна 0,1. Найти вероятность того, что при попадании в самолет он будет сбит.

Решение. Пусть событие A — попали, B — самолет сбит. Тогда $P(A) = 0,4$, $P(B) = 0,1$ и $P(B|A) = P(AB)/P(A) = P(B)/P(A) = 0,25$.

Напомним определение независимых событий из предыдущей лекции:

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

Тогда из определения условной вероятности сразу следует

$$P(A|B) = P(A)$$

Теперь понятно, почему именно такие события называют независимыми — вероятность события A не изменилась от того, что событие B произошло.

Введенное определение позволяет вывести несколько важных и часто применяемых формул.

Пусть некоторое событие может осуществиться несколькими способами. Формально, будем считать, что пространство элементарных исходов разбито на несколько непересекающихся подмножеств H_1, \dots, H_n , объединение которых совпадает с Ω . Эти подмножества принято называть *гипотезами*.

Пусть для события A известны вероятности гипотез $P(H_1), \dots, P(H_n)$ и условные вероятности $P(A|H_1), \dots, P(A|H_n)$. Тогда справедлива формула

$$A = AH_1 \cup AH_2 \cup \dots \cup AH_n,$$

из которой легко следует формула

$$P(A) = P(A|H_1) \cdot P(H_1) + \dots + P(A|H_n) \cdot P(H_n),$$

называемая формулой полной вероятности.

Иллюстрация приведена на рисунке 4.

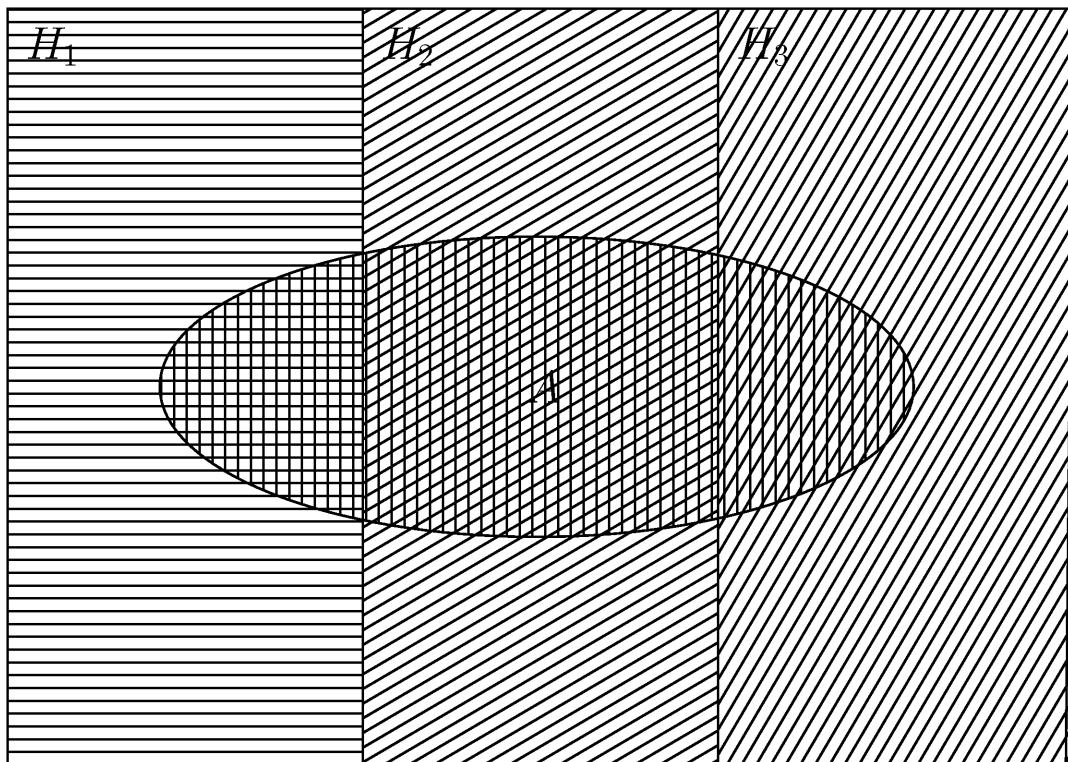


Рисунок 4. Формула полной вероятности.

Эта часто используемая формула аналогична формуле взвешенного среднего.

Пример 16. В фирме работают 40% мужчин. В проекте заняты 25% мужчин фирмы и 15% женщин. Сколько процентов сотрудников фирмы занято в проекте?

Решение. Эта задача на формулу взвешенного среднего, но ей легко придать вероятностный смысл: какова вероятность, что наугад выбранный сотрудник занят в проекте? Запишем вероятности:

$$P(H_1) = 0,4, P(H_2) = 0,6,$$

$$P(A|H_1) = 0,25, P(A|H_2) = 0,15.$$

Поэтому по формуле полной вероятности $P(A) = 0,19$.

Пример 17. Пусть по вражескому самолету выпущено три ракеты. Вероятности попаданий равны:

ни одного попадания — 0,2;

ровно одно попадание — 0,3;

ровно два попадания — 0,3;

ровно три попадания — 0,2.

При одном попадании самолет будет сбит с вероятностью 0,2, при двух — с вероятностью 0,6, а при трех — с вероятностью 1. Найти вероятность того, что самолет будет сбит.

Решение. Обозначим H_0 событие, состоящее в том, что не было ни одного попадания, H_1 — было ровно одно попадание, и т. д. Эти события не пересекаются, и никакие другие исходы невозможны. Стало быть, мы имеем дело с набором гипотез. Применим формулу полной вероятности:

$$\begin{aligned} P(A) &= P(A|H_0) \cdot P(H_0) + P(A|H_1) \cdot P(H_1) + \\ &\quad + P(A|H_2) \cdot P(H_2) + P(A|H_3) \cdot P(H_3) = \\ &= 0 \cdot 0,2 + 0,2 \cdot 0,3 + 0,6 \cdot 0,3 + 1 \cdot 0,2 = 0,44. \end{aligned}$$

Еще одной важной формулой, служащей для переоценки вероятностей гипотез по результатам прошедших испытаний, является формула Байеса.

Из равенств

$$P(AH_1) = P(A|H_1) \cdot P(H_1) = P(H_1|A) \cdot P(A)$$

следует

$$P(H_1|A) = \frac{P(A|H_1) \cdot P(H_1)}{P(A)}.$$

Пример 18. Тест на туберкулез (реакция Манту) дает положительный результат для больных с вероятностью 0,95, а для здоровых — с вероятностью 0,1. Среди тестируемых

0,1 % больных. Какова вероятность, что человек здоров, если тест дал для него положительный результат?

Решение. Обозначим H_0 — событие “тестируемый болен”, H_1 — событие “тестируемый здоров”, A — “тест дал положительный результат”. Тогда по условию задачи:

$$P(H_0) = 0,001, P(H_1) = 0,999;$$

$$P(A|H_0) = 0,95, P(A|H_1) = 0,1.$$

По формуле полной вероятности

$$P(A) = 0,001 \cdot 0,95 + 0,999 \cdot 0,1 = 0,10085.$$

По формуле Байеса

$$P(H_1|A) = \frac{0,1 \cdot 0,999}{0,10085} = 99,06\%.$$

Эта формула часто используется, например, в задачах распознавания образов. В этих случаях гипотезы непосредственно не наблюдаемы, и о них можно судить только по косвенным данным. Пусть, например, мы пытаемся распознать сигнал (букву, команду и т. п.) из перечня (алфавита), состоящего из n сигналов (букв). Тогда H_1, \dots, H_n — это просто известные априорные вероятности, то есть частоты встречаемости букв. То, что мы измерили, назовем событием A . Тогда формула Байеса позволяет нам на основании известных вероятностей $P(A|H_i)$ вычислить апостериорные вероятности $P(H_i|A)$, то есть точнее распознать, какой был сигнал.

Описанное выше называют байесовским методом распознавания (или байесовским подходом, и т. д.). Такие методы используются часто, и вообще это направление в последнее время интенсивно развивается.

Задачи к лекции 3.

Задача 50. В семье двое детей, причем один из них мальчик. Какова вероятность того, что второй тоже мальчик?

Задача 51. Доказать, что если события A и B независи-

мы и \bar{B} непусто, то

$$P(A|\bar{B}) = P(A).$$

Задача 52. Независимы ли события при бросании кости:
1) число выпавших очков делится на 2; 2) делится на 3?

Задача 53. Сумма очков при бросании нескольких игральных костей может делиться на 2 и может делиться на 3. Независимы ли эти события?

Задача 54. В ведро, содержащее 10 литров воды температуры $24^{\circ}C$, наливают 2 литра кипятка температуры $100^{\circ}C$. Определить температуру полученной смеси.

Задача 55. Рабочий обслуживает 3 станка, на которых обрабатываются однотипные детали. Вероятность брака для первого станка равна 0,02, для второго — 0,03, для третьего — 0,04. Обработанные детали складываются в один ящик. Производительность первого станка в три раза больше, чем второго, а третьего в два раза меньше, чем второго. Определить вероятность того, что взятая наудачу деталь будет бракованной.

Задача 56. На овощехранилище поступает продукция от трёх хозяйств. Продукция первого хозяйства составляет 20% , второго — 46% и третьего — 34% . Известно, что средний процент нестандартных овощей для первого хозяйства равен 3% , для второго — 2% , для третьего — 1% . Найти вероятность того, что наудачу взятый овощ оказался нестандартным.

Задача 57. 30% изделий данного предприятия — это продукция высшего сорта. Некто приобрел 6 изделий, изготовленных на этом предприятии. Чему равна вероятность того, что ровно 4 из них высшего сорта? Чему равна вероятность того, что не менее 4 из них высшего сорта?

Задача 58. На заводе, производящем болты, машины А,

В, С производят соответственно 25, 35 и 40% всех изделий. В их продукции брак составляет соответственно 5, 4 и 2%. Какова вероятность, что случайно выбранный из продукции болт окажется дефектным?

Задача 59. В некоторых сельских местностях России существовало когда-то следующее гадание. Девушка зажимает в руке шесть травинок так, чтобы концы травинок торчали сверху и снизу; подруга связывает эти травинки попарно между собой сверху и снизу в отдельности. Если при этом все шесть травинок оказываются связанными в кольцо, то это должно было означать, что девушка в текущем году выйдет замуж. Найти вероятность того, что травинки при завязывании наудачу образуют кольцо.

Задача 60. В каждой из трёх урн содержится 6 чёрных и 4 белых шара. Из первой урны наудачу извлечён один шар и переложен во вторую урну, после чего из второй урны наудачу извлечён один шар и переложен в третью урну. Найти вероятность того, что шар, наудачу извлечённый из третьей урны, окажется белым.

Задача 61. В некоторой области профессиональный и бытовой травматизм составляет 1 случай на 60 тыс. человек, в том числе в областном центре — 1 случай на 80 тыс. человек, а в остальной части области — 1 случай на 40 тыс. человек. Известно, что во всей области проживают 2 миллиона 400 тысяч человек. Сколько из них проживают в областном центре, а сколько — в остальной части области?

Задача 62. По вражескому самолёту было выпущено 3 ракеты. Вероятность попадания первой из них равна 0,2, второй — 0,5, а третьей — 0,8. Если в самолёт попадут три ракеты, то он обязательно будет сбит, если попадут две, то он будет сбит с вероятностью 0,6, а если попадёт одна — с вероятностью 0,2. Какова вероятность того, что самолёт будет сбит?

Задача 63. Отец, желая промотивировать занятия сына теорией вероятностей, предлагает ему такую игру. Сын должен десять купюр по 100\$ и десять купюр по 1\$ разложить по двум шляпам. Затем сын с завязанными глазами выбирает шляпу, а затем — одну купюру из выбранной шляпы. Если будет выбрана купюра в 100\$, сын получит её в подарок. Купюры на ощупь не отличаются. Как сын должен распределить купюры по шляпам, чтобы увеличить вероятность успеха, и чему равна эта вероятность?

Задача 64. Разыскивая специальную книгу, студент решил обойти три библиотеки. Для каждой библиотеки одинаково вероятно, есть в её фондах книга или нет. Если книга есть, то одинаково вероятно, занята она другим читателем или нет. Какова вероятность того, что студент достанет книгу?

Задача 65. По вражескому самолёту было выпущено 3 ракеты. Вероятность того, что ни одна из них не попадёт в цель равна 0,2, вероятность того, что в цель попадёт ровно одна — 0,3, а того, что в цель попадут ровно две — также 0,3. Если в самолёт попадут три ракеты, то он обязательно будет сбит, если попадут две, то он будет сбит с вероятностью 0,6, а если попадёт одна — с вероятностью 0,2. Какова вероятность того, что самолёт будет сбит?

Задача 66. На шахматную доску случайным образом ставят белого короля и какую-то черную фигуру. Какова вероятность, что черная фигура окажется под боем?

Задача 67. Вероятность для изделий некоторого производства соответствовать стандарту равна 0,96. Предлагается упрощённая система контроля качества, дающая положительный результат с вероятностью 0,98 для изделий, удовлетворяющих стандарту, а для изделий, которые не удовлетворяют стандарту, с вероятностью 0,05. Какова вероятность, что изделие, выдержавшее испытание, удовле-

творяет стандарту?

Задача 68. Профессор культурологии К., обедая в институтской столовой, обнаружил у себя в порции плова всего 1 кусочек мяса. Вызванный для объяснений шеф-повар пояснил, что мясо в котёл с пловом положено достаточно, и вероятность того, что в тарелке окажется всего 1 кусок, равна 0,01. Правда, есть ещё два котла — для студентов и для подшефной свинофермы, для которых вероятность попадания всего одного куска в порцию равны соответственно 0,2 и 0,5, но он, шеф-повар, уверен, что профессор получил свою порцию из профессорского котла. Знакомый профессора, главный инженер, рассказал затем профессору, что когда раздатчица пьяная, ей становится всё равно, из какого котла накладывать порции. Зная, что, по мнению главного инженера, вероятность пьянства раздатчицы равна $1/4$, оцените с помощью формулы Байеса справедливость высказывания шеф-повара.

Лекция 4. Дискретные случайные величины.

Пример 19. Проводится матч из 4 партий между двумя игроками, причем вероятность победы первого игрока в каждой партии одинакова и равна $3/5$, а ничьих не бывает. Каков будет счет? В частности, какова вероятность того, что матч закончится вничью?

Решение. Вероятность ничьей равна

$$P(2 : 2) = C_4^2 \left(\frac{3}{5}\right)^2 \left(\frac{2}{5}\right)^2 = \frac{216}{625}.$$

Можно рассчитать вероятности для каждого возможного исхода матча. Мы получим такую таблицу, где в первой строке указано число побед первого игрока, а во второй — соответствующие вероятности.

X	0	1	2	3	4
P	0,0256	0,1536	0,3456	0,3456	0,1296

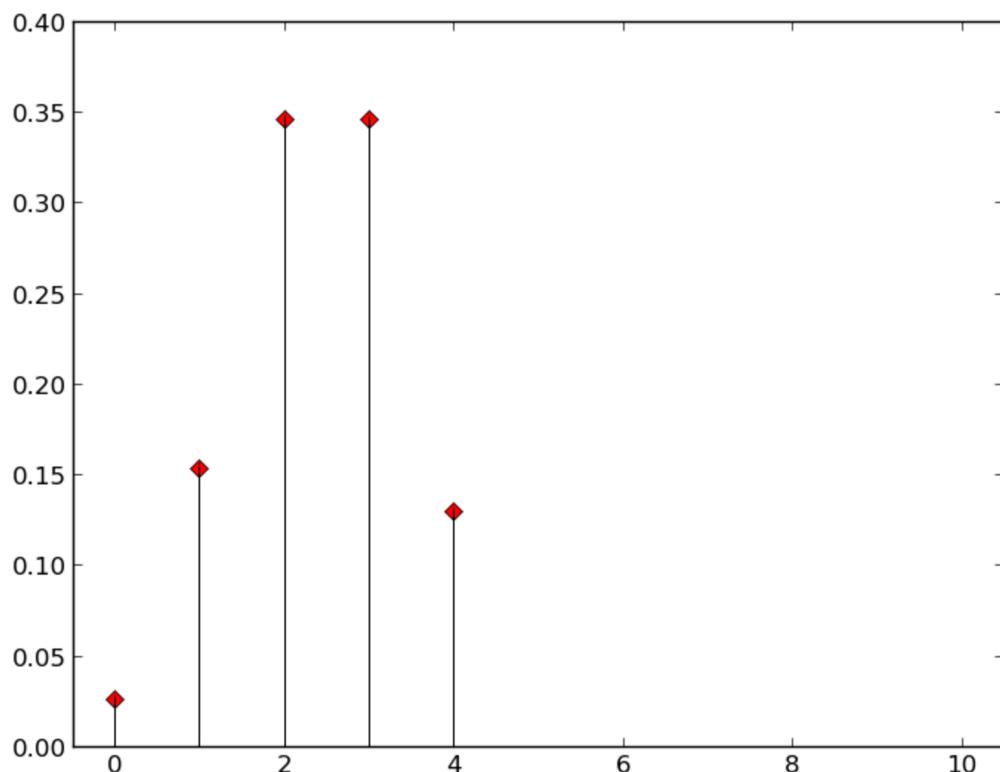


Рисунок 5. Распределение Бернулли с $n = 4, p = 0,6$.

Общий случай принято называть *последовательностью независимых испытаний* или *схемой Бернулли*. Пусть проводится n испытаний, в каждом из которых независимо от других может быть или успех, или неудача. Обозначим вероятность успеха в одном испытании через p , а вероятность неудачи — через $q = 1 - p$. Сколько будет успехов?

Вероятность того, что будет ровно k успехов, равна

$$P_n(X = k) = C_n^k p^k q^{n-k}.$$

Здесь мы впервые в курсе встречаемся с понятием случайной величины — ключевым понятием современной теории вероятностей и математической статистики. Знание вероятностей позволяет строить модели, прогнозировать, предугадывать, проверять гипотезы и т. д.

Более того. Во многих случаях вид распределения известен заранее, и при моделировании или анализе достаточно просто уточнить параметры распределения.

Переходим к изучению наиболее часто встречающихся распределений. Сегодня речь пойдет о дискретных распределениях: случайная величина принимает значения из дискретного множества — обычно из множества неотрицательных целых чисел.

Говорят, что определенная выше случайная величина k имеет *распределение Бернулли* (или — подчиняется распределению Бернулли). Часто используется также название *биномиальное распределение*. Это распределение возникает тогда, когда есть некоторый конечный набор испытаний, причем число этих испытаний заранее известно, и вероятность успеха не меняется от испытания к испытанию. Несколько примеров таких ситуаций:

- матч, как в примере;
- количество брака в партии;
- результаты голосования;

— и многие другие.

Общее определение дискретной случайной величины.

Дискретной случайной величиной, или законом распределения дискретной случайной величины принято называть правило, которое позволяет определить, какие значения и с какими вероятностями эта величина принимает. Часто такое правило можно представить в виде таблицы.

X	x_1	x_2	...	x_n
P	p_1	p_2	...	p_n

Здесь x_1, \dots, x_n — значения, которые может принимать с. в., а p_1, \dots, p_n — вероятности, с какими эти значения принимаются. Должны выполняться два важных свойства:

1. $p_i \geq 0$;
2. $p_1 + \dots + p_n = 1$.

Вместо событий мы теперь имеем дело с числами. Поэтому теперь можно ввести числовые характеристики случайной величины. Наиболее часто используются математическое ожидание и дисперсия. Рассмотрим определения этих понятий для дискретной случайной величины.

Математическое ожидание — усредненное значение случайной величины. Оно по определению равно

$$MX = x_1 p_1 + x_2 p_2 + \dots + x_n p_n.$$

Сумма отклонений от среднего равна нулю, поэтому для характеристики разброса все отклонения возводят в квадрат. Разброс относительно среднего характеризует дисперсия случайной величины. Она равна по определению

$$DX = M(X - MX)^2.$$

Стандартное (среднеквадратическое) отклонение — квадратный корень из дисперсии

$$\sigma(X) = \sqrt{DX}.$$

Надо иметь в виду, что для того, чтобы понять, каков примерный масштаб уклонения от среднего, надо использовать именно стандартное отклонение. Дело в том, что оно имеет ту же размерность, что и случайная величина, а размерность дисперсии — квадрат размерности случайной величины.

Случайные величины можно складывать, умножать на числа и друг на друга.

Свойства математического ожидания и дисперсии:

$$M(aX) = a \cdot MX;$$

$$M(X + Y) = MX + MY;$$

$$D(aX) = a^2 \cdot DX.$$

Независимые случайные величины — такие, для которых значение, принятое одной из них не зависит от того, какое значение приняла другая. Для независимых случайных величин

$$M(XY) = MX \cdot MY;$$

$$D(X + Y) = DX + DY.$$

Еще одна полезная формула, верная для любой случайной величины:

$$DX = MX^2 - (MX)^2.$$

Для распределения Бернулли

$$MX = np; \quad DX = npq.$$

Пример 20. Случайная величина X — число очков при бросании одной игральной кости. Найти математическое ожидание и дисперсию.

Решение. $MX = 3,5; DX = 35/12$.

Распределение Пуассона

$$P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \quad k = 0, 1, \dots$$

Оно получается из распределения Бернулли при $n \rightarrow \infty$,

$p \rightarrow 0$ так, что при этом $np = \lambda$. Для этого распределения

$$MX = \lambda; \quad DX = \lambda.$$

Распределение Пуассона часто используется в задачах массового обслуживания: поток вызовов часто представляет собой стационарный пуассоновский поток. Для того, чтобы поток вызовов был Пуассоновским, достаточно выполнения следующих условий:

- среднее число событий за некоторый интервал времени пропорционально длительности интервала и не зависит от момента начала этого интервала (стационарность);
- числа событий в непересекающиеся интервалы времени независимы (отсутствие последействия);
- события появляются поодиночке, то есть вероятность появления двух и более событий в малом интервале Δt есть бесконечно малая функция $\bar{o}(\Delta t)$ (ординарность).

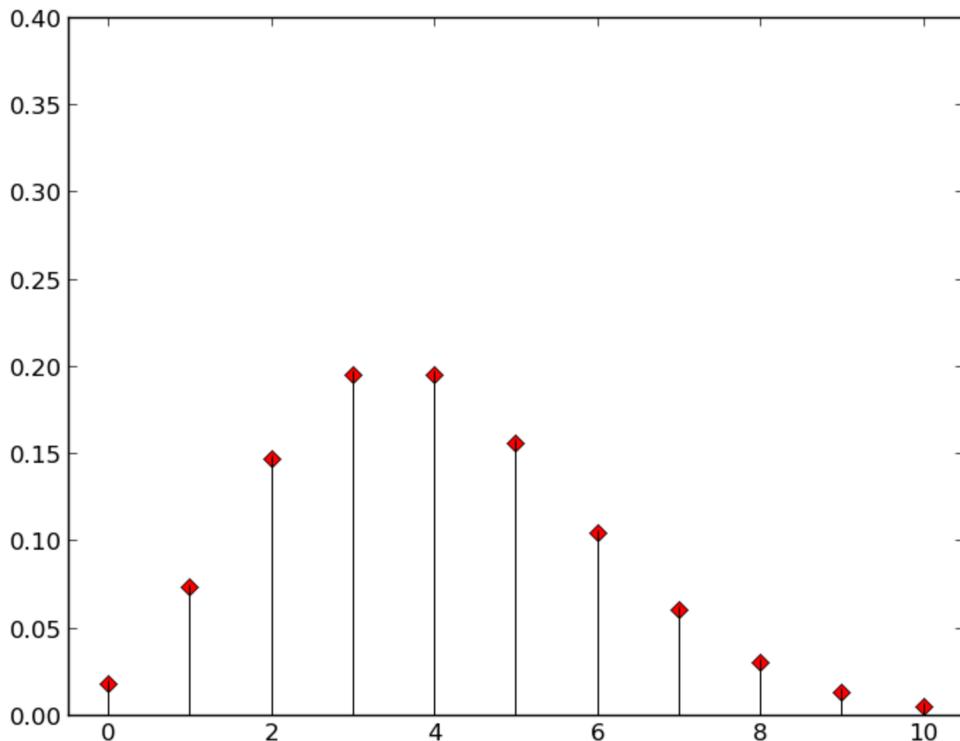


Рисунок 6. Распределение Пуассона с $\lambda = 4$.

Геометрическое распределение (число испытаний Бернул-

ли до первого успеха):

$$P(X = k) = pq^k, \quad k = 0, 1, \dots$$

Для этого распределения

$$MX = \frac{q}{p}; \quad DX = \frac{q}{p^2}.$$

Задачи к лекции 4.

Задача 69. В семье 5 детей. Считая, что вероятность рождения мальчика равна 0,51 и что пол каждого ребенка не зависит от пола остальных детей в семье, найти закон распределения случайной величины — числа мальчиков в семье.

Задача 70. Случайная величина принимает только значения 1 и 2, причем 1 — с вероятностью 0,3. Найти математическое ожидание и дисперсию.

Задача 71. Два стрелка независимо друг от друга делают по одному выстрелу в мишень. Первый попадает с вероятностью p_1 , второй — с вероятностью p_2 . Случайная величина — число попаданий. Найти ее закон распределения, математическое ожидание и дисперсию.

Задача 72. Двое равносильных партнеров играют в игру без ничьих матч из 4 партий. Найти вероятность того, что будет ничья 2:2.

Задача 73. Устройство состоит из восьми независимо работающих элементов. Вероятность отказа одинакова для каждого из этих элементов и равна 0,2. Элементы отказывают независимо друг от друга. Найти вероятность отказа устройства, если для этого нужно, чтобы отказали не менее трех элементов.

Задача 74. В сентябре в Подмосковье в среднем бывает 12 дождливых дней. Найти вероятность того, что из восьми наугад взятых дней будут ровно два дождливых; ровно три дождливых. Какая из этих вероятностей больше?

Задача 75. Вероятность попадания в мишень постоянна и равна p . Стрелку выдают патроны до тех пор, пока он не промахнется. Пусть X — число выданных патронов. Найти математическое ожидание и дисперсию случайной величины X .

Задача 76. Испытания Бернулли с вероятностью успеха в одном испытании p продолжаются до получения ровно k успехов. Найти закон распределения случайной величины — числа проведенных испытаний.

Задача 77. В тесто для 200 сдобных булочек положили 1000 изюминок. Какова вероятность того, что в одной отдельно взятой булочке оказалось ровно 5 изюминок? не оказалось изюма вообще?

Задача 78. Застраховано 10000 лиц одного возраста и одной социальной группы. Вероятность страхового случая в течение года равна $p = 0,006$. Ежегодный страховой взнос — \$120, выплата — \$10 000. Найти вероятность того, что — страховая компания потерпит убыток;
— страховая компания получит прибыль больше \$500 000.

Задача 79. В страховом обществе застраховано 10000 лиц одного возраста и одной социальной группы. Вероятность смерти в течение года для каждого лица равна 0,006. Страховая выплата составляет \$10 000. Каким должен быть страховой взнос, чтобы с вероятностью 0,99 страховое общество получило прибыль не менее \$500 000?

Задача 80. Вероятность попадания в самолет из стрелкового оружия равна 0,001. По самолету было произведено 4000 выстрелов. Какова вероятность того, что он будет сбит, если для этого нужно 3 попадания?

Задача 81. Среднее число вызовов, поступающих на АТС в минуту, равно 120. Найти вероятности того, что а) за две секунды не поступит ни одного вызова;

- б) за две секунды поступит не менее двух вызовов;
- в) за одну секунду поступит ровно три вызова;
- г) за три секунды поступит не менее трех вызовов.

Задача 82. (*) В подготовленной к корректорской проверке книге число опечаток на странице есть случайная величина, подчиняющаяся распределению Пуассона с параметром λ . Корректор обнаруживает каждую опечатку независимо от других с вероятностью p . Определить закон распределения числа опечаток на странице после корректорской проверки.

Лекция 5. Непрерывные случайные величины.

Непрерывная случайная величина X — случайная величина, которая может принимать значения из интервала или даже на всей числовой оси. Такие величины уже нельзя охарактеризовать таблицей значений и вероятностей, поскольку вероятность каждого конкретного значения равна нулю, а смысл имеет только вероятность попадания в какой-то интервал. Для таких величин вводится функция $\rho(x)$, которая называется плотностью вероятностей. При этом должны выполняться следующие условия (аналогичные условиям для дискретной случайной величины):

1. $\rho(x) \geq 0$.
2. $\int_{-\infty}^{\infty} \rho(x)dx = 1$.

Вероятность попадания в интервал (a, b) при этом равна

$$P(a < X < b) = \int_a^b \rho(x)dx.$$

При этом то, открытый интервал или замкнутый, то есть попадают ли в него точки a и b , — несущественно, поскольку вероятности того, что случайная величина примет значение a или b , равна нулю.

Математическое ожидание и дисперсия непрерывной случайной величины определяются следующим образом:

$$MX = \int_{-\infty}^{\infty} x\rho(x)dx;$$

$$DX = M(X - MX)^2 = \int_{-\infty}^{\infty} (x - MX)^2 \rho(x)dx.$$

Все формулы для математического ожидания и дисперсий, которые были выписаны для дискретных случайных величин на прошлой лекции, справедливы и для непрерывных

случайных величин. Напомним их.

$$M(aX) = a \cdot MX;$$

$$M(X + Y) = MX + MY;$$

$$D(aX) = a^2 \cdot DX.$$

Для независимых случайных величин

$$M(XY) = MX \cdot MY;$$

$$D(X + Y) = DX + DY.$$

И еще одна полезная формула:

$$DX = MX^2 - (MX)^2.$$

Для непрерывных случайных величин полезным является понятие функции распределения. По определению, функция распределения случайной величины X — это функция

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

Связь между функцией распределения и плотностью вероятностей сразу следует из формулы вероятности попадания в интервал:

$$F_X(x) = P(X < x) = \int_{-\infty}^x \rho(t) dt.$$

А отсюда следует еще одна важная формула:

$$\rho(x) = F'_X(x).$$

Для различных распределений часто возникает задача нахождения такого значения x_0 , что вероятность того, что случайная величина примет значение, меньшее x_0 , равно заданному числу α . Такое значение называется α -квантилем распределения (ударение на “и”).

