

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

МИРЭА – РОССИЙСКИЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ

И. С. Пулькин

МАТЕМАТИЧЕСКАЯ
СТАТИСТИКА

Учебное пособие

Москва 2020

Этот текст представляет собой электронную версию пособия по математической статистике. Бумажный вариант был выпущен издательством МИРЭА в 2915 году. В данном тексте исправлены некоторые неточности.

© Пулькин И. С.

Оглавление

Предисловие	4
1. Основные понятия статистики	5
2. Случайные величины	18
3. Нормальное распределение	43
4. Порядковые статистики	58
5. Проверка статистических гипотез	70
6. Гипотезы о математическом ожидании и дисперсии	88
7. Дисперсионный анализ	99
8. Корреляция и регрессия	110
Таблицы	130
Литература	135

Предисловие

Это пособие представляет собой введение в математическую статистику. Оно основано на лекциях, читавшихся автором на протяжении ряда лет студентам, обучающимся по специальности “Управление в технических системах”.

Малый объем курса — а учебным планом предусмотрено всего 16 лекционных часов на изучение дисциплины — вынудил автора отказаться от включения в пособие многих тем, обычно рассматриваемых в математической статистике. Так, в пособие не вошли ранговые критерии и другие вопросы непараметрической статистики, элементы теории информации, а доверительное оценивание упомянуто лишь вкратце. Основное внимание уделено проверке статистических гипотез, а также оценкам параметров распределений.

На отбор материала повлиял в первую очередь инженерно-технический характер указанной специальности. Автор хотел добиться того, чтобы студенты после изучения данного курса могли свободно работать с ГОСТами и другими нормативными документами, содержащими описания процедур статистической обработки. С этим же связан и “инженерный” уровень строгости в математических рассуждениях. Впрочем, в нескольких местах, там, где подчеркивается роль ортогональных преобразований в многомерной статистике, приведены и полные доказательства.

В пособии приведено большое количество задач, как теоретических, так и расчетных. При выполнении расчетов следует ориентироваться на использование электронных таблиц, и поэтому в пособии приведены описания некоторых часто используемых встроенных статистических функций.

§1. Основные понятия статистики

Основная задача 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 температуры 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 1. На сайте <http://ru.wikinews.org/> приведены средние суточные температуры в феврале 2013 года в Москве.

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

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 среднее 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*. Так же 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 2 (из книги М. Гарднера “А ну-ка, догадайся” [4]). 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 3. “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 “децильный 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

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

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 $n/(n - 1)$.

В Libre Office для вычисления выборочной дисперсии используется стандартная функция VAR, дающая правильное значение. 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 в Microsoft Office можно воспользоваться стандартной функцией СТАНДОТКЛОН, а в Libre Office — функцией STDEV.

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 используют 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 интервалы от -12 до -9 , от -9 до -6 , и так далее. Подсчитав, сколько раз были именно такие температуры, получим следующую таблицу:

$-12 - -9$	$-9 - -6$	$-6 - -3$	$-3 - 0$	$0 - 3$
3	6	3	9	7

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. Эта гистограмма изображена на рисунке 1.

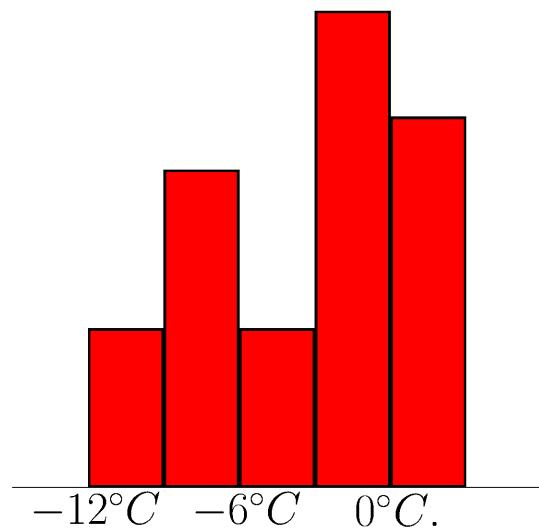


Рис. 1. Гистограмма температур воздуха в Москве в феврале 2013 года

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 при $n = 28$ наблюдениях 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 в Microsoft Office или COUNTIF в Libre Office. А если данные (в виде группированной выборки, то есть таблицы вроде приведенной выше) готовы, то в офисных пакетах 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 должны 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 дли-

на 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 — это один из методов наглядного представления данных. Такие методы используются довольно часто, 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 ознакомиться с некоторыми из них, следует порекомендовать недавно вышедшую книгу Нейтана Яу "Искусство визуализации в бизнесе"[10].

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 отличаются. 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.

Рассмотрим теперь еще один пример на построение гистограммы.

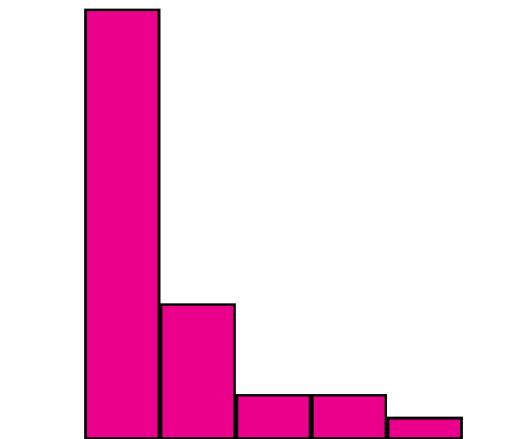
Пример 4. Пусть дана следующая выборка.

3,89	1,22	2,47	1,43	4,05	1,56
1,90	5,94	7,07	1,56	3,12	8,45
0,94	0,90	1,10	0,26	1,12	0,84
0,75	0,21	0,73	3,15	0,11	2,65
1,60	7,80	1,03	0,12	2,67	0,39

Для того, чтобы наглядно представить себе эти данные, следует сначала построить по ним группированную выборку. Все числа в выборке положительны, а максимальное значение равно 8,45. Можно построить гистограмму, выбрав отрезки 0 – 2, 2 – 4, и так далее до 8 – 10. Получим такую группированную выборку

0 – 2	2 – 4	4 – 6	6 – 8	8 – 10
19	6	2	2	1

Построим гистограмму по этой выборке.



Оказывается, мы имеем дело с сильно асимметричной выборкой, рисунок позволяет судить об особенностях этих данных. В дальнейшем мы еще вернемся к этому примеру, а сейчас вычислим средние и разброс. Вычисления показывают, что

- выборочное среднее $\bar{x} = 2,301$;
- выборочная медиана $m = 1,495$;
- группированное выборочное среднее $\bar{x}_{\text{групп}} = 2,33$;

- выборочная дисперсия $s^2 = 5,241$;
- группированная выборочная дисперсия $s_{\text{групп}}^2 = 4,782$.

Задачи

Задача 1. 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 2. 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?

В задачах 3 — 6 следует по заданной выборке построить группированную выборку, а по ней — гистограмму, а также вычислить средние и дисперсии.

Задача 3.

9,46	4,57	11,31	10,24	11,08	5,97
13,43	9,39	15,67	13,21	6,16	5,88
16,56	12,46	10,15	8,17	7,84	9,38
11,96	8,29	6,21	8,25	6,93	3,16
9,05	7,94	8,09	9,63	11,69	3,53

Задача 4.

6,21	5,17	5,82	10,59	9,64	8,38
5,98	2,69	1,43	1,50	1,79	7,47
4,30	5,16	11,48	1,52	4,55	8,35
1,01	8,32	4,47	4,58	8,80	6,59
9,71	3,97	0,97	16,41	8,72	0,84

Задача 5.

5,78	4,37	5,23	15,07	12,56	9,68
5,45	1,99	1,24	1,27	1,42	7,89
3,38	4,36	17,67	1,28	3,65	9,62
1,04	9,57	3,56	3,67	10,60	6,38
12,75	3,05	1,02	36,82	10,43	0,97

Задача 6.

7,90	1,44	12,66	9,72	11,98	2,54
20,09	7,75	30,45	19,23	2,73	2,45
35,25	16,44	9,50	5,41	4,88	7,72
14,72	5,61	2,78	5,54	3,60	0,75
7,03	5,04	5,27	8,27	13,85	0,89

§2. Случайные величины

Рассмотрение построенной гистограммы позволяет нам строить прогнозы и предсказания. Например, глядя на гистограмму февральских температур, мы можем более или менее точно предсказывать, сколько, скорее всего, в феврале будет морозных дней, дней с плюсовой температурой, и так далее.

Во многих случаях, однако, мы можем предсказать заранее, даже не сделав ни одного наблюдения, как приблизительно будет выглядеть гистограмма. Рассмотрим, например, такой случай. Пусть мы производим измерение некоторой величины. Это может быть электрическое сопротивление, или угол между двумя направлениями, или площадь земельного участка — можете представить себе что-нибудь еще на ваш выбор. Важно только, что речь должна идти о реальной величине, и ее измерение — не чье-то мнение или экспертный прогноз, а некоторая объективная процедура.

Всякое измерение всегда совершается с некоторой ошибкой — приборов с нулевой погрешностью не бывает. Мы, однако, можем принять два достаточно естественных предположения: 1) наш прибор намеренно не врет, то есть в среднем его показания точны, и при этом он может ошибаться как в ту, так и в другую сторону; 2) мы приблизительно представляем себе точность нашего прибора.

В метрологии для принятых нами предположений используется такая терминология:

- 1) измерительный прибор не имеет систематической погрешности;
- 2) измерительный прибор характеризуется некоторой точностью.

Тогда, если мы много раз повторим наше измерение (одной и той же величины), а затем построим гистограмму, то мы

должны получить картинку, похожую на ту, что изображена на рисунке 2.

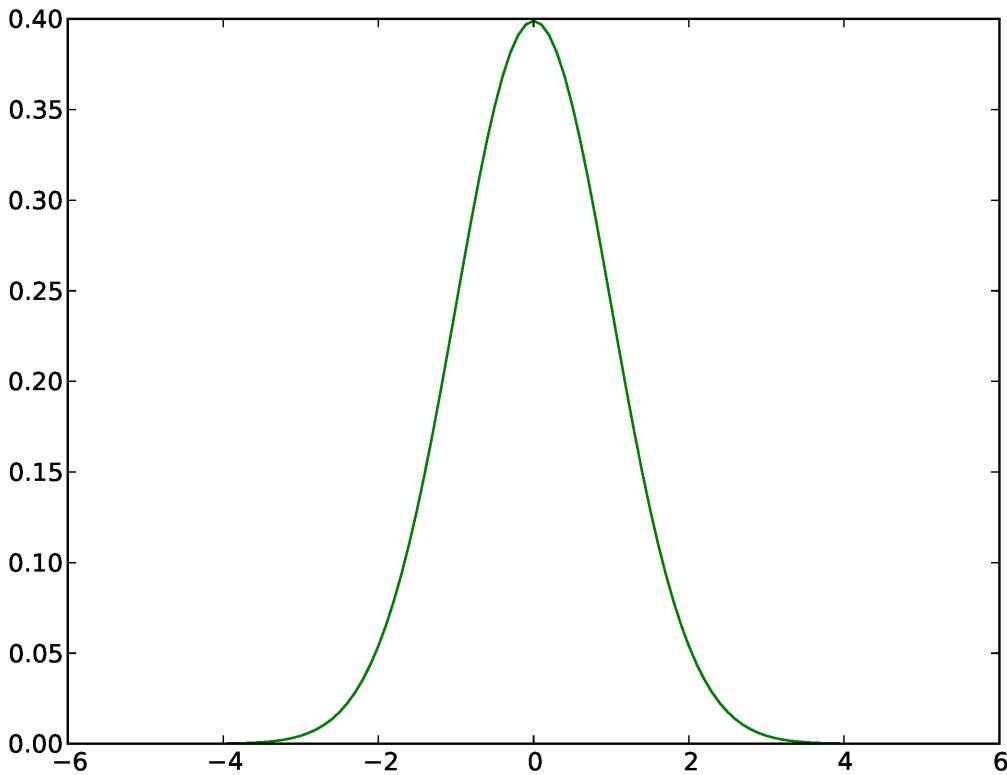


Рис. 2. График погрешности измерений

Из рисунка ясно, что чаще всего ошибки в одном конкретном измерении не слишком велики, более грубые промахи случаются значительно реже. Далее мы увидим, что повторение измерений приводит к повышению точности.

Кривая, изображенная на рисунке, называется кривой нормального распределения (ее часто называют еще гауссовой кривой или гауссианой), и говорят, что ошибки измерений имеют нормальное (гауссово) распределение. Как мы увидим чуть позже, нормальное распределение встречается и во многих других случаях.

На этом примере мы видим, что в некоторых случаях можно не гадать и не собирать примеры, а сразу воспользоваться готовым распределением, подходящим к нашему случаю. Далее

мы разберем некоторые из распределений, представленных в ГОСТ 50779-2000 “Статистические понятия и термины”, и рассмотрим, когда они применяются.

Для начала скажем, что распределения описывают *случайную величину*.

Понятие случайной величины — ключевое понятие 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.

Случайными величинами будут, например:

- результат измерения сопротивления;
- число посетителей в офисе завтра;
- температура воздуха завтра в 12-00;
- число очков, набранных командой в чемпионате.

Из этого перечисления ясно, что случайные величины бывают двух видов. Одни из них могут быть только целыми (число посетителей, число очков), а другие могут быть и дробными (сопротивление, температура). Первые случайные величины называют *дискретными*, а вторые — *непрерывными*.

Познакомимся сначала с дискретными случайными величинами.

Распределение Бернулли

Пример 5. 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

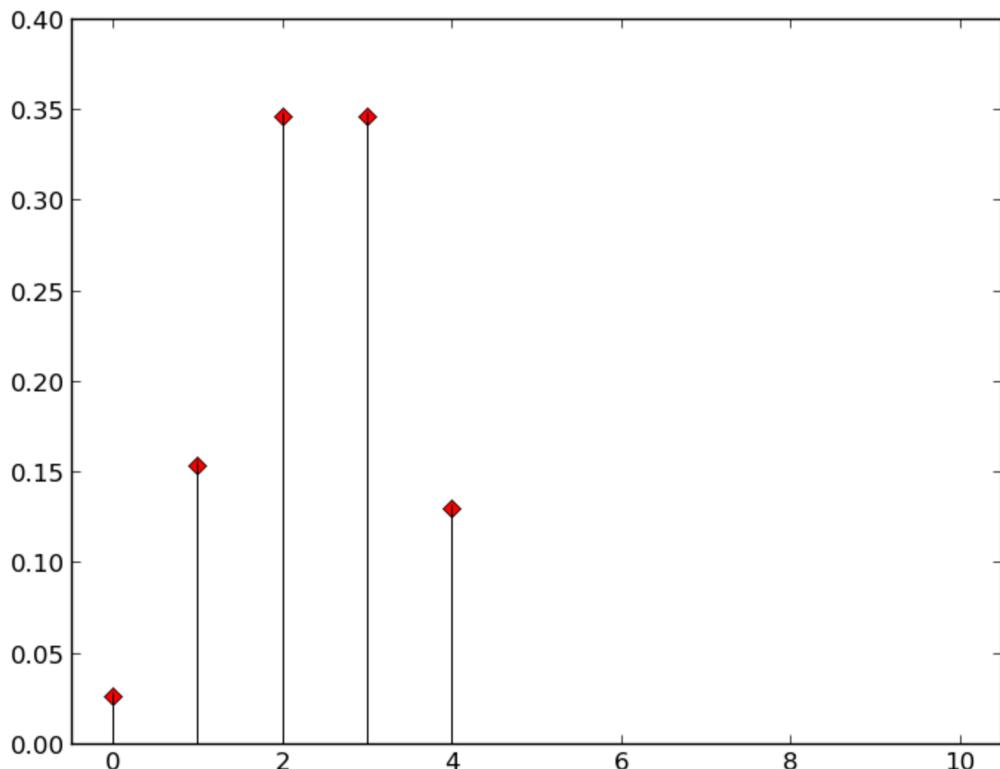


Рис. 3. Распределение Бернулли с $n = 4, p = 0,6$

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

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

Напомним здесь, что число C_n^k принято называть числом сочетаний из n по k . Оно равно числу способов, которыми можно из n предметов выбрать k , и вычисляется по формуле

$$C_n^k = \frac{n!}{k!(n - k)!},$$

где восклицательный знак означает факториал числа, то есть произведение всех целых чисел от 1 до этого числа. Например, $5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$. В Microsoft Excel для вычисления числа сочетаний можно воспользоваться стандартной функцией ЧИСЛКОМБ. В Libre Office число комбинаций вычисляет стандартная функция СОМБИН.

Говорят, что определенная выше случайная величина k имеет распределение Бернулли (или — подчиняется распределению Бернулли). 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.

Это распределение также называют биномиальным распределением. Для нахождения вероятностей по указанной выше формуле в Microsoft Excel можно воспользоваться стандартной функцией БИНОМРАСП, а в Libre Office — стандартной функцией В.

Общее определение и числовые характеристики 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	\dots	x_n
P	p_1	p_2	\dots	p_n

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

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

Введем теперь числовые характеристики случайной величины. Наиболее часто используются математическое ожидание и дисперсия, характеризующие соответственно среднее значение и разброс случайной величины. Рассмотрим определения этих понятий.

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

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

Стандартное (среднеквадратическое) отклонение — квадратный корень из дисперсии

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

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, принятное 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:

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

Для распределения Бернулли

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

Пример 6. Случайная величина X — число очков при бросании одной игральной кости. Найти математическое ожидание и дисперсию.

Решение. Подставив значения в формулы, получим: $MX = 3,5$; $DX = 35/12$.

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 достаточно естественных 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).

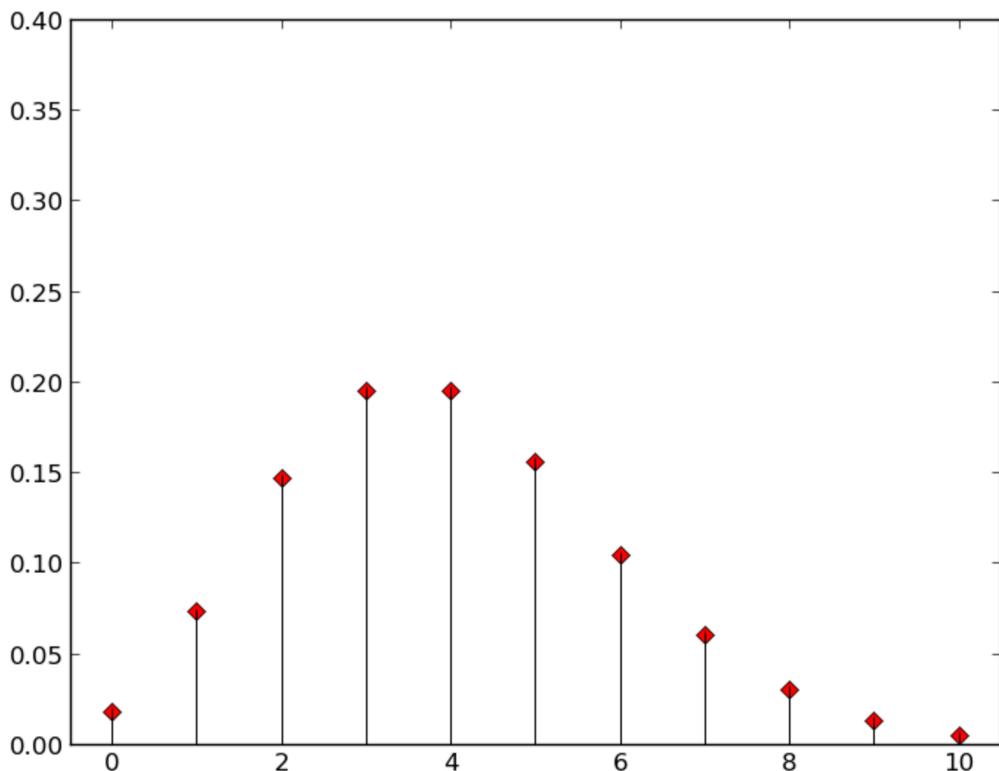


Рис. 4. Распределение Пуассона с $\lambda = 4$

Разумеется, существуют и другие дискретные распределения, но в жизни они встречаются реже, поэтому перейдем к рассмотрению непрерывных случайных величин.

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 в точности a или b , равна нулю.

Математическое ожидание и дисперсия непрерывной случайной величины определяются следующим образом:

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

Все формулы для математического ожидания и дисперсии, которые были выписаны ранее для дискретных случайных ве-

личин, будут справедливы и для непрерывных случайных величин. Напомним эти формулы.

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

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

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

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

Для независимых случайных величин

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

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

Еще одной характеристикой непрерывной случайной величины является медиана. Это такое значение x_0 , что $P(X < x_0) = 1/2$. Для симметричных распределений медиана совпадает с математическим ожиданием.

Для непрерывных случайных величин, наряду с плотностью вероятностей, часто используется *функция распределения* случайной величины. 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”).

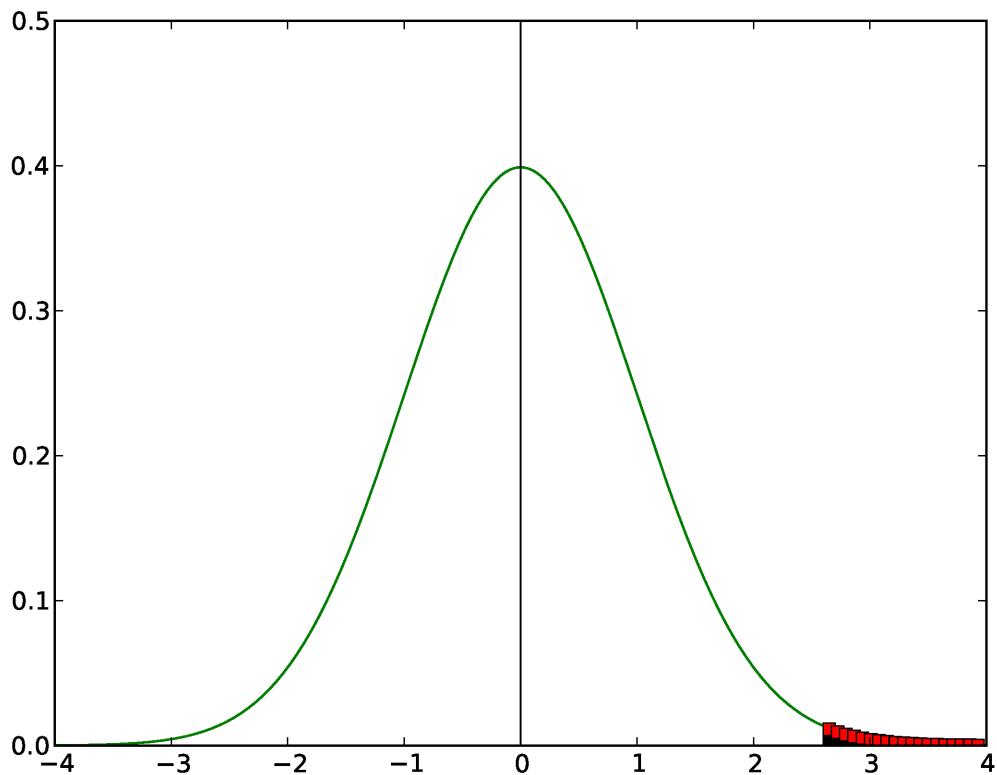


Рис. 5. Квантиль. Площадь незаштрихованной части под кривой равна α , заштрихованной части — $1 - \alpha$

Таким образом, α -квантиль распределения — это решение x_0 уравнения

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

В терминах функции распределения α -квантиль — это решение уравнения

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

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.

Ясно, что $1/2$ -квантиль — это медиана распределения. Свое специальное название имеют еще некоторые квантили. $1/4$ и $3/4$ -квантили называются *квартили* (от слова квартал — четверть). Аналогично, $1/10$ и $9/10$ -квантили — это децили. В словах квартиль и дециль — ударение на “и”.

Переходим к примерам часто используемых распределений.

Равномерное распределение

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

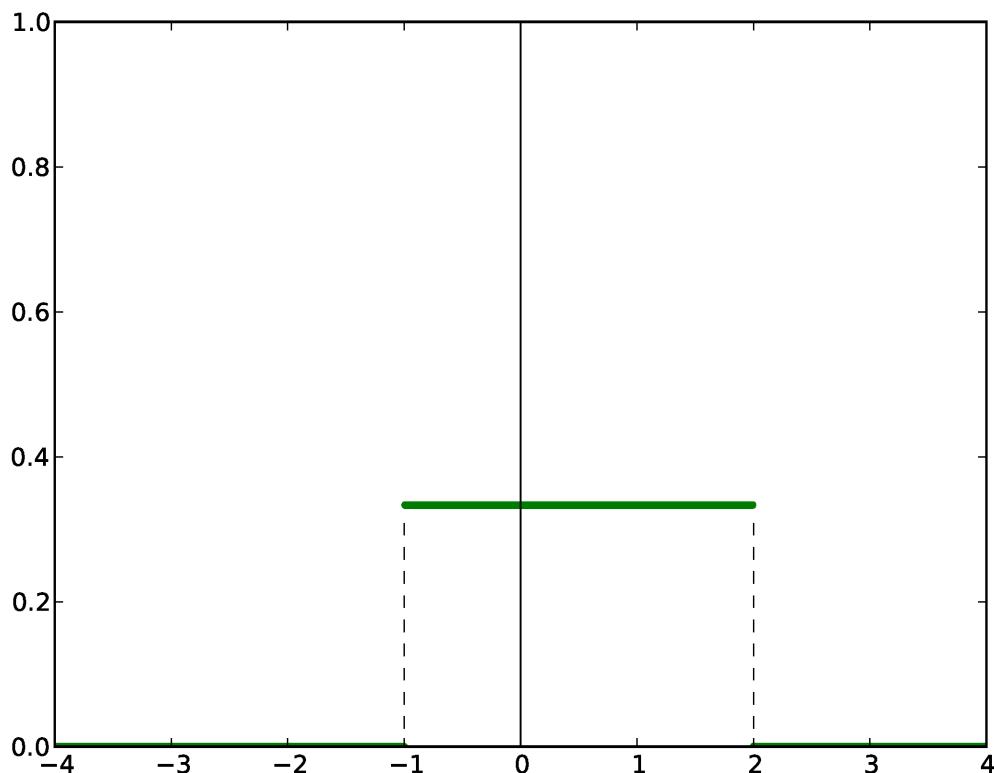


Рис. 6. Графики плотности вероятности равномерного на отрезке $[-1; 2]$ распределения

Это — чисто теоретическое, модельное распределение. В реальной жизни автору встречаться с ним не приходилось, толь-

ко 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

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

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.