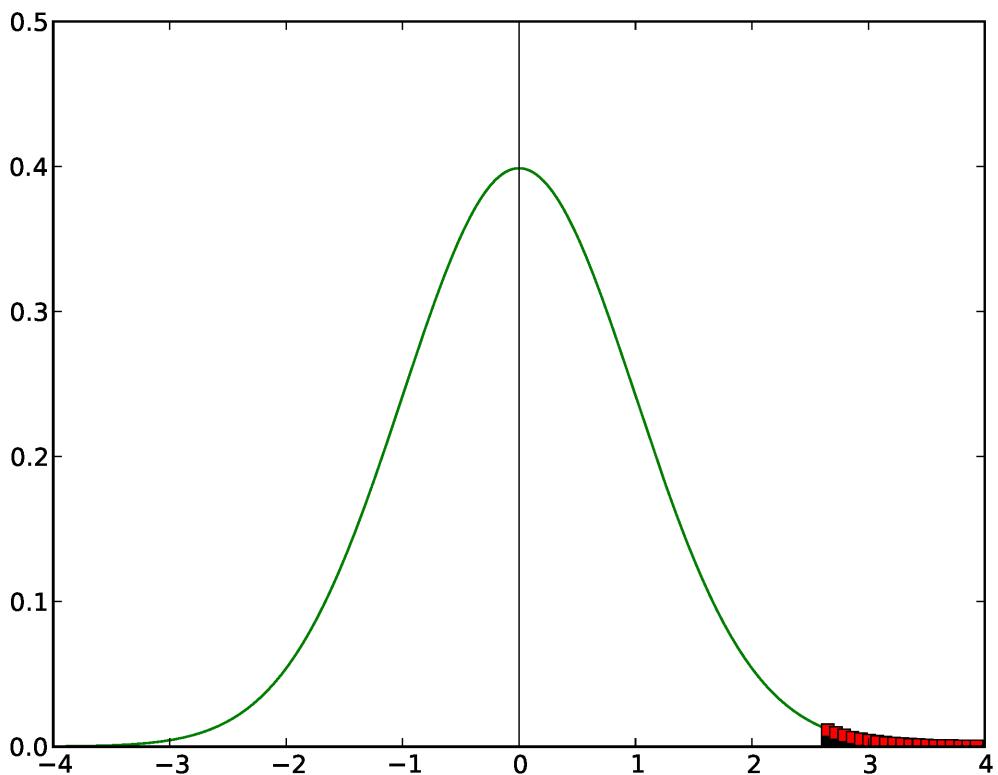


Рисунок 7. Квантиль.

Таким образом, α -квантиль распределения — это решение уравнения

$$\int_{-\infty}^{x_0} \rho(t)dt = \alpha.$$

Если же воспользоваться понятием функции распределения, то можно записать такое уравнение

$$F_X(x_0) = \alpha.$$

Слово “квантиль” обычно относят к мужскому роду: так предписывает, например, сайт *gramota.ru* или действующий в России ГОСТ Р 50779.10-2000 “Вероятность и основы статистики. Термины и определения”. Однако в математической литературе это слово часто встречается в женском роде. Это — один из примеров профессионального жаргона.

Переходим к примерам часто используемых распределе-

ний.

Равномерное распределение.

Плотность вероятности для этого распределения равна константе внутри отрезка $[a, b]$ и нулю вне этого отрезка. Функция распределения линейна на отрезке, 0 — левее и 1 — правее отрезка. Чему равна эта константа и каков коэффициент наклона функции распределения — попробуйте догадаться сами.

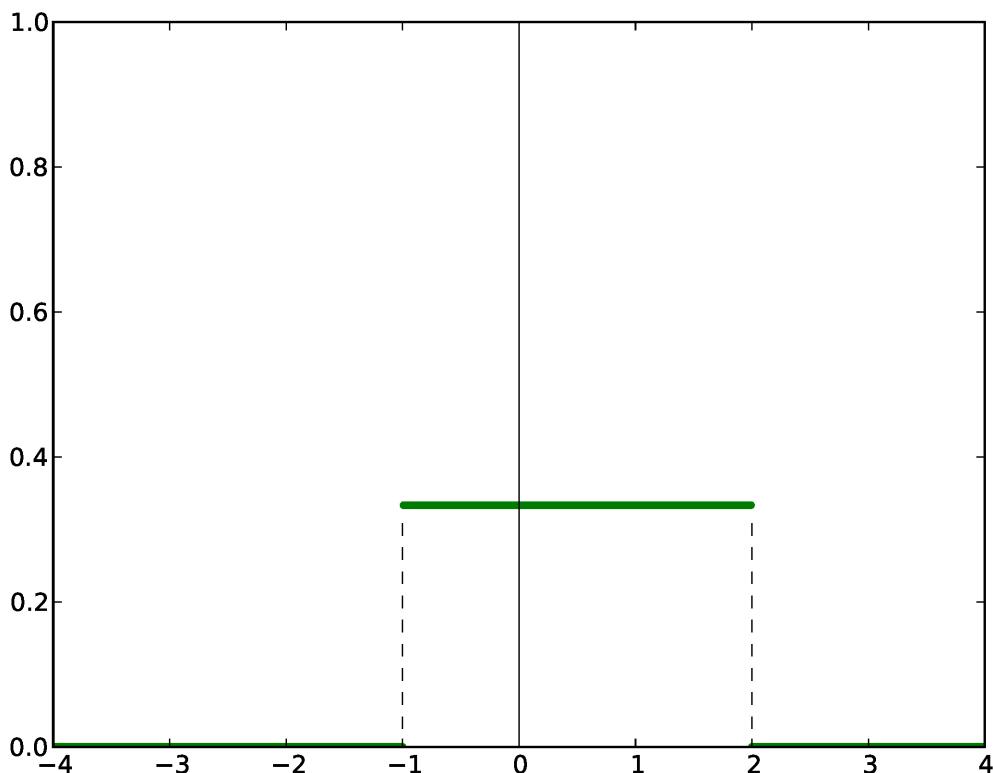


Рисунок 8. Плотность равномерного распределения на $[-1, 2]$.

Это — чисто теоретическое, модельное распределение. В реальной жизни автору встречаться с ним не приходилось, только в абстрактных задачах. Программисты, однако, могут вспомнить стандартную функцию `random()` в C++, возвращающую случайное число, равномерно распределенное на $[0, 1]$.

Для этого распределения математическое ожидание и дисперсия равны соответственно

$$MX = \frac{b-a}{2}; \quad DX = \frac{(b-a)^2}{12}.$$

Показательное, или экспоненциальное, распределение.

Плотность вероятностей для этого распределения равна нулю для отрицательных x , а для положительных — задается формулой

$$\rho(x) = \lambda e^{-\lambda x}$$

с некоторым положительным параметром λ . Функция распределения при $x > 0$ равна

$$F_X(x) = 1 - e^{-\lambda x}.$$

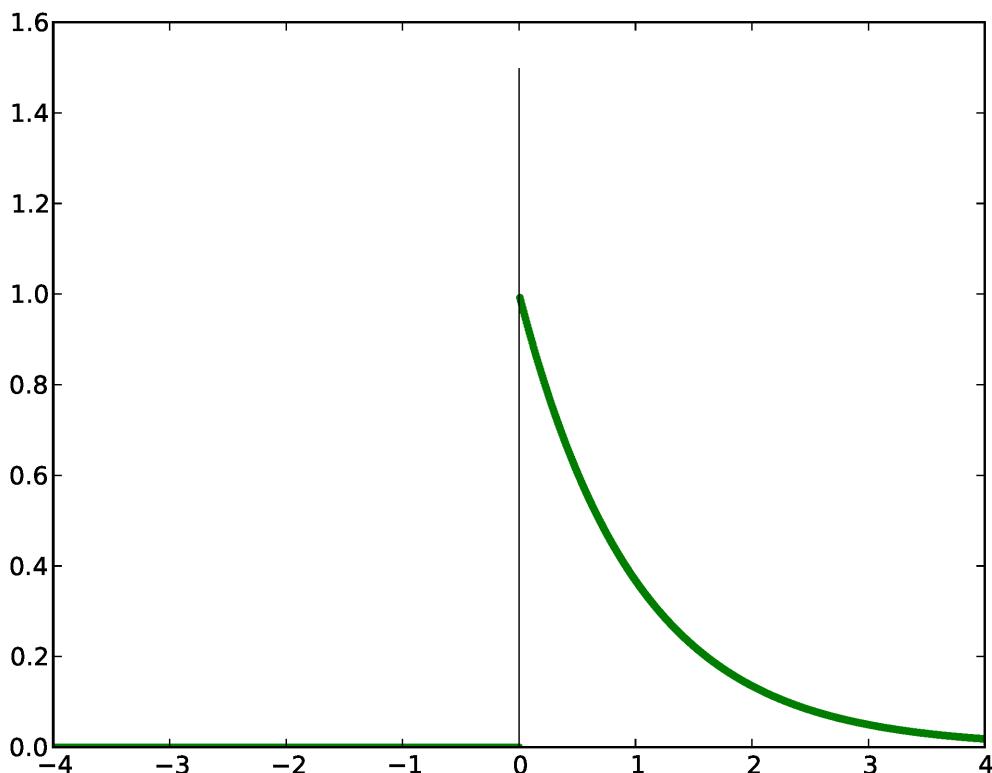


Рисунок 9. Плотность показательного распределения.

Это распределение встречается в теории массового обслуживания: для пуассоновского потока вызовов промежутки

времени между последовательными вызовами имеют показательное распределение. Также оно применяется в теории надежности как распределение времен безотказной работы для каких-нибудь простых устройств или комплектующих.

Для этого распределения математическое ожидание и дисперсия равны соответственно

$$MX = \frac{1}{\lambda}; \quad DX = \frac{1}{\lambda^2}.$$

Нормальное распределение.

Это распределение встречается, пожалуй, наиболее часто. Впервые оно появилось в трудах Гаусса в связи с распределением ошибок измерений. Поэтому его еще называют Гауссовым распределением. Впоследствии оказалось, что роль этого распределения не ограничивается только ошибками. Оно появляется как предельное распределение для суммы случайных величин. Более подробно об этом будет сказано ниже, когда речь пойдет о центральной предельной теореме.

Плотность вероятностей для нормального распределения задается формулой

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

Как видно из этой формулы, нормальное распределение имеет два параметра — a и σ — равные соответственно математическому ожиданию и среднеквадратическому отклонению. Дисперсия, стало быть, равна σ^2 .

Стандартным нормальным распределением называется такое нормальное распределение, у которого среднее равно 0, а дисперсия равна 1. Плотность вероятности для него равна

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

Функция распределения равна интегралу от этой плотности вероятности. Этот интеграл не берется в элементарных

функциях, но встречается часто, и поэтому удостоился специального названия. Функция

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

часто называется интегралом вероятностей, или интегралом ошибок.

Значения этой функции приводятся в специальных таблицах. Такие таблицы есть в любом сколько-нибудь серьезном учебнике по теории вероятностей. При работе с электронными таблицами, в частности, с Microsoft Excel, можно использовать стандартную функцию СТНОРМОБР.

В таблицах приводятся значения интеграла ошибок только для положительных x . Для отрицательных x из того, что под интегралом стоит четная функция и из того, что $\Phi(0) = 1/2$, следует формула $\Phi(-x) = 1 - \Phi(x)$.

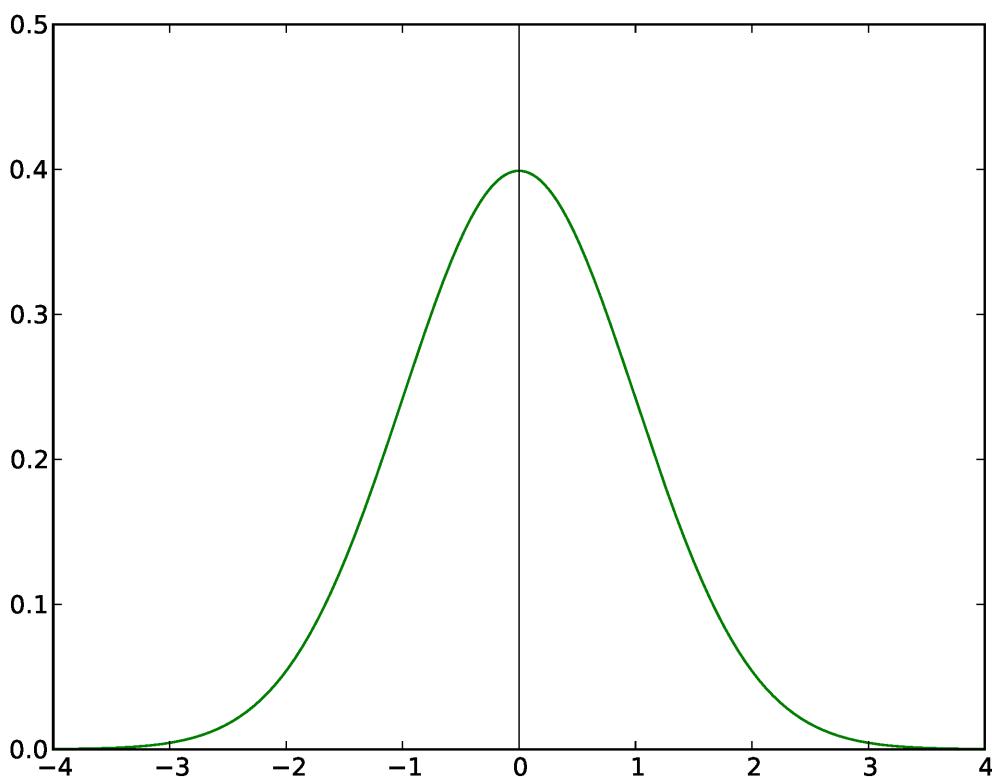


Рисунок 10. Плотность нормального распределения.

Во многих задачах, как мы увидим чуть позже, требуется находить квантили распределений по заданным вероятностям α , или наоборот, значения вероятностей α по заданным квантилям. Для плотностей вероятностей, заданных простыми формулами, применяют, например, интегральное исчисление. Но (к радости не одного поколения студентов), во многих часто встречающихся случаях, как и для нормального распределения, интегралы не берутся, и надо пользоваться специальными таблицами.

Для нормального распределения верна такая теорема. Если случайная величина X имеет нормальное распределение со средним a и дисперсией σ^2 , то случайная величина

$$\frac{X - a}{\sigma}$$

имеет стандартное нормальное распределение. Желающий доказать эту теорему может сделать это самостоятельно: для этого надо только знать, как делается замена переменных в определенном интеграле.

Несмотря на простоту этой теоремы, она используется довольно часто: в частности, она позволяет пользоваться таблицами только для стандартного нормального распределения.

Пример 21. Найти вероятность того, что случайная величина, нормально распределенная со средним $a = 8$ и дисперсией $\sigma^2 = 16$, принимает значения в интервале от 0 до 12.

Решение. Обозначим искомую вероятность через p и запишем это формулой:

$$p = P(0 < X < 12).$$

Далее преобразуем эту формулу:

$$p = P\left(\frac{0 - 8}{4} < \frac{X - a}{\sigma} < \frac{12 - 8}{4}\right);$$

$$p = P\left(-2 < \frac{X - a}{\sigma} < 1\right).$$

Собственно, мы ничего не сделали, только вычли среднее и разделили на σ . Но теперь в центре двойного неравенства стоит случайная величина, имеющая стандартное нормальное распределение, и можно пользоваться таблицей. Окончательно получим

$$p = \Phi(1) - \Phi(-2) = \Phi(1) + \Phi(2) - 1 = 0,8413 + 0,9772 - 1 = 0,8385.$$

Пример 22. Найти вероятность того, что нормально распределенная случайная величина отклонится от среднего больше, чем на 3σ .

Решение. Проще найти вероятность противоположного события.

$$\begin{aligned} P(|X - a| < 3\sigma) &= P(-3\sigma < X - a < 3\sigma) = \\ &= P\left(-3 < \frac{X - a}{\sigma} < 3\right) = \Phi(3) - \Phi(-3) = 2\Phi(3) - 1 = 0,9974. \end{aligned}$$

Таким образом, вероятность отклонений, больших 3σ , составляет примерно 0,26%. На этом основано часто применяемое правило трех сигм. Если мы, скажем, проводим какие-то измерения, то мы можем столкнуться с грубыми промахами. Как узнать, является ли данный результат измерения грубым промахом? Если оно дальше от среднего, чем три сигмы, то да, а если ближе, то нет.

Пример 23. В каких пределах с заданной вероятностью $1 - \alpha$ будет находиться значение нормально распределенной с. в. с заданными средним a и дисперсией σ^2 ? Найти минимальный из таких интервалов.

Решение. Будем искать симметричный интервал $(a - z, a +$

z) — он будет минимальным по длине. Напишем уравнение.

$$P(a - z < X < a + z) = 1 - \alpha.$$

Далее действуем уже традиционным способом:

$$\begin{aligned} P\left(-\frac{z}{\sigma} < \frac{X - a}{\sigma} < \frac{z}{\sigma}\right) &= 1 - \alpha. \\ 2\Phi\left(\frac{z}{\sigma}\right) - 1 &= 1 - \alpha. \\ \Phi\left(\frac{z}{\sigma}\right) &= 1 - \alpha/2. \end{aligned}$$

Обозначим через $t_{1-\alpha/2}$ соответствующий квантиль стандартного нормального распределения, и получим выражение для искомого интервала:

$$X \in (a - \sigma \cdot t_{1-\alpha/2}, a + \sigma \cdot t_{1-\alpha/2}) \text{ с вероятностью } 1 - \alpha.$$

Функции случайной величины.

Если X — случайная величина, и f — некоторая функция, то $Y = f(X)$ также будет случайной величиной. Каково будет ее распределение?

Общую формулу можно получить для случая, когда f — монотонная функция. Тогда, если $y = f(x)$, то $x = g(y)$, где $g = f^{-1}$ обозначена обратная функция. Пусть это — монотонно возрастающая функция. Тогда

$$F_Y(y) = P(Y < y) = P(X < g(y)) = F_X(g(y)) = \int_{-\infty}^{g(y)} \rho(t) dt.$$

Дифференцируя по верхнему пределу и применяя формулу дифференцирования сложной функции, получим

$$\rho_Y(y) = \rho_X(g(y))g'(y).$$

В общем случае для подобных задач следует использовать функцию распределения.

Пример 24. Плотность вероятности случайной величины X равна

$$\rho(x) = \frac{1}{\pi(1+x^2)}.$$

Найти плотность вероятности случайной величины

$$Y = |X|^3 - 4.$$

Решение. Будем искать функцию распределения Y :

$$\begin{aligned} F_Y(y) &= P(Y < y) = P(|X|^3 - 4 < y) = \\ &= P(-\sqrt[3]{y+4} < X < \sqrt[3]{y+4}). \end{aligned}$$

Последний переход справедлив для $y \geq -4$. Для $y < -4$ функция распределения, очевидно, равна нулю.

Продолжим вычисления:

$$\begin{aligned} F_Y(y) &= \frac{1}{\pi} \int_{-\sqrt[3]{y+4}}^{\sqrt[3]{y+4}} \frac{dx}{1+x^2} = \frac{1}{\pi} \arctg x \Big|_{-\sqrt[3]{y+4}}^{\sqrt[3]{y+4}} = \\ &= \frac{2}{\pi} \arctg \sqrt[3]{y+4}. \end{aligned}$$

Для того, чтобы найти плотность вероятности, надо полученное выражение продифференцировать. Получим

$$\rho_Y(y) = \frac{2}{3\pi \left(1 + (y+4)^{2/3}\right)(y+4)^{2/3}} \text{ при } y \geq -4.$$

Часто требуется определить не всю функцию от случайной величины, а только ее среднее значение. В этом случае можно применить формулу

$$M(f(X)) = \int_{-\infty}^{\infty} f(x)\rho(x)dx.$$

Задачи к лекции 5.

Задача 83. Плотность вероятности случайной величины задана формулой

$$\rho(x) = \frac{A}{1+x^2}.$$

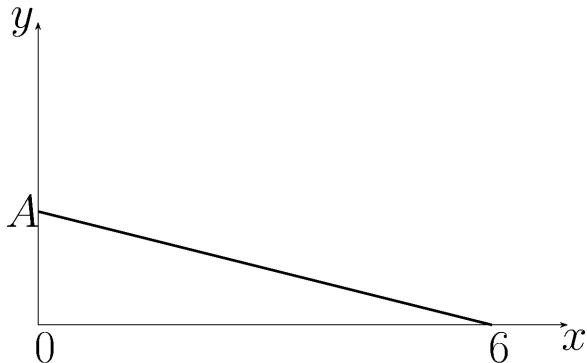
Найти константу A и вероятность того, что случайная величина примет значение больше 1.

Задача 84. Плотность вероятности случайной величины задана формулой

$$\rho(x) = \frac{A}{\sqrt{1-x^2}} \text{ при } x \in (-1, 1).$$

Найти константу A , математическое ожидание, дисперсию и вероятность того, что случайная величина примет значение, по модулю меньшее $1/2$.

Задача 85. Случайная величина отлична от нуля на интервале $(0, 6)$. График ее плотности вероятности изображен на рисунке.



Найти константу A , математическое ожидание, дисперсию и вероятность того, что случайная величина примет значение, меньшее 2.

Задача 86. Плотность вероятности случайной величины задана формулой

$$\rho(x) = A(4-x^2) \text{ при } x \in (-2, 2).$$

Найти константу A , математическое ожидание, дисперсию

и вероятность того, что случайная величина примет значение, по модулю меньшее 1.

Задача 87. Плотность вероятности случайной величины задана формулой

$$\rho(x) = A \sin x \text{ при } x \in (0, \pi).$$

Найти константу A , математическое ожидание, дисперсию и вероятность того, что случайная величина примет значение, меньшее $\pi/2$.

Задача 88. Распределение Максвелла — распределение скоростей молекул. Для этого распределения плотность вероятности равна

$$\rho(x) = \sqrt{\frac{2}{\pi}} \beta^{3/2} x^2 e^{-\beta x^2/2}$$

для положительных x и 0 для отрицательных. Как распределены энергии? Чему равна средняя энергия?

Задача 89. Найти функцию распределения квадрата стандартной нормальной случайной величины. Найти математическое ожидание этого квадрата.

Задача 90. Найти среднюю длину хорды, соединяющую случайную точку на верхней половине единичной окружности с центром в начале координат с точкой $(1, 0)$.

Задача 91. Автомат выпускает детали, размер которых представляет собой нормально распределенную случайную величину со средним 81 мм и стандартным отклонением 3 мм. Деталь считается годной, если ее размер от 78 до 82 мм, в противном случае она бракуется. Каков процент брака у автомата?

Задача 92. Случайные величины X и Y независимы, $DX = 4$, $DY = 3$. Найти $D(X - Y)$.

Задача 93. Угол A недоступен для непосредственного измерения. Чтобы измерить этот угол, его включили в пяти-

угольник $ABCDE$, где остальные углы можно измерить (в геодезии применяется термин *геодезический ход*). Точность измерения углов B, C, D и E составляет 1° . Какова точность определения угла A ?

Задача 94. Случайные величины X_1, \dots, X_n — результаты независимых измерений некоторой величины. Все измерения равноточные и характеризуются некоторой точностью (стандартным отклонением) σ . Для определения значения измеряемой величины используется среднее арифметическое

$$X = \frac{X_1 + \dots + X_n}{n}.$$

Найти точность (стандартное отклонение) случайной величины X .

Задача 95. Случайные величины X_1, \dots, X_n — результаты независимых измерений некоторой величины. При этом измерения неравноточные — i -е измерение характеризуются точностью (стандартным отклонением) σ_i . Для определения значения измеряемой величины используется линейная комбинация данных измерений

$$X = w_1 X_1 + \dots + w_n X_n$$

с некоторыми весами w_1, \dots, w_n . Определить, при каких значениях весов точность (стандартное отклонение) величины X будет минимальным.

Лекция 6. Теорема Муавра — Лапласа.

На “бытовом языке” теорему Муавра — Лапласа можно сформулировать так: при больших n число успехов в n испытаниях Бернулли имеет приблизительно нормальное распределение.

Приведенный ниже рисунок иллюстрирует теорему Муавра — Лапласа. Действительно, при больших n распределение числа успехов в n испытаниях похоже на нормальное.

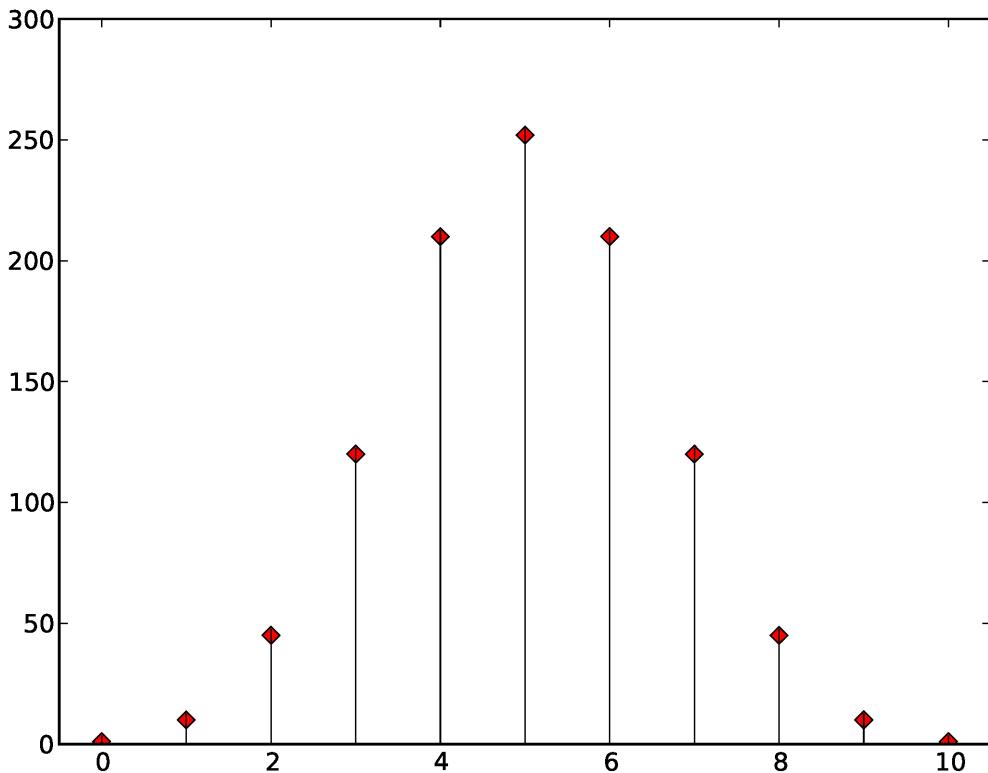


Рисунок 11. Распределение Бернулли с $n = 10$, $p = 1/2$.

В большинстве учебников приводится строгая формулировка, которую понять довольно трудно. Например, формулировка наиболее важного частного случая (называемого интегральной теоремой Лапласа) в классическом учебнике Гмурмана [3] выглядит так.

Теорема. Если вероятность p наступления события A в каждом испытании постоянна и отлична от нуля и единицы,

то вероятность $P_n(k_1, k_2)$ того, что событие A появится в n независимых испытаниях от k_1 до k_2 раз, приближенно равна определенному интегралу

$$P_n(k_1, k_2) = \frac{1}{\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{t^2}{2}} dt,$$

где $x_1 = (k_1 - np)/\sqrt{npq}$ и $x_2 = (k_2 - np)/\sqrt{npq}$.

Чтобы понять, в чем суть дела, разберем использование этой теоремы на нескольких примерах.

Пример 25. Вероятность заболеть гриппом во время эпидемии равна 0,4. Какова вероятность того, что заболевших на предприятии, на котором работают 600 человек, будет от 210 до 270?

Пусть k — число “успехов”, то есть случайная величина, обозначающая число заболевших. Мы знаем, что эта случайная величина имеет распределение Бернулли. Собственно, этим можно воспользоваться: подсчитать для каждого из 61 значений (от 210 до 270) его вероятность, затем эти вероятности просуммировать... Дел хватит надолго.

Вместо этого воспользуемся теоремой Муавра — Лапласа. Нам нужно оценить вероятность

$$P = P_{600}(210, 270) = P(210 \leq k \leq 270).$$

На основании “бытовой” формулировки поступим с этой вероятностью так, как если бы k имело нормальное распределение:

$$P = P\left(\frac{210 - np}{\sqrt{npq}} \leq \frac{k - np}{\sqrt{npq}} \leq \frac{270 - np}{\sqrt{npq}}\right).$$

Математическое ожидание и дисперсия для распределения Бернулли нам известны, и для нашего примера легко подсчитать, что

$$np = 240; \quad npq = 144; \quad \sqrt{npq} = 12.$$

Поэтому

$$\begin{aligned} P &= P\left(\frac{210 - 240}{12} \leq \frac{k - np}{\sqrt{npq}} \leq \frac{270 - 240}{12}\right) = \\ &= P(-2,5 \leq \frac{k - np}{\sqrt{npq}} \leq 2,5). \end{aligned}$$

Дробь в середине этого двойного неравенства имеет, как нетрудно догадаться, стандартное нормальное распределение. Поэтому осталось просто воспользоваться таблицей:

$$P = \Phi(2,5) - \Phi(-2,5) = 2 \cdot 0,9938 - 1 = 0,9876.$$

Воспользовавшись формулой Бернулли и электронными таблицами, например Microsoft Excel, можно найти точное значение этой вероятности. Оно равно 0,9890. Если теорему Муавра — Лапласа использовать для оценки, а не для точного нахождения вероятностей, то точность вполне приемлема.

При рассмотрении этого примера мы не обращали внимания на такой аспект: входят ли граничные точки, то есть 210 и 270? Оказывается, это неважно. Вероятность каждого значения близка к 0, а теорема Муавра — Лапласа используется не для точного подсчета, а для приближенной оценки вероятности.