Для нормального распределения плотность вероятности выражается следующей простенькой формулой

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

Из этой формулы видно, что нормальное распределение характеризуется двумя параметрами — a и σ . Оказывается, первый из этих параметров равен среднему значению, а второй — стандартному отклонению. Какое именно среднее? Поскольку нормальное распределение симметрично, для него математическое ожидание и медиана равны.

На рисунке изображены графики плотности вероятности для двух нормальных распределений. У обоих этих распределений среднее равно нулю, но у одного из них стандартное отклонение равно 1, а у другого — равна 2. График первого из них мы уже видели на рисунке 2.

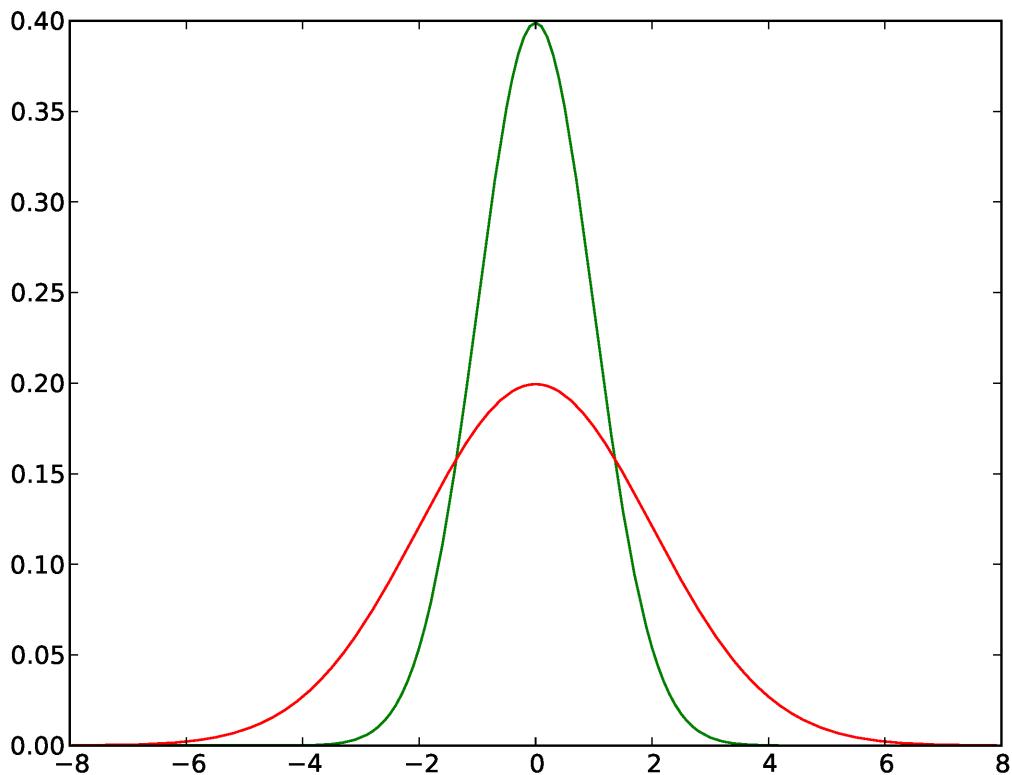


Рис. 7. Графики плотности вероятности двух нормальных распределений

Стандартным нормальным распределением называется такое нормальное распределение, у которого среднее равно 0, а дисперсия равна 1. Функция распределения для случайной величины, подчиняющейся этому распределению, выражается интегралом

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

Это, по сути, стандартная функция, которая используется настолько часто, что заслужила особое наименование: интеграл ошибок. Кстати, буква, ее обозначающая — это не русская “Ф”, как могло бы показаться, а заглавная греческая буква “фи”.

Для нахождения значений плотности вероятности нормального распределения в Microsoft Excel можно использовать стандартную функцию НОРМРАСП, а в Libre Office — стандарт-

ную функцию NORMDIST. Эти же функции используются для нахождения функции распределения. Один из аргументов этих функций — аргумент “интегральный” — может принимать значение 0 и 1. В первом случае вычисляется плотность вероятности, а во втором — функция распределения.

Для нахождения интеграла ошибок — функции распределения стандартного нормального распределения — в Microsoft Excel и в Libre Office предусмотрены стандартные функции НОРМСТРАСП и NORMSDIST соответственно.

Гораздо чаще, чем значение функций, требуется находить квантили нормального распределения. В Microsoft Excel для этого предусмотрены стандартные функции НОРМРАСПОБР и НОРМСТРАСПОБР. Первая из них — для нахождения квантилей нормального распределения с заданными средним и дисперсией, а вторая — для нахождения квантилей стандартного нормального распределения. В Libre Office соответствующие функции называются NORMINV и NORMSINV.

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. Одно из применений отметим сразу: она позволяет пользоваться таблицами только для стандартного нормального распределения. Именно такие таблицы и приведены во всех учебниках по теории вероятностей и математической статистике. Надо отметить, что в них приведены зна-

чения интеграла ошибок только для положительных значений x . Для отрицательных следует пользоваться формулой $\Phi(-x) = 1 - \Phi(x)$.

Пример 7. Вычислить вероятность того, что случайная величина, нормально распределенная со средним 8 и дисперсией 16, принимает значения в интервале от 0 до 12.

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

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

Далее преобразуем эту формулу:

$$z = P\left(\frac{0-8}{4} < \frac{X-8}{4} < \frac{12-8}{4}\right).$$

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

$$\begin{aligned} z &= P\left(-2 < \frac{X-8}{4} < 1\right) = \Phi(1) - \Phi(-2) = \\ &= 0,8413 - 0,0228 = 0,8186. \end{aligned}$$

Показательное распределение

Также его называют экспоненциальным распределением. 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

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

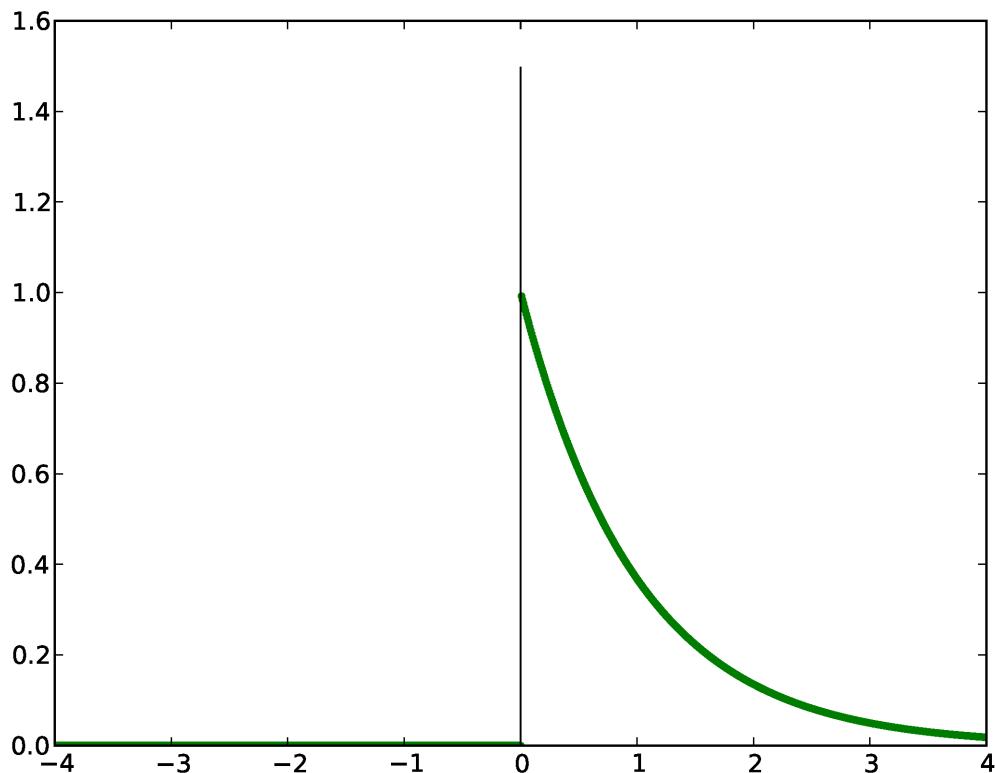


Рис. 8. Плотность показательного распределения

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

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

По видимому, мы уже встречались с этим распределением: построенная в примере 4 гистограмма позволяет выдвинуть вполне правдоподобное предположения о том, что эти данные подчиняются именно показательному распределению.

Распределение Вейбулла

Функция распределения задается формулой

$$F(x) = 1 - \exp\left(-\left(\frac{x-a}{b}\right)^k\right),$$

а плотность вероятности — формулой

$$\rho(x) = \frac{k}{b^k} (x-a)^{k-1} \exp\left(-\left(\frac{x-a}{b}\right)^k\right).$$

Распределение Вейбулла зависит от трех параметров:

k — параметр формы кривой распределения;

b — параметр масштаба;

a — параметр сдвига (его, однако, часто принимают равным нулю).

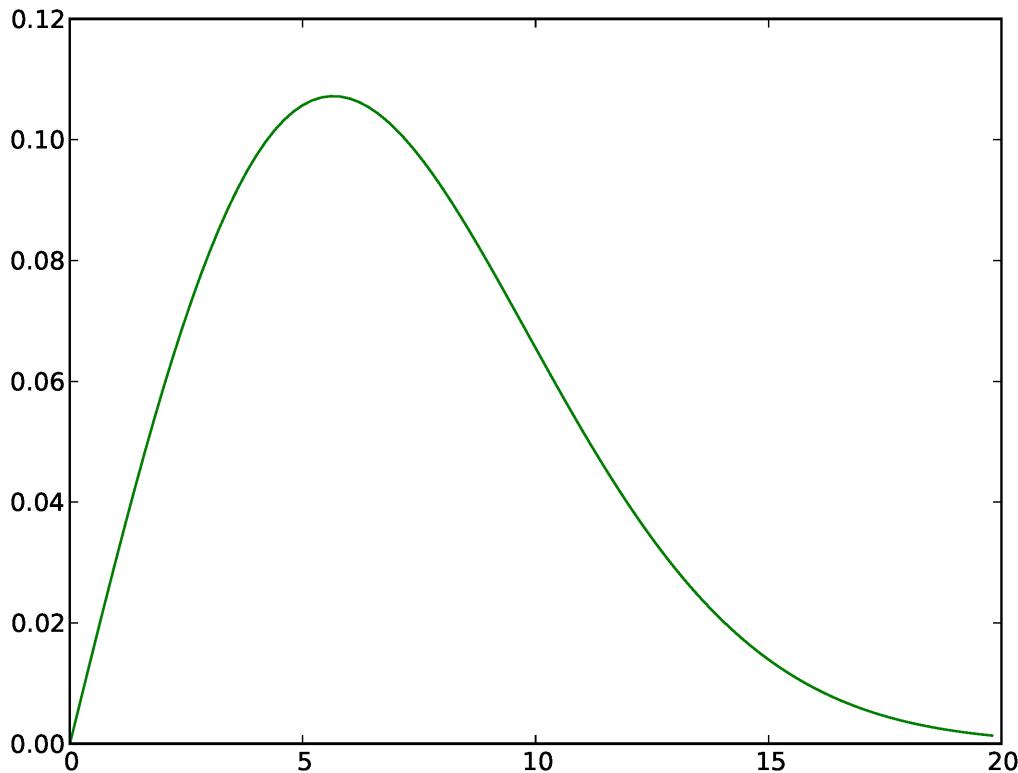


Рис. 9. Графики плотности вероятности распределения Вейбулла при $a = 0, b = 8, k = 2$

Это распределение также используется в теории надежности как время работы агрегата до первого отказа. В отличие от

показательного распределения, распределение Вейбулла описывает наработку на отказ для достаточно сложных агрегатов. Оно позволяет учесть такие тонкие моменты, как “приработку” деталей агрегата и уменьшение вероятности отказа при такой приработке. Впрочем, показательное распределение является частным случаем распределения Вейбулла при $k = 1, a = 0$.

Математическое ожидание и дисперсия для этого распределения равны:

$$MX = b\Gamma\left(1 + \frac{1}{k}\right), \quad DX = b^2 \cdot \left(\Gamma\left(1 + \frac{2}{k}\right) - \Gamma^2\left(1 + \frac{1}{k}\right)\right).$$

Распределение Парето

Это распределение задается плотностью вероятностей

$$\rho(x) = \frac{\alpha}{x_0} \left(\frac{x_0}{x}\right)^{\alpha+1}$$

при $x > x_0$. При $x < x_0$ плотность вероятности равна нулю.

Это распределение широко применяется в экономической статистике. Оно описывает распределение доходов в обществе, при условии, что все эти доходы превышают некоторый базовый уровень x_0 .

Его математическое ожидание и дисперсия равны:

$$MX = \frac{\alpha x_0}{\alpha - 1}, \quad DX = \left(\frac{x_0}{\alpha - 1}\right)^2 \frac{\alpha}{\alpha - 2}.$$

При $\alpha \leq 2$ дисперсия бесконечна, а при $\alpha \leq 1$ бесконечно и математическое ожидание.

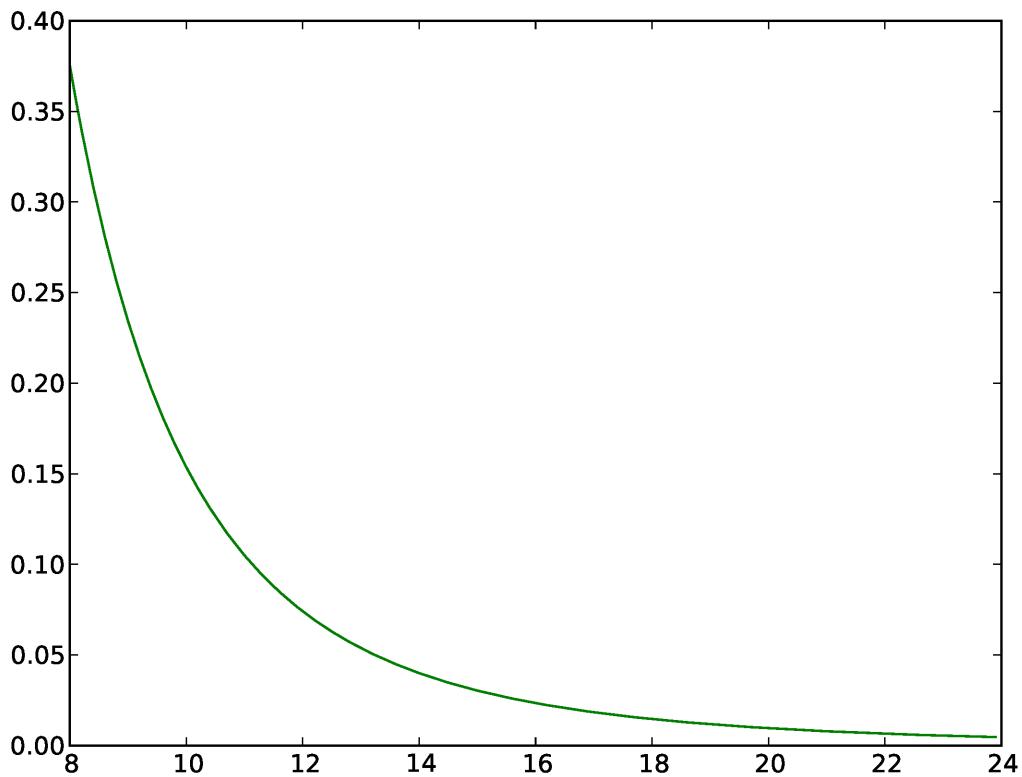


Рис. 10. Графики плотности вероятности распределения Парето при $x_0 = 8, \alpha = 3$

Про еще несколько непрерывных распределений, часто используемых в статистике, будет рассказано ниже, где пойдет речь о многомерном нормальном распределении.

Как уже было сказано, если есть основания считать, что случайная величина имеет заданное распределение, то достаточно по выборке оценить параметры этого распределения. Такой подход принято называть *параметрической статистикой*.

Метод максимального правдоподобия

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*. Эту оценку часто можно получить, просто проинтегрировав 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 8. 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. Как и предлагалось ранее, для упрощения задачи 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 9. 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 10. 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:

$$a = \frac{x_1 + \cdots + x_n}{n};$$

$$\sigma^2 = \frac{(x_1 - a)^2 + \cdots + (x_n - a)^2}{n}.$$

Пример 11. 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

Задача 7. В задачах 3 — 6 были построены гистограммы. Можете ли Вы выдвинуть предположение — какому из уже известных распределений подчиняются данные.

Задача 8. В классическом задачнике Гмурмана [5] приведено несколько другое определение интеграла ошибок:

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

Как связаны между собой функции $\Phi(x)$ и $\tilde{\Phi}(x)$?

Задача 9. Что больше, медиана или математическое ожидание, для вытянутых вправо асимметричных распределений, например, для показательного или распределения Парето?

Задача 10. Для случайной величины, равномерно распределенной на отрезке $[a, b]$, написать формулы для плотности вероятностей и функции распределения.

Задача 11. Доказать, что распределение Вейбулла при $k = 1, a = 0$ — показательное распределение.

Задача 12. Доказать, что сумма двух независимых случайных величин, распределенных по Пуассону с параметрами λ_1 и λ_2 , также распределена по Пуассону с параметром $\lambda_1 + \lambda_2$.

Задача 13. Получена выборка k_1, \dots, k_n , подчиняющаяся распределению Пуассона. Найти оценку максимального правдоподобия для неизвестного параметра λ .

Задача 14. Получена выборка x_1, \dots, x_n , подчиняющаяся распределению Вейбулла. Найти оценку максимального правдоподобия для неизвестных параметров k и b , считая, что параметр сдвига a равен нулю.

Задача 15. Получена выборка x_1, \dots, x_n , подчиняющаяся распределению Парето. Найти оценку максимального правдоподобия для неизвестного параметра α , если параметр x_0 известен.

Задача 16. Случайная величина X имеет геометрическое распределение, если $P(X = k) = pq^k$, где $q = 1 - p$, $k = 0, 1, \dots$. Получена выборка k_1, \dots, k_n , подчиняющаяся геометрическому распределению. Найти оценку максимального правдоподобия для неизвестного параметра p .

Задача 17. В условиях предыдущей задачи оценка максимального правдоподобия может быть получена другим способом. Известно, что случайная величина, имеющая геометрическое распределение — не что иное, как число испытаний Бернулли до первого успеха. Таким образом, в условиях предыдущей задачи, достигнуто n успехов, при этом было проведено $k = k_1 + \dots + k_n + n$ испытаний. Совпадают ли полученные оценки?

Задача 18. Распределение Максвелла, зависящее от параметра β , задается функцией плотности вероятностей

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

Получена выборка x_1, \dots, x_n , подчиняющаяся этому распределению. Найти оценку максимального правдоподобия для неизвестного параметра β .

Задача 19. Логнормальное распределение, зависящее от двух параметров, a и σ , задается при $x > 0$ функцией плотности вероятностей

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

Получена выборка x_1, \dots, x_n , подчиняющаяся этому распределению. Найти оценки максимального правдоподобия для неизвестных параметров a и σ .

Задача 20. Распределение Рэлея, зависящее от параметра a , задается функцией плотности вероятностей

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

Получена выборка x_1, \dots, x_n , подчиняющаяся этому распределению. Найти оценку максимального правдоподобия для неизвестного параметра a .

Задача 21. Распределение Накагами, зависящее от двух параметров, a и ω , задается функцией плотности вероятности

$$\rho(x) = \frac{2a^a}{\Gamma(a)\omega^a}x^{2a-1}\exp\left(-\frac{ax^2}{\omega}\right).$$

Получена выборка x_1, \dots, x_n , подчиняющаяся этому распределению. Считая параметр a известным, найти оценку максимального правдоподобия для неизвестного параметра ω .

§3. Нормальное распределение

Нормальное распределение и предельные теоремы

Нормальное распределение мы уже рассматривали. Здесь мы рассмотрим некоторые важные применения этого распределения.

Пример 12. 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? Обычно применяют такой критерий: если оно дальше от среднего, чем три сигмы, то это — промах, а если ближе, то нет.

Нормальное распределение является предельным распределением, и с этим связаны многие примеры его использования. В частности, классическая теорема Муавра — Лапласа утверждает, что при больших n число успехов в n испытаниях Бернулли имеет приблизительно нормальное распределение.

Следующий пример показывает, как эта теорема может применяться в статистике.

Пример 13. Экзамен по теории вероятностей с первого раза сдает примерно половина студентов. В каких пределах с веро-

ятностью 0,99 будет лежать число студентов, сдавших сразу экзамен, если сдает 100 человек?

Заместитель декана сообщил студентам, что экзамен с первого раза “чисто случайно” сдали 35 человек. Следует ли считать это отклонение случайным?

Как мы узнаем чуть позже, в математической статистике принята такая терминология. Если число 35 лежит внутри вычисленных пределов, то говорят, что данные с доверительной вероятностью 0,95 согласуются с гипотезой (о случайности отклонения), в противном случае — противоречат гипотезе.

Решение. Мы должны узнать, в каких пределах с вероятностью 0,99 лежит число “успехов”, то есть число студентов, не сдавших экзамен. Можно это записать в виде уравнения с неизвестным z :

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

Случайная величина X имеет распределение Бернулли с параметрами $n = 100$, $p = 1/2$. На основании теоремы Муавра — Лапласа мы можем считать, что X имеет приближенно нормальное распределение. При этом среднее и дисперсия будут такими же, как и для исходного распределения Бернулли: $MX = np = 50$, $DX = npq = 25$. Тогда, приводя к стандартному нормальному распределению, получим:

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

$$2\Phi\left(\frac{z}{\sqrt{npq}}\right) - 1 = 0,99;$$

$$\Phi\left(\frac{z}{\sqrt{npq}}\right) = 0,995;$$

$$\frac{z}{\sqrt{npq}} = 2,58; z = 2,58 \cdot \sqrt{25} \approx 12,9.$$

Таким образом, почти всегда — с вероятностью 99% — число сдавших экзамен будет находиться в пределах от 37 до 63 человек, и всего 35 сдавших — это не случайность.

Забегая вперед, скажем, что мы провели проверку статистической гипотезы. На основании этой проверки мы должны сделать вывод о том, что предположение было неправильным — оно противоречит наблюдаемому результату. Скорее всего, придется отвергнуть предположение о том, что в этот раз было $p = 1/2$.

Мы вернемся к подобным задачам после того, как рассмотрим общую схему проверки статистических гипотез.

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_i , а слу-

чайной величины X — через σ , то из последнего соотношения получим

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

то есть разброс растет пропорционально квадратному корню из числа слагаемых.

Центральная предельная теорема также часто применяется в статистике.

Пример 14. Игровую кость подбрасывали 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.

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:

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

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. Следовательно, компоненты независимы, и мы приходим к такому важному выводу: для многомерного нормального распределения некоррелированность и независимость — эквивалентны.

Пример 15. 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 корреляция.

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 Махаланобиса, зависит, будет ли отнесен этот объект к этому кластеру, или нет, то есть решение задачи классификации.

На рисунке 11 изображены три сгенерированных с помощью датчика случайных чисел кластера не сферической формы.

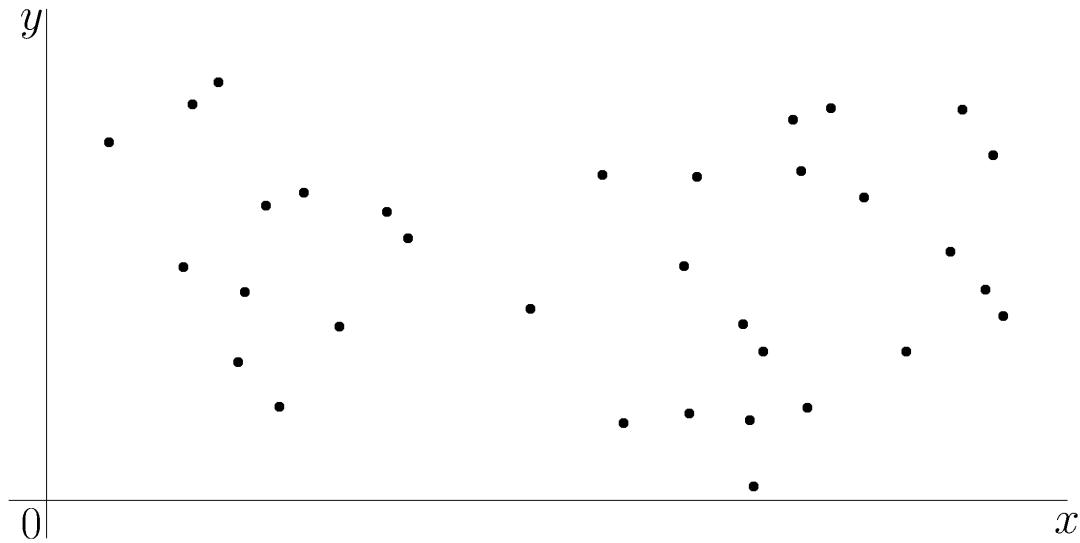


Рис. 11. Можно ли на этом рисунке разглядеть три кластера?

Если же распределение объектов в кластере явно не сферическое, а 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, лучшим 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 эллипса в направлении заданной точки.