Точность на самом деле можно и повысить. Мы заменяем дискретное распределение Бернулли непрерывным нормальным распределением. Тогда получается, что мы каждое целое значение заменяем отрезком, и на каждое целое значение будет приходиться отрезок длины 1. Поэтому каждому возможному значению k из распределения Бернулли будет соответствовать интервал $[k - 1/2, k + 1/2]$ для нормального распределения. Тогда для оценки надо брать не интервал $[210, 270]$, а интервал $[209,5, 270,5]$. Проведя вычисления с этим интервалом, мы получим значение вероятности 0,9890, то есть с точностью до четырех знаков совпада-

дающее с верным значением.

Эта идея позволяет получить еще один важный результат, оценив вероятность каждого отдельного значения. Как и ранее, каждому возможному значению k из распределения Бернулли будет соответствовать интервал $[k - 1/2, k + 1/2]$ для нормального распределения. Тогда

$$\begin{aligned} P_n(k) &\approx P(k - 1/2 < X < k + 1/2) = \\ &= P\left(\frac{k - np - 1/2}{\sqrt{npq}} < \frac{X - np}{\sqrt{npq}} < \frac{k - np + 1/2}{\sqrt{npq}}\right). \end{aligned}$$

Вероятность попадания в интервал, как мы знаем, равна интегралу от плотности вероятностей. Здесь же мы можем считать, что интервал маленький, и вероятность попадания в него приблизительно равна значению плотности вероятностей в середине интервала, умноженной на длину интервала. Окончательно получим:

$$P_n(k) \approx \frac{1}{\sqrt{npq}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \text{ где } x = \frac{k - np}{\sqrt{npq}}.$$

Этот результат носит название *локальной теоремы Муавра — Лапласа*. Как видно, вероятность каждого конкретного значения действительно стремится к 0 с ростом n .

Теорема Муавра — Лапласа применяется для решения различных задач. Рассмотрим еще несколько примеров.

Пример 26. Театр вместимостью 1000 мест имеет два входа, при каждом из них есть гардероб. Какова должна быть вместимость каждого из этих гардеробов, чтобы все зрители с вероятностью 0,99 могли раздеться в гардеробе того входа, через который вошли? Предполагается, что каждый из входов зрители выбирают с вероятностью 0,5 независимо друг от друга, а театр каждый день полон.

Решение. Мы имеем дело с испытаниями Бернулли с параметрами: $n = 1000$, $p = 1/2$. Требуется узнать, в каких пределах с вероятностью 0,99 лежит число успехов, то есть

число людей, выбравших первый, скажем, вход. Можно это записать в виде уравнения с неизвестным z :

$$P(np - z < X < np + z) = 0.99.$$

Двойное неравенство здесь потому, что нас не устраивает как слишком большое, так и слишком малое число “успехов”: так как театр полон, то малое число посетителей у одного гардероба означает перегрузку второго.

Дальнейшее просто:

$$\begin{aligned} P\left(-\frac{z}{\sqrt{npq}} < \frac{X - np}{\sqrt{npq}} < \frac{z}{\sqrt{npq}}\right) &= 0,99; \\ 2\Phi\left(\frac{z}{\sqrt{npq}}\right) - 1 &= 0,99; \\ \Phi\left(\frac{z}{\sqrt{npq}}\right) &= 0,995; \\ \frac{z}{\sqrt{npq}} &= 2,58; z = 2,58 \cdot \sqrt{250} \approx 41. \end{aligned}$$

Следовательно, в гардеробе надо предусмотреть 541 место.

Еще один пример

Пример 27. В поселке живут 180 человек. Каждый из них пять раз в месяц ездит на автобусе в город, выбирая дни поездок случайным образом независимо от других. Какой вместимости должен быть автобус, чтобы он не переполнялся в 99% случаев? Считать, что в месяце 30 дней.

Решение. И вновь мы имеем дело с испытаниями Бернулли, на сей раз с параметрами $n = 180$, $p = 1/6$. Но, в отличие от предыдущего примера, здесь нам надо найти несимметричный интервал с заданной вероятностью: если автобус будет недогружен — ничего страшного. Поэтому можно записать такое уравнение: $P(X < z) = 0,99$. Решение его аналогично:

$$P\left(\frac{X - np}{\sqrt{npq}} < \frac{z - np}{\sqrt{npq}}\right) = 0,99;$$

$$\Phi\left(\frac{z - np}{\sqrt{npq}}\right) = 0,99; \frac{z - np}{\sqrt{npq}} = 2,33;$$

$$z = np + 2,33 \cdot \sqrt{npq} = 30 + 2,33 \cdot \sqrt{25} \approx 42.$$

Если в n испытаниях Бернулли было достигнуто k успехов, то относительная доля успехов (или частота успехов) будет равна k/n . Зададимся вопросом, велико ли отклонение этой частоты от p — вероятности успеха в каждом испытании. Точнее, какова вероятность того, что эта частота отклонится от p не более, чем на ε .

Рассчитаем эту вероятность:

$$\begin{aligned} P\left(\left|\frac{k}{n} - p\right| < \varepsilon\right) &= P\left(-n\varepsilon < k - np < n\varepsilon\right) = \\ &= P\left(-\frac{n\varepsilon}{\sqrt{npq}} < \frac{k - np}{\sqrt{npq}} < \frac{n\varepsilon}{\sqrt{npq}}\right) = 2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right) - 1. \end{aligned}$$

Эта формула используется для решения задач, а мы обратим внимание на одно из ее следствий. Из формулы видно, что, какой бы малый ни был ε , вероятность еще меньших отклонений стремится к единице. Это важное наблюдение часто называют *законом больших чисел* для испытаний Бернулли. В статистических исследованиях этому закону придается глубокий мировоззренческий смысл. Вот что, например, написано в одном из учебников по экономической статистике.

“Исключительно важное значение для статистической методологии имеет закон больших чисел. Его содержание таково: в массе индивидуальных явлений общая закономерность проявляется тем полнее и точнее, чем больше их охвачено наблюдением. В числах, суммирующих результат масштабного наблюдения, выступают определенные закономерности, которые не могут быть обнаружены на небольшом числе фактов. Закон больших чисел выражает диалектику случайного и необходимого.”

И еще одна цитата (Маркс К., Энгельс Ф. Соч. Т. 25, ч. 11. С. 396.): "...внутренний закон, прокладывающий себе дорогу через эти случайности и регулирующий их, становится видимым лишь тогда, когда они охватываются в больших массах..."

Пример 28. За лейбористов и консерваторов голосуют примерно поровну избирателей. Проводится опрос с целью уточнить долю избирателей, голосующих за лейбористов. Сколько человек надо опросить, чтобы ошибка оценки была не более 4% с вероятностью 99%?

Решение. Из предыдущей формулы следует

$$\varepsilon \sqrt{n} \geq 2,58 \sqrt{pq}.$$

Поскольку $pq \approx 1/4$, получим

$$\sqrt{n} \approx 2,58 \cdot \frac{\sqrt{pq}}{\varepsilon} \approx 32,2; \quad n \approx 1040.$$

Еще одной теоремой, объясняющей важность и широту применения нормального распределения, является центральная предельная теорема. Мы будем использовать такую формулировку этой теоремы: если случайные величины $X_1, X_2 \dots X_n$ независимы, имеют одинаковые математические ожидания и дисперсии

$$MX_i = a; DX_i = D,$$

то при достаточно больших n сумма этих случайных величин

$$X = X_1 + X_2 + \dots + X_n$$

является случайной величиной, имеющей приблизительно нормальное распределение.

На практике центральную предельную теорему используют (не для точного вычисления вероятности, а для ее оценки), когда $n \geq 10$.

В соответствии с правилами сложения математических

ожиданий и дисперсий для независимых случайных величин получим

$$MX = na; DX = nD.$$

В частности, если обозначить среднеквадратическое отклонение (разброс) каждой из случайных величин X_i через σ , а случайной величины X — через s , то из последнего соотношения получим

$$s = \sigma\sqrt{n},$$

то есть разброс растет пропорционально квадратному корню из числа слагаемых.

Пример. При составлении статистического отчета надо было сложить 10 000 чисел, каждое из которых было округлено до ближайшего целого. Предполагая, что ошибки, возникшие от округления чисел, взаимно независимы и равномерно распределены на интервале $(-0,5; 0,5)$, найти пределы, в которых с вероятностью 0,997 будет лежать суммарная ошибка.

Решение. Ошибка суммы X есть случайная величина, равная сумме случайных величин $X_i, i = 1 \dots 10000$, равных ошибке каждого слагаемого. Поскольку

$$MX_i = 0; DX_i = 1/12,$$

получим, что

$$MX = 0; DX = s^2 = 833,33; s = 28,87.$$

Будем искать симметричный промежуток $(-z; z)$, значения в котором случайная величина X принимает с вероятностью 0,997. Запишем уравнение

$$P\{-z < X < z\} = 0,997.$$

Переходя к стандартному нормальному распределению, получим

$$P\left\{\frac{-z}{28,87} < \frac{X}{28,87} < \frac{z}{28,87}\right\} = 0,997;$$

$$2\Phi\left(\frac{z}{28,87}\right) - 1 = 0,997.$$

Из таблицы находим

$$\frac{z}{28,87} = 2,96,$$

следовательно, $z = 85,45$.

Ответ: $\pm 85,45$.

Пример 29. Игровую кость подбросили 1000 раз и просуммировали число выпавших очков. В каких пределах с вероятностью 0,99 лежит эта сумма (для “правильной” кости)?

Вася подбросил украденную из казино кость 1000 раз и насчитал в сумме 3298 очков. Правильная ли это кость?

Решение. Число очков при одном бросании — случайная величина, ее среднее значение (математическое ожидание) было подсчитано ранее, оно равно 3,5. Дисперсию ее тоже можно подсчитать — она равна $35/12$.

В соответствии с центральной предельной теоремой суммарное число выпавших очков должно быть распределено приблизительно нормально со средним

$$a = 3500$$

и дисперсией

$$\sigma^2 = 1000 \cdot \frac{35}{12}.$$

Пусть X — суммарное число выпавших очков. Составим уравнение

$$P(a - z < X < a + z) = 0.99.$$

Далее действуем уже привычным способом:

$$P\left(-\frac{z}{\sigma} < \frac{X - a}{\sigma} < \frac{z}{\sigma}\right) = 0.99.$$

Отсюда найдем по таблицам

$$\frac{z}{\sigma} = 2.58.$$

Следовательно:

$$z = 2.58 \cdot \sqrt{1000 \cdot \frac{35}{12}} = 139.$$

Поэтому сумма выпавших очков почти наверняка (с вероятностью 99 %) лежит в интервале от $a - z$ до $a + z$, то есть от 3361 до 3639. Число 3298 в этот интервал не попадает, поэтому кость на правильную не похожа.

Задачи к лекции 6.

Задача 96. Вероятность появления события в каждом из независимых испытаний равна 0,8. Сколько нужно произвести испытаний, чтобы с вероятностью 0,9 можно было ожидать, что событие появится не менее 75 раз?

Задача 97. Сколько раз нужно бросить игральную кость, чтобы вероятность неравенства

$$\left| \frac{k}{n} - \frac{1}{6} \right| < 0,01$$

была не меньше, чем вероятность противоположного неравенства? Здесь k — число выпавших шестерок.

Задача 98. Вероятность наступления события в каждом из одинаковых и независимых испытаний равна 0,8. Найти вероятность того, что в 225 испытаниях событие наступит не менее 170 и не более 185 раз.

Задача 99. Вероятность наступления события в каждом из независимых испытаний равна 0,2. Найти вероятность того, что в 100 испытаниях событие наступит не менее 20 раз и не более 30 раз.

Задача 100. Вероятность появления события в каждом из 10 000 независимых испытаний равна 0,75. Найти вероятность того, что относительная частота появления события отклонится от его вероятности по абсолютной величине не более чем на 0,01.

Задача 101. Вероятность появления события в каждом

из независимых испытаний равна 0,2. Найти число испытаний, при котором с вероятностью 0,9876 можно ожидать, что относительная частота появления события отклонится от его вероятности по абсолютной величине не более чем на 0,04.

Задача 102. Вероятность появления события в каждом из 2100 независимых испытаний равна 0,7. Найти математическое ожидание числа появлений события и вероятность того, что событие появится не менее 1440 и не более 1500 раз.

Задача 103. Вероятность изготовления детали высшего сорта на данном станке равна 40%. Найти вероятность того, что среди наудачу взятых 50 деталей ровно 20 окажутся высшего сорта.

Задача 104. Вероятность появления события в каждом из 10 000 независимых испытаний равна 0,75. Найти вероятность того, что относительная частота появления события отклонится от его вероятности по абсолютной величине не более чем на 0,01.

Задача 105. Театр вместимостью 1000 мест имеет два входа, при каждом из них есть гардероб. Какова должна быть вместимость каждого из этих гардеробов, чтобы все зрители с вероятностью 0,99 могли раздеться в гардеробе того входа, через который вошли? Предполагается, что каждый из входов зрители выбирают с вероятностью 0,5 независимо друг от друга, а театр каждый день полон. В примере 26 рассмотрен случай, когда зрители приходят поодиночке. Что изменится, если зрители будут приходить парами, и сколько тогда понадобится мест?

Задача 106. Вероятность появления события в каждом из независимых испытаний равна 0,75. Найти число испытаний, которые нужно провести, чтобы с вероятностью 0,99

относительная частота появления события отклонилась от значения 0,75 по абсолютной величине не более чем на 0,01.

Задача 107. Вероятность появления события в каждом из независимых испытаний равна 0,2. Найти число испытаний, при котором с вероятностью 0,9876 можно ожидать, что относительная частота появления события отклонится от его вероятности по абсолютной величине не более чем на 0,04.

Задача 108. Отдел технического контроля проверяет 475 изделий на брак. Вероятность того, что изделие бракованное, равно 0,05. Найти с вероятностью 0,9426 границы, в которых будет заключено число бракованных изделий среди проверенных.

Задача 109. Вероятность безотказной работы телевизора "Рубин" в течение гарантийного срока равна 0,8. В каких пределах лежит количество бракованных телевизоров в партии из 1000 штук, если эти пределы должны быть гарантированы с вероятностью 0,98?

Задача 110. Вероятность прохождения инструментального контроля только что выпущенным автомобилем "Москвич" равна 0,6. Сколько надо выпустить таких автомобилей, чтобы с вероятностью 0,99 среди них нашлось бы не менее 100 прошедших инструментальный контроль?

Задача 111. Число поступивших в некий университет юношей и девушек одинаково. Какова вероятность того, что во вновь набранной группе из 30 человек юношей и девушек окажется поровну?

Задача 112. Вероятность появления события в каждом из 2000 независимых испытаний равна 0,75. Найти математическое ожидание числа появлений события и вероятность того, что отклонение числа событий от математического ожидания составит не более 2% .

Задача 113. На фирме работают 900 сотрудников. Администрация предполагает организовать на новогодние каникулы коллективный отдых сотрудников в Египте. Для этого планируется зафрахтовать авиалайнер, а билеты на оставшиеся свободными места продать на сторону. Допустим, что вероятность того, что каждый отдельно взятый сотрудник желает лететь в Египет, равна 0,2. Сколько должно быть посадочных мест в лайнере, чтобы с вероятностью 0,99 никто из желающих лететь сотрудников не остался без места?

В задачах 114 — 116 используется следующая терминология из математической статистики. Наблюдаемое в опыте значение лежит внутри вычисленных пределов, то говорят, что данные с доверительной вероятностью 0,95 согласуются с гипотезой (о случайности отклонения), в противном случае — противоречат гипотезе.

Задача 114. С 1871 по 1900 год в Швейцарии родилось 2 644 757 детей. Если принять гипотезу, что вероятность рождения мальчика равна 0,5, то в каких пределах будет с вероятностью 0,997 находиться число мальчиков?

С 1871 по 1900 год в Швейцарии родилось 1 359 671 мальчик и 1 285 086 девочек. Согласуются ли эти данные с гипотезой о том, что вероятность рождения мальчика равна 0,5?

Задача 115. Каждый из студентов потока ходит на лекцию с вероятностью $1/5$. В каких пределах с вероятностью 0,95 будет лежать число студентов на лекции, если на потоке 100 человек?

Заместитель декана, проводя проверку посещаемости, обнаружил, что на лекцию пришло 30 человек. Следует ли считать это отклонение случайным или сведения о проверке просочились из деканата?

Задача 116. Многие ботаники делали опыты по скрещиванию жёлтого (гибридного) гороха. По известной гипотезе Менделя вероятность появления зелёного гороха в таких опытах равна 0,25. Если предположить справедливость этой гипотезы, то в каких пределах с вероятностью 0,95 будет лежать число появлений зелёного гороха при 34153 опытах скрещивания?

В указанных 34153 экспериментах в 8506 случаях был получен зелёный горох. Согласуются ли эти эксперименты с гипотезой Менделя?

Примечание. Задача (с небольшими изменениями) взята из задачника по теории вероятностей Л. Д. Мешалкина, выпущенного издательством Московского университета в 1963 году, когда генетика ещё не была общепризнанной наукой.

Задача 117. Топливозаправщик заливает на нефтезаводе 10 тысяч литров солярки, после чего развозит эту солярку потребителям. Каждому потребителю следует отпускать ровно 200 литров, но топливо отпускается на глазок с точностью $\sigma = 10$ литров. Какова вероятность того, что последнему потребителю вообще не достанется топлива? достанется не более 100 литров?

Задача 118. Каков будет средний дневной заработка профессионального нищего в московском метро, если предположить, что он успевает за рабочий день обойти 80 вагонов, 10 рублей ему подадут в среднем в каждом 30-м вагоне, а полученная в каждом вагоне сумма является случайной величиной, подчиняющейся геометрическому распределению? Какова вероятность, что он заработает больше 600 рублей?

Задача 119. Импортный картофель в магазине фасуется в пакеты по 3 кг. Одна картофелина весит в среднем 120 грамм, со среднеквадратическим отклонением 20 грамм.

Ленивый фасовщик просто кладет в пакет 25 случайно выбранных картофелин. Торговая инспекция взвешивает два произвольно взятых пакета и штрафует магазин, если хотя бы в одном из них меньше 2900 грамм. С какой вероятностью магазин будет оштрафован?

Задача 120. Электрики подсчитали, что лампочки в офисе фирмы должны работать 120 000 часов в течение декабря. Срок службы одной лампочки представляет собой случайную величину, распределенную по показательному закону со средним 1000 часов. На 1 декабря осталась 121 лампочка. Какова вероятность, что их не хватит до конца месяца?

Задача 121. Срок работы электрической лампочки является случайной величиной, подчиняющейся показательному распределению со средним значением 4000 часов. В каких пределах с вероятностью 0,99 будет лежать число электрических лампочек, вышедших из строя за год на предприятии, на котором используется 1000 лампочек, каждая из которых должна быть включена в среднем в течение 2500 часов в год?

Задача 122. Электрики подсчитали, что лампочки в офисе фирмы должны работать 120 000 часов в течение декабря. Срок службы одной лампочки представляет собой случайную величину, распределенную по показательному закону со средним 1000 часов. Сколько лампочек должно быть в наличии, чтобы с вероятностью 99 процентов их хватило до конца месяца?

Задача 123. Родион Раскольников забирает у каждой из зарубленных топором старушек по 20 ± 5 копеек. Сколько старушек ему надо зарубить топором, чтобы с вероятностью 99 процентов ему хватило на бутылку водки, которая стоит 3 рубля 62 копейки?

Задача 124. В торгово-развлекательном центре “Мега” 170 магазинов. Эвелина Эммануиловна тратит на покупки в каждом из них 1000 ± 200 рублей. Какова вероятность, что перед последним магазином у нее не останется ни копейки, если сначала у нее было 175500 рублей?

Задача 125. Автопредприятие, вывозя грунт из котлована, выполнило 100 рейсов грузовиков. За один рейс грузовик перевозит $5 \pm 0,5 \text{м}^3$ грунта. В каких пределах с вероятностью 0,98 лежит общее количество вывезенного грунта?

Задача 126. В результате 16 измерений расстояния на местности получено среднее значение этого расстояния 502 метра. Выдвинута гипотеза, что расстояние на самом деле равно 500 метрам, а разница объясняется ошибками измерений. Считать, что точность одного измерения (среднеквадратическое отклонение) равно 3 метрам, а систематическая ошибка отсутствует.

Если предположить, что истинное значение расстояния равно 500 метрам, то в каких пределах с вероятностью 0,95 будут находиться средние значения результатов 16 измерений?

В задачах 127 – 131 используется следующая терминология из математической статистики. Если наблюдаемое в опыте значение лежит внутри вычисленных пределов, то говорят, что данные с доверительной вероятностью 0,95 (0,98, 0,99) согласуются с гипотезой (о случайности отклонения), в противном случае – противоречат гипотезе.

Задача 127. Покупатель коммерческого лотка в среднем тратит на покупку 100 рублей, причем эта трата подчиняется геометрическому распределению вероятностей. В каких пределах с вероятностью 0,95 лежит дневная выручка лотка, если за день приходит 100 покупателей?

В один из дней новый продавец лотка сдал выручку в

размере 9400 рублей. Следует ли считать такое снижение выручки случайностью?

Задача 128. Проверка обнаружила, что у оптового строительного склада имеется в наличии 240 м^3 песка. Завоз песка осуществлялся большегрузными автомобилями, каждый из которых за один рейс перевозит $8 \pm 0,5 \text{ м}^3$ песка. По документам, было выполнено 36 таких рейсов. Следует ли признать факт наличия недостачи или такое отклонение можно объяснить случайностью?

Задача 129. Средняя урожайность яблони в пору полного плодоношения обычно составляет от 100 до 150 килограммов. Примем, что эта урожайность является случайной величиной со средним 125 кг и среднеквадратическим отклонением 25 кг. В каких пределах тогда с вероятностью 0.98 будет находиться урожай яблок в саду, в котором 100 яблонь? Может ли урожай “случайно” составить 100 центнеров? 120 центнеров?

Задача 130. Маша выкуриивает в день 30 ± 5 сигарет, причем все их она таскает из папиных запасов. В каких пределах с вероятностью 0,95 лежит число выкуренных Машей за неделю сигарет?

Как-то раз пapa обнаружил, что у него в течение недели пропало всего 180 сигарет. Следует ли считать эту флюктуацию случайностью, или Маша действительно задумывается о том, чтобы бросить курить?

Задача 131. Мама наварила много банок с клубничным и сливовым вареньем, причем с тем и другим поровну. Если она проверит случайным образом выбранные 20 банок, сколько среди них с вероятностью 0,98 будет банок с клубничным вареньем?

Малышу одинаково нравится и то, и другое варенье, а Карлсон категорически предпочитает клубничное. Если из

проверенных мамой 20 банок было только 6 с клубничным вареньем, то следует ли считать такое отклонение случайным, или Карлсон все-таки прилетал?

Лекция 7. Случайные векторы.

Системы случайных величин, они же случайные векторы — это вектор (X_1, \dots, X_n) , каждый из компонентов которого является случайной величиной.

Мы рассмотрим случайные векторы на примере двумерных.

Закон распределения дискретной двумерной случайной величины представляет собой двумерную таблицу. Для каждой пары значений (X_1, X_2) указана вероятность того, что принимается именно эта пара значений. Разумеется, все вероятности неотрицательны, а их сумма по всей таблице равна 1. Например, в приведенной ниже таблице вероятность того, что и X_1 , и X_2 примут значение 1, равна $1/6$.

$X_1 \setminus X_2$	1	2	3
1	1/6	0	1/12
2	0	1/4	1/4
3	1/12	1/12	1/12

Непрерывная двумерная (а также многомерная) случайная величина, задается плотностью вероятностей, как и в одномерном случае. Это функция двух переменных (или n переменных) $\rho(x_1, x_2)$, всюду неотрицательная, и при этом

$$\iint_{\mathbb{R}^2} \rho(x_1, x_2) dx_1 dx_2 = 1.$$

Вероятность попадания в некоторую область Ω для непрерывной случайной величины выражается формулой:

$$P((x_1, x_2) \in \Omega) = \iint_{\Omega} \rho(x_1, x_2) dx_1 dx_2.$$

Для многомерных случайных величин также определяется функция распределения:

$$F(x_1, x_2) = P(X_1 < x_1, X_2 < x_2).$$

Функция распределения и плотность вероятностей связаны следующими соотношениями:

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y \rho(x, y) dx dy; \quad \rho(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}.$$

Из двумерных распределений первым мы встретимся с равномерным распределением. Плотность вероятности случайной величины, равномерно распределенной в некоторой ограниченной области, равна константе внутри области и нулю вне ее. Чему равна константа? Нужно, чтобы интеграл был равен единице, поэтому константа будет равна 1, деленной на площадь области.

Законы распределения компонент случайного вектора.

Для дискретных случайных величин определить закон распределения компонент довольно просто: надо просуммировать соответствующий столбец или соответствующую строку. Например, для приведенной выше таблицы закон распределения компоненты X_2 таков:

X_2	1	2	3
P	1/4	1/3	5/12

В непрерывном случае вместо суммирования надо выполнить интегрирование. Получим следующую формулу:

$$\rho_X(x) = \int_{-\infty}^{\infty} \rho(x, y) dy$$

Следующим важным понятием будет понятие условного закона распределения компонента случайного вектора. И снова формулы будут различны в случае дискретных и непрерывных случайных величин. В дискретном случае

$$P(x_i | y_j) = \frac{P(x_i, y_j)}{P(y_j)}.$$

В непрерывном же случае для одного из компонент определяется плотность вероятности:

$$\rho(x|y) = \frac{\rho(x,y)}{\int\limits_{-\infty}^{\infty} \rho(x,y)dx}.$$

Две компоненты случайной величины называются независимыми, если условное распределение каждой из них не зависит от значения второго компонента. По сути, это то же самое определение независимости, что и раньше.

Имеет место следующая важная теорема о независимости компонент случайного вектора. Эти компоненты независимы тогда и только тогда, когда функция распределения двумерной случайной величины равна произведению функций распределения компонент.

Доказательство основано на определении функции распределения. Если компоненты независимы, то независимы события $X < x$ и $Y < y$, и поэтому

$$F(x,y) = P(X < x, Y < y) = P(X < x) \cdot P(Y < y) = F_X(x) \cdot F_Y(y).$$

Обратно, если $F(x,y) = F_X(x) \cdot F_Y(y)$, то легко получить равенство, выраждающее независимость событий $X < x$ и $Y < y$.

Следствие 1. Непрерывные случайные величины независимы тогда и только тогда, когда плотность двумерного распределения равна произведению плотностей компонент.

Следствие 2. Если двумерная случайная величина равномерно распределена в некоторой области, то компоненты независимы тогда и только тогда, когда область представляет собой прямоугольник со сторонами, параллельными осям координат.

Теперь пришло время ввести новые и очень важные понятия, не имеющие аналога для одномерных случайных ве-

личин. *Коэффициентом ковариации* компонент двумерного случайного вектора или пары случайных величин называется

$$\text{cov}(X, Y) = M((X - MX)(Y - MY)).$$

Это — мера связи, или взаимной зависимости, между компонентами. Этот коэффициент размерный, и поэтому его значение зависит от применяемых единиц измерения. Поэтому чаще используется несколько другой показатель — *коэффициент корреляции*. Это тот же коэффициент ковариации, только нормированный. Его определение

$$r(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{M((X - MX)(Y - MY))}{\sqrt{DX \cdot DY}}.$$

Коэффициент корреляции принимает значения от -1 до 1 и показывает меру линейной связи между компонентами. Например, если коэффициент корреляции равен -1, то это означает, что компонента Y линейно зависит от компоненты X , причем коэффициент линейной зависимости отрицательный. Иными словами, при возрастании переменной X переменная Y будет убывать, и при этом значения Y будут полностью определяться значениями X .

Поскольку коэффициент корреляции нормирован, он не зависит от применяемых систем единиц, а также от начала отсчета. Более общее утверждение: коэффициент корреляции двух случайных величин не изменяется при линейных преобразованиях этих величин.

Приведем формулу для вычисления коэффициента ковариации. Вычислить коэффициент корреляции после этого труда не составит, поскольку вычислять дисперсии мы уже умеем. Для дискретных случайных величин

$$\text{cov}(X, Y) = \sum_{i=1}^n \sum_{j=1}^m (x_i - MX)(y_j - MY)p(x_i, y_j),$$

где

x_i , $i = 1, \dots, n$ — значение компоненты X ;

y_j , $j = 1, \dots, m$ — значение компоненты Y ;
 $p(x_i, y_j)$ — вероятность, с которой принимается эта пара значений.

Для непрерывных случайных величин формула аналогична:

$$\text{cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - MX)(y - MY)\rho(x, y)dxdy.$$

Важное утверждение. Если случайные величины независимы, то коэффициент корреляции между ними равен нулю. Такие величины называют некоррелированными. Обратное утверждение неверно. Примеры некоррелированных, но зависимых величин встречаются среди задач к этой лекции.

Коэффициент корреляции очень широко используется в статистике, о чем мы поговорим позже.

Задачи к лекции 7.

Задача 132. Двумерная случайная величина равномерно распределена в квадрате с вершинами $(1, 0)$, $(0, 1)$, $(-1, 0)$, $(0, -1)$, а $F(x, y)$ — ее функция распределения. Найти $F(0, 0)$.

Задача 133. Доказать, что $\text{cov}(X, X) = DX$.

Задача 134. Дано распределение двумерного случайного вектора (X, Y) .

$X \setminus Y$	1	2	3
1	1/18	1/18	1/18
2	1/9	1/9	1/9
3	1/6	1/6	1/6

Найти законы распределения компонент X и Y , их математические ожидания, дисперсии и коэффициент корреляции. Верно ли, что X и Y независимы?

Задача 135. Дано распределение двумерного случайного

вектора (X, Y) .

$X \setminus Y$	1	2	3
1	0	$1/4$	0
2	$1/4$	0	$1/4$
3	0	$1/4$	0

Найти законы распределения компонент X и Y , их математические ожидания, дисперсии и коэффициент корреляции. Верно ли, что X и Y независимы?

Задача 136. Дано распределение двумерного случайного вектора (X, Y) .

$X \setminus Y$	1	2	3
1	$1/16$	$1/8$	$1/6$
2	$1/3$	0	$1/24$
3	$1/12$	$1/12$	$3/16$

Найти законы распределения компонент X и Y , их математические ожидания, дисперсии и коэффициент корреляции. Верно ли, что X и Y независимы?

Задача 137. Дано распределение двумерного случайного вектора (X, Y) .