Расстояние Махalanобиса было сформулировано в 1936 году индийским статистиком Прасанта Чандра Махalanобисом 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 слу-

чайного 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 Q , то есть получаем новые случайные величины Y_1, \dots, Y_n по формуле

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

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 Q — это матрица ортогональной замены базиса. Тогда для преобразованного случайного вектора (Y_1, \dots, Y_n) плотность вероятности в точке с координатами $\vec{y} = Q\vec{x}$, 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:

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

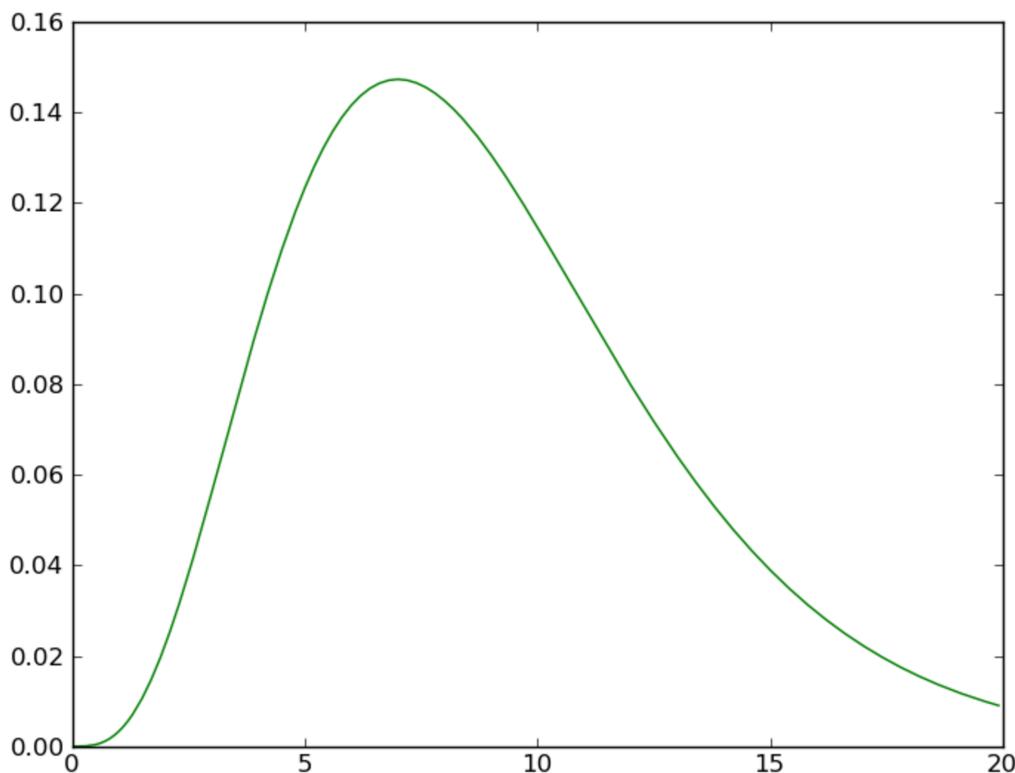


Рис. 12. Плотность распределения χ^2 с 9 степенями свободы

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.

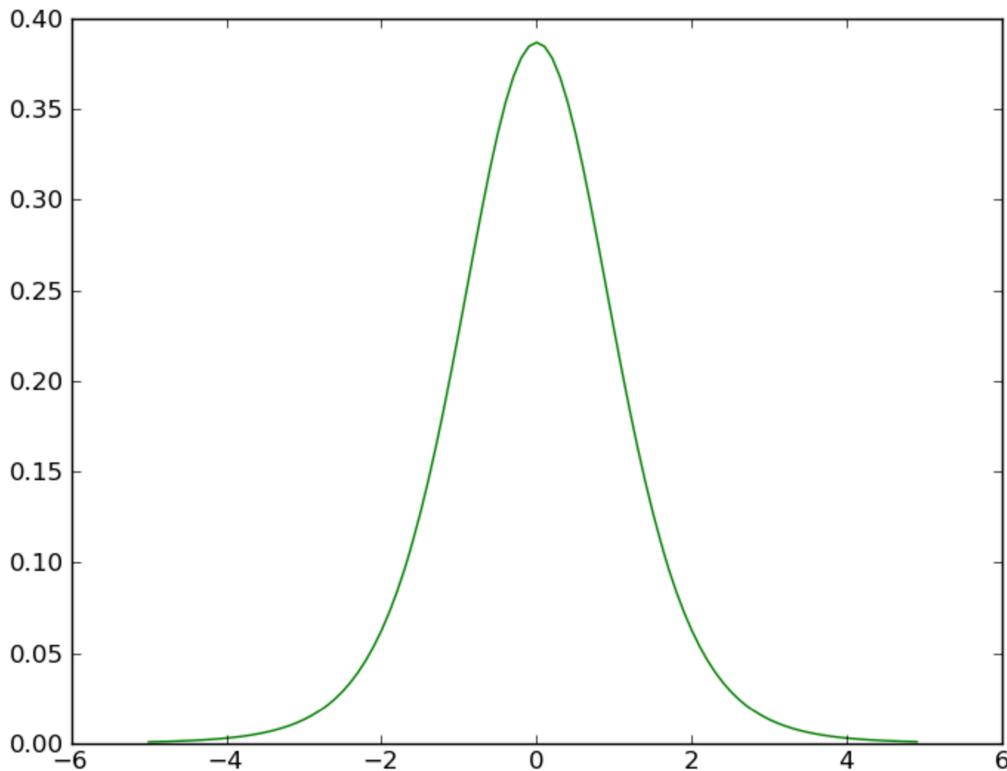


Рис. 13. Плотность распределения Стьюдента с 8 степенями свободы

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. Определим новую случайную величину в соответствии с формулой

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

Такая случайная величина называется распределенной по Фишеру с (m, n) степенями свободы. Для нее

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

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

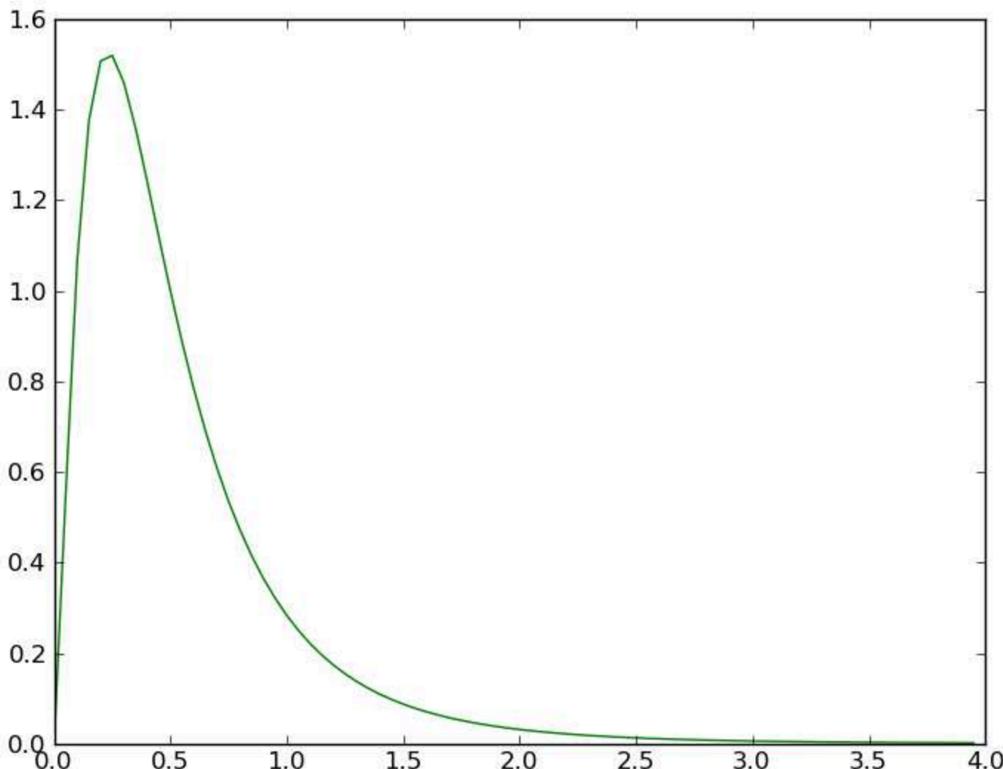


Рис. 14. Плотность распределения Фишера с 5 степенями свободы числителя и 11 степенями свободы знаменателя

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 ХИ2ОБР (в Microsoft Office) или CHIINV (в Libre Office). Для распределения Стьюдента встроенные функции называются соответственно СТЬЮДРАСПОБР 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,строенная 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 версийстроенная 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. Более подробные таблицы приведены в конце пособия.

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

Задача 22. Чему равна медиана распределения Стьюдента?

Задача 23. Найти 96% квантиль распределения Стьюдента с 9 степенями свободы.

Задача 24. Найти 98% квантиль распределения Стьюдента с 7 степенями свободы.

Задача 25. Найти 99% квантиль распределения Стьюдента с 11 степенями свободы.

Задача 26. Найти 96% квантиль распределения хи-квадрат с 9 степенями свободы.

Задача 27. Найти 98% квантиль распределения хи-квадрат с 7 степенями свободы.

Задача 28. Найти 99% квантиль распределения хи-квадрат с 11 степенями свободы.

Задача 29. Найти 96% квантиль распределения Фишера с (4,5) степенями свободы.

Задача 30. Найти 98% квантиль распределения Фишера с (7,8) степенями свободы.

Задача 31. Найти 99% квантиль распределения Фишера с (18,12) степенями свободы.

Задача 32. Найти вероятность того, что случайная величина, распределенная по Стьюденту с 5 степенями свободы, не превысит значения 2.

Задача 33. Найти вероятность того, что случайная величина, распределенная по хи-квадрат с 6 степенями свободы, не превысит значения 10.

Задача 34. Найти вероятность того, что случайная величина, распределенная по Фишеру с (5, 11) степенями свободы, не превысит значения 4.

Задача 35. Найти функцию распределения случайной величины, распределенной по χ^2 с одной степенью свободы.

Задача 36. Случайная величина распределена по Стьюденту с n степенями свободы. Как распределен квадрат этой случайной величины?

Задача 37. Пусть Z — α -квантиль распределения Фишера с (m, n) степенями свободы. Чему равен $(1 - \alpha)$ -квантиль распределения Фишера с (n, m) степенями свободы?

§4. Порядковые статистики

Рассмотрим такую задачу. В аудитории 10 абитуриентов сдают экзамены. 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 экзамены?

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, использовав 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 $x(n)$ — смещенная.

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 эф-

фективности 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 неизвестного 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:

$$P(x(n) \leq i) = P(x_1 \dots x_n \leq i) = \frac{C_i^n}{C_N^n};$$

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

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.

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

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

Ясно, что это — несмешенная оценка. Нетрудно также найти дисперсию этой оценки. Она равна

$$D\bar{x} = \frac{\sigma^2}{n}.$$

Рассмотрим теперь оценку дисперсии

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

Найдем математическое ожидание этой статистики. Для этого преобразуем выражение

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

Теперь воспользуемся формулой

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

справедливой для любой случайной величины. Из нее следует, что

$$M(\bar{x}^2) = D\bar{x} + (M\bar{x})^2 = \frac{\sigma^2}{n} + a^2;$$

$$M(x_i^2) = Dx_i + (Mx_i)^2 = \sigma^2 + a^2.$$

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 несмешенную 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” ([3]).

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 объяс-

нение 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”).

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.

Задачи

Задача 38. Найти распределение k -й порядковой статистики из равномерного распределения на $[0, 1]$.

Задача 39. 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 40. Первая из моделей движения лифтов из предыдущей задачи также допускает различные толкования:

1а) 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 1а) 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 41. Какова вероятность, что лифт, вызванный со второго этажа, придет снизу, если лифтов в здании два? Использовать все модели из двух предыдущих задач.

Задача 42(*). 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?

Задача 43. Пусть x_1, \dots, x_n — выборка из нормального распределения с известным средним a . Доказать, что оценка дисперсии

$$\tilde{s}^2 = \frac{(x_1 - a)^2 + \dots + (x_n - a)^2}{n}$$

является несмешенной.

Задача 44. Пусть x_1, \dots, x_n — выборка из нормального распределения с известным средним a и неизвестной дисперсией σ^2 .

а) Доказать, что оценка

$$T = \sqrt{\frac{\pi}{2}} \frac{1}{n} \sum_{i=1}^n |x_i - a|$$

— несмешенная оценка параметра σ .

б) (*) Найти дисперсию этой оценки.

Задача 45. Величина X оценивается по данным нескольких измерений этой величины x_1, \dots, x_n , причем дисперсии x_i известны, равны σ_i^2 и, вообще говоря, могут быть различны. (Такая ситуация часто встречается в геодезии — так называемые неравноточные измерения). Строится оценка вида

$$\hat{X} = w_1 x_1 + \dots + w_n x_n,$$

где w_i — некоторые положительные числа (веса).

а) Доказать, что условие несмешенности этой оценки:

$$w_1 + \dots + w_n = 1.$$

б) Какие должны быть веса, чтобы эта оценка была эффективной (то есть имела наименьшую дисперсию)?

§5. 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 должны 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 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 известно, что критерий подчиняется заданному распределению, то вероятности ошибок для разных областей можно 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. Вероят-

нность области принятия гипотезы (площадь под графиком), равна $1 - \alpha$, вероятность критической области равна α . На втором из них изображена двусторонняя критическая область.

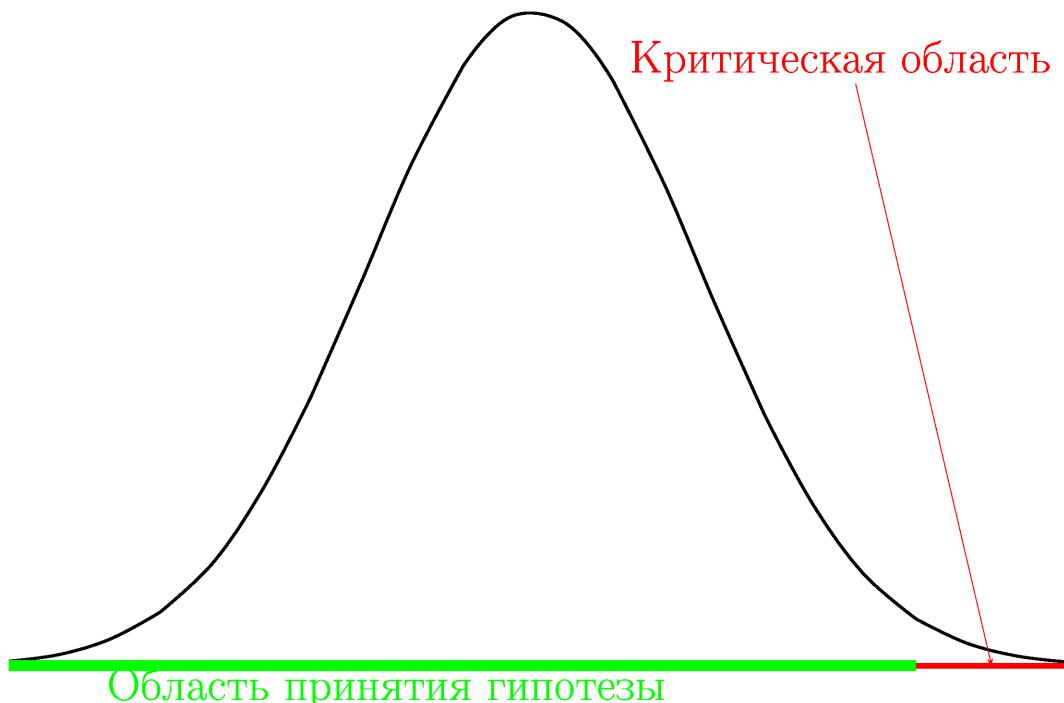


Рис. 15. Односторонняя критическая область



Рис. 16. 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 достаточно. 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.