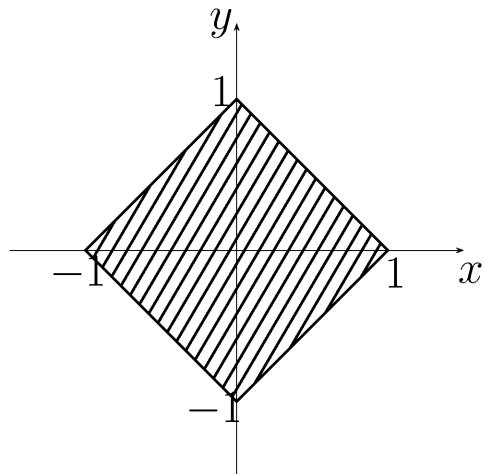
$X \setminus Y$	1	2	3
1	$1/15$	$1/10$	$1/6$
2	0	$1/3$	0
3	$1/6$	$1/10$	$1/15$

Найти законы распределения компонент X и Y , их математические ожидания, дисперсии и коэффициент корреляции. Верно ли, что X и Y независимы?

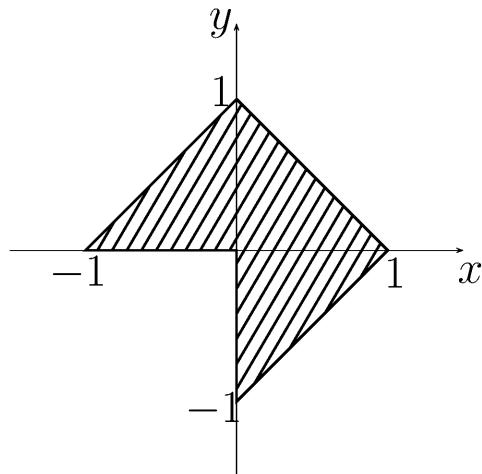
В задачах 138 — 141 случайный вектор равномерно распределен в области, изображенной на рисунке. Найти плотности вероятности компонент X и Y , их математические ожидания, дисперсии и коэффициент корреляции. Верно

ли, что X и Y независимы?

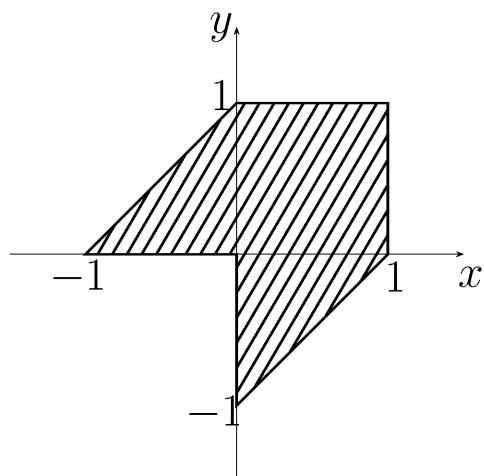
Задача 138.



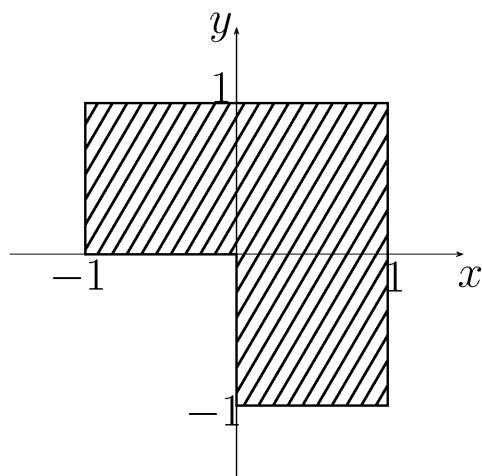
Задача 139.



Задача 140.



Задача 141.



Задача 142. (*) Случайный вектор (X, Y) равномерно распределен в единичном круге. Найти плотность вероятности случайной величины $Z = X/Y$.

Лекция 8. Производящие функции.

Производящие функции определяются для целочисленных неотрицательных дискретных случайных величин.

Пусть X — случайная величина, принимающая только целые неотрицательные значения. К таким относятся, в частности, случайные величины, подчиняющиеся распределению Пуассона, биномиальному, геометрическому распределению и другие. Для таких целых неотрицательных случайных величин можно определить так называемую производящую функцию, полезную во многих приложениях.

Определение. Пусть случайная величина X имеет следующий закон распределения

X	0	1	...	n	...
P	p_0	p_1	...	p_n	...

Тогда производящей функцией этой случайной величины называется сумма ряда

$$f_X(z) = p_0 + p_1 z + \dots + p_n z^n + \dots$$

Здесь z — некоторая формальная переменная.

Пользуясь мажорантным признаком сходимости Вейерштрасса, можно доказать, что этот ряд абсолютно сходится при $|z| < 1$.

Свойства производящей функции

1. $f_X(1) = 1$.
2. $\frac{1}{n!} f_X^{(n)}(0) = p_n$, в частности, $f_X(0) = p_0$.
3. $f'_X(1) = MX$.
4. $f''_X(1) + f'_X(1) - [f'_X(1)]^2 = DX$.
5. $f''_X(1) + f'_X(1) - [f'_X(1)]^2 = DX$.

Как видно из перечисленных выше свойств, знание производящей функции позволяет восстановить закон распределения случайной величины. Кроме того, с использованием

производящей функции можно вычислить математическое ожидание, дисперсию и другие числовые характеристики случайной величины, причем часто гораздо проще, чем напрямую.

6. Пусть X_1 и X_2 — две независимые целые неотрицательные случайные величины и $X = X_1 + X_2$ — их сумма. Тогда $f_X(z) = f_{X_1}(z) \cdot f_{X_2}(z)$.

Пример 30. Пусть случайная величина X принимает значение 1 с вероятностью p и значение 0 с вероятностью $q = 1 - p$ (то есть X — число успехов в единичном испытании). По определению, производящая функцию X равна $f(z) = pz + q$. Пользуясь этой формулой и свойством 6, получим, что производящая функция для случайной величины, подчиняющейся распределению Бернулли (число успехов в n независимых испытаниях), вычисляется по формуле

$$f_n(z) = (pz + q)^n.$$

Используя свойства 3 и 5, получим, что математическое ожидание и дисперсию для распределения Бернулли равны соответственно

$$MX = np, \quad DX = npq.$$

Пример 31. Производящая функция для случайной величины, подчиняющейся распределению Пуассона с параметром λ , вычисляется по формуле

$$f_{Poisson}(z) = e^{\lambda(z-1)}.$$

Математическое ожидание и дисперсию для распределения Пуассона равны соответственно

$$MX = \lambda, \quad DX = \lambda.$$

Пример 32. Производящая функция для случайной величины, подчиняющейся геометрическому распределению

с параметром p , вычисляется по формуле

$$f_{geom}(z) = \frac{p}{1 - qz}.$$

Математическое ожидание и дисперсия для геометрического распределения равны соответственно

$$MX = \frac{q}{p}, \quad DX = \frac{q}{p^2}.$$

Задачи к лекции 8.

Задача 143. Доказать, что ряд для производящей функции абсолютно сходится при $|z| < 1$.

Задача 144. Доказать свойства 1 – 5.

Задача 145. Пользуясь формулами для коэффициентов произведения степенных рядов, а также равенством

$$\begin{aligned} P(X = n) = & P(X_1 = 0)P(X_2 = n) + P(X_1 = 1)P(X_2 = n-1) + \\ & \cdots + P(X_1 = n-1)P(X_2 = 1) + P(X_1 = n)P(X_2 = 0), \end{aligned}$$

справедливым для независимых целых неотрицательных случайных величин X_1 и X_2 , доказать свойство 6.

Задача 146. Проводится n независимых испытаний, вероятности успеха в них равны соответственно p_1, \dots, p_n , а вероятности неудач соответственно q_1, \dots, q_n . Найти производящую функцию для числа успехов в этих испытаниях.

Задача 147. Орудие стреляет по цели до трех попаданий. Сначала вероятность попадания при каждом выстреле равна 0,6. После первого попадания цель начинает защищаться, и вероятность попадания при единичном выстреле падает до 0,3. Найти математическое ожидание числа потраченных снарядов.

Лекция 9. Многомерное нормальное распределение.

Это распределение является, пожалуй, самым важным примером многомерных случайных величин. Его плотность вероятностей на n -мерном векторе

$$\vec{x} = \begin{pmatrix} x_1 \\ \dots \\ x_n \end{pmatrix}$$

задается формулой

$$\rho(\vec{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^n \sqrt{\det(B)}} \cdot \exp\left\{-\frac{1}{2}(\vec{x} - \vec{a})^t B^{-1}(\vec{x} - \vec{a})\right\},$$

где

$$\vec{a} = \begin{pmatrix} a_1 \\ \dots \\ a_n \end{pmatrix}$$

— вектор средних значений компонент,

$$B = \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \dots & \dots & \dots \\ b_{n1} & \dots & b_{nn} \end{pmatrix}$$

— симметричная положительно определенная матрица, называемая ковариационной матрицей. Как принято, верхний индекс t означает операцию транспонирования, B^{-1} — обратная матрица, $\det(B)$ — определитель. Напомним, что выражение в показателе экспоненты называется квадратичной формой.

Часто многомерное нормальное распределение возникает в задачах, связанных с измерением. Тогда компоненты случайного вектора — это результаты измерений, возможно, различных величин, причем эти результаты получены при совместных измерениях. Тогда величины могут быть связаны между собой, то есть коррелированы.

Параметры a_1, \dots, a_n — это, как уже было сказано, средние значения компонент. Коэффициенты матрицы B также

имеют ясную интерпретацию: коэффициент b_{ij} — это коэффициент ковариации между i -м и j -м компонентами. В частности, диагональные элементы матрицы B — дисперсии компонент.

Воспользуемся теперь известной теоремой из линейной алгебры — теоремой о приведении к главным осям. Она утверждает, что для любой квадратичной формы найдется ортонормированный базис, в котором ее матрица диагональна.

Пусть y_1, \dots, y_n — координаты, в которых эта матрица диагональна. Допустим также, что все средние значения компонент y_i равны нулю — этого тоже нетрудно достичь линейным преобразованием. Тогда в этих координатах формула для плотности вероятности примет вид:

$$\rho(\vec{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^n \sigma_1 \cdots \sigma_n} \cdot \exp\left\{-\frac{1}{2}\left(\frac{y_1^2}{\sigma_1^2} + \cdots + \frac{y_n^2}{\sigma_n^2}\right)\right\},$$

Здесь σ_i^2 — дисперсии компонент.

Мы видим, что все ковариации между компонентами равны нулю, то есть новые переменные — некоррелированы. Но в правой части последней формулы, как легко видеть, стоит произведение плотностей вероятностей нормально распределенных величин. Таким образом, по ранее доказанной теореме, компоненты независимы, и для многомерного нормального распределения некоррелированность и независимость — эквивалентны.

Пример 33. В различных статистических исследованиях, проводимых, например, биологами, медиками, историками, и т. д., каждый объект исследования представляется набором значений признаков, то есть n -мерным вектором (x_1, \dots, x_n) . При этом переменные часто коррелированы. Например, если один признак — это рост человека или животного, а другой — его вес, то, скорее всего, между этими двумя переменными будет большая положительная корреля-

ция.

В исследованиях, например, с целью классификации, часто требуется определить меру сходства, или расстояние, между объектами. Обычное евклидово расстояние здесь не годится, как из-за разницы в разбросах переменных, так и из-за корреляций между ними. Поэтому в такой ситуации часто используют обобщение евклидова расстояния — расстояние Махаланобиса. Оно отличается от расстояния Евклида тем, что учитывает корреляции между переменными и инвариантно к масштабу. Оно определяется формулой длины вектора (ведь расстояние между векторами — это длина вектора — их разности):

$$\|\vec{x}\|_M = \sqrt{(\vec{x} - \vec{a})^t B^{-1} (\vec{x} - \vec{a})}.$$

Как видно, под корнем стоит то же выражение, что и в показателе степени в формуле плотности многомерного нормального распределения, и переменные имеют тот же смысл. Обычно, правда, в реальной жизни ковариационная матрица B точно не известна, и ее оценивают, исходя из результатов предыдущих исследований. Считается, что каждая единица классификации (кластер) имеет приблизительно нормальное распределение. И тогда от расстояния от объекта до центра кластера, измеренного с помощью расстояния Махалонобиса, зависит, будет ли отнесен этот объект к этому кластеру, или нет, то есть решение задачи классификации.

Если же распределение объектов в кластере явно не сферическое, а, например, эллипсоидальное, то было бы естественным учитывать не только расстояние до центра масс, но и направление на него. В направлении короткой оси эллипса заданная точка должна быть ближе к центру масс, чтобы принадлежать кластеру, в то время как в направлении длинной оси она может быть дальше.

Для записи этого в математическом виде эллипсоид, луч-

шим образом представляющий вероятностное распределение множества, может быть задан матрицей ковариаций множества. Расстояние Махalanобиса — это просто расстояние между заданной точкой и центром масс, делённое на ширину эллипсоида в направлении заданной точки.

Расстояние Махalanобиса было сформулировано в 1936 году индийским статистиком Прасанта Чандра Махалонобисом во время работы над идентификацией сходства черепов, основанной на измерениях 1927 года. Оно широко используется в кластерном анализе и других методах классификации.

Таким образом, если случайный вектор имеет многомерное нормальное распределение, то можно подобрать такую замену переменных, что в этих новых переменных компоненты случайного вектора станут независимыми. Более того, такую замену можно выбрать ортогональной.

Еще одним “усилением” этого утверждения служит теорема, которая нам пригодится в дальнейшем. Пусть случайные величины X_1, \dots, X_n имеют стандартное нормальное распределение и независимы. Пусть мы случайный вектор с этими компонентами подвергаем ортогональному преобразованию, то есть получаем новые случайные величины Y_1, \dots, Y_n по формуле

$$\begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix} = B \begin{pmatrix} X_1 \\ \dots \\ X_n \end{pmatrix}.$$

Тогда эти случайные величины Y_1, \dots, Y_n также имеют стандартное нормальное распределение и независимы.

Доказательство. Плотность вероятности случайного вектора $(X_1, \dots, X_n)^t$ в точке $\vec{x} = (x_1, \dots, x_n)^t$ равна

$$\rho(\vec{x}) = \frac{1}{(\sqrt{2\pi})^n} \cdot \exp\left\{-\frac{1}{2}\left(x_1^2 + \dots + x_n^2\right)\right\}.$$

Мы можем считать, что преобразованный случайный вектор (Y_1, \dots, Y_n) — это тот же вектор, просто записанный в новых координатах, а матрица B — это матрица ортогональной замены базиса. Тогда для преобразованного случайного вектора (Y_1, \dots, Y_n) плотность вероятности в точке с координатами $\vec{y} = B\vec{x}$, разумеется, такая же (поскольку это та же самая точка). Но ортогональное преобразование сохраняет длины, то есть

$$y_1^2 + \dots + y_n^2 = x_1^2 + \dots + x_n^2.$$

Поэтому плотность вероятности случайного вектора $(Y_1, \dots, Y_n)^t$ в точке $\vec{y} = (y_1, \dots, y_n)^t$ также равна

$$\rho(\vec{y}) = \frac{1}{(\sqrt{2\pi})^n} \cdot \exp\left\{-\frac{1}{2}\left(y_1^2 + \dots + y_n^2\right)\right\}$$

и, как и ранее, распадается в произведение одномерных нормальных плотностей, что и требовалось доказать.

Распределения хи-квадрат, Стьюдента, Фишера.

Если независимые случайные величины X_1, \dots, X_n независимы и имеют стандартное нормальное распределение, то можно с их помощью определить новые случайные величины, играющие важную роль в математической статистике.

Рассмотрим случайную величину

$$Z = X_1^2 + \dots + X_n^2.$$

Распределение, которому подчиняется такая случайная величина, называется распределением χ^2 (хи-квадрат) с n степенями свободы. Математическое ожидание и дисперсия для распределения χ^2 равны соответственно:

$$MZ = n; \quad DZ = 2n.$$

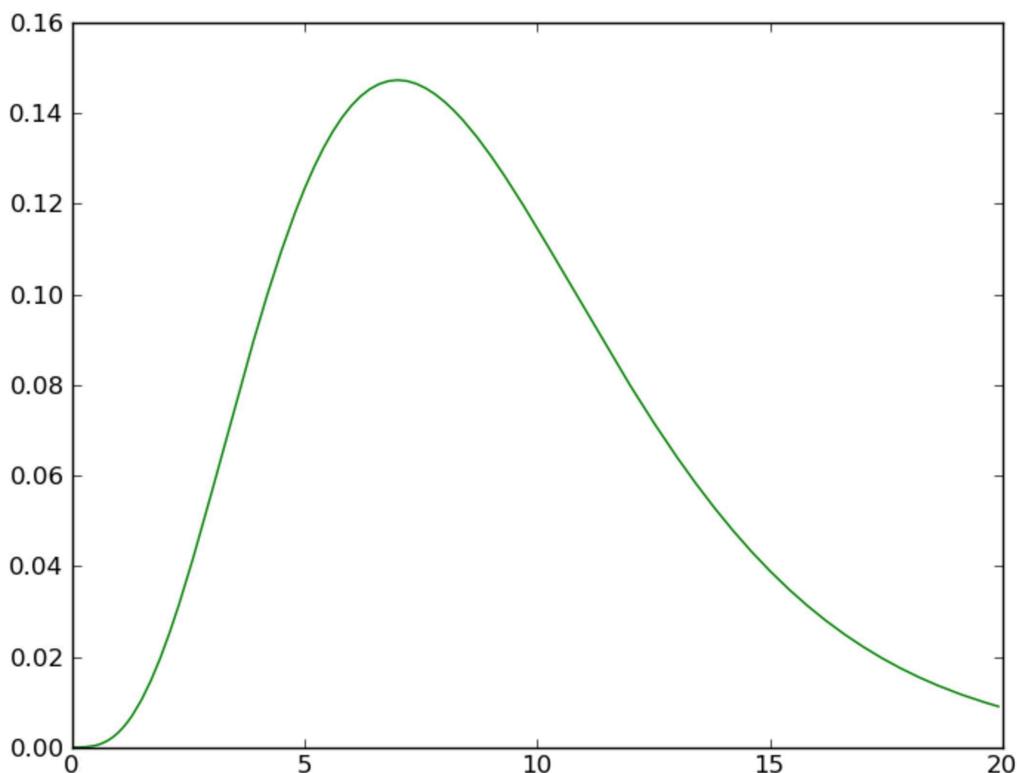


Рисунок 12. Плотность распределения χ^2 с 9 степенями свободы.

Пусть X и Z — независимые случайные величины, причем X распределена стандартно нормально, а Z распределена по χ^2 с n степенями свободы. Образуем новую случайную величину:

$$T = \frac{X}{\sqrt{Z/n}}.$$

Такую случайную величину называют распределенной по Стьюденту с n степенями свободы. Математическое ожидание и дисперсия для распределения Стьюдента равны

$$MT = 0; \quad DT = \frac{n}{n - 2},$$

если же $n \leq 2$, то дисперсия бесконечна.

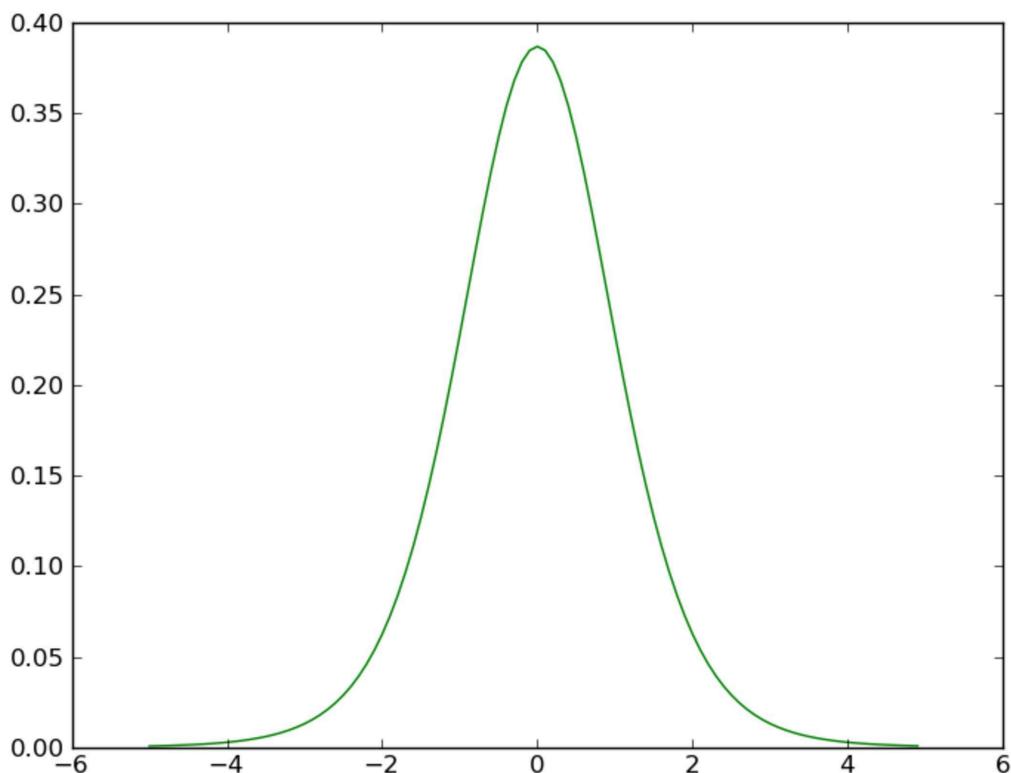


Рисунок 13. Плотность распределения Стьюдента с 8 степенями свободы.

Еще одно важное распределение называется распределением Фишера. Пусть Z_m и Z_n — две независимые случайные величины, распределенные по закону χ^2 с m и n степенями свободы соответственно. Тогда случайная величина

$$F = \frac{Z_m/m}{Z_n/n}$$

называется распределенной по Фишеру с (m, n) степенями свободы. Для этой случайной величины

$$MF = \frac{n}{n-2}; \quad DZ = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}.$$

При $n \leq 4$ ее дисперсия бесконечна, а при $n \leq 2$ бесконечно и математическое ожидание.

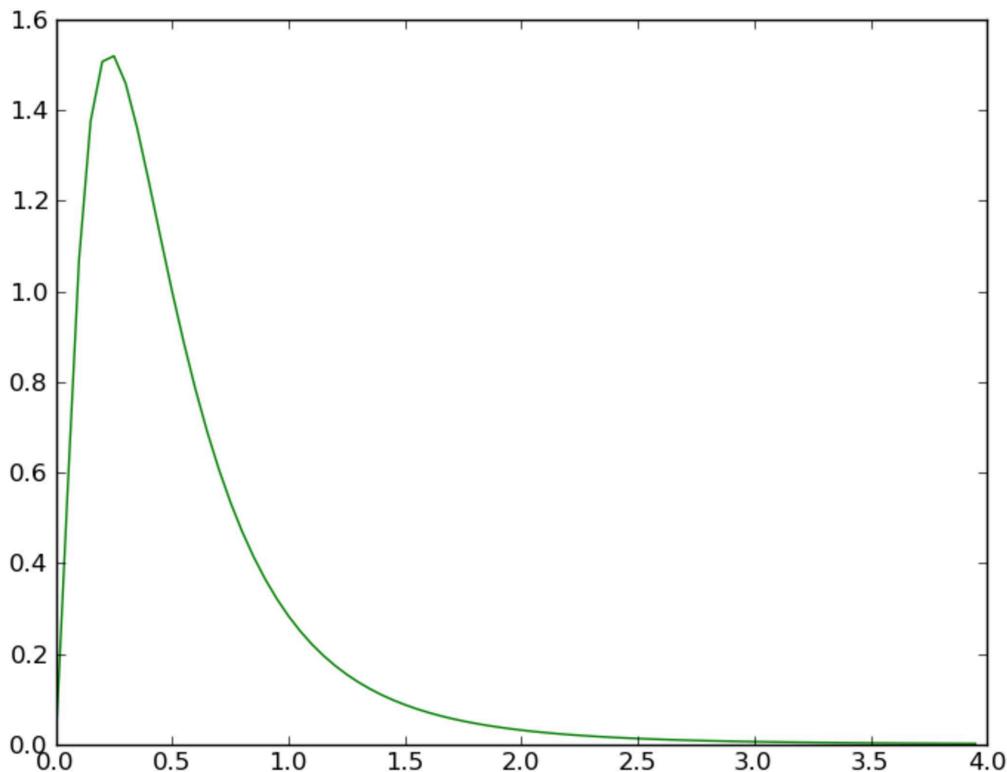


Рисунок 14. Плотность распределения Фишера с 5 степенями свободы числителя и 11 степенями свободы знаменателя.

Все три эти распределения широко применяются в математической статистике. Важной и часто встречающейся задачей является нахождение квантилей этих распределений. Мы встретимся с этим позднее, при проверке статистических гипотез.

Далеко не во всех книгах по статистике есть достаточно подробные таблицы квантилей этих распределений. В настоящее время для нахождения квантилей используются электронные таблицы. Встроенные функции для вычисления квантилей есть во всех достаточно распространенных офисных пакетах, как для Windows, так и для других платформ. Для распределения хи-квадрат можно пользоваться функциями СТЬЮДРАСПОБР (в Microsoft Office) или CHIINV (в Libre Office). Для распределения Стьюдента встроенные

функции называются соответственно СТЪЮДРАСПОБР и TINV, а для распределения Фишера — FPACПОБР и FINV.

К сожалению, сразу работать с этими функциями вряд ли получится. Они не только не соответствуют российскому ГОСТу, но и могут без предупреждения меняться от версии к версии.

Например, 95% квантиль распределения хи-квадрат с 4 степенями свободы равен 9,488. Для того, чтобы получить это значение в Libre Office, следует набрать CHIINV(0,05;4) вместо естественного CHIINV(0,95;4). Иными словами, встроенная функция показывает не квантиль — точку, левее которой случайная величина окажется с заданной вероятностью, а точку, правее которой с заданной вероятностью окажется случайная величина. Аналогично во многих версиях работают электронные таблицы с распределением Фишера.

Еще меньше повезло распределению Стьюдента. В большинстве версий встроенная функция “без объявления войны” показывает границу двусторонней симметричной области, внутри которой случайная величина окажется с заданной вероятностью. Например, в Libre Office, чтобы получить 95% квантиль распределения Стьюдента с 10 степенями свободы (равный 1,812), следует набрать TINV(0,1;10). Если же набрать наиболее естественное TINV(0,05;10), то ответом будет 2,228 — 97,5% квантиль.

Поэтому для успешной работы с электронными таблицами следует разобраться, как именно работают встроенные статистические функции в вашей версии. Ниже приведена небольшая справочная таблица, которая позволит это понять.

Распределение	Ст. свободы	Вероятность p	p -квантиль
Стьюдента	5	0,95	2,015
Стьюдента	5	0,99	3,365
Стьюдента	10	0,98	2,359
Стьюдента	10	0,99	2,764
Хи-квадрат	5	0,95	11,070
Хи-квадрат	5	0,98	13,388
Хи-квадрат	10	0,98	21,161
Хи-квадрат	10	0,99	23,209
Фишера	5, 10	0,95	3,326
Фишера	5, 10	0,98	4,555
Фишера	12, 10	0,98	3,868
Фишера	12, 10	0,99	4,706

Задачи к лекции 9.

Задача 148. Найти 96% квантиль распределения Стьюдента с 9 степенями свободы.

Задача 149. Найти 98% квантиль распределения Стьюдента с 7 степенями свободы.

Задача 150. Найти 99% квантиль распределения Стьюдента с 11 степенями свободы.

Задача 151. Найти 96% квантиль распределения хи-квадрат с 9 степенями свободы.

Задача 152. Найти 98% квантиль распределения хи-квадрат с 7 степенями свободы.

Задача 153. Найти 99% квантиль распределения хи-квадрат с 11 степенями свободы.

Задача 154. Найти 96% квантиль распределения Фишера с (4,5) степенями свободы.

Задача 155. Найти 98% квантиль распределения Фишера с (7,8) степенями свободы.

Задача 156. Найти 99% квантиль распределения Фишера с (18,12) степенями свободы.

Задача 157. Найти распределения квадрата случайной величины, распределенной по Стьюденту с n степенями свободы.

Задача 158. Пусть Q — p -квантиль распределения Фишера с (n, n) степенями свободы. Чему равен $(1 - p)$ -квантиль того же распределения?

Лекция 10. Основные понятия математической статистики

Изучение математики начинается с чисел. Изучение статистики мы начнем с вероятностного аналога числа — со *случайной величиной*.

Никто из нас не знает будущего. Мы не знаем, скажем, какая завтра будет температура воздуха, или сколько посетителей придет в нашу организацию. Все это — случайные величины.

Но, тем не менее, мы можем довольно успешно планировать свою деятельность на ближайшее время. Из предыдущего опыта мы знаем, скажем, что ежедневно к нам приходят от 20 до 30 человек. Или, если завтра, например, 1 февраля, то вряд ли температура воздуха на улице будет $+20^{\circ}\text{C}$.

Когда завтра наступит, к нам придет 27 человек, а термометр покажет -5°C . Это — реализации случайной величины.

Все возможные реализации случайной величины составляют *генеральную совокупность*. Нам эта совокупность обычно известна только приблизительно. Однако мы можем судить о том, какие значения случайной величины довольно вероятны, какие — маловероятны, а какие — практически невозможны. Наблюдая все новые и новые реализации случайной величины, мы увеличиваем свои знания о ней.

Те реализации, которые мы уже смогли наблюдать, составляют *выборку*. Основная задача математической статистики — по выборке сделать какие-то выводы о генеральной совокупности.

Генеральная совокупность может состоять из конечного или бесконечного числа возможностей. Так, число посетителей всегда конечно, а вот возможных значений темпера-

туры воздуха — бесконечное число (если не ограничивать точность измерений). А вот выборка — всегда конечный набор значений. Мы будем обозначать эти значения так: x_1, \dots, x_n , а число n будем называть объемом выборки.

Пример 34. На сайте <http://ru.wikinews.org/> приведены средние суточные температуры в феврале 2013 года в Москве.

Таблица 1.

01	02	03	04	05	06	07
-2,3	0,3	0,6	-1,3	-4,0	0,4	-0,1
08	09	10	11	12	13	14
-1,8	-1,2	0,1	1,2	0,2	-1,8	-4,0
15	16	17	18	19	20	21
-6,3	-8,8	-7,1	-7,9	-8,6	-9,1	-10,6
22	23	24	25	26	27	28
-9,2	-7,7	-5,2	-2,1	-1,4	-0,1	1,3

Средние температуры в феврале 2013 года в Москве.

Эти числа и представляют собой выборку. Хотя и считается, что февраль 2013 года был немного теплее обычного, тем не менее эти числа характеризуют февральскую погоду в Москве.

Среднее и разброс выборки

Видимо, самой важной характеристикой выборки является вреднее значение. Его можно вычислить по формуле

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}.$$

В Microsoft Excel для вычисления среднего можно воспользоваться стандартной функцией СРЗНАЧ, а в Libre Office — функцией AVERAGE.

Средняя суточная температура в феврале 2013 года составляла -3,45 градуса.

То, что мы получили, называется *выборочным средним*.

Также его называют выборочным средним арифметическим. По-видимому, чем больше наша выборка, тем больше это выборочное среднее похоже на истинное среднее всей генеральной совокупности.

И здесь же нас ждет первая тонкость.

“Средним” можно с полным основанием назвать и другое значение. Если мы все числа из нашей выборки выстроим в порядке возрастания, то среднее — то значение, которое стоит в самой середине. Такое значение называют *выборочной медианой*. Скажем, если всего значений 31, то медиана — это 16-ое “по росту”. А если всего значений 28, то тогда медиана — среднее между 14 и 15 значениями. В Microsoft Excel для вычисления медианы используется стандартная функция МЕДИАНА, а в Libre Office — функция MEDIAN (кто бы мог подумать!).

Для февральских температур, как легко проверить, медиана составит -1,95 градуса.

Медиана, как видно, часто не совпадает с выборочным средним, поэтому когда мы слышим слова “в среднем”, стоит понимать, какое именно среднее имеется в виду. В случае температуры воздуха или числа посетителей разница, как мы уже убедились, невелика, но есть и куда менее безобидные случаи.

Пример 35 (из книги М. Гарднера “А ну-ка, догадайся”). Том решил устроиться на работу. На собеседовании начальник сказал ему: средняя зарплата у нас — 600 долларов в неделю.

Проработав неделю, Том обратился к начальнику: — Я поговорил со всеми рабочими, и выяснил, что никто из них не получает больше 200 долларов в неделю. Как может средняя зарплата быть 600?

Все правильно, Том — ответил начальник — и сейчас я

это докажу. Вот смотри:

Я (начальник) получаю	\$4800
Мой заместитель получает	\$2000
Каждый из 6 моих родственников в правлении получает	\$500
Каждый из 5 бригадиров получает	\$400
Каждый из 10 рабочих получает	\$200
Всего у нас работает 23 человека, и получают они	\$13800

Так что среднее арифметическое — 600 долларов.

Вот если бы ты спросил про медиану, тогда бы я сказал, что она составляет 400 долларов. А рабочие у меня получают 200. Теперь понятно?

Понятно — ответил Том — ищите дурачков в другом месте!

Пример 36. “Российская газета” опубликовала 25 мая 2010 года статью под заголовком “Средняя зарплата в апреле составила 20 383 рубля”. В этой статье приведены данные Росстата Российской Федерации о зарплатах в России. Какое среднее имеется в виду?

Оказывается (и об этом прямо написано в статье) — среднее арифметическое. К сожалению, этот показатель не очень хорош для описания жизни страны. Дело тут вот в чем. Если, к примеру, какой-то один россиянин заработает, скажем, 10 миллиардов рублей, то средняя арифметическая зарплата увеличится примерно на 200 рублей. Иными словами, средняя арифметическая зарплата может сильно измениться, если изменятся доходы у малой части населения.

Другой показатель — медиана — свободен от этого недостатка. Какова же медиана зарплат россиян? Та же статья дает ответ и на этот вопрос. Там написано: “Зарплата примерно половины россиян недотягивает до 16 тысяч рублей в месяц”. То есть медиана — примерно 16 тысяч рублей.

Во многих случаях разница между средними может быть, как мы видели, не очень маленькой. Поэтому следует понимать, какое среднее имеется в виду. Недопонимание может быть причиной ошибок и даже манипуляций.

Вот, например, цитата с известного сайта zadolba.li

8072 - Всё взять и поделить (4 мая 2012, 10:15)

Когда я впервые прочитал что-то типа “они складывают доходы олигарха и уборщицы, берут среднее арифметическое и так получают уровень благосостояния”, мне было смешно. Когда я это увидел в десятый раз, то уже не смеялся.

На сотый раз меня это люто, бешено задолбало. Хочется взять такого человека за голову и, аккуратно ударяя ей об стол прорычать на ухо: “Запомни, кретин, не используют среднее арифметическое в подобной статистике!” Иногда берут медианные показатели, но чаще разбивают на группы по уровню дохода, да ещё и замеряют отношение доходов между крайними группами. Смешно, но это называется “десильный коэффициент”.

Понимаю, что для многих это слишком сложно, что они знают только “среднее арифметическое”, но сколько же можно тиражировать свое невежество!

Как видно из цитированной статьи, именно среднее арифметическое Росстат и использует.

Кроме среднего значения, наша генеральная совокупность должна характеризоваться еще и тем, насколько велик разброс значений вокруг этого среднего. Для этого чаще всего используется *выборочная дисперсия*. Она вычисляется по формуле

$$s^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}.$$

Почему в знаменателе стоит $n - 1$ вместо ожидаемого n ,

будет объяснено позже.

Для нахождения дисперсии в Microsoft Excel можно воспользоваться стандартной функцией ДИСП. Эта функция, впрочем, делит на n , а не на $n - 1$, поэтому для получения правильного значения нужно результат умножить на дробь $(n - 1)/n$.

В Libre Office для вычисления выборочной дисперсии используется стандартная функция VAR. Есть еще функция VARP, аналогичная ДИСП.

Важно знать, что дисперсия — это не средний разброс, а квадрат среднего разброса. Чтобы получить разброс, нужно из дисперсии извлечь квадратный корень, и тогда получится *выборочное стандартное отклонение*. Его еще часто называют среднеквадратическим отклонением. Для его вычисления можно воспользоваться стандартной функцией СТАНДОТКЛОН.

Для характеристик совокупности используется слово “выборочная”. Это слово означает, что значение получено только по данным, представленным в выборке. Здесь важно понимать, что у генеральной совокупности тоже есть и среднее, и медиана, и дисперсия, но мы их не знаем. По выборке, то есть по тем только данным, которыми мы располагаем, мы и стараемся определить эти истинные значения. Поэтому для этих оценок и используется слово “выборочная”.

Добавим еще, что любое значение, полученное по данным выборки, любую функцию элементов выборки принято называть *статистикой*.

Гистограммы

Выборочные среднее и дисперсия — далеко не вся информация, которую можно извлечь из выборки. Еще одной важной вещью будет приблизительное представление о том, как выглядит распределение. Для этого использу-

ют способ наглядного представления данных — построение гистограммы.

Построение гистограммы начинается с построения группированной выборки. Все данные из выборки разбиваются на несколько интервалов, после чего подсчитывается, сколько значений лежит в каждом интервале. Например, для февральской погоды можно взять такие интервалы.

Таблица 2.

-12 — -9	-9 — -6	-6 — -3	-3 — 0	0 — 3
3	6	3	9	7

Средние температуры в феврале 2013 года в Москве:
группированная выборка.

Далее, построим рисунок. Разобьем горизонтальную ось на 5 отрезков: от -12 до -9, от -9 до -6, и так далее до 0 — 3. Над каждым отрезком изобразим столбик соответствующей высоты: над первым — 3 единицы, над вторым — 6 единиц, и так далее. Полученный график и называется гистограммой. Он изображен на рисунке 15.

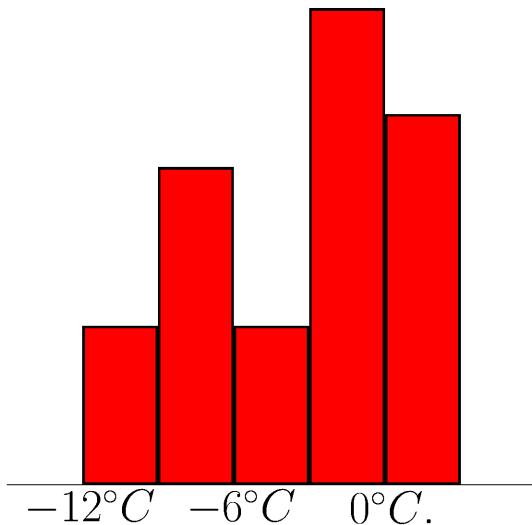


Рисунок 15. Гистограмма температур воздуха в Москве в феврале 2013 года.

Длины интервалов при этом должны быть одинаковы, то есть весь промежуток, в который попадают наблюдения выборки, должен быть разбит с помощью одинакового шага. Каким должен быть этот шаг разбиения? Общих правил не существует, только некоторые эмпирические правила. Чаще всего рекомендуют пользоваться эмпирической “формулой Стерджесса”: число интервалов разбиения k должно

быть примерно равно

$$k = \log_2 n + 1,$$

где n — число наблюдений в выборке. В приведенном примере по этой формуле получается 5 — 6 интервалов, у нас их 5.

Еще одно соображение проистекает из психологии: гистограммы используются для наглядного представления данных, чтобы читающий мог “с одного взгляда” ухватить изображенную закономерность. Поэтому интервалов не должно быть намного больше семи: большинство людей не может удерживать в памяти более 7 вещей одновременно. Это соображение, впрочем, не работает, когда данных много: тогда правильно выбранный шаг (по формуле Стерджесса) позволяет построить довольно гладкую кривую, которая воспринимается не по частям, а целиком.

Что значит “много данных”? В соответствии с “формулой Стерджесса”, если число наблюдений больше 100, то интервалов будет больше 7.

И последнее. Шаг должен быть целым и достаточно круглым числом просто для удобства интерпретации. Гораздо легче понять “между -6 и -3”, чем “между -2,45 и 0,55”.

В Microsoft Excel нет встроенных средств для подготовки данных для гистограммы, но можно воспользоваться стандартной функцией СЧЕТЕСЛИ. А если данные (в виде группированной выборки, то есть таблицы вроде приведенной выше) готовы, то в Microsoft Excel среди возможностей построения графиков есть и гистограммы.

Можно ли было выбрать шаг как-то по-другому? Наверное, да, но давайте посмотрим. Сначала рассчитаем размах нашей выборки: это разница между максимальным (1,3) и минимальным (-10,6) значениями. Таким образом, размах выборки равен 11,9. Стало быть, все наши интервалы долж-

ны умещаться на отрезке длины чуть больше, чем размах выборки. Например, можно взять отрезок длины 12, в то время как мы взяли отрезок длины 15. Из формулы Стерджесса следует, что у нас должно быть 5 или 6 интервалов. Пусть их будет 6, тогда длина каждого из них составит $12:6 = 2$. Надо тогда разбивать на такие интервалы: от -10,6 до -8,6, и так далее.

Теперь видны недостатки нашего нового разбиения:

- некруглые и поэтому не слишком наглядные границы интервалов;
- поскольку речь идет о температурах воздуха, есть смысл сделать температуру в 0 градусов границей интервала, чтобы сразу видеть, сколько было плюсовых и минусовых температур. Этого в новом разбиении нет;
- февраль частенько бывает более холодным: -14 — -15 градусов никого не удивят. Если для старого разбиения можно просто добавить лишний интервал, и не возникнет противоречий с формулой Стерджесса, то для нового разбиения такого запаса нет, и придется все пересчитывать.

Таким образом, можно сделать такой вывод: построение гистограмм является своего рода искусством, и удачное разбиение требует некоторого опыта.

Гистограммы используются довольно часто, однако это далеко не единственный метод наглядного представления данных. Такие методы часто используются как для того, чтобы объяснить ситуацию, скажем, руководителям, которые обычно не владеют статистическими методами, так и для того, чтобы выдвинуть какую-то гипотезу, которую можно проверить.

Разумеется, существуют и другие методы наглядного представления данных, кроме построения гистограмм. В результате бурного развития компьютерной техники и программирования появляются все новые и новые методы. Тому,

кто хотел бы ознакомится с некоторыми из них, следует по-рекомендовать недавно вышедшую книгу Нейтана Яу "Искусство визуализации в бизнесе".

Точечные оценки математического ожидания и дисперсии.

По группированной выборке можно построить другие оценки математического ожидания и дисперсии. Для этого мы просто считаем, что все значения выборки, попавшие в интервал, равны середине этого интервала. Таким образом, пусть у нас есть группированная выборка вида

c_1	\dots	c_n
k_1	\dots	k_n

Здесь

c_1, \dots, c_n — середины интервалов;

k_1, \dots, k_n — число наблюдений в интервалах;

$n = k_1 + \dots + k_n$ — общее число наблюдений.

Тогда мы получаем такие формулы

$$\bar{x}_{\text{групп}} = \frac{c_1 k_1 + \dots + c_n k_n}{n}$$

$$s_{\text{групп}}^2 = \frac{(c_1 - \bar{x}_{\text{групп}})^2 k_1 + \dots + (c_n - \bar{x}_{\text{групп}})^2 k_n}{n - 1}.$$

Для февральской погоды получим

$$\bar{x}_{\text{групп}} = 3,32; \quad s_{\text{групп}}^2 = 16,89.$$

Как видим, оценки математического ожидания и дисперсии по группированной выборке могут отличаться от оценки по исходной выборке, и чаще всего, действительно отличается. Это не ошибка. Каждое из этих чисел — оценка неизвестного параметра — истинного математического ожидания или дисперсии генеральной совокупности, и еще

неизвестно, какая лучше. Обычно обе оценки достаточно точны. Впрочем, мы еще вернемся к этому вопросу, когда научимся сравнивать оценки.

Задачи к лекции 10

Задача 159. В деревне Большие Ухабы проживает 20 человек. Из них 19 имеют среднемесячный доход 5 тысяч рублей, а один — 500 тысяч рублей. Каков средний доход жителей деревни Большие Ухабы? Как изменится этот средний доход, если самый богатый житель будет получать не 500, а 900 тысяч рублей?

Задача 160. Все предприятия России обязаны ежеквартально сдавать утвержденную статистическую отчетность. Форма Государственной статистической отчетности П-4 “СВЕДЕНИЯ О ЧИСЛЕННОСТИ, ЗАРАБОТНОЙ ПЛАТЕ И ДВИЖЕНИИ РАБОТНИКОВ” содержит следующие подпункты:

- Средняя численность работников за отчетный месяц;
- Фонд начисленной заработной платы работников за отчетный месяц.

Какое среднее значение можно получить на основании этих данных?

Лекция 11. Порядковые статистики.

Рассмотрим такую задачу. В аудитории 10 абитуриентов сдают экзамен. Номера их экзаменационных листов:

$$188; 1; 54; 141; 158; 193; 120; 190; 242; 92.$$

Экзамен проходит не только в этой аудитории. Можно ли по приведенным данным оценить, сколько всего абитуриентов сдают экзамен?

Таким образом, мы имеем дело со следующей задачей: экзамен пришли сдавать N человек, соответственно было выдано N экзаменационных листов. В нашу аудиторию попали случайно выбранные $n = 10$ человек, и их номера экзаменационных билетов x_1, \dots, x_n нам известны. Чему равно N ?

Для оценки неизвестного параметра N следует подобрать какую-то функцию от известных нам номеров, то есть от выборки. Как мы знаем, любую функцию выборки называют *статистикой*.

Напрашивается идея оценить N как удвоенное выборочное среднее, то есть использовать статистику

$$N_1 = 2 \cdot \frac{x_1 + \dots + x_n}{n}.$$

В нашем случае мы получим $N_1 = 275.8$. Как мы увидим чуть позже, эта оценка не очень точна.

Здесь можно раскрыть один секрет. Разумеется, автор знает, чему равно N . Автор сам получил эти номера, используя генератор случайных чисел, а на самом деле $N = 256$.

Так можно ли подобрать статистику получше?

Еще одна статистика могла бы быть такой: сумма максимального и минимального номеров. Давайте расположим номера по возрастанию, получим такой набор:

$$1; 54; 92; 120; 141; 158; 188; 190; 193; 242.$$

Для общего случая введем обозначения $x(1), \dots, x(n)$: это те же самые наши x_1, \dots, x_n , только расположенные в порядке возрастания: $x(1) < \dots < x(n)$. Вновь введенные величины $x(1), \dots, x(n)$ называются *порядковыми статистиками*.

Новая статистика для оценки N будет равна

$$N_2 = x(1) + x(n) = 243.$$

В нашем случае она оказалась лучше, чем N_1 . Но может быть, это случайность?

А нельзя ли в качестве оценки N использовать $x(n)$ — максимальный номер в выборке? На первый взгляд это плохо — оценка почти всегда будет меньше истинного значения. Но может быть, с этим можно что-нибудь сделать?

Для ответа на эти вопросы придется понять, как сравнивают статистики. Самое главное — любая статистика есть случайная величина, ее значение меняется от опыта к опыту. Поэтому у статистики, как и у всякой случайной величины, должны быть числовые характеристики: функция распределения, плотность вероятности, а также математическое ожидание и дисперсия.

Теперь можно сообразить, какие требования мы должны предъявлять к статистикам, чтобы они были хорошими оценками неизвестных параметров. Во-первых, математическое ожидание статистики должно совпадать с истинным значением оцениваемого параметра. Во-вторых, дисперсия статистики должна быть как можно меньше.

Статистика, обладающая свойством “во-первых”, называется *несмешенной*. Чуть позже мы выясним, что обе статистики N_1 и N_2 являются несмешенными, а статистика $x(n)$ — смешенная.

Для статистик, обладающих свойством “во-вторых”, вводится термин “эффективная” оценка. Точнее, *эффективной*

называется такая несмешенная оценка, у которой дисперсия минимальна (при заданном объеме выборки). Доказательство эффективности обычно является трудной задачей, и мы не будем здесь ей заниматься. Однако из тех статистик, которые мы будем сравнивать, мы найдем самую эффективную.

Теперь займемся нахождением математического ожидания и дисперсии рассмотренных статистик. Для этого примем, что абитуриентов в нашу аудиторию выбирали случайным образом из всех N возможных независимо друг от друга. Иными словами, будем считать, что x_1, \dots, x_n — независимые случайные величины, равномерно распределенные на промежутке $[0; N]$.

Заметим, что мы слегка изменили условие задачи — вместо бесповторной выборки из набора $\{1, \dots, N\}$ мы рассматриваем выборку из равномерного распределения на отрезке $[0, N]$. Позже мы рассмотрим и другую, более точную формулировку. Забегая вперед, скажем, что решение будет намного более сложным, а ответ почти не изменится.

Математическое ожидание и дисперсия для равномерного распределения нам известны: они равны

$$Mx_i = \frac{N}{2}; \quad Dx_i = \frac{N^2}{12}.$$

Следовательно, по формулам для математического ожидания и дисперсии суммы случайных величин получим:

$$MN_1 = N; \quad DN_1 = \frac{N^2}{12n}.$$

Случайная величина $x(n)$ уже не будет, разумеется, распределена равномерно. Для нее, однако, можно найти функцию распределения. Действительно, событие $x(n) < x$ означает в точности то, что каждая из n случайных величин x_1, \dots, x_n меньше, чем x . Так как эти n случайных величин равномерно распределены и независимы, вероятность

этого события равна $(x/N)^n$. Таким образом, функция распределения $x(n)$ равна

$$F_{x(n)}(x) = P(x(n) < x) = \left(\frac{x}{N}\right)^n;$$

плотность вероятности

$$\rho(x) = F'_{x(n)}(x) = \frac{n}{N} \left(\frac{x}{N}\right)^{n-1};$$

математическое ожидание

$$Mx(n) = \int_0^N x \rho(x) dx = \frac{Nn}{n+1};$$

дисперсия

$$Dx(n) = \int_0^N (x - Mx(n))^2 \rho(x) dx = \frac{nN^2}{(n+1)^2(n+2)}.$$

Математическое ожидание и дисперсию первой порядковой статистики $x(1)$ можно легко найти таким образом: если на отрезке $[0; N]$ выбрано n случайных точек, то, проходя вдоль отрезка от 0 до N , мы сначала наткнемся на $x(1)$, а последней будет $x(n)$. Если же мы пойдем в обратном направлении, от N до 0, то на $x(n)$ мы наткнемся первой. Поэтому $x(1)$ распределена так же, как $N - x(n)$, а, следовательно, ее математическое ожидание и дисперсия равны:

$$Mx(1) = \frac{N}{n+1}; \quad Dx(1) = \frac{nN^2}{(n+1)^2(n+2)}.$$

Отсюда следует, что N_2 — также несмешенная оценка, а ее дисперсия равна

$$DN_2 = \frac{2nN^2}{(n+1)^2(n+2)}.$$

Кроме того, мы можем устраниТЬ смещение у статистики $x(n)$ и построить таким образом новую оценку:

$$N_3 = \frac{n+1}{n} x(n).$$

Это также несмешенная оценка, а ее дисперсия равна

$$DN_3 = \frac{N^2}{n(n+2)}.$$

Из трех рассмотренных статистик: N_1 , N_2 и N_3 , оценивающих параметр N , эта — наилучшая, потому что, как легко убедиться, ее дисперсия минимальна. Действительно, при больших n $DN_1 \sim n^{-1}$, $DN_2 \sim n^{-2}$, а DN_3 почти в 2 раза меньше, чем DN_2 .

Для нашего примера $N_3 = 266.2$. Хотя это значение ближе всех к истинному, не следует думать, что так будет всегда. Однако так будет в большинстве случаев, потому что разброс (а именно его характеризует дисперсия) у N_3 минимальный из трех рассмотренных статистик.

Здесь стоит напомнить, что есть причина, почему оценка $x(n)$ с самого начала заслуживала внимательного рассмотрения: эта оценка была получена ранее как оценка максимального правдоподобия.

Рассмотрим теперь другую, более точную формулировку нашей задачи. Пусть из набора $\{1, \dots, N\}$ взята бесповторная выборка $\{x_1, \dots, x_n\}$ объема n . Как оценить неизвестный параметр N ? Мы здесь ограничимся получением и “доведением до ума” оценки максимального правдоподобия.

Пусть, как и ранее, $x(1) < \dots < x(n)$ — порядковые статистики. Тогда, если $N \geq x(n)$, то вероятность получить из набора $\{1, \dots, N\}$ нашу выборку одинакова для всех выборок объема n и равна

$$G = \frac{1}{C_N^n} = \frac{1}{N \cdot (N-1) \cdot \dots \cdot (N-n+1)}.$$

Нетрудно сообразить, что, каково бы ни было n , эта функция — а это не что иное как функционал правдоподобия — является монотонно убывающей функцией от N . Поэтому оценкой максимального правдоподобия для неизвестно-

го параметра N будет максимальная порядковая статистика $x(n)$.

Поскольку значение $x(n)$, скорее всего, будет меньше истинного значения N , это — смещенная оценка. Чтобы устранить смещение этой оценки, надо найти ее математическое ожидание. Займемся этим. Сначала вычислим вероятности:

$$\begin{aligned} P(x(n) \leq i) &= P(x_1 \dots x_n \leq i) = \frac{C_i^n}{C_N^n}; \\ P(x(n) = i) &= P(x(n) \leq i) - P(x(n) \leq i-1) = \\ &= \frac{C_i^n - C_{i-1}^n}{C_N^n} = \frac{C_{i-1}^{n-1}}{C_N^n}. \end{aligned}$$

По определению математического ожидания

$$\begin{aligned} M(x(n)) &= \sum_{i=n}^N P(x(n) = i) \cdot i = \sum_{i=n}^N \frac{C_{i-1}^{n-1}}{C_N^n} \cdot i = \\ &= \frac{1}{C_N^n} \sum_{i=n}^N C_{i-1}^{n-1} \cdot i = \frac{n}{C_N^n} \sum_{i=n}^N C_i^n. \end{aligned}$$

Мы воспользовались формулой

$$C_{i-1}^{n-1} \cdot i = C_i^n \cdot n,$$

которую легко доказать. А далее мы воспользуемся формулой

$$\sum_{i=n}^N C_i^n = C_{N+1}^{n+1},$$

которая также верна, хотя доказать ее несколько труднее. Впрочем, когда-то она была доказана — в первой лекции в примере 7.

Продолжаем выкладки:

$$M(x(n)) = \frac{n}{C_N^n} \cdot C_{N+1}^{n+1} = \frac{n}{n+1} \cdot (N+1).$$

Теперь мы можем устраниТЬ смещение. Окончательно получаем, что несмещенная оценка неизвестного параметра

N такова:

$$\hat{N} = \frac{n+1}{n} \cdot x(n) - 1.$$

Как видим, разница невелика.

Вернемся теперь к выборочным оценкам математического ожидания и дисперсии. Пусть есть выборка x_1, \dots, x_n из случайных величин, имеющих одинаковое распределение со средним a и дисперсией σ^2 . Для математического ожидания использовалась статистика

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}.$$

Ясно, что это — несмешенная оценка. Нетрудно также найти дисперсию этой оценки. Она равна

$$D\bar{x} = \frac{\sigma^2}{n}.$$

Рассмотрим теперь оценку дисперсии

$$\hat{s}^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}.$$

Найдем математическое ожидание этой статистики. Для этого преобразуем выражение

$$\hat{s}^2 = \frac{x_1^2 + \dots + x_n^2}{n} - \bar{x}^2.$$

Теперь воспользуемся формулой

$$DX = MX^2 - (MX)^2,$$

справедливой для любой случайной величины. Из нее следует, что

$$\begin{aligned} M(\bar{x}^2) &= D\bar{x} + (M\bar{x})^2 = \frac{\sigma^2}{n} + a^2; \\ M(x_i^2) &= Dx_i + (Mx_i)^2 = \sigma^2 + a^2. \end{aligned}$$

Следовательно

$$M\hat{s}^2 = \sigma^2 + a^2 - \frac{\sigma^2}{n} - a^2 = \frac{n-1}{n}\sigma^2.$$

Таким образом, оценка \hat{s}^2 — смещенная, и для оценки дисперсии следует пользоваться формулой исправленной выборочной дисперсии

$$s^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1},$$

которая дает несмещенную оценку.

Порядковые статистики часто возникают во многих случаях, в частности, в жизненных ситуациях. В качестве примера приведем цитату из книги Г. Гамова и М. Стерна “Занимательные задачи” (М.: Изд-во УРСС, 2003).

Летом 1956 года одному из нас (Г. Г.) приходилось часто бывать в Сан-Диего (Калифорния) в качестве консультанта авиастроительной фирмы “Конвэр”, в которой в качестве постоянного сотрудника работал другой из авторов (М. С.). Нам приходилось обсуждать множество (секретнейших!) проблем, а поскольку рабочий кабинет одного из нас (М. С.) находился на шестом этаже Главного здания и был более комфортабельным, другой из нас (Г. Г.) обычно садился в лифт на втором этаже, где находился его рабочий кабинет. Для этого один из нас (Г. Г.) шел к лифту на втором этаже и нажимал кнопку, и первым обычно приходил лифт, который шел не в том направлении, которое было нужно, т. е. шел вниз. Примерно в пяти случаях из шести первым приходил лифт, который шел вниз, и только в одном случае — лифт, который шел вверх.

— Послушайте — сказал один из нас (Г. Г.) другому (М. С.), — вы что, непрерывно изготавливаете на крыше новые лифты и спускаете их на склад в подвале?

— Что за нелепая идея! — возмутился другой (М. С.). — Разумеется, ничего такого мы не делаем. Предлагаю вам подсчитать, сколько раз первым приходит лифт, идущий в нужном вам направлении, когда вы покинете мой кабинет на шестом этаже и будете возвращаться к себе на второй

этаж.

Через несколько недель разговор снова зашел о лифтах, и один из нас (Г. Г.) вынужден был признать, что его первое замечание относительно лифтов было лишено смысла. Ожидая вызванного лифта на шестом этаже, он обнаружил, что примерно в пяти случаях из шести первым приходил лифт, идущий вверх, а не вниз. И Г. Г. быстро предложил объяснение этому загадочному явлению, противоположное первому: должно быть, компания “Конвэр” строила лифты в подвале и посыпала готовые лифты на крышу. откуда производимые компанией самолеты доставляли их к месту назначения.

— Позвольте, — прервал его другой (М. С.) — я и не знал, что наша компания занимается производством лифтов... Разумеется, — продолжал он, — правильное объяснение очень просто. Но разрешите мне прежде заметить, что если бы я и не знал, сколько этажей в этом здании, то теперь, располагая той информацией, которую вы мне сообщили, смог бы сказать, что в здании семь этажей.

— Но я ничего не говорил о высоте здания. Я только сообщил вам о том, с какими трудностями сталкиваюсь, поджидая лифт, идущий в нужном мне направлении.

— Верно, но разве вы не понимаете, что это классическая задача, которая лишь наглядно показывает, чем частота отличается от фазы?

Поразмыслив немного, мы нашли решение задачи (его вы найдете в истории “Проходящие поезда”, с. 73 — 78).

В задаче о поездах приводится история машиниста на пенсии, который часто приходил к железнодорожному переезду и смотрел на проходящие поезда. Спустя некоторое время он заметил, что на восток поезда шли гораздо чаще, чем на запад. В книге приводится следующее объяснение,

с замечанием, что оно применимо и к задаче о лифтах.

Возьмем, например, один-единственный поезд “Суперчиф”, курсирующий между Чикаго и Лос-Анджелесом. Предположим, что мы находимся в пятистах милях от Чикаго и в тысяче пятистах милях от Лос-Анджелеса, и что вы приходите к переезду в случайно выбранные моменты времени. Где с наибольшей вероятностью находится в этот момент поезд?

Так как до Лос-Анджелеса втрое дальше, чем до Чикаго, то шансы 3:1 за то, что поезд находится к западу от вас, а не к востоку! А коль скоро он находится к западу от вас, то впервые поезд пройдет мимо вас, двигаясь на восток. Разумеется, если между Чикаго и Калифорнией курсирует не один, а много поездов, как это и происходит в действительности, то ситуация не изменится, и первый поезд, которым проследует мимо нашего городка в любой момент времени, вероятнее всего будет двигаться на восток.