Отметим один принципиальный момент, связанный 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 выглядели 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 неизвестно-

го 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 16. 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 нормального 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 $n = 767$.

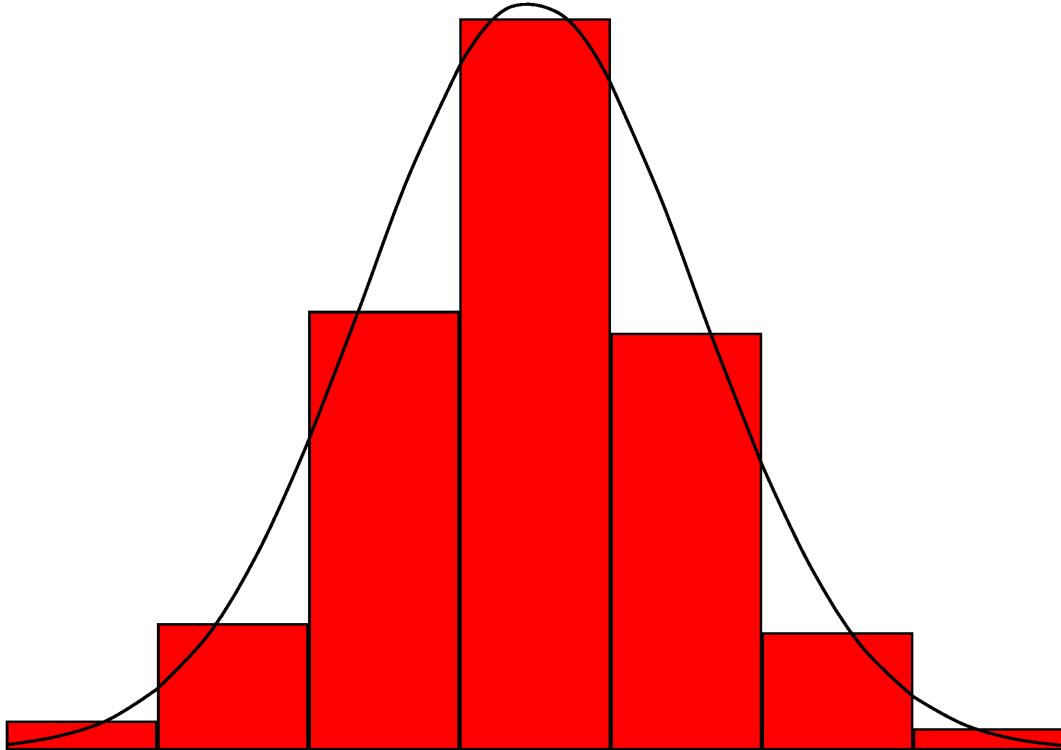


Рис. 17. Гистограмма экспериментальных данных и теоретическая кривая

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 интервала;

$n_1, \dots n_m$ — 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 перекрывают-ся, 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 , 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 $n_i - pr_i$ не слишком велики, то гипотеза похожа на правильную. Эту идею мы и используем. Дальней-

шее 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” [2].

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, понижающая 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.

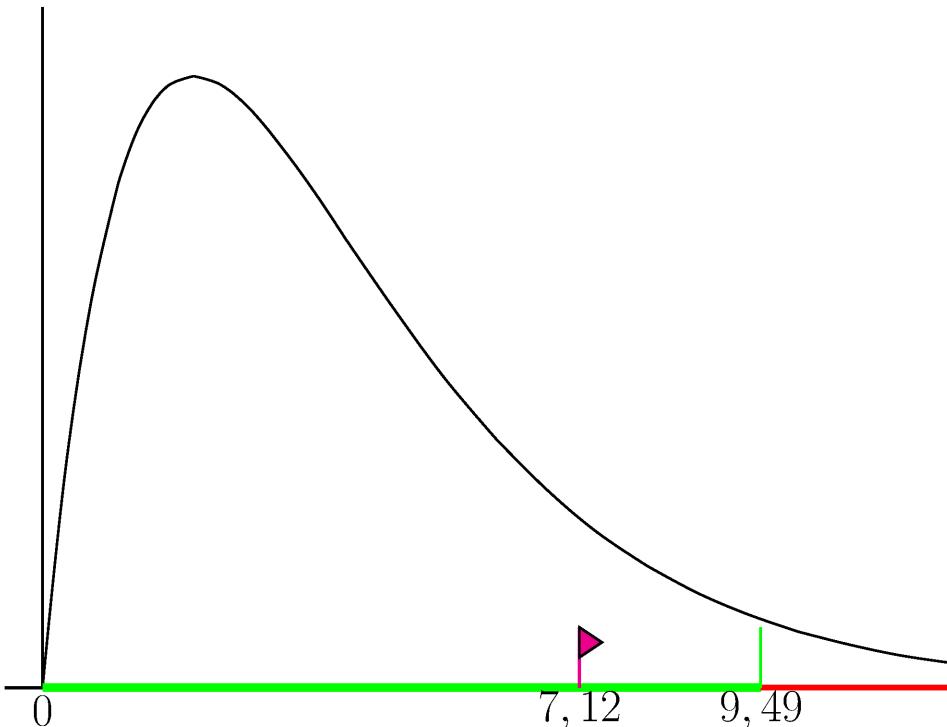


Рис. 18. Сравнение критерия и критического значения

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.

Рассмотрим еще один пример.

Пример 17. Ранее в примере 4 была построена гистограмма, а затем было отмечено, что эти данные, похоже, подчиняются показательному распределению. Можно ли проверить эту гипотезу?

Выполним процедуру проверки статистической гипотезы с использованием критерия хи-квадрат. Вычисляем теоретические частоты, предполагая, что данные подчиняются показательному распределению со средним $\bar{x} = 2,301$ — это среднее было вычислено ранее. В нашем случае функция распределения равна

$$F(x) = 1 - e^{-\lambda x}, \text{ где } \lambda = \frac{1}{\bar{x}}.$$

Вероятность интервала, например, от 6 до 8, равна $F(8) - F(6)$, а теоретическая частота равна произведению этой вероятности на объем выборки, то есть на 30. Как и ранее, продолжим крайний интервал на всю ось, то есть вместо интервала 8 — 10 будем рассматривать интервал 8 — $+\infty$. Тогда вероятность этого интервала будет равна не $F(10) - F(8)$, а $F(+\infty) - F(8)$, то есть $1 - F(8)$. Разница, впрочем, несущественна.

Заполняем таблицу, добавив 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
ЭЧ	19	6	2	2	1
ТЧ	17,42	7,30	3,066	1,28	0,93
χ^2	0,14	0,23	0,37	0,40	0,01

Суммируя элементы последней строки, находим значение критерия

$$\chi^2 = 1,15.$$

В отличие от нормального распределения, у показательного по выборке определяется только $r = 1$ параметр, поэтому для $m = 5$ интервалов число степеней свободы будет равно

$$\nu = m - 1 - r = 3.$$

Из таблиц находим значение 95% квантиля распределения хи-квадрат:

$$\chi^2_{0,95}(3) = 7,815.$$

Наблюдаемое значение меньше критического, так что выдвинутая ранее гипотеза вполне правдоподобна.

Задачи

Задача 46. Проверить с помощью критерия хи-квадрат гипотезы, выдвинутые в задаче 7 на основании гистограмм, построенных в задачах 3 – 6.

Задача 47. В $N = 900$ независимых испытаниях было получено $k = 300$ успехов. При уровне значимости $\alpha = 0.05$ проверить гипотезу о том, что вероятность p успеха в каждом из испытаний равна $p_0 = 0,3$, если альтернативная гипотеза — $H_1 = p > p_0$.

Задача 48. В $N = 1200$ независимых испытаниях было получено $k = 450$ успехов. При уровне значимости $\alpha = 0.05$ проверить гипотезу о том, что вероятность p успеха в каждом из испытаний равна $p_0 = 0,33$, если альтернативная гипотеза — $H_1 = p \neq p_0$.

Задача 49. В $N = 700$ независимых испытаниях было получено $k = 295$ успехов. При уровне значимости $\alpha = 0.02$ проверить гипотезу о том, что вероятность p успеха в каждом из испытаний равна $p_0 = 0,4$, если альтернативная гипотеза — $H_1 = p > p_0$.

Задача 50. В $N = 500$ независимых испытаниях было получено $k = 289$ успехов. При уровне значимости $\alpha = 0.02$ проверить гипотезу о том, что вероятность p успеха в каждом из испытаний равна $p_0 = 0,6$, если альтернативная гипотеза — $H_1 = p < p_0$.

Задача 51. В $N = 420$ независимых испытаниях было получено $k = 198$ успехов. При уровне значимости $\alpha = 0.01$ проверить гипотезу о том, что вероятность p успеха в каждом из испытаний равна $p_0 = 0,5$, если альтернативная гипотеза — $H_1 = p \neq p_0$.

Задача 52. В $N = 400$ независимых испытаниях было получено $k = 103$ успеха. При уровне значимости $\alpha = 0.01$ проверить гипотезу о том, что вероятность p успеха в каждом из испытаний равна $p_0 = 0,3$, если альтернативная гипотеза — $H_1 = p < p_0$.

В задачах 53 — 58 приведены данные группированной выборки, то есть экспериментальные частоты для заданных интервалов. Требуется:

- построить гистограмму;
- найти точечные оценки математического ожидания и дисперсии;
- построить теоретическую кривую нормального распределения с найденными значениями математического ожидания и дисперсии;
- с помощью критерия χ^2 (хи-квадрат) на уровне значимости α проверить гипотезу о нормальности распределения.

Здесь введены следующие обозначения:

α — уровень значимости;

n_1 — число наблюдений, попавших в интервал $[0, 2]$;

n_2 — число наблюдений, попавших в интервал $[2, 4]$;

n_3 — число наблюдений, попавших в интервал $[4, 6]$;

n_4 — число наблюдений, попавших в интервал $[6, 8]$;

- n_5 — число наблюдений, попавших в интервал $[8, 10]$;
 n_6 — число наблюдений, попавших в интервал $[10, 12]$;
 n_7 — число наблюдений, попавших в интервал $[12, 14]$;
 n_8 — число наблюдений, попавших в интервал $[14, 16]$.
 n_9 — число наблюдений, попавших в интервал $[16, 18]$.

Задача 53.

α	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9
0,01	11	24	85	141	183	162	60	23	13

Задача 54.

α	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9
0,05	14	29	79	151	193	161	64	28	10

Задача 55.

α	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9
0,05	11	24	69	176	202	156	85	22	14

Задача 56.

α	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9
0,02	12	22	71	161	196	170	77	28	13

Задача 57.

α	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9
0,02	12	28	60	121	210	176	80	21	14

Задача 58.

α	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9
0,01	12	22	74	151	205	179	82	24	10

§6. Гипотезы 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 в первую очередь 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 выведенные ранее несмешанные статистики

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

для математического ожидания a , и

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

для дисперсии 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 математического ожидания и дисперсии

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \text{ и } S^2 = \frac{(X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n - 1}$$

тоже являются случайными величинами. Чтобы подчеркнуть этот момент, случайные величины, а также их оценки, будут обозначаться заглавными буквами.

Как было показано ранее, эти оценки несмещенные 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

$$S^2 = \frac{((X_1 - a) - (\bar{X} - a))^2 + \dots + ((X_n - a) - (\bar{X} - a))^2}{n - 1}.$$

Преобразуем дальше:

$$S^2 = \frac{(X_1 - a)^2 + \dots + (X_n - a)^2}{n - 1} - \frac{n}{n - 1}(\bar{X} - a)^2.$$

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

$$\begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = Q \cdot \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

с некоторыми новыми случайными величинами Z_2, \dots, Z_n и некоторой ортогональной матрицей Q .

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

$$Z_1^2 + \cdots + Z_n^2 = Y_1^2 + \cdots + Y_n^2.$$

Но тогда

$$\frac{(n-1)S^2}{\sigma^2} = Z_2^2 + \cdots + Z_n^2,$$

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.

Отметим, что все рассматриваемые здесь задачи играют важную роль в теории измерений. По этой причине мы будем считать, что все выборки есть результаты измерений, то есть извлечены из нормально распределенных генеральных совокупностей.