Объяснение это неверно! Точнее, оно верно, когда лифт или поезд один. А когда поездов много, это объяснение уже не проходит. Положение каждого поезда можно считать случайной величиной, имеющей равномерное распределение. Но положение поезда, ближайшего к переезду — это случайная величина, имеющая распределение первой порядковой статистики. Это распределение зависит от числа поездов, и чем больше поездов, тем сильнее оно сконцентрировано вблизи нуля. Поэтому при большом числе поездов обе вероятности — появится первый из них с востока или с запада — будут стремиться к 1/2.

Задача 161. Найти распределение k -й порядковой статистики из равномерного распределения на $[0, 1]$.

Задача 162. Для лифтов возможно рассмотрение различных моделей их движения. Самые простые:

- 1) время перемещения лифта между этажами пренебрежимо мало по сравнению со временем стоянки на этаже;
- 2) время стоянки на этаже пренебрежимо мало по сравнению с временем перемещения лифта между этажами.

Какая из этих моделей подразумевается в приведенной цитате? В обоих случаях принять, что лифт останавливается на каждом этаже (хотя во втором случае это, разумеется, несущественно).

Задача 163. Первая из моделей движения лифтов из предыдущей задачи также допускает различные толкования:

- 1a) лифт останавливается на всех этажах, в том числе на втором, и есть вероятность подойти и обнаружить, что лифт уже стоит;
- 1б) второй этаж — на особом положении: лифт на нем не останавливается, если не нажата кнопка вызова.

Казалось бы, модель 1a) вполне соответствует условию в приведенной цитате. Это не совсем так: в этой модели *не надо нажимать кнопку!*

Какова вероятность, что вызванный со второго этажа лифт придет снизу, для каждой из этих моделей?

Задача 164. Какова вероятность, что лифт, вызванный со второго этажа, придет снизу, если лифтов в здании два? Использовать все модели из двух предыдущих задач.

Задача 165(*). Сможете ли Вы решить задачу о поездах в общем случае? Дано: двухколейная железная дорога, длина пути к западу от начала отсчета равна L , а к востоку — равна l ($L > l$). По этой железной дороге курсируют N поездов, причем каждый из них доезжает до конца дороги и только там поворачивает назад. Положения всех поездов — независимые равномерно распределенные случайные величины, а скорости всех поездов равны. Какова вероятность, что наблюдатель, в случайный момент подошедший к точке

начала отсчета, первым увидит поезд, идущий с запада?

Лекция 12. Метод максимального правдоподобия.

Если есть основания считать, что случайная величина имеет заданное распределение, то достаточно по выборке оценить параметры этого распределения. Такой подход принято называть *параметрической статистикой*.

Опишем теперь способ получения оценок параметров распределения, который часто (но далеко не всегда) приводит к хорошим результатам. Это *метод максимального правдоподобия*. Он заключается в следующем. Пусть нам дана выборка x_1, \dots, x_n , из n величин, подчиняющихся заданному распределению с неизвестным параметром θ . Этот параметр может быть как числом, так и вектором, то есть набором чисел. Требуется оценить этот параметр.

Поскольку вид распределения известен, можно записать вероятность того, что в n независимых реализациях случайной величины будут получены значения x_1, \dots, x_n , как функционал от неизвестного пока параметра θ :

$$G(x_1, \dots, x_n | \theta) = P(x_1 | \theta) \cdot \dots \cdot P(x_n | \theta).$$

Этот функционал принято называть функционалом правдоподобия (likelihood). Для непрерывных случайных величин вместо вероятности надо использовать плотность вероятностей.

То значение параметра θ , которое дает максимальное значение функционала правдоподобия, называется *оценкой максимального правдоподобия*. Этую оценку часто можно получить, просто продифференцировав функционал. Сразу добавим, что часто упрощает задачу переход к логарифмическому функционалу правдоподобия по формуле

$$L(x_1, \dots, x_n | \theta) = \ln G(x_1, \dots, x_n | \theta),$$

поскольку у логарифма функции максимум находится там же, где и у самой функции.

Переходим к примерам.

Пример 37. Для распределения Бернулли обычно требуется оценить неизвестную вероятность p успеха в одном испытании. Пусть было проведено n испытаний, и в них было достигнуто k успехов. Вероятность этого равна

$$G(k|p) = C_n^k p^k (1-p)^{n-k}.$$

Это и есть функционал правдоподобия для нашего случая. Перейдем к логарифмическому функционалу:

$$L(k|p) = C + k \ln p + (n - k) \ln (1 - p).$$

Здесь сразу написано C вместо $\ln C_n^k$, поскольку этот член все равно пропадет при дифференцировании. Дифференцируем:

$$\frac{dL(k|p)}{dp} = \frac{k}{p} - \frac{n - k}{1 - p} = \frac{k - np}{p(1 - p)}.$$

Производная обращается в нуль при

$$p = \frac{k}{n}.$$

Нетрудно доказать, что это действительно максимум функционала правдоподобия. Таким образом, полученная оценка вероятности p является оценкой максимального правдоподобия.

Пример 38. Для показательного распределения функционал правдоподобия выглядит так:

$$G(x_1, \dots, x_n | \lambda) = \lambda^n \cdot e^{-\lambda(x_1 + \dots + x_n)}.$$

Переходя к логарифмам и дифференцируя, получаем

$$\frac{1}{\lambda} = \frac{x_1 + \dots + x_n}{n}.$$

Пример 39. Для нормального распределения нужно определить два неизвестных параметра — математическое ожидание a и стандартное отклонение σ . Функционал правдоподобия имеет вид:

$$G(x_1, \dots, x_n | a, \sigma) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left(-\frac{(x_1 - a)^2 + \dots + (x_n - a)^2}{2\sigma^2} \right).$$

Логарифм этого функционала равен, с точностью до слагаемого, не зависящего ни от a , ни от σ :

$$L(x_1, \dots, x_n | a, \sigma) = -n \ln \sigma - \frac{(x_1 - a)^2 + \dots + (x_n - a)^2}{2\sigma^2}.$$

Частные производные равны

$$\begin{aligned}\frac{\partial L}{\partial a} &= \frac{1}{\sigma^2}((x_1 - a) + \dots + (x_n - a)); \\ \frac{\partial L}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{(x_1 - a)^2 + \dots + (x_n - a)^2}{\sigma^3}.\end{aligned}$$

Приравнивая обе эти производные к нулю, получаем оценки максимального правдоподобия:

$$\begin{aligned}a &= \frac{x_1 + \dots + x_n}{n}; \\ \sigma &= \sqrt{\frac{(x_1 - a)^2 + \dots + (x_n - a)^2}{n}}.\end{aligned}$$

Пример 40. Для равномерного распределения также нужно определить два неизвестных параметра — a и b — левую и правую границу интервала. Здесь, в отличие от предыдущих случаев, метод максимального правдоподобия приводит к не очень удачным оценкам. Поскольку плотность вероятности для равномерного распределения обратно пропорциональна длине отрезка, максимальной она будет, если взять самый маленький из возможных отрезков. Это приводит к оценкам

$$a = \min(x_1, \dots, x_n); b = \max(x_1, \dots, x_n).$$

Ясно, что такие оценки будут похожи на правильные только при больших n . Позже мы вернемся к этому вопросу.

Задачи к лекции 12.

Во всех задачах подразумевается, что получена выборка x_1, \dots, x_n , подчиняющаяся заданному распределению, и неизвестный параметр следует выразить через эту выборку.

Задача 166. Дискретная случайная величина, имеющая распределение Пуассона, принимает значения $0, 1, \dots$ с вероятностями

$$P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}.$$

Найти оценку максимального правдоподобия для неизвестного параметра λ распределения Пуассона.

Задача 167. Случайная величина X имеет геометрическое распределение, если $P(X = k) = pq^k$, где $q = 1 - p$, $k = 0, 1, \dots$. Найти оценку максимального правдоподобия для неизвестного параметра p .

Задача 168. Случайная величина X имеет распределение Вейбулла с нулевым сдвигом, если ее плотность вероятности задана формулой

$$\rho(x) = \frac{k}{b^k} x^{k-1} \exp\left(-\left(\frac{x}{b}\right)^k\right).$$

Найти оценку максимального правдоподобия для неизвестного параметра b .

Задача 169. Распределение Максвелла, зависящее от параметра β , задается функцией плотности вероятностей

$$\rho(x) = C\beta^{3/2} x^2 e^{-\frac{\beta x^2}{2}}.$$

Найти оценку максимального правдоподобия для неизвестного параметра β .

Задача 170. Логнормальное распределение, зависящее от двух параметров, a и σ , задается при $x > 0$ функцией плотности вероятностей

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

Найти оценки максимального правдоподобия для неизвестных параметров a и σ .

Задача 171. Распределение Рэлея, зависящее от парамет-

ра a , задается функцией плотности вероятностей

$$\rho(x) = \frac{x}{a^2} e^{-\frac{x^2}{2a^2}}.$$

Найти оценку максимального правдоподобия для неизвестного параметра a .

Задача 172. Распределение Накагами, зависящее от двух параметров, a и ω , задается функцией плотности вероятности

$$\rho(x) = \frac{2a^a}{\Gamma(a)\omega^a} x^{2a-1} \exp\left(-\frac{ax^2}{\omega}\right).$$

Считая параметр a известным, найти оценку максимального правдоподобия для неизвестного параметра ω .

Задача 173. (*) Распределение Парето задается плотностью вероятностей

$$\rho(x) = \frac{\alpha}{x_0} \left(\frac{x_0}{x}\right)^{\alpha+1} \text{ при } x > x_0.$$

Найти оценку максимального правдоподобия для неизвестного параметра α распределения Парето. Доказать, что этим методом невозможно получить оценку для x_0 : соответствующая частная производная в нуль не обращается.

Примечание. Если производная всюду дифференцируемой функции не обращается в ноль на отрезке, то эта функция достигает максимального значения на одном из краев отрезка. Это соображение приводит к оценке максимального правдоподобия

$$\hat{x}_0 = \min(x_1, \dots, x_n).$$

Лекция 13. Проверка статистических гипотез.

Мы, собственно говоря, уже проверяли статистические гипотезы. Например, напомним такую задачу: игральная кость была подброшена 1000 раз, и в сумме получилось 3298 очков. Правильная ли это игральная кость?

Понятно, что сумма очков должна быть примерно равна 3500. Если бы было 3499 очков — ничего страшного, результат вполне правдив: отклонение всего на одно очко вполне возможно. А вот если бы сумма была равна 1000 очков? Тогда это означало бы, что при каждом из тысячи бросаний выпадало бы ровно по 1 очку. Вряд ли такое могло произойти случайно.

Теперь ясно, что вопрос в том, где провести границу, как отделить правдоподобный результат от неправдоподобного. Для этого нужно ответить на такой вопрос: в каких пределах почти наверняка лежит число выпавших очков? Сделать это можно на основании известной в теории вероятностей центральной предельной теоремы.

Эта теорема утверждает, что сумма независимых одинаково распределенных случайных величин с конечной дисперсией имеет приближенно нормальное распределение. Число очков при одном бросании — случайная величина, ее среднее значение (математическое ожидание) равно 3,5. Дисперсию ее тоже можно подсчитать — она равна $35/12$.

В соответствии с центральной предельной теоремой суммарное число выпавших очков должно быть распределено приблизительно нормально со средним

$$a = 3500$$

и дисперсией

$$\sigma^2 = 1000 \cdot \frac{35}{12}.$$

Обозначим суммарное число выпавших очков через X и

найдем, в каких пределах находится это число с вероятностью 99 %. Для этого составим уравнение

$$P(a - z < X < a + z) = 0.99.$$

Теперь преобразуем его так, чтобы можно было воспользоваться таблицами для стандартного нормального распределения:

$$P\left(-\frac{z}{\sigma} < \frac{X - a}{\sigma} < \frac{z}{\sigma}\right) = 0.99.$$

Отсюда найдем по таблицам

$$\frac{z}{\sigma} = 2.58.$$

Следовательно:

$$z = 2.58 \cdot \sqrt{1000 \cdot \frac{35}{12}} = 139.$$

Поэтому сумма выпавших очков почти наверняка (с вероятностью 99 %) лежит в интервале от $a - z$ до $a + z$, то есть от 3361 до 3639.

Число 3298 в этот интервал не попадает, поэтому кость на правильную не похожа.

Перейдем теперь к общей схеме проверки статистических гипотез. Это схема такова:

- 0). Допустим, что гипотеза верна.
- 1). Выбираем статистический критерий, имеющий (если 0) верно) заданное распределение.
- 2). Задаем доверительную вероятность $1 - \alpha$.
- 3). Выбираем область принятия гипотезы (ее вероятность $1 - \alpha$) и критическую область (ее вероятность α).
- 4). Вычисляем значение критерия.
- 5). Если критерий попадает в область принятия гипотезы — гипотеза принимается, если попадает в критическую область — отвергается.

В нашем случае значение критерия вычислять не надо: оно известно с самого начала — это число 3298. Однако

можно проверить ту же гипотезу немного иначе (хотя по сути — так же), чтобы эта проверка была больше похожа на общий случай.

Вновь обозначим число выпавших очков через X . Тогда по теореме о нормальном распределении величина

$$\xi = \frac{X - a}{\sigma}$$

должна подчиняться стандартному нормальному распределению. Эту величину мы и будем считать статистическим критерием. В нашем случае ее значение равно $\xi = -3,74$.

Область принятия гипотезы, отвечающую вероятности 99 % для стандартного нормального распределения находим по таблице: это интервал $(-2,58; 2,58)$. Поскольку значение критерия $-3,74$ не попало в эту область, гипотезу отвергаем.

Догадливый читатель уже, видимо, почувствовал, что все не так просто, и готов задать парочку недоуменных вопросов.

1. А откуда берется параметр α ?
2. Область принятия гипотезы можно выбрать по-разному. В приведенной ранее задаче, например, вероятности $1 - \alpha = 99\%$ соответствует не только интервал $(-2,58; 2,58)$, но и, например, интервал $(-\infty; 2,33)$. В эту область число $-3,74$ попадает. Так надо ли отвергать гипотезу?

Попробуем ответить на эти вопросы. Во-первых, отметим, что сумма очков 3298 даже на правильной кости является не абсолютно невозможным событием, а только очень маловероятным. Так что же нам делать? Даже если в каждом из тысячи бросаний выпадет 1 очко, сказать: ну, это случайно, бывает? Нет! Если произошло такое маловероятное событие, то пожалуй, все-таки кость была неправильной.

Из сказанного, однако, можно сделать вывод, что мы долж-

ны быть готовы в некоторых случаях ошибиться: отвергнуть нашу гипотезу, хотя она верна. Надо только, чтобы это происходило не слишком часто.

Оказывается (и нетрудно понять — почему), что α — не что иное, как вероятность допустить такую ошибку. Соответственно доверительная вероятность $1 - \alpha$ — вероятность не допустить этой ошибки, то есть вероятность принять нашу гипотезу, если она на самом деле верна.

Эта доверительная вероятность всегда задается извне. В практических задачах эта вероятность может быть предписана, например, нормативными документами. Скажем, в задачах об испытании образцов грунта на прочность, в соответствии с действующим ГОСТ 20522-96 принята равной 95 %. Такая же доверительная вероятность принята, например, еще и в ГОСТ 17.1.5.05-85 для проб вод, льдов и атмосферных осадков, а также в других нормативных документах. А в других случаях, когда ошибки допускать нельзя, например при испытаниях новых лекарств, доверительная вероятность может быть иной.

Вероятность α — вероятность допустить ошибку первого рода — называют еще *уровнем значимости* критерия.

Ответим теперь на второй недоуменный вопрос. Мы можем ошибиться и по-другому: принять нашу гипотезу, хотя она и неверна. Эти две возможные ошибки называются ошибками соответственно первого и второго рода.

С ошибкой первого рода мы уже разобрались: она хоть и неизбежна, но происходит не слишком часто, с вероятностью α . А можно ли избежать ошибки второго рода?

Понятно, что избежать ее тоже нельзя. Ведь неправильная кость тоже может когда-нибудь дать сумму 3500 при 1000 бросаниях, то есть иногда давать такие же результаты, как правильная.

Поэтому надо поставить вопрос о том, как уменьшить вероятность ошибки второго рода. Вот тут-то мы и воспользуемся той единственной свободой, которая у нас осталась в рассматриваемой задаче: выбором области принятия гипотезы. Сформулируем такое правило: из всех возможных областей с ошибкой первого рода α область принятия гипотезы надо выбирать так, чтобы вероятность ошибки второго рода была минимальна. Однако в отличие от случая с ошибкой первого рода здесь мы вынуждены встать на скользкий путь предположений и эмпирических правил.

Действительно, если мы знаем, что критерий подчиняется заданному распределению, то вероятности ошибок для разных областей мы можем подсчитать. А как быть, если мы знаем только, что критерий заданному распределению не подчиняется?

Во многих случаях простой выход состоит в том, чтобы область принятия гипотезы была минимальной. В разобранном примере это как раз интервал $(-2, 58; 2, 58)$.

Чаще всего применяется такой подход. Нам следует рассмотреть так называемую конкурирующую гипотезу. Например, в случае с игральной костью мы фактически рассматривали такую конкурирующую гипотезу: среднее число очков для одного броска не равно 3,5. А могли бы, скажем, рассмотреть такую: среднее число очков для одного броска больше 3,5. Тогда было бы уместно использовать другую область принятия гипотезы, и в этом случае, наверное, основная гипотеза была бы принята.

Сформулируем еще раз правило выбора области принятия гипотезы. Среди всех возможных областей с доверительной вероятностью α выбирается такая, для которой вероятность ошибки второго рода минимальна.

Введенные понятия иллюстрируют рисунки. На первом из

них изображена односторонняя критическая область. Вероятность области принятия гипотезы равна $1 - \alpha$, вероятность критической области равна α . На втором изображен двусторонняя критическая область.

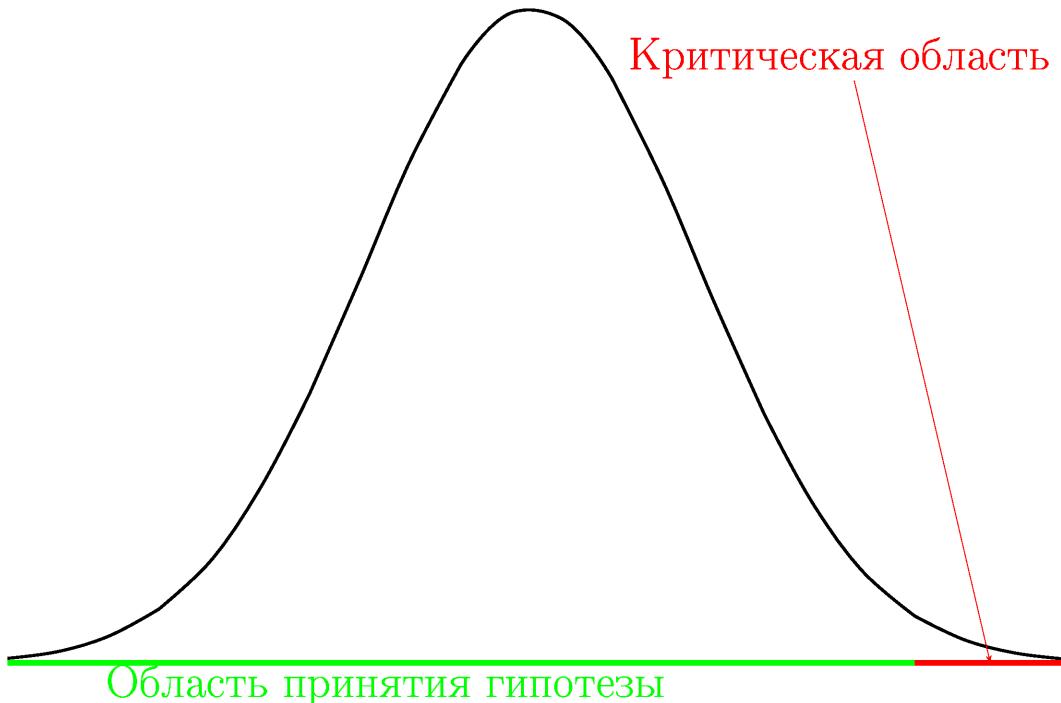


Рисунок 16. Односторонняя критическая область.

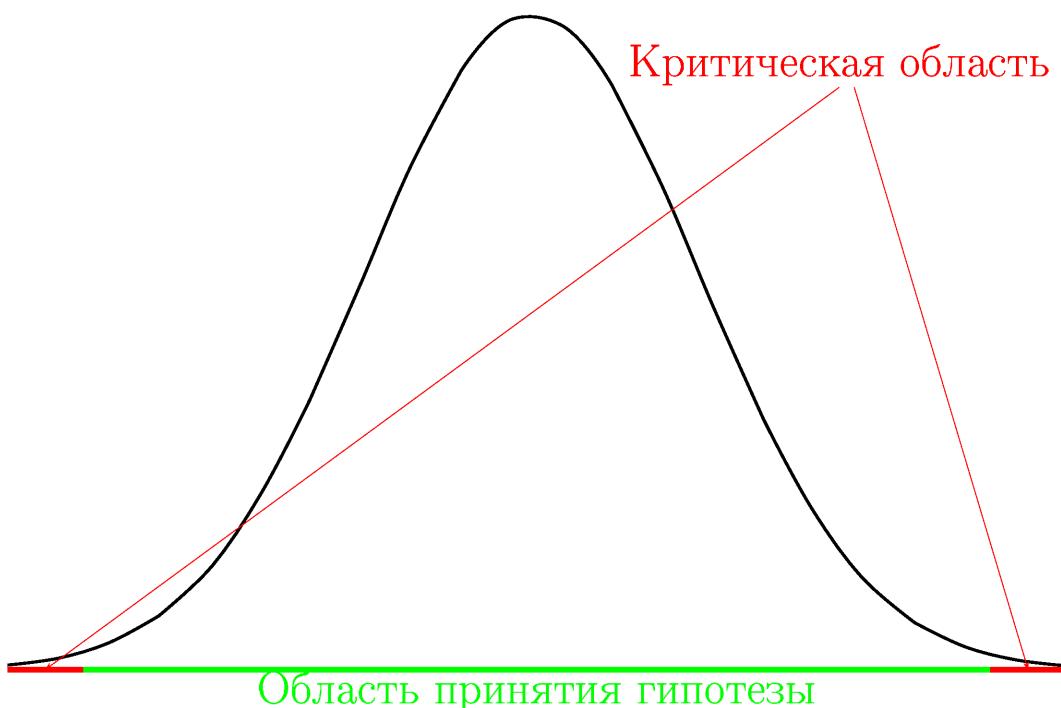


Рисунок 17. Двусторонняя критическая область.

В реальных задачах часто ясно, какую альтернативную гипотезу надо рассматривать. Поясним сказанное на примере, взятом из одного из разделов метрологии — теории измерений. Например, здесь могут возникать две такие задачи:

- 1) сравнить показания поверяемого измерительного средства с показаниями эталонного измерительного средства с целью определить, обеспечивает ли поверяемое средство достаточную точность (проверка точности средства измерений);
- 2) сравнить показания двух измерительных средств с целью определить, одинакова ли точность их измерений (проверка единства измерений для двух измерительных средств).

Проведя несколько измерений, мы сможем получить выборочные дисперсии для обоих средств измерения и сравнить, достаточно ли близки эти дисперсии. Найдем их отношение: если дисперсии близки, то оно будет близко к единице. Как мы вскоре узнаем, статистический критерий в обеих задачах основан на распределении Фишера. Однако правила построения критических областей — разные. В самом деле, в первом случае слишком малое отношение дисперсий нас не беспокоит: точность поверяемого средства выше, чем у эталонного, то есть достаточна. Во втором же случае слишком малое значение отношения говорит о том, что точность у измерительных средств различна: второе менее точно.

Таким образом, если мы обозначим выборочные дисперсии измерительных средств как s_1^2 и s_2^2 , то против нулевой гипотезы $s_1^2 = s_2^2$ в первом случае должна рассматриваться альтернативная гипотеза $s_1^2 > s_2^2$, а во втором случае — гипотеза $s_1^2 \neq s_2^2$. Соответственно и области принятия гипотез будут различны.

Важно подчеркнуть один принципиальный момент, свя-

занный с проверкой статистических гипотез. Если гипотеза не отвергается, то это на самом деле вовсе не означает, что она справедлива. Это означает только, что данные не противоречат гипотезе.

Пусть, например, в примере с игральной костью, мы насчитали 3451 очко. Тогда гипотеза о том, что кость правильная, отвергнута не будет. Означает ли это, что кость правильная? Нет, не означает. Например, кость может быть “не совсем” правильной, скажем, со средним значением не 3,5, а 3,45. Такую маленькую разницу при 1000 бросаниях уловить довольно трудно. Поэтому результат проверки следует трактовать не как доказательство, а только как подтверждение гипотезы, и утверждать, что “гипотеза не противоречит данным”.

В связи с этим можно привести один забавный случай, взятый с сайта *anekdot.ru*:

Историограф королевского двора адресовал знаменитому математику, президенту Королевского статистического общества сэру Франку Йейтсу (1902 — 1994) следующий вопрос. В преамбуле запроса сообщалось, что короли Генрихи, принадлежавшие четырем различным правящим династиям, непременно умирали по пятницам. В подтверждении этого приводились точные даты смерти восьми английских королей, принадлежавших Нормандской династии, а также династиям Плантагенетов, Ланкастеров и Тюдоров. За развернутой преамбулой следовал лапидарный вопрос: «Не является ли пятница роковым днем для Генрихов английских?».

Ответ сэра Йейтса вошел в историю науки: «Дорогой сэр! Представленные Вами статистические данные не противоречат сформулированной Вами статистической гипотезе. Королевское статистическое общество рекомендует Вам продолжать наблюдения». Размышления сэра Йейтса выгля-

дели особенно эпично в свете того факта, что последний из правящих Генрихов умер в 1547 году.

Часто вместо проверки гипотез используют несколько другую интерпретацию описанной выше техники. Можно говорить о том, что мы ищем оценку значения некоторого неизвестного параметра распределения, но не точечную, как раньше, а интервальную. Тогда тот интервал, который мы называли областью принятия гипотезы, будет называться *доверительным интервалом*, а вероятность $1 - \alpha$ — доверительной вероятностью или уровнем доверия. Иными словами: доверительный интервал — это интервал, который накрывает истинное значение неизвестного параметра с заданной вероятностью.

Пример 41. В качестве примера проверки гипотез рассмотрим так называемый критерий согласия хи-квадрат. Эта статистическая процедура разработана для того, чтобы получить ответ на вопрос: соответствуют ли экспериментальные данные теоретическому распределению.

Рассмотрим такой пример с числами.

В таблице ниже указаны интервалы, а также количество наблюдений, попавших в каждый из этих интервалов. Как мы помним, это называется группированной выборкой. Нас будет интересовать вопрос: соответствуют ли эти данные тому предположению, что они получены из нормальной совокупности.

Инт.	0 — 2	2 — 4	4 — 6	6 — 8	8 — 10	10 — 12	12 — 14
ЭЧ	12	51	179	298	170	48	9

Построим по этим данным гистограмму. Вычислим выборочные среднее и дисперсию:

$$\bar{x}_{\text{групп}} = 6,94; \quad s_{\text{групп}}^2 = 4,87.$$

Теперь мы можем построить теоретическую кривую нор-

мального распределения с вычисленными средним и дисперсией. Изобразим эту кривую вместе с гистограммой на одном рисунке. Для того, чтобы можно было сравнивать гистограмму и теоретическую кривую, их надо изобразить в одном масштабе. Для этого плотность вероятности нормального распределения следует умножить на общее число наблюдений, в нашем случае $n = 767$.

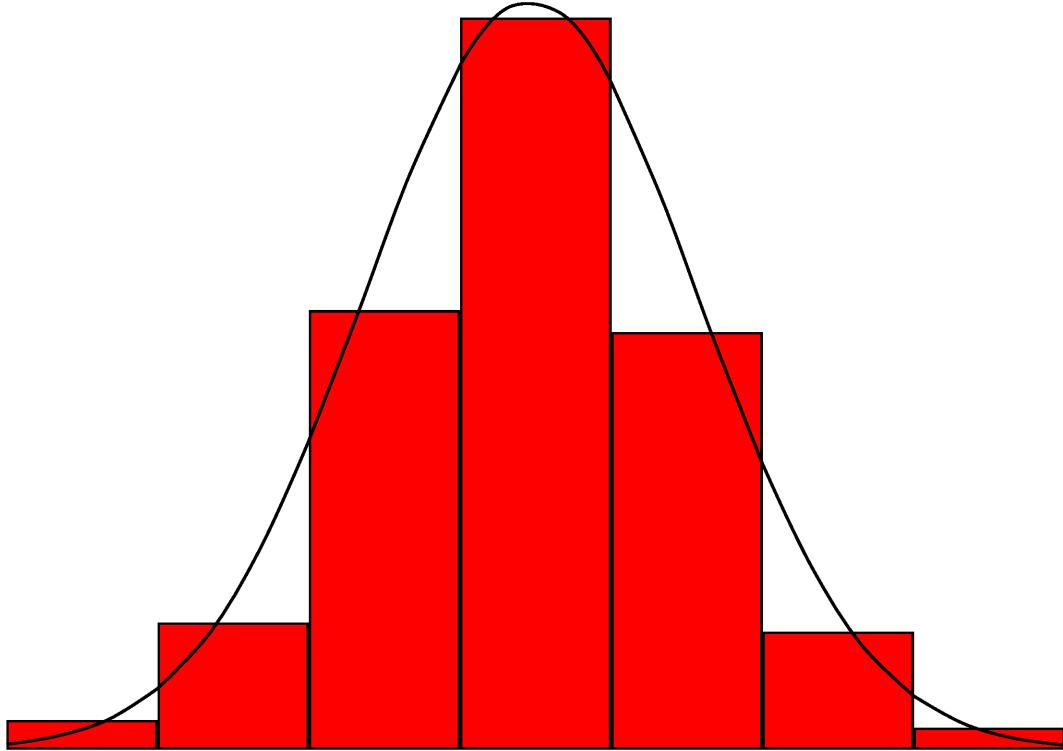


Рисунок 18. Гистограмма экспериментальных данных и теоретическая кривая.