Также во всех этих задачах мы будем предполагать, что альтернативная гипотеза — “не равно заданному” или “не равны между собой”, поэтому во всех случаях будут строиться двусторонние критические области.

Пример 18.

№	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 $\alpha = 0,05$ гипотезу о том, что дисперсия генеральной совокупности $\sigma^2 = 2$.

Решение. Пусть n — объем выборки (в нашем случае $n = 9$). Если гипотеза верна, то величина

$$\frac{(n-1) \cdot s^2}{\sigma^2},$$

как это следует из только что доказанной теоремы, имеет распределение хи-квадрат с $(n-1)$ степенью свободы.

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 области принятия гипотезы находим по таблице распределения $\chi^2(8)$, они равны 2,18 и 17,53. Следовательно, нет оснований отвергнуть гипотезу.

Пример 19.

№	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 $\alpha = 0,05$ 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

$$\frac{\sqrt{n} \cdot (\bar{x} - a)}{\sigma} = 2,51.$$

Область принятия гипотезы находим по таблице нормального распределения. Это интервал $(-1,96; 1,96)$. Поэтому гипотезу следует отвергнуть.

Пример 20.

№	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 $\alpha = 0,05$ 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 область принятия гипотезы — это интервал $(-2,31; 2,31)$, поэтому нет оснований отвергнуть гипотезу.

Пример 21.

№	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 $\alpha = 0,05$ 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 величина

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(m - 1)s_1^2 + (n - 1)s_2^2}} \cdot \sqrt{\frac{mn(m + n - 2)}{m + n}}$$

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. Область принятия гипотезы — интервал $(-2, 12; 2, 12)$, поэтому гипотеза не отвергается.

Пример 22.

№	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 $\alpha = 0,05$ 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 величины

$$\frac{(m - 1)s_1^2}{\sigma^2} \text{ и } \frac{(n - 1)s_2^2}{\sigma^2}$$

распределены по $\chi^2(m - 1)$ и $\chi^2(n - 1)$ соответственно. 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 область принятия гипотезы — интервал $(0, 23; 4, 43)$, поэтому гипотеза не отвергается.

Задачи

Задача 59. По имеющимся данным проверить при уровне значимости $\alpha = 0,05$ гипотезу о том, что дисперсия генеральной совокупности равна $\sigma^2 = 2,25$.

№	1	2	3	4	5	6	7	8	9
Данные	4,52	4,82	3,33	5,22	2,83	3,52	3,57	5,07	4,34

Задача 60. По имеющимся данным проверить при уровне значимости $\alpha = 0,02$ гипотезу о том, что дисперсия генеральной совокупности равна $\sigma^2 = 1,96$.

№	1	2	3	4	5	6	7	8	9
Данные	5,11	5,32	4,69	5,49	5,00	4,61	4,96	4,70	4,65

Задача 61. По имеющимся данным проверить при уровне значимости $\alpha = 0,05$ гипотезу о том, что среднее генеральной совокупности равна $\mu = 3$. Считать известным, что дисперсия генеральной совокупности равна $\sigma^2 = 3,24$.

№	1	2	3	4	5	6	7	8	9
Данные	4,56	1,14	6,93	5,65	3,57	1,40	2,50	1,52	5,09

Задача 62. По имеющимся данным проверить при уровне значимости $\alpha = 0,02$ гипотезу о том, что среднее генеральной совокупности равна $\mu = 2,8$. Считать известным, что дисперсия генеральной совокупности равна $\sigma^2 = 3$.

№	1	2	3	4	5	6	7	8	9
Данные	5,98	1,68	0,29	5,78	0,33	3,38	1,36	2,09	1,60

Задача 63. По имеющимся данным проверить при уровне значимости $\alpha = 0,05$ гипотезу о том, что среднее генеральной

совокупности равна $\mu = 4,5$. Дисперсию генеральной совокупности считать неизвестной.

№	1	2	3	4	5	6	7	8	9
Данные	7,49	5,20	6,17	5,32	2,80	6,69	4,62	4,67	6,12

Задача 64. По имеющимся данным проверить при уровне значимости $\alpha = 0,02$ гипотезу о том, что среднее генеральной совокупности равна $\mu = 5,4$. Дисперсию генеральной совокупности считать неизвестной.

№	1	2	3	4	5	6	7	8	9
Данные	4,25	6,93	6,43	5,82	6,71	4,65	5,06	5,42	6,73

Задача 65. По имеющимся данным проверить при уровне значимости $\alpha = 0,05$ гипотезу о том, что средние двух генеральных совокупностей, из которых взяты выборки, равны. Дисперсии генеральных совокупностей считать неизвестными, но равными.

№	1	2	3	4	5	6	7	8	9
Данные	6,99	6,68	3,74	1,70	6,46	3,91	1,33	3,21	6,57
Данные	5,52	4,81	3,54	6,81	5,96	2,22	6,93	6,19	1,72

Задача 66. По имеющимся данным проверить при уровне значимости $\alpha = 0,02$ гипотезу о том, что средние двух генеральных совокупностей, из которых взяты выборки, равны. Дисперсии генеральных совокупностей считать неизвестными, но равными.

№	1	2	3	4	5	6	7	8	9
Данные	3,23	3,51	3,94	3,28	2,88	3,34	2,20	2,24	2,58
Данные	3,72	2,87	3,20	2,96	3,04	2,84	4,33	3,72	4,04

Задача 67. По имеющимся данным проверить при уровне значимости $\alpha = 0,05$ гипотезу о том, что дисперсии двух генеральных совокупностей, из которых взяты выборки, равны.

№	1	2	3	4	5	6	7	8	9
Данные	3,64	1,75	2,68	2,12	2,33	4,26	1,14	3,37	3,29
Данные	7,25	5,36	6,39	4,65	6,79	4,05	4,90	4,04	4,82

Задача 68. По имеющимся данным проверить при уровне значимости $\alpha = 0,02$ гипотезу о том, что дисперсии двух генеральных совокупностей, из которых взяты выборки, равны.

№	1	2	3	4	5	6	7	8	9
Данные	1,65	3,73	5,51	2,37	5,87	3,75	4,95	0,72	1,04
Данные	7,85	7,87	8,44	3,67	6,20	7,32	4,20	6,73	4,66

Задача 69. По выборке объемом $n = 14$ получено, что исправленная выборочная дисперсия равна $s^2 = 1.44$. Проверить на уровне значимости $\alpha = 0.05$ гипотезу о том, что дисперсия генеральной совокупности равна 1.

Задача 70. По выборке объемом $n = 9$ получено, что выборочное среднее равно $\bar{x} = 4.45$. Проверить на уровне значимости $\alpha = 0.05$ гипотезу о том, что среднее генеральной совокупности равно 3, если известно, что дисперсия генеральной совокупности равна $\sigma^2 = 3$.

Задача 71. По выборке объемом $n = 9$ получено, что выборочное среднее равно $\bar{x} = 5.11$, а исправленная выборочная дисперсия равна $s^2 = 2.11$. Проверить на уровне значимости $\alpha = 0.05$ гипотезу о том, что среднее генеральной совокупности равно 6. Считать, что дисперсия генеральной совокупности неизвестна.

Задача 72. По двум выборкам объемом $n = 9$ получено, что их выборочные средние равны $\bar{x}_1 = 11.21$ и $\bar{x}_2 = 10.24$, а исправленные выборочные дисперсии равны $s_1^2 = 2.06$ и $s_2^2 = 1.94$. Проверить на уровне значимости $\alpha = 0.05$ гипотезу о том, что средние этих генеральных совокупностей равны. Считать, что дисперсии обеих совокупностей неизвестны, но равны.

Задача 73. По двум выборкам объемом $n = 12$ получено, что их исправленные выборочные дисперсии равны $s_1^2 = 4.32$ и $s_2^2 = 5.78$. Проверить на уровне значимости $\alpha = 0.05$ гипотезу о том, что дисперсии этих генеральных совокупностей равны.

§7. Дисперсионный анализ

Рассмотрим теперь следующий подход к определению степени влияния фактора на результат. Пусть мы имеем результаты измерений некоторого признака (отклика) на разных уровнях фактора, предположительно, влияющего на этот отклик. Мы хотим определить, действительно ли этот фактор влияет, или разницу можно списать на ошибки измерений.

Описываемый здесь метод называется *однофакторный дисперсионный анализ*, поскольку исследуется влияние изменений одного фактора. Возможен и многофакторный дисперсионный анализ, но он выходит за рамки настоящего издания.

Появление дисперсионного анализа связано с именем английского статистика Рональда Фишера и его исследованиями 20-х годов XX века. Фишер отвечал за статистическую обработку данных на сельскохозяйственной станции в Рочестере, близ Лондона. До сих пор в литературе часто употребляется термин “обработка” вместо “уровень фактора”, что выдает сельскохозяйственное происхождение термина. Мы также будем использовать эту терминологию.

Самый простой случай возникает, если число наблюдений одинаково для всех обработок. Тогда мы имеем дело со следующей задачей: дано nk чисел x_{ij} , где:

i — номер наблюдения, $i = 1, \dots, n$;

j — номер уровня фактора, $j = 1, \dots, k$.

Таким образом, у нас есть данные n измерений при каждом из k уровней фактора. Нас интересует такой вопрос: будут ли отличаться средние значения при разных уровнях фактора?

Мы будем рассматривать более общий случай, когда число наблюдений различно при разных уровнях фактора. Тогда по-прежнему x_{ij} — данные, причем

i — номер наблюдения, $i = 1, \dots, n_j$;

j — номер уровня фактора, $j = 1, \dots, k$;

n_j — число наблюдений при j -м уровне фактора.

Прежде всего ответим на такой вопрос: а почему нельзя просто сравнивать средние при разных уровнях фактора попарно? Процедура сравнения средних подробно описана в предыдущем параграфе.

Оказывается, если мы будем сравнивать средние попарно, то такая процедура приведет к слишком частому появлению различий между измерениями. Действительно, пусть доверительная вероятность выбрана равной 95%. Тогда при 5, скажем, уровнях фактора надо провести $(4 \cdot 5)/2 = 10$ попарных проверок, и вероятность несовпадения какой-то пары средних будет равна $1 - 0,95^{10} \approx 0,40$ — и это в том случае, если нулевая гипотеза о равенстве средних справедлива.

Множественные попарные сравнения — это серьезная ловушка, которая может поджидать неподготовленного человека. Вот какой любопытный пример приведен в книге Стентона Гланца “Медико-биологическая статистика” [6].

К чему может привести вольная группировка данных, полученных в безупречно выполненнном рандомизированном исследовании, было убедительно показано Ли и соавторами в статье K. Lee, F. McNeer, F. Starmer, P. Harris, R. Rosati. *Clinical judgement and statistics: lessons from a simulated randomized trial in coronary artery disease. Circulation, 61:508–515, 1980.*

Они воспроизвели достаточно типичное исследование. Взяв истории болезни 1073 больных ишемической болезнью сердца, они случайным образом разделили их на две группы. Одну группу назвали контрольной, а другую экспериментальной (представим себе, что попавшие в нее получали волшебный препарат “рандомизин”). Между группами не было обнаружено значимых различий по таким признакам, как возраст, пол, число пораженных коронарных артерий и т. д. По одному при-

знаку — сократимости левого желудочка — статистически значимое различие наблюдалось. Несомненно, пытливый исследователь не преминул бы связать это различие с использованием “рандомизина”. Однако, увы, по самому важному признаку — выживаемости — различие было статистически не значимым. В этой ситуации исследователь наверняка продолжил бы поиск различий, разделив больных на более мелкие группы. Так и поступил Ли. Больные были разделены (стратифицированы) по двум признакам: числу пораженных коронарных артерий (1, 2 или 3) и сократимости левого желудочка (нормальной или сниженной). В результате получилось 6 подгрупп. Влияние рандомизина на выживаемость изучалось в каждой из этих подгрупп. Но этого мало. Каждая подгруппа была разделена еще на две в зависимости от наличия или отсутствия сердечной недостаточности. В каждой из получившихся 12 подгрупп вновь оценивалась эффективность рандомизина. Упорные усилия были вознаграждены. В одной из подгрупп (больные с поражением 3 коронарных артерий и сниженной сократимостью левого желудочка) рандомизин оказался эффективен: различия выживаемости “леченных” и “нелеченых” были статистически значимыми.

Рандомизин — выдумка. Но многочисленные препараты, эффективность которых была доказана совершенно таким же способом, существуют в действительности. Секрет их “эффективности” очень прост — это множественность сравнений. В исследовании рандомизина было построено 18 пар подгрупп и выполнено 18 сравнений. Чему равна вероятность получить хотя бы один значимый результат в 18 сравнениях, уровень значимости в каждом из которых равен 0,05? Находим:

$$\alpha' = 1 - (1 - \alpha)^k = 1 - (1 - 0,05)^{18} \approx 1 - 0,40 = 0,60.$$

Таким образом, истинная вероятность ошибки I рода оказалась в 12 раз выше той, о которой доложил бы исследователь.

С тем, чтобы избежать множественных попарных сравнений, и связано появление статистической процедуры, которая проводит не попарное сравнение средних, а сравнение сразу нескольких средних. Эта процедура носит название *дисперсионный анализ*. Идея его, а также происхождение его названия, связаны с тем, что дисперсии (разбросы) внутри групп сравниваются с общим разбросом.

Могут проводиться исследования с несколькими факторами, но мы ограничимся одним, то есть опишем однофакторный дисперсионный анализ.