При разглядывании рисунка гипотеза кажется правдоподобной. Визуально, однако, мы можем только выдвинуть гипотезу, а как ее проверить? Разработкой соответствующей теории и выводом критерия мы сейчас и займемся.

Пусть мы имеем дело с группированной выборкой вида

$c(1)_l - c(1)_r$	\dots	$c(m)_l - c(m)_r$
n_1	\dots	n_m

Здесь

$c(1)_l - c(1)_r, \dots, c(m)_l - c(m)_r$ — интервалы, причем $c(k)_l$ и $c(k)_r$ — соответственно левая и правая границы k -го интер-

вала;

$n_1, \dots n_m$ — число наблюдений, попавших в этот интервал;

m — число интервалов;

$n = n_1 + \dots + n_m$ — общее число наблюдений.

При этом мы будем считать, что интервалы не перекрываются, хотя граничные точки должны совпадать — правая граница предыдущего интервала совпадает с левой границей следующего. Тогда все эти интервалы будут покрывать целый отрезок от $c(1)_l$ до $c(m)_r$.

Нас будет интересовать такая гипотеза: верно ли, что данные согласуются с тем предположением, что выборка взята из генеральной совокупности, имеющей заданное распределение.

Для такой группированной выборки ранее приводились оценки выборочного среднего и дисперсии. Мы рассмотрим пример с нормальным распределением. При этом мы будем считать, что параметры нормального распределения — математическое ожидание и дисперсия — совпадают с полученными оценками.

Мы должны получить один критерий — случайную величину, имеющую заданное распределение. (Почему критерий должен быть один — этот важный вопрос будет обсуждаться позже, в главе о дисперсионном анализе).

Для нормального распределения следует распространить интервал наблюдений на всю числовую ось — ведь его плотность вероятности нигде в нуль не обращается.

Пусть гипотеза верна.

Для каждого из интервалов можно получить тогда теоретическое значение вероятности попадания в этот интервал — ведь математическое ожидание и дисперсия нам известны. Умножив эти вероятности (обозначим их p_i) на число наблюдений n , получим теоретические частоты pr_i , соот-

всегда соответствующие каждому из этих интервалов. Введенные ранее числа n_i будем называть экспериментальными частотами.

Заметим, что, поскольку интервалы у нас теперь покрывают всю числовую ось, то сумма всех теоретических частот, как и сумма всех экспериментальных, равна n .

Если разности $n_i - np_i$ не слишком велики, то гипотеза похожа на правильную. Эту идею мы и используем. Дальнейшее изложение будет вестись на уровне идей, опуская тонкости строгого математического вывода (требующие не один десяток страниц). Подробности можно посмотреть в книге Б. Л. Ван дер Вардена “Математическая статистика”.

Если мы будем считать n_i случайными величинами, то легко поймем, что они имеют распределение Бернулли числа успехов в n испытаниях с вероятностью успеха p_i . Отсюда следует, в соответствии с теоремой Муавра — Лапласа, что случайные величины

$$\frac{n_i - np_i}{\sqrt{np_i q_i}}$$

имеют приблизительно стандартное нормальное распределение. Здесь, как всегда, $q_i = 1 - p_i$.

Если число интервалов m достаточно велико, то все q_i близки к 1. Поэтому приблизительно стандартное нормальное распределение будут иметь величины

$$\frac{n_i - np_i}{\sqrt{np_i}}.$$

Рассмотрим сумму квадратов этих случайных величин. Это величина

$$\chi^2 = \sum_{i=1}^m \frac{(n_i - np_i)^2}{np_i}$$

имела бы распределение χ^2 с m степенями свободы, если

бы n_1, \dots, n_m были бы независимы. Но это, разумеется не так. Одну из зависимостей можно указать сразу:

$$n_1 + \dots + n_m = n.$$

Тем не менее можно доказать, что квадратичная форма χ^2 действительно имеет распределение хи-квадрат, но с меньшим числом степеней свободы, равным рангу этой формы.

Одна из зависимостей, поникающая ранг, приведена выше. Оказывается, можно доказать, что остальные зависимости — не что иное, как параметры распределения, определяемые по выборке. В случае нормального распределения их два — математическое ожидание и дисперсия. Поэтому в этом случае величина χ^2 имеет распределение хи-квадрат с $m - 3$ степенями свободы.

Подводя итог, сформулируем рецепт проверки, называемый *критерием согласия хи-квадрат*. Пусть по заданному набору интервалов и экспериментальных частот n_i нужно проверить гипотезу о том, соответствуют ли эти данные заданному теоретическому распределению. Тогда следует:

- вычислить параметры теоретического распределения;
- найти теоретические частоты np_i ;
- вычислить значение статистического критерия

$$\chi^2 = \sum_{i=1}^m \frac{(n_i - np_i)^2}{np_i}.$$

Если гипотеза верна, то эта случайная величина имеет распределение χ^2 с $m - 1 - r$ степенями свободы, где:

m — число интервалов;

r — число параметров, определяемых по выборке.

Вернемся теперь к нашему числовому примеру.

Пусть надо на уровне значимости $\alpha = 0,05$ проверить гипотезу о том, соответствуют ли указанные экспериментальные частоты тому, что данные получены из нормальной генеральной совокупности.

В нашем случае число интервалов $m = 7$. Поскольку, как уже было сказано, для нормального распределения по выборке были найдены два параметра — математическое ожидание и дисперсия — то в нашем случае $r = 2$. Таким образом, число степеней свободы будет равно 4.

Для нормального распределения с вычисленными средним и дисперсией вычисляем теоретические частоты. Продолжим таблицу, добавим в нее строку теоретических частот и строку величин $(n_i = np_i)^2 / np_i$.

Инт.	0 — 2	2 — 4	4 — 6	6 — 8	8 — 10	10 — 12	12 — 14
ЭЧ	12	51	179	298	170	48	9
ТЧ	9,71	60.59	187.06	267.92	178.30	55.04	8.73
χ^2	0.54	1.52	0.35	3.38	0.39	0.90	0.05

Суммируя элементы последней строки, находим значение критерия

$$\chi^2 = 7,12.$$

Из таблиц находим значение 95% квантиля распределения хи-квадрат с 4 степенями свободы:

$$\chi^2_{0,95}(4) = 9,49.$$

Поскольку наблюдаемое значение меньше критического, нет оснований отвергать гипотезу.

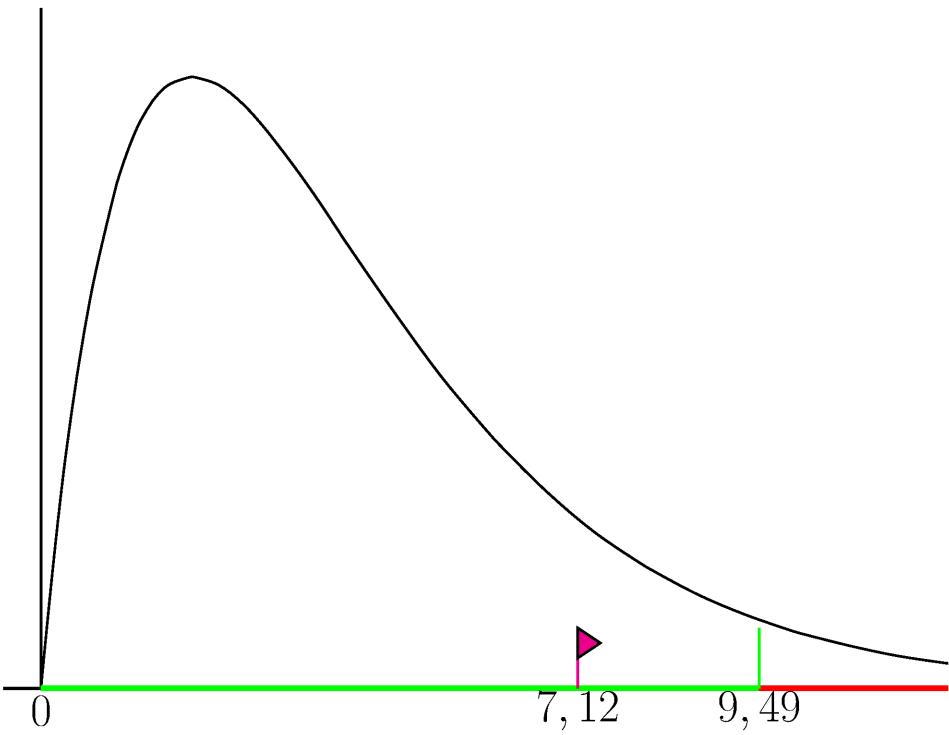


Рисунок 19. Сравнение критерия и критического значения.

Процедуру проверки статистической гипотезы иллюстрирует рисунок.

Остается пояснить, почему была выбрана односторонняя критическая область. Близкое к нулю значение критерия означает, что теоретические частоты очень мало отличаются от экспериментальных. По-видимому, это не должно быть причиной отказа от гипотезы.

Лекция 14. Проверка гипотез о математическом ожидании и дисперсии.

В этой лекции речь пойдет о случае, когда выборка x_1, \dots, x_n представляет собой n результатов независимых измерений одной и той же величины. Такой подход к измерениям был развит, начиная с трудов К. Гаусса в первой половине XIX века в связи с потребностями астрономии и геодезии. Однако большая часть результатов, о которых здесь пойдет речь, относятся к более позднему периоду — к началу XX века и связаны в первую очередь с именами Р. Фишера, К. Пирсона и У. Госсета.

Сформулируем задачу таким образом: проводятся измерения с целью определить истинное значение некоторой величины a , эти измерения характеризуются некоторой точностью σ , причем чаще всего оба этих параметра заранее неизвестны. В результате n независимых измерений получены значения x_1, \dots, x_n . В теории измерений принято считать, что эти значения распределены нормально с математическим ожиданием a и дисперсией σ^2 . Требуется оценить эти параметры.

Для определения этих параметров можно использовать статистики

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}$$

для параметра a , и

$$s^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}$$

для σ^2 .

Здесь нужно обратить внимание на одно важное обстоятельство. Если такое n -кратное измерение проводится один раз, то мы имеем дело с выборкой x_1, \dots, x_n . Если же нас интересуют статистические свойства этой выборки, то нам следует считать эти числа одной из возможных реализаций

набора случайных величин X_1, \dots, X_n . Тогда соответствующие оценки

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \text{ и}$$

$$S^2 = \frac{(X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n - 1}$$

тоже являются случайными величинами. Чтобы подчеркнуть этот момент, случайные величины, а также их оценки, будут обозначаться заглавными буквами.

Как было показано ранее, эти оценки несмешенные даже в том случае, если случайные величины X_1, \dots, X_n не подчиняются нормальному распределению. Если же эти величины — нормально распределенные, то тогда можно доказать следующий важный результат.

Теорема (Р. Фишер). Пусть X_1, \dots, X_n — независимые случайные величины, распределенные нормально с математическими ожиданиями a и дисперсиями σ^2 . Тогда для ранее определенных величин \bar{X} и S^2 справедливо следующее:

- 1) \bar{X} распределена нормально со средним a и дисперсией σ^2/n ;
- 2) $(n - 1)S^2/\sigma^2$ распределено по закону χ^2 с $n - 1$ степенями свободы;
- 3) \bar{X} и S^2 независимы.

Доказательство. Пункт 1 сразу следует из того, что сумма нормально распределенных независимых случайных величин также нормально распределена. Математическое ожидание и дисперсия \bar{X} были получены ранее.

Для доказательства пунктов 2) и 3) сначала преобразуем выражение для S^2 , заменив везде

$$(X_i - \bar{X})^2$$

на

$$((X_i - a) - (\bar{X} - a))^2.$$

Получим

$$S^2 = \frac{((X_1 - a) - (\bar{X} - a))^2 + \cdots + ((X_n - a) - (\bar{X} - a))^2}{n - 1}.$$

Преобразуем дальше:

$$S^2 = \frac{(X_1 - a)^2 + \cdots + (X_n - a)^2}{n - 1} - \frac{n}{n - 1}(\bar{X} - a)^2.$$

Введем теперь вспомогательные величины

$$Y_i = \frac{X_i - a}{\sigma}.$$

Ясно, что это независимые стандартные нормальные величины. Кроме того, их среднее арифметическое равно

$$\bar{Y} = \frac{Y_1 + \cdots + Y_n}{n} = \frac{\bar{X} - a}{\sigma}.$$

Домножим теперь последнее выражение для S^2 на $(n - 1)$ и разделим на σ^2 , получим:

$$\frac{(n - 1)S^2}{\sigma^2} = Y_1^2 + \cdots + Y_n^2 - n\bar{Y}^2 = Y_1^2 + \cdots + Y_n^2 - Z_1^2.$$

Здесь обозначено

$$Z_1 = \frac{Y_1}{\sqrt{n}} + \cdots + \frac{Y_n}{\sqrt{n}}.$$

Поскольку вектор

$$\left(\frac{1}{\sqrt{n}}; \dots; \frac{1}{\sqrt{n}} \right)$$

имеет, как нетрудно убедиться, длину 1, можно выражение для Z_1 продолжить до ортогонального преобразования

$$\begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = B \cdot \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

с некоторыми новыми случайными величинами Z_2, \dots, Z_n и некоторой ортогональной матрицей B .

Как следует из теоремы об ортогональном преобразовании, случайные величины Z_1, \dots, Z_n , как и Y_1, \dots, Y_n , — независимые стандартные нормальные величины. Так как всякое ортогональное преобразование сохраняет длины, то

$$Z_1^2 + \dots + Z_n^2 = Y_1^2 + \dots + Y_n^2.$$

Но тогда

$$\frac{(n-1)S^2}{\sigma^2} = Z_2^2 + \dots + Z_n^2,$$

следовательно, эта величина распределена по $\chi^2(n-1)$ и независима от Z_1 .

Тем самым теорема доказана. Она будет использована далее для решения задач: для проверки гипотез о значениях среднего и дисперсии, а также для проверки гипотез о равенстве средних и дисперсий для различных выборок. Во всех этих задачах предполагается, что альтернативная гипотеза — “не равно заданному” или “не равны между собой”, поэтому во всех случаях будут строиться двусторонние критические области.

Пример 42.

№	1	2	3	4	5	6	7	8	9
Данные	3,96	5,58	5,23	6,01	5,83	4,01	4,48	6,36	4,33

По имеющимся данным проверить на уровне значимости 0,05 гипотезу о том, что дисперсия генеральной совокупности $\sigma^2 = 2$.

Решение. Пусть n — объем выборки (в нашем случае $n = 9$). По только что доказанной теореме, если гипотеза верна, то величина

$$\frac{(n-1) \cdot s^2}{\sigma^2}$$

имеет распределение $\chi^2(n-1)$.

Находим оценки среднего и дисперсии: $\bar{x} = 5,09$, $s^2 =$

0,83. Вычисляем значение критерия:

$$\frac{(n - 1) \cdot s^2}{\sigma^2} = 3,33.$$

Границы критической области находим по таблице распределения $\chi^2(8)$, они равны 2, 18 и 17, 53. Следовательно, нет оснований отвергнуть гипотезу.

Пример 43.

№	1	2	3	4	5	6	7	8	9
Данные	1,01	5,85	5,90	6,09	5,63	2,77	1,50	5,51	5,80

По имеющимся данным проверить на уровне значимости 0,05 гипотезу о том, что среднее генеральной совокупности равно 3. Считать известным, что дисперсия генеральной совокупности $\sigma^2 = 3$.

Решение. Поскольку σ известно, величина

$$\frac{\bar{x} - a}{\sigma}$$

имеет нормальное распределение со средним 0 и дисперсией $1/n$, если наша гипотеза верна (n — снова объем выборки). Легко составить выражение, имеющее стандартное нормальное распределение:

$$\frac{\sqrt{n} \cdot (\bar{x} - a)}{\sigma}.$$

Подставляя значения из выборки, находим $\bar{x} = 4,45$. Поэтому значение критерия равно

$$\frac{\sqrt{n} \cdot (\bar{x} - a)}{\sigma} = 2,51.$$

Критическую область находим по таблице нормального распределения. Это интервал $(-1,96; 1,96)$. Поэтому гипотезу следует отвергнуть.

Пример 44.

№	1	2	3	4	5	6	7	8	9
Данные	5,47	4,63	4,63	2,64	4,36	5,15	5,94	5,99	5,45

По имеющимся данным проверить на уровне значимости 0,05 гипотезу о том, что среднее генеральной совокупности равно 5. Дисперсию генеральной совокупности считать неизвестной.

Решение. В этом случае σ неизвестно, поэтому разброс следует оценивать по этой же выборке. В следующих формулах $n = 9$, $a = 5$, а в самих формулах, как и ранее, оставлены буквы, чтобы их можно было применять в общем случае. Величина

$$\sqrt{n} \cdot \frac{\bar{x} - a}{\sigma}$$

имеет стандартное нормальное распределение, а величина

$$\frac{(n - 1)s^2}{\sigma^2}$$

имеет распределение $\chi^2(n - 1)$.

Нам нужно получить такое выражение, которое имело бы заданное распределение и при этом не содержало бы неизвестных параметров. В нашем случае неизвестен один параметр — это σ^2 . Чтобы σ сократилось, надо первое выражение разделить на квадратный корень из второго, и тогда получится, как легко убедиться, что величина

$$\frac{\sqrt{n}(\bar{x} - a)}{s}$$

имеет распределение Стьюдента с $(n - 1)$ степенями свободы.

В нашем случае $\bar{x} = 4,92$, $s^2 = 1,06$, значение критерия равно $-0,24$, а критическая область — это интервал $(-2,31; 2,31)$, поэтому нет оснований отвергнуть гипотезу.

Пример 45.

№	1	2	3	4	5	6	7	8	9
Данные	5,24	3,59	4,72	6,95	6,05	5,27	5,25	3,93	5,86
Данные	4,65	5,08	4,69	4,15	5,29	4,52	3,96	5,73	5,20

По имеющимся данным проверить на уровне значимости 0,05 гипотезу о том, что средние двух генеральных совокупностей, из которых взяты выборки, равны. Дисперсии генеральных совокупностей считать неизвестными, но равными.

Примечание. Равенство дисперсий (хотя и неизвестных) для многих практических задач можно интерпретировать как то, что измерения проводились по одинаковой методике, например, одним и тем же инструментом.

Решение. В общем случае эту задачу можно решить и для случая, когда число наблюдений в выборках не одинаково. Пусть в первой выборке m чисел, а во второй — n . Обозначим также через \bar{x}_1 и \bar{x}_2 выборочные средние по каждой из выборок, а через s_1^2 и s_2^2 — оценки дисперсии. Как и ранее, a и σ^2 — неизвестные математическое ожидание и дисперсия обеих выборок.

Если гипотеза о равенстве средних верна, то обе случайные величины $(\bar{x}_1 - a)/\sigma$ и $(\bar{x}_2 - a)/\sigma$ имеют нормальное распределение с нулевым средним и дисперсиями $1/m$ и $1/n$ соответственно. Отсюда следует, что их разность $(\bar{x}_1 - \bar{x}_2)/\sigma$ также нормальна со средним 0 и дисперсией

$$\frac{1}{m} + \frac{1}{n} = \frac{m+n}{mn}.$$

Поэтому величина

$$\frac{\bar{x}_1 - \bar{x}_2}{\sigma} \cdot \sqrt{\frac{mn}{m+n}}$$

имеет стандартное нормальное распределение.

Случайные величины $(m-1)s_1^2/\sigma^2$ и $(n-1)s_2^2/\sigma^2$, как и ранее, имеют распределения χ^2 с $m-1$ и $n-1$ степенями свободы соответственно. Поэтому их сумма $((m-1)s_1^2 + (n-1)s_2^2)/\sigma^2$ распределена по $\chi^2(m+n-2)$.

Как и в прошлом примере, разделим первое выражение на квадратный корень из второго. Тогда получим, что ве-

личина

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(m-1)s_1^2 + (n-1)s_2^2}} \cdot \sqrt{\frac{mn(m+n-2)}{m+n}}$$

имеет распределение Стьюдента с $(m+n-2)$ степенями свободы.

В нашем случае $\bar{x}_1 = 5,21$, $\bar{x}_2 = 4,81$, $s_1^2 = 1,08$, $s_2^2 = 0,32$, $m = n = 9$. Следовательно, значение критерия для нашей выборки равно 1,01. Критическая область — интервал $(-2, 12; 2, 12)$, поэтому гипотеза не отвергается.

Пример 46.

№	1	2	3	4	5	6	7	8	9
Данные	1,81	1,65	3,26	1,89	1,84	0,80	2,23	4,66	4,53
Данные	6,25	4,10	6,71	3,77	4,35	7,08	7,14	3,93	4,36

По имеющимся данным проверить на уровне значимости 0,05 гипотезу о том, что дисперсии двух генеральных совокупностей, из которых взяты выборки, равны.

Решение. И эту задачу можно решить для случая, когда число наблюдений в выборках различно. Пусть в первой выборке m чисел, а во второй — n . Обозначим оценки дисперсий в выборках как s_1^2 и s_2^2 .

Предположим, что обе генеральные совокупности имеют одинаковые дисперсии σ^2 . Тогда величина

$$\frac{(m-1)s_1^2}{\sigma^2}$$

распределена по $\chi^2(m-1)$, а величина

$$\frac{(n-1)s_2^2}{\sigma^2}$$

распределена по $\chi^2(n-1)$. Отсюда следует, что величина

$$\frac{s_1^2}{s_2^2}$$

подчиняется распределению Фишера с $(m-1, n-1)$ степенями свободы.

В нашем случае $m = n = 9$. Подставляя числа из наших выборок, получим, что $s_1^2 = 1,79$, $s_2^2 = 2,11$. Значение критерия равно 0,85, а критическая область — интервал $(0, 23; 4, 43)$, поэтому гипотеза не отвергается.

Лекция 15. Выборочная корреляция и регрессия.

Коэффициент корреляции

Мы приступаем теперь к исследованию связи и зависимости между переменными в статистике. Допустим, что у нас есть массив данных, характеризующийся двумя переменными. Нас будет интересовать вопрос: как связаны эти переменные.

Например, нас может интересовать зависимость между ценой земли в Подмосковье и расстоянием до МКАД. Или связь между температурой образца и его электрическим сопротивлением. Или, скажем, как связаны плотность населения в районе и засоленность почв этого района.

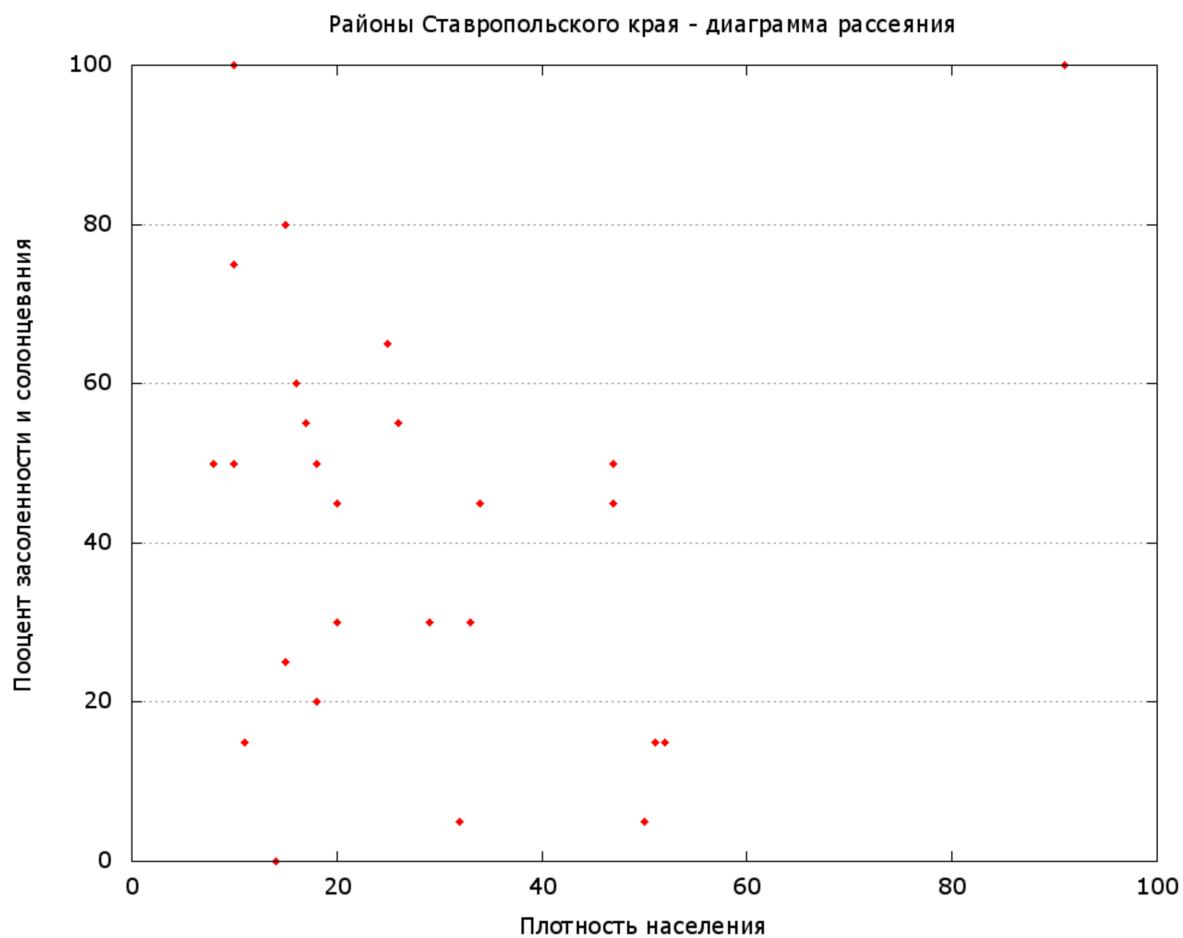


Рисунок 20. Диаграмма рассеяния — связь засоленности почв и плотности населения районов Ставропольского края.

Если каждый элемент выборки характеризуется двумя числами, то нашу выборку мы можем представить в виде двухмерной картинки, где каждому элементу соответствует точка на плоскости. Мы получим “облако точек”. Такая картинка носит название *диаграмма рассеяния*.

Если точки в диаграмме рассеяния расположены хаотично, то, скорее всего, связь между переменными отсутствует. Если же точки имеют тенденцию выстраиваться вдоль определенных наклонных линий, то переменные связаны более или менее значительной зависимостью.

Мы рассмотрим самый простой случай — как выяснить, связаны ли переменные линейной зависимостью. Для определения такой связи между двумя переменными обычно используется (выборочный) *коэффициент корреляции*.

Напомним, что в теории вероятностей при рассмотрении многомерных случайных величин вводились коэффициенты ковариации и корреляции по следующим формулам:

$$\text{cov}(X, Y) = M((X - MX)(Y - MY))$$

для коэффициента ковариации, и

$$r(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{DXDY}}$$

для коэффициента корреляции.

Теперь мы имеем дело с выборками, а не с генеральными совокупностями, и требуется другое определение. Если каждый элемент выборки характеризуется парой чисел (x_i, y_i) , $i = 1 \dots n$, то коэффициент выборочной корреляции определяется по формуле

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y},$$

где, как обычно

\bar{x} и \bar{y} — выборочные средние;

s_x и s_y — выборочные стандартные отклонения. Для вычисления коэффициента корреляции в Microsoft Excel используется стандартная функция КОРРЕЛ. В Libre Office то же самое делает стандартная функция CORREL.

Коэффициент корреляции — это число, которое может находиться в пределах от -1 до 1 . Это — безразмерная величина, он не меняется при линейных преобразованиях переменных, то есть при изменении масштабов переменных, в частности:

- при изменении единиц измерения (доллары вместо рублей или мили вместо километров);
- при изменении начала отсчета (например, если мерять расстояние не от МКАД, а от центра).

Коэффициент корреляции может быть положительным и отрицательным. Если он положителен, то это означает, что переменные связаны прямой зависимостью: чем больше одна, тем больше и другая. Если же он отрицателен, то зависимость обратная: если одна переменная увеличивается, то вторая уменьшается. В рассматриваемом случае с ценами на землю с увеличением расстояния МКАД цена земли, скорее всего, будет уменьшаться.

Важно подчеркнуть, что если коэффициент корреляции не равен -1 или 1 , то им нельзя пользоваться для предсказаний в каждом конкретном случае: он выражает только общую тенденцию. Допустим, он равен $-0,52$ для какого-то Подмосковного района. Это означает, что при увеличении расстояния до МКАД цена падает в большинстве случаев, но далеко не во всех случаях.

Обычно принято так интерпретировать значение коэффициента корреляции:

от -1 до $-0,7$	сильная отрицательная корреляция
от $-0,7$ до $-0,3$	средняя отрицательная корреляция
от $-0,3$ до $0,3$	слабая корреляция
от $0,3$ до $0,7$	средняя положительная корреляция
от $0,7$ до 1	сильная положительная корреляция

Остановимся на графической интерпретации коэффициента корреляции.

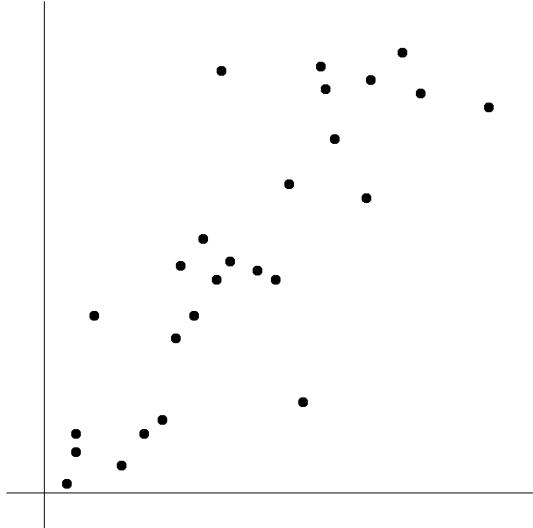


Рис. 21. Сильная положительная корреляция. Коэффициент корреляции равен 0,82

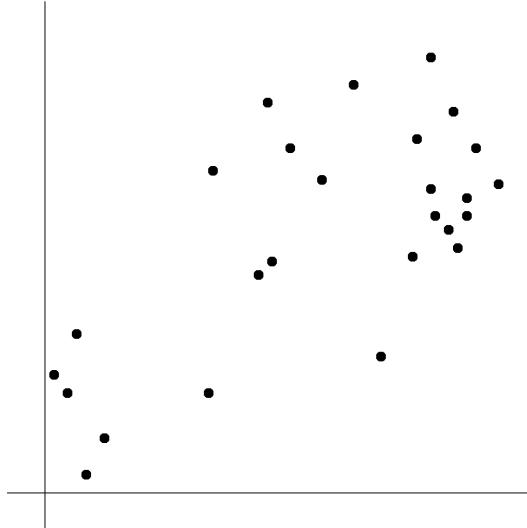


Рис. 22. Средняя положительная корреляция. Коэффициент корреляции равен 0,68

Графические иллюстрации, показывающие, что означает то или иное значение коэффициента корреляции, приведены на рисунках 21 — 27. Важно также отметить, что коэффициент корреляции показывает наличие именно линейной связи между переменными, а не какой-либо другой функциональной зависимости. Это иллюстрирует рисунок 28.

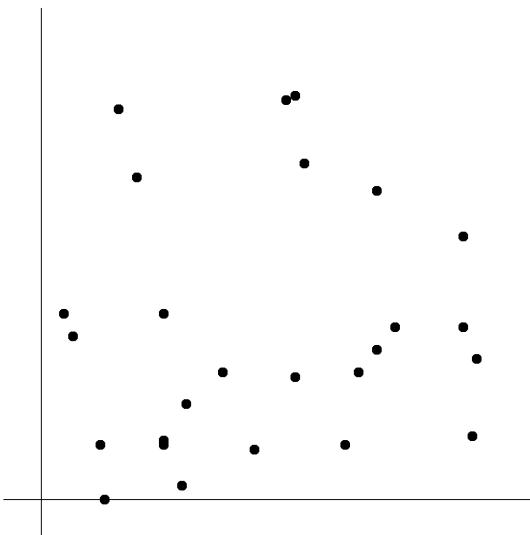


Рис. 23. Слабая корреляция. Коэффициент корреляции равен 0,11

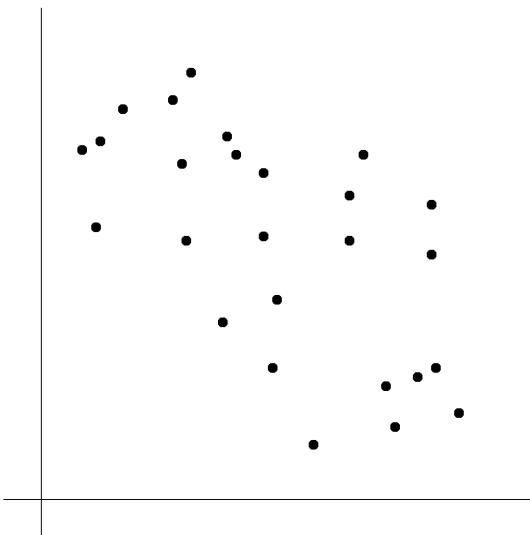


Рис. 24. Средняя отрицательная корреляция. Коэффициент корреляции равен $-0,62$

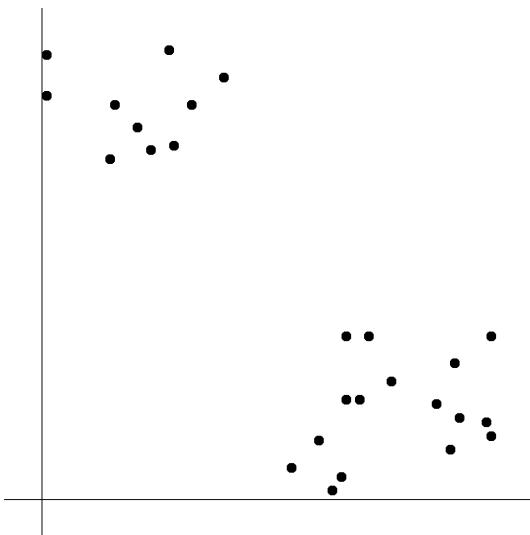


Рис. 25. Сильная отрицательная корреляция. Коэффициент корреляции равен $-0,84$

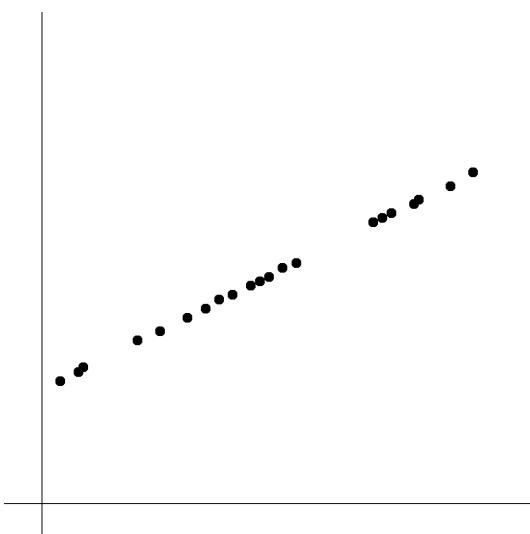


Рис. 26. Положительная линейная зависимость. Коэффициент корреляции равен 1

Часто при первом знакомстве думают, что коэффициент корреляции указывает на наклон некоторой усредненной прямой, проходящей через наше “облако точек” и в каком-то смысле описывающей наши данные. Это не так. Коэффициент корреляции описывает степень сгруппированно-

сти данных около этой прямой.

Про эту прямую — линию регрессии — речь пойдет чуть позже, а сейчас подчеркнем, что наклон зависит от масштаба, в частности от единиц измерения, а коэффициент корреляции от масштаба не зависит.

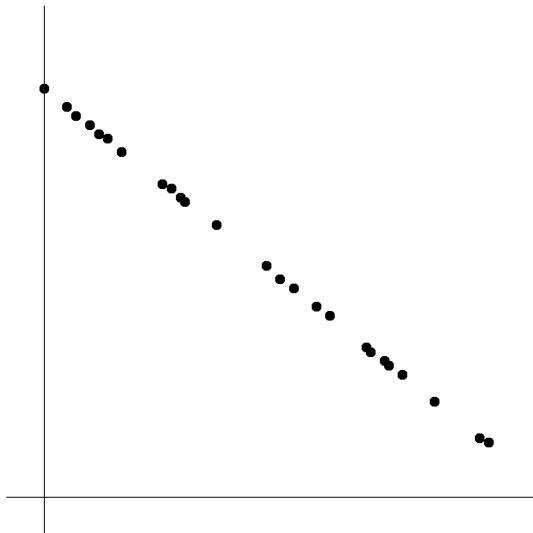


Рис. 27. Отрицательная линейная зависимость. Коэффициент корреляции равен -1

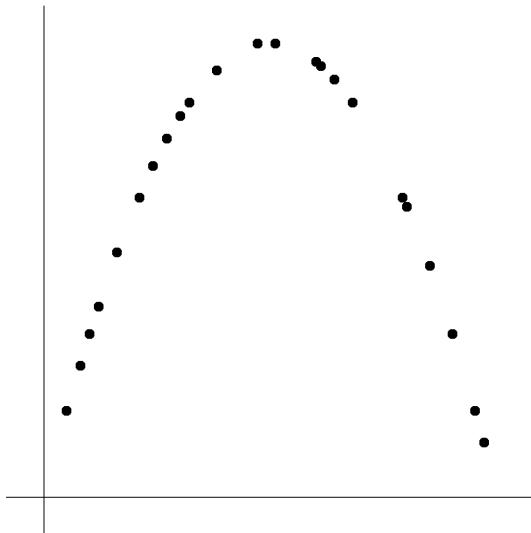


Рис. 28. Нелинейная функциональная зависимость. Коэффициент корреляции равен 0

Проверка гипотез о значении коэффициента корреляции.

Займемся теперь задачей о проверке значимости коэффициента корреляции. Пусть на самом деле корреляция равна нулю. Какие значения может принимать выборочный коэффициента корреляции?

Вновь мы имеем дело с двумерной выборкой $(x_i, y_i), i = 1 \dots n$. Будем считать, что эта выборка из двумерной нормальной совокупности, причем переменные некоррелированы и, следовательно, независимы. Числа из выборки, как x_i , так и y_i тоже будем считать независимыми друг от друга. При выяснении того, какое значение может тогда принимать выборочный коэффициент корреляции мы дважды

воспользуемся теоремой об ортогональном преобразовании многомерного нормально распределенного вектора.

Поскольку коэффициент корреляции не меняется при линейных преобразованиях переменных, то мы с самого начала можем считать, что как x_i , так и y_i имеют стандартное нормальное распределение. Переходим к новым переменным:

$$u_1 = \frac{x_1 + \cdots + x_n}{\sqrt{n}} = \bar{x}\sqrt{n}$$

и подбираем остальные переменные u_2, \dots, u_n так, чтобы преобразование было ортогональным. Тогда для y_i получим

$$v_1 = \frac{y_1 + \cdots + y_n}{\sqrt{n}} = \bar{y}\sqrt{n},$$

кроме того, как нетрудно доказать, коэффициент корреляции будет равен

$$r = \frac{u_2 v_2 + \cdots + u_n v_n}{\sqrt{\sum_{i=2}^n u_i^2 \cdot \sum_{i=2}^n v_i^2}}.$$

Еще отметим, что

$$(n-1)s_x^2 = u_2^2 + \cdots + u_n^2,$$

$$(n-1)s_y^2 = v_2^2 + \cdots + v_n^2,$$

и оба выражения в левой части имеют распределение $\chi^2(n-1)$.

Введем теперь новое ортогональное преобразование. Пусть

$$w_2 = b_{22}v_2 + \cdots + b_{2n}v_n,$$

где

$$b_{2i} = \frac{u_i}{\sqrt{\sum_{i=2}^n u_i^2}}.$$

Это выражение также можно дополнить до ортогонального преобразования к новым переменным w_2, \dots, w_n . Поскольку

ортогональное преобразование сохраняет длины, получим

$$r = \frac{w_2}{\sqrt{v_2^2 + \dots + v_n^2}} = \frac{w_2}{\sqrt{w_2^2 + \dots + w_n^2}}.$$

Обозначим для удобства

$$w = w_2; \quad \zeta^2 = w_3^2 + \dots + w_n^2,$$

и заметим, что по теореме об ортогональном преобразовании новые переменные w и ζ^2 независимы, первая из них имеет стандартное нормальное распределение, а вторая распределена по $\chi^2(n - 2)$. Теперь получим

$$r^2 = \frac{w^2}{w^2 + \zeta^2}.$$

Отсюда можно выразить ζ :

$$\zeta = w \frac{\sqrt{1 - r^2}}{r}.$$

Кроме того, как мы знаем, переменная

$$t = \frac{w}{\zeta / \sqrt{n - 2}}$$

распределена по Стьюденту с $(n - 2)$ степенями свободы. Окончательно получаем, что в случае правильности нулевой гипотезы, то есть при отсутствии корреляции, переменная

$$t = \frac{r \sqrt{n - 2}}{\sqrt{1 - r^2}}$$

имеет распределение Стьюдента с $(n - 2)$ степенями свободы. Теперь мы можем проверять гипотезы, строить критические области и т. д.

Пример 47. Даны двумерная выборка $(x_i, y_i), i = 1 \dots 9$. Найти выборочный коэффициент корреляции между x и y . При уровне значимости $\alpha = 0,05$ проверить гипотезу о значимости коэффициента корреляции.

x_i	11	14	25	27	18	10	16	20	21
y_i	7	8	14	15	10	11	9	14	18

Решение. Расчет, проведенный с помощью встроенной функции CORREL пакета Libre Office, дает результат $r = 0,743$. В нашем случае $n = 9$. Подставив эти значения в формулу, найдем значение t -критерия: $t = 2,936$.

Находим теперь 95% квантиль распределения Стьюдента с 7 степенями свободы: $t_{\text{кр}} = 2,365$. Поскольку $t > t_{\text{кр}}$, на этом уровне значимости мы обязаны отклонить гипотезу о том, что коэффициент корреляции не значим.

Линейная регрессия.

Рассмотрим теперь такую задачу. Пусть мы опять имеем дело с двумерной выборкой $(x_i, y_i), i = 1 \dots n$. Можно ли провести прямую $y = kx + b$, в каком-то смысле оптимально аппроксимирующую наши данные?

Решение этой задачи, то есть, по сути, определение неизвестных параметров k и b , было предложено Гауссом в связи с проблемой обработки наблюдений, в первую очередь астрономических. Введенный им *метод наименьших квадратов* и по сей день широко используется в различных областях, как в технических, так и в экономических, биологических, психологических и других науках.

Назовем разность $(y_i - kx_i - b)$ между измеренным и предсказанным значениями *невязкой*. Потребуем теперь, следуя Гауссу, чтобы сумма квадратов невязок была минимальной.

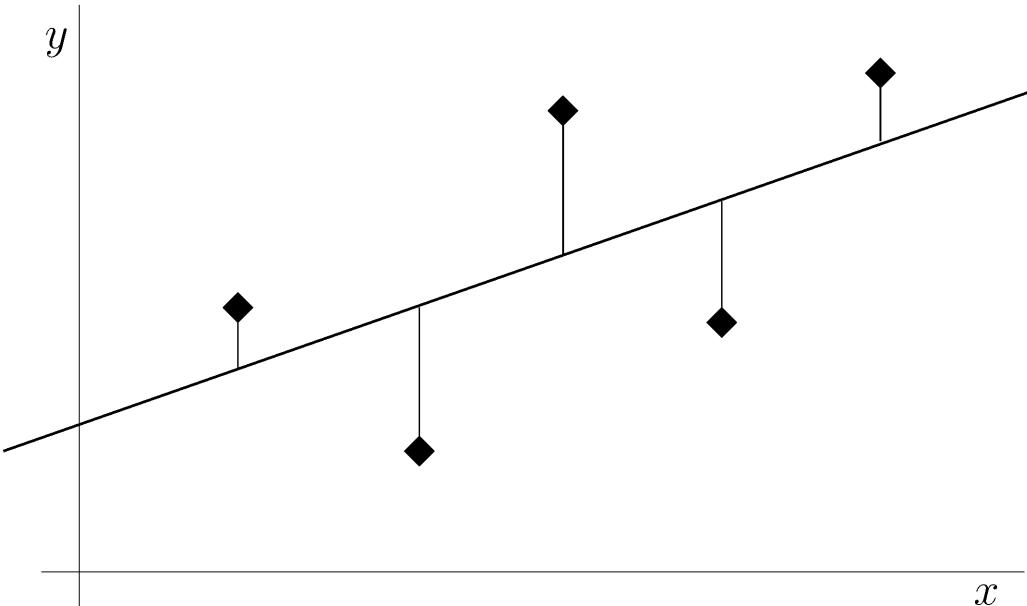


Рисунок 29. Линейная регрессия.

На рисунке 29 невязки представлены вертикальными отрезками, соединяющими экспериментальные точки с прямой линией регрессии.

Сразу сделаем два замечания. Во-первых, казалось бы, логичнее было бы минимизировать сумму модулей, а не сумму квадратов. Оказывается, однако, что при этом мы лишаемся возможности пользоваться средствами дифференциального исчисления: модуль — не везде дифференцируемая функция. Не последнюю роль играет и эстетические соображения, например, простота и изящество получаемых формул.

Во вторых, мы минимизируем квадраты разностей между игреками, а не квадраты расстояний до прямой. Это означает, по сути, что иксы нам известны точно.

Вывод нормальных уравнений.

Какм надо действовать дальше — понятно. Надо записать выражение для суммы квадратов невязок, продифференцировать его по k и по b , а затем приравнять к нулю получившиеся частные производные. Из этой пары уравнений

найдем интересующие нас коэффициенты k и b . В нашем случае получатся такие уравнения

$$\begin{cases} b \cdot n + k \cdot \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ b \cdot \sum_{i=1}^n x_i + k \cdot \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i. \end{cases}$$

Эти уравнения называют нормальными уравнениями регрессии, коэффициенты k и по b — коэффициентами регрессии, а саму найденную прямую $y = kx + b$ — линией регрессии y на x . При этом часто y называют переменной отклика, а x — факторной переменной.

Разумеется, расчеты следует проводить в электронных таблицах, да и сами коэффициенты системы нормальных уравнений вычислять не обязательно. В пакете Libre Office коэффициент наклона k вычисляет стандартная функция SLOPE, а свободный член b — стандартная функция INTERCEPT. В пакете Microsoft Office аналогичные функции называются НАКЛОН и ОТРЕЗОК соответственно.

Вывод формул же проведен совсем не для того, чтобы заставить студентов считать вручную, а для того, чтобы показать, как действовать в более общих случаях, когда подбираются коэффициенты для других видов зависимости, например, с нелинейными формулами или для большего числа факторных переменных. В таких случаях говорят о нелинейной или многомерной регрессии. Метод вывода нормальных уравнений в этих случаях будет аналогичен.

Пример 48. Для данных примера 47 построить линию регрессии y на x .

Решение. Электронные таблицы дают такой ответ:

$$y = 1,19x + 3,983.$$

В нашем случае простой линейной регрессии систему нормальных уравнений можно упростить. Разделив первое урав-

нение на n , получим

$$\bar{y} = k\bar{x} + b,$$

а это означает, что при простой линейной регрессии прямая проходит через точку (\bar{x}, \bar{y}) , то есть через центр облака точек. Поэтому мы можем сразу записать уравнение регрессии в виде

$$y - \bar{y} = k(x - \bar{x}).$$

Тогда после преобразований получим систему нормальных уравнений с диагональной матрицей коэффициентов

$$\begin{cases} b \cdot n = \sum_{i=1}^n y_i \\ k \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})y_i. \end{cases}$$

Для невязки мы получим такое выражение:

$$z = (y - \bar{y}) - k(x - \bar{x}).$$

Заметим, что невязка имеет нулевое среднее, а сумма квадратов невязок пропорциональна ее дисперсии. Найдем дисперсию этой невязки, воспользовавшись определением коэффициента корреляции $\text{cov}(x, y) = r\sigma_x\sigma_y$. При этом в выкладках, не упоминая об этом особо, мы будем пользоваться уже давно известной нам формулой $D(x + c) = Dx$:

$$Dz = Dy - 2k\text{cov}(x, y) + k^2Dx = \sigma_y^2 - 2kr\sigma_x\sigma_y + k^2\sigma_x^2.$$

Поскольку коэффициент k в уравнении регрессии подбирался так, чтобы эта дисперсия была минимальной, мы можем получить связь между коэффициентами корреляции и регрессии. Действительно, хорошо известная школьная формула координаты вершины графика квадратного трехчлена дает

$$k = r \frac{\sigma_y}{\sigma_x}.$$

Подставляя значение k в формулу для дисперсии невязки (или, если угодно, пользуясь известной школьной формулой минимума квадратного трехчлена), получим, что

$$Dz = Dy(1 - r^2).$$

Несложная выкладка показывает, что факторная переменная x и невязка z некоррелированы:

$$\begin{aligned} \text{cov}(z, x) &= \text{cov}(y - kx, x) = \text{cov}(y, x) - k\sigma_x^2 = \\ &= r\sigma_x\sigma_y - r\frac{\sigma_y}{\sigma_x}\sigma_x^2 = 0. \end{aligned}$$

Тогда получим, что

$$\begin{aligned} Dy &= D(kx + z) = k^2 Dx + Dz = r^2 \frac{Dy}{Dx} Dx + Dz = \\ &= Dyr^2 + Dy(1 - r^2), \end{aligned}$$

и это разложение, хоть и представляет собой тождество, имеет свою важную интерпретацию. Первый член в правой части называют объясненной дисперсией, а второй — остаточной дисперсией. Квадрат коэффициента корреляции показывает, какая доля изменчивости данных объясняется уравнением регрессии. Поэтому этот квадрат часто называют *коэффициентом детерминации* (от английского to determine — объяснять).

Еще более важную роль этот коэффициент играет в множественной регрессии.

Термин “множественная регрессия” был впервые использован в работе английского статистика Карла Пирсона (Carl Pearson), опубликованной в 1908 году. Для этого статистического метода также применяется название “многомерный регRESSIONНЫЙ анализ”.

В общественных и естественных науках процедуры множественной регрессии чрезвычайно широко используются в исследованиях. В общем, множественная регрессия поз-

воляет исследователю задать вопрос (и, вероятно, получить ответ) о том, “что является лучшим предсказывающим фактором для …”. Например, исследователь в области образования мог бы пожелать узнать, какие факторы являются лучшими предсказывающими факторами успешной учебы в средней школе. А психолога мог быть заинтересовать вопрос, какие индивидуальные качества позволяют лучше предсказать степень социальной адаптации индивида. Социологи, вероятно, хотели бы найти те социальные индикаторы, которые лучше других предсказывают результат адаптации новой группы мигрантов и степень ее слияния с обществом. Заметим, что термин “множественная” указывает на наличие нескольких факторов, которые используются в модели.

В статистических исследованиях наиболее часто применяется множественная линейная регрессия, когда зависимость между влияющими факторами x_1, x_2, \dots, x_k и результирующим фактором (результатом) y выражается формулой линейной зависимости

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k.$$

Вычисление неизвестных коэффициентов $b_0, b_1, b_2, \dots, b_k$ производится с помощью метода наименьших квадратов.

Общее назначение множественной регрессии состоит в анализе связи между несколькими независимыми переменными (обычно их называют факторами) и зависимой переменной.

Специалисты по кадрам обычно используют процедуры множественной регрессии для определения вознаграждения адекватного выполненной работе. Можно определить некоторое количество факторов или параметров, таких, как “размер ответственности” (x_1) или “число подчиненных” (x_2), которые, как ожидается, оказывают влияние на стоимость работы. Кадровый аналитик затем проводит исследование

размеров окладов (y) среди сравнимых компаний на рынке, записывая размер жалования и соответствующие характеристики (т.е. значения параметров) по различным позициям. Эта информация может быть использована при анализе с помощью множественной регрессии для построения регрессионного уравнения в следующем виде:

$$y = b_0 + b_1 x_1 + b_2 x_2.$$

Как только эта так называемая линия регрессии определена (то есть вычислены неизвестные коэффициенты b_1 и b_2), аналитик оказывается в состоянии построить график ожидаемой (предсказанной) оплаты труда и реальных обязательств компании по выплате жалования. Таким образом, аналитик может определить, какие позиции недооценены (лежат ниже линии регрессии), какие оплачиваются слишком высоко (лежат выше линии регрессии), а какие оплачены адекватно.

Линия регрессии выражает наилучшее предсказание зависимой переменной (y) по независимым переменным (x_i). Однако, природа редко (если вообще когда-нибудь) бывает полностью предсказуемой и обычно имеется существенный разброс наблюдаемых точек относительно подогнанной прямой. Отклонение отдельной точки от линии регрессии (от предсказанного значения) называется остатком.

Как и в случае одномерной регрессии, коэффициент b_0 совпадает со средним значением переменной отклика y , $\bar{y} = b_0$.

Можно показать, что общую сумму квадратов остатков SS (Sum of Squares, ее называют общей дисперсией, она характеризует разброс значений относительно среднего) можно разложить на две составляющие: дисперсию, обусловленную регрессией $SS_{\text{пер}}$ и “необъясненную”, или остаточную дисперсию $SS_{\text{ост}}$. Мы с этим уже встречались ранее,

когда речь шла о дисперсионном анализе. Там такое разложение называлось теоремой Штейнера. Запишем его в виде формулы:

$$SS = SS_{\text{пер}} + SS_{\text{ост.}}$$

Чем меньше отношение необъясненной дисперсии (необъясненного разброса) к общей дисперсии (общему разбросу), тем, очевидно, лучше прогноз. Например, если связь между переменными x и y отсутствует, то отношение остаточной дисперсии переменной y к исходной дисперсии равно 1. Если x и y жестко связаны, то остаточная изменчивость отсутствует, и отношение дисперсий будет равно 0. В остальных случаях отношение дисперсий будет между 0 и 1.

Коэффициентом детерминации называется

$$R^2 = \frac{SS_{\text{пер}}}{SS}.$$

Это значение непосредственно интерпретируется следующим образом. Если коэффициент детерминации равен 0,4, то изменчивость значений переменной y около линии регрессии составляет $1 - 0,4 = 0,6$ от исходной дисперсии; другими словами, 40% от исходной изменчивости могут быть объяснены, а 60% остаточной изменчивости остаются необъясненными. В идеале желательно иметь объяснение если не для всей, то хотя бы для большей части исходной изменчивости. Значение R^2 является индикатором степени подгонки модели к данным (значение R^2 близкое к 1 показывает, что модель объясняет почти всю изменчивость соответствующих переменных, то есть, по-видимому, модель является адекватной).

Надо отметить, что указанная выше формула для коэффициента детерминации верна только для случая линейной регрессии. Для более сложных случаев иногда могут помочь специализированные статистические пакеты, например, Statistica, SPSS и другие.

- Обычно рекомендуют провести тест на значимость коэффициентов множественной корреляции. Предположим, что:
- всего исследуются n наблюдений и k независимых факторов;
 - все случайные ошибки имеют нормальное распределение с нулевым средним и одной и той же дисперсией σ^2 ;
 - на самом деле никакой зависимости нет, то есть все коэффициенты b_1, b_2, \dots, b_k равны нулю.

В этом случае можно доказать, что случайная величина $SS_{\text{рег}}/\sigma^2$ имеет распределение χ^2 с k степенями свободы, а случайная величина $SS_{\text{ост}}/\sigma^2$ имеет распределение χ^2 с $n - k - 1$ степенями свободы. Поэтому величина

$$F = \frac{SS_{\text{рег}}/k}{SS_{\text{ост}}/(n - k - 1)}$$

имеет распределение Фишера с $(k, n - k - 1)$ степенями свободы. Это дает возможность провести тест на значимость коэффициентов регрессии.

Значение критерия, как легко убедиться, можно выразить через коэффициент детерминации:

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)},$$

или

$$F = \frac{R^2(n - k - 1)}{(1 - R^2)k}.$$

Далее следует провести обычную процедуру сравнения вычисленного критерия с квантилем F-распределения. Для вычисления этого квантиля в Microsoft Excel нужно использовать стандартную функцию FPACПОБР, а в Libre Office — стандартную функцию FINV. Например, =FPACПОБР(0,05; 7; 52) вычислит значение 3,49. В этой формуле 0,05 — уровень значимости;

7, 52 — число степеней свободы числителя и знаменателя соответственно.

Следует иметь в виду, что формула =FPACСПОБР(0,05; 7; 52) вычисляет не 0,05, как могло бы показаться, а 0,95-квантиль распределения Фишера. Как и в других случаях, программа Microsoft Excel делает не так, как было бы разумно при знакомстве с российским ГОСТом.

Надо иметь в виду, что если значение критерия оказалось больше значения квантиля (и надо отвергнуть гипотезу), то это вовсе не значит, что построенная модель адекватна и может использоваться для предсказаний. Это означает только, что какой-то из коэффициентов b_1, b_2, \dots, b_k значим, то есть не равен нулю. Иными словами, должна быть отвергнута гипотеза $y = b_0$. Для суждений об адекватности модели следует опираться на величину коэффициента детерминации.

Если модель признана достаточно адекватной, может быть поставлена задача определения того, какие факторы в этой модели являются наиболее существенными, а также задача исключения малозначащих факторов.

Множественная регрессия предоставляет пользователю “сомнение” включить в качестве предикторов все переменные, какие только можно, в надежде, что некоторые из них окажутся значимыми. Это происходит из-за того, что извлекается выгода из случайностей, возникающих при простом включении возможно большего числа переменных, рассматриваемых в качестве факторов другой, представляющей интерес переменной. Эта проблема возникает тогда, когда к тому же и число наблюдений относительно мало. Интуитивно ясно, что едва ли можно делать выводы из анализа вопросника со 100 пунктами на основе ответов 10 респондентов. Большинство авторов советуют использовать, по крайней мере, от 10 до 20 наблюдений (респондентов) на одну переменную, в противном случае оценки регрессионной линии будут, вероятно, очень ненадежными и, скорее

всего, невоспроизводимыми для желающих повторить это исследование.

Если дело дойдет до того, что факторов будет больше, чем наблюдений (и они будут независимы), то остаточная дисперсия окажется равной нулю, а коэффициент детерминации станет равным единице. Можно ли такую модель использовать для предсказания? Вряд ли.

Высокие значения коэффициента детерминации, вообще говоря, не свидетельствуют о наличии причинно-следственной зависимости между переменными (так же как и в случае обычного коэффициента корреляции). Основное концептуальное ограничение всех методов регрессионного анализа состоит в том, что они позволяют обнаружить только числовые зависимости, а не лежащие в их основе причинные связи. Например, можно обнаружить сильную положительную связь (корреляцию) между разрушениями, вызванными пожаром, и числом пожарных, участвующих в борьбе с огнем. Следует ли заключить, что пожарные вызывают разрушения?

Конечно, наиболее вероятное объяснение этой корреляции состоит в том, что размер пожара (внешняя переменная, которую забыли включить в исследование) оказывает влияние, как на масштаб разрушений, так и на привлечение определенного числа пожарных (т.е. чем больше пожар, тем большее количество пожарных вызывается на его тушение). Хотя этот пример довольно прозрачен, в реальности при исследовании корреляций альтернативные причинные объяснения часто даже не рассматриваются.

Задача 174. Пусть двумерная генеральная совокупность конечна. Тогда, представив всю эту совокупность как выборку, можно определить выборочную корреляцию между компонентами. Доказать, что она совпадает с корреляцией между компонентами, определенной для генеральной сово-

купности.

Задача 175. Написать систему нормальных уравнений для определения коэффициентов квадратичной зависимости вида

$$y = ax^2 + bx + c.$$

Задача 176. Доказать, что коэффициент детерминации при множественной линейной регрессии переменной y по переменным x_1, x_2, \dots, x_k не может быть меньше, чем квадрат частного коэффициента корреляции между y и любым из x_i .