Надо сказать, что его иногда неправильно называют “факторным анализом”. Это совершенно разные вещи. Факторный анализ полностью выходит за рамки данного курса, но с ним можно ознакомиться по литературе или по интернету.

Для наших целей рассмотрим следующую модель:

$$x_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

где

μ — общее среднее;

τ_i — эффект для i -го уровня фактора;

ε_{ij} — случайная ошибка.

Будем считать, что ошибки независимы, имеют нормальное распределение с нулевым средним и дисперсией σ^2 , которая, впрочем, нам неизвестна. Можно считать, что для эффектов уровней фактора (эффектов обработок) выполнено:

$$\tau_1 + \cdots + \tau_k = 0.$$

Нас интересует проверка гипотезы

$$\tau_1 = \cdots = \tau_k = 0$$

при альтернативной гипотезе

$$\tau_i \neq 0 \text{ хотя бы для одного } i.$$

Введем следующие обозначения:

$N = n_1 + \dots + n_k$ — общее число наблюдений;

$\bar{x}_j = \frac{x_{1j} + \dots + x_{n_k j}}{n_k}$ — среднее при j -м уровне фактора;

$$\bar{x} = \frac{\sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij}}{N} — общее среднее.$$

Отметим, кстати, что μ — это истинное среднее генеральной совокупности (нам неизвестное), а \bar{x} — его выборочная оценка.

Рассмотрим сумму квадратов отклонений от выборочного среднего:

$$SS_{\text{общ}} = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2.$$

Представив ее в виде

$$SS_{\text{общ}} = \sum_{j=1}^k \sum_{i=1}^{n_j} ((x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x}))^2$$

и раскрыв скобки, мы увидим, что член с перекрестными произведениями равен нулю, и следовательно, общая сумма квадратов представляется в виде суммы

$$SS_{\text{общ}} = SS_{\text{обр}} + SS_{\text{ош}},$$

где введены следующие обозначения:

$SS_{\text{обр}} = \sum_{j=1}^k \sum_{i=1}^{n_j} n_j (\bar{x}_j - \bar{x})^2$ — сумма квадратов, обусловленная обработками;

$SS_{\text{ош}} = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$ — сумма квадратов, обусловленная ошибками внутри обработок.

Только что доказанная теорема является очень важной в математической статистике. Следовало бы дать ей какое-то название, однако до сих пор такого общепринятого названия нет.

В литературе, однако, подмечена глубокая связь между этой теоремой и известной из механики теоремой Штейнера о сумме моментов. Поэтому для краткости будем называть доказанное выше разбиение суммы квадратов теоремой Штейнера.

Для того, чтобы сформулировать статистический критерий для проверки интересующей нас гипотезы, следует определить, как распределены слагаемые полученного разбиения.

Из теоремы Фишера следует, что сумма квадратов ошибок внутри любой обработки

$$\frac{SS_j}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

имеет распределение χ^2 с $n_j - 1$ степенями свободы. Поэтому, поскольку суммы квадратов в разных обработках независимы, общая нормированная сумма квадратов ошибок внутри обработок $SS_{\text{общ}}/\sigma^2$ имеет распределение χ^2 с $N - k$ степенями свободы. Кроме того, общее среднее и средние по обработкам имеют нормальные распределения:

$$\bar{x}_j \sim N(\mu, \frac{\sigma^2}{n_j}); \quad \bar{x} \sim N(\mu, \frac{\sigma^2}{N}).$$

Для того, чтобы найти распределение $SS_{\text{общ}}$, перейдем теперь к новым переменным, имеющим стандартное нормальное распределение:

$$u_j = \frac{\bar{x}_j - \mu}{\sigma} \cdot \sqrt{n_j}, \quad j = 1, \dots, k, \quad w = \frac{\bar{x} - \mu}{\sigma} \cdot \sqrt{N}.$$

Случайные величины u_j , разумеется, независимы, поскольку относятся к разным обработкам. Общее среднее μ нам, конечно, неизвестно, но вычитание его из каждой переменной не повлияет на сумму квадратов отклонений, потому что

$$(\bar{x}_j - \bar{x})^2 = ((\bar{x}_j - \mu) - (\bar{x} - \mu))^2.$$

Из формулы

$$\bar{x} = \frac{\bar{x}_1 \cdot n_1 + \cdots + \bar{x}_k \cdot n_k}{N}$$

сразу следует, что

$$w = \frac{u_1\sqrt{n_1} + \cdots + u_k\sqrt{n_k}}{N}.$$

Поскольку сумма квадратов коэффициентов в правой части равна 1, существует ортогональное преобразование

$$\begin{pmatrix} w \\ w_2 \\ \vdots \\ w_k \end{pmatrix} = Q \cdot \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{pmatrix}$$

с некоторыми новыми стандартными нормальными независимыми случайными величинами w_2, \dots, w_n и некоторой ортогональной матрицей Q . Ортогональное преобразование сохраняет длины, поэтому

$$u_1^2 + u_2^2 + \cdots + u_k^2 = w^2 + w_2^2 + \cdots + w_k^2.$$

Преобразуем теперь выражение для $SS_{\text{обр}}$, приведя его к новым переменным:

$$\begin{aligned} SS_{\text{обр}} &= \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2 = \\ &= \sigma^2 \sum_{j=1}^k \left(\frac{\sqrt{n_j}}{\sigma} (\bar{x}_j - \mu) - \frac{\sqrt{n_j} \cdot \sqrt{N}}{\sigma \cdot \sqrt{N}} (\bar{x} - \mu) \right)^2 = \\ &= \sigma^2 \sum_{j=1}^k (u_j - \frac{\sqrt{n_j}}{\sqrt{N}} w)^2 = \sigma^2 \left(\sum_{j=1}^k u_j^2 - w^2 \right). \end{aligned}$$

Поэтому для $SS_{\text{обр}}$ получим

$$\frac{SS_{\text{обр}}}{\sigma^2} = w_2^2 + \cdots + w_k^2$$

для вновь введенных переменных w_2, \dots, w_n . Следовательно $SS_{\text{обр}}/\sigma^2$ имеет распределение χ^2 с $k - 1$ степенями свободы.

По теореме Фишера средние по обработкам \bar{x}_j и суммы квадратов ошибок внутри обработок SS_j независимы. Следовательно, независимы и $SS_{\text{обр}}$ и $SS_{\text{ош}}$, потому что первая из них зависит только от средних, а вторая — только от сумм квадратов.

Отсюда следует, что отношение

$$F = \frac{SS_{\text{обр}}/(k - 1)}{SS_{\text{ош}}/(N - k)}$$

имеет распределение Фишера с $(k - 1, N - k)$ степенями свободы.

Для проверки интересующей нас гипотезы следует применять односторонний критерий, поскольку слишком малое отношение $SS_{\text{обр}}$ и $SS_{\text{ош}}$ означает, что разница между обработками мала по сравнению с разбросами внутри обработок, а такая ситуация вполне согласуется с нулевой гипотезой.

Для расчетов в том случае, когда число наблюдений на каждом уровне фактора одинаково, в литературе, например, в книге Д. Монтгомери “Планирование эксперимента и анализ данных” [9], приведены удобные расчетные формулы:

$$SS_{\text{общ}} = \sum_{j=1}^k \sum_{i=1}^n x_{ij}^2 - N\bar{x}^2;$$

$$SS_{\text{обр}} = n \sum_{j=1}^k \bar{x}_j^2 - N\bar{x}^2.$$

Сумма квадратов, обусловленная ошибками, находится вычитанием:

$$SS_{\text{ош}} = SS_{\text{общ}} - SS_{\text{обр}}.$$

В различных электронных таблицах существуют стандартные функции, вычисляющие сумму квадратов отклонений от

среднего. В Microsoft Excel это КВАДРОТКЛ, а в Libre Office – DEVSQ.

Пример 23. Пусть в результате пяти измерений при различных уровнях фактора были получены следующие результаты. Эти данные приведены в таблице. Проверить на уровне значимости $\alpha = 0,05$ гипотезу о том, что фактор не влияет на результат.

№	1	2	3	4	5	6	7	8
ряд 1	7,35	8,15	9,44	5,03	7,28	6,11	5,94	7,02
ряд 2	5,19	8,36	10,87	5,40	11,19	6,53	7,16	6,51
ряд 3	9,54	5,57	11,44	9,91	9,22	7,07	7,56	9,96
ряд 4	5,55	5,44	8,70	5,31	6,30	6,35	5,03	7,66
ряд 5	8,46	6,11	6,72	8,40	9,93	7,78	5,17	7,28

Решение. Вычисляем средние для каждого уровня фактора: $\bar{x}_1 = 7,04$, $\bar{x}_2 = 7,65$, $\bar{x}_3 = 8,78$, $\bar{x}_4 = 6,29$, $\bar{x}_5 = 7,48$. Общее среднее $\bar{x} = 7,45$. Для сумм квадратов получим $SS_{\text{обр}} = 26,64$, $SS_{\text{ош}} = 102,98$. Число степеней свободы: числителя – 4, знаменателя – 35. По приведенной выше формуле вычисляем значение критерия Фишера: $F = 2,26$. По таблицам находим значение 95-процентного квантиля распределения Фишера: $F_{\text{кр}} = 2,64$. Поскольку вычисленное значение меньше критического, нет основания отвергнуть гипотезу о том, что фактор на результат не влияет.

Задачи

Задача 74. В случае, когда уровней фактора всего два, то есть попарное сравнение всего одно, можно пользоваться как дисперсионным анализом, так и сравнением средних при неизвестной дисперсии из предыдущего параграфа. Какой из способов лучше?

Указание. Воспользоваться результатом решения задачи 36.

Задача 75. Для каждого из $m = 5$ уровней фактора получено $n = 7$ результатов измерений. При этом $SS_{\text{общ}} = 75,3$, $SS_{\text{ош}} = 21,1$. Проверить на уровне значимости $\alpha = 0.05$ гипотезу о том, что фактор не влияет на результат.

Задача 76. Для каждого из $m = 3$ уровней фактора получено $n = 11$ результатов измерений. При этом $SS_{\text{общ}} = 114,4$, $SS_{\text{ош}} = 36,6$. Проверить на уровне значимости $\alpha = 0.05$ гипотезу о том, что фактор не влияет на результат.

В задачах 77 — 83 для каждого из четырех значений уровня фактора получено восемь значений случайной величины. Требуется: с помощью однофакторного дисперсионного анализа при уровне значимости α проверить гипотезу о равенстве средних, то есть о том, что значения случайной величины не зависят от уровня фактора.

Задача 77. $\alpha = 0,01$.

Уровни	Данные наблюдений							
1	6,51	6,58	6,49	6,74	6,72	6,78	6,72	5,98
2	5,23	4,84	4,22	4,22	4,34	4,63	5,36	4,37
3	4,72	4,31	4,45	4,07	4,43	3,81	4,12	4,97
4	5,14	5,60	5,83	6,50	5,28	5,55	6,31	6,01

Задача 78. $\alpha = 0,02$.

Уровни	Данные наблюдений							
1	3,59	4,45	3,87	4,20	3,72	4,36	3,77	3,92
2	3,89	3,62	4,23	4,47	4,26	4,36	3,96	3,74
3	3,75	4,15	4,46	4,12	4,17	4,48	3,88	3,67
4	3,70	4,15	4,24	4,24	4,18	4,49	3,74	4,09

Задача 79. $\alpha = 0,02$.

Уровни	Данные наблюдений							
1	3,32	2,24	3,11	3,48	2,91	3,32	2,49	3,89
2	4,00	3,04	3,38	3,95	4,76	3,75	3,49	3,53
3	5,29	4,38	3,63	4,13	5,19	4,08	5,23	4,60
4	4,46	4,94	3,39	3,45	4,94	3,46	4,55	3,88

Задача 80. $\alpha = 0,05$.

Уровни	Данные наблюдений							
1	7,11	8,21	7,34	9,42	6,24	5,93	6,39	7,83
2	6,12	6,55	9,11	7,44	7,38	7,14	8,02	9,15
3	6,99	5,87	7,13	7,35	5,22	7,39	5,48	6,49
4	8,16	8,83	6,14	6,51	7,09	6,94	7,17	7,37

Задача 81. $\alpha = 0,05$.

Уровни	Данные наблюдений							
1	8,04	8,34	7,86	8,44	7,68	5,65	6,51	7,03
2	6,17	6,87	8,16	6,29	7,89	7,75	8,42	8,05
3	7,44	5,79	7,14	7,05	5,72	7,19	5,99	6,04
4	8,00	8,19	6,08	6,15	7,99	6,04	7,99	7,21

Задача 82. $\alpha = 0,05$.

Уровни	Данные наблюдений							
1	8,51	8,82	7,62	8,39	7,14	5,95	6,22	7,32
2	6,84	6,71	8,96	6,11	7,39	7,77	8,43	8,91
3	7,79	5,55	7,18	7,63	5,69	7,95	5,71	6,47
4	8,29	8,71	6,42	6,39	7,140	6,73	7, 12	7,74

Задача 83. $\alpha = 0,05$.

Уровни	Данные наблюдений							
1	7,24	6,32	7,98	6,21	8,37	8,19	8,48	8,14
2	4,57	5,95	6,47	4,65	5,28	4,91	5,88	7,11
3	5,81	5,34	5,32	6,35	5,49	5,51	5,65	5,95
4	6,42	6,68	4,52	6,93	6,38	7,31	5,51	6,83

§8. 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, к примеру, какова зависимость между ростом и весом человека. Или, скажем, как связаны плотность населения в районе и засоленность почв этого района.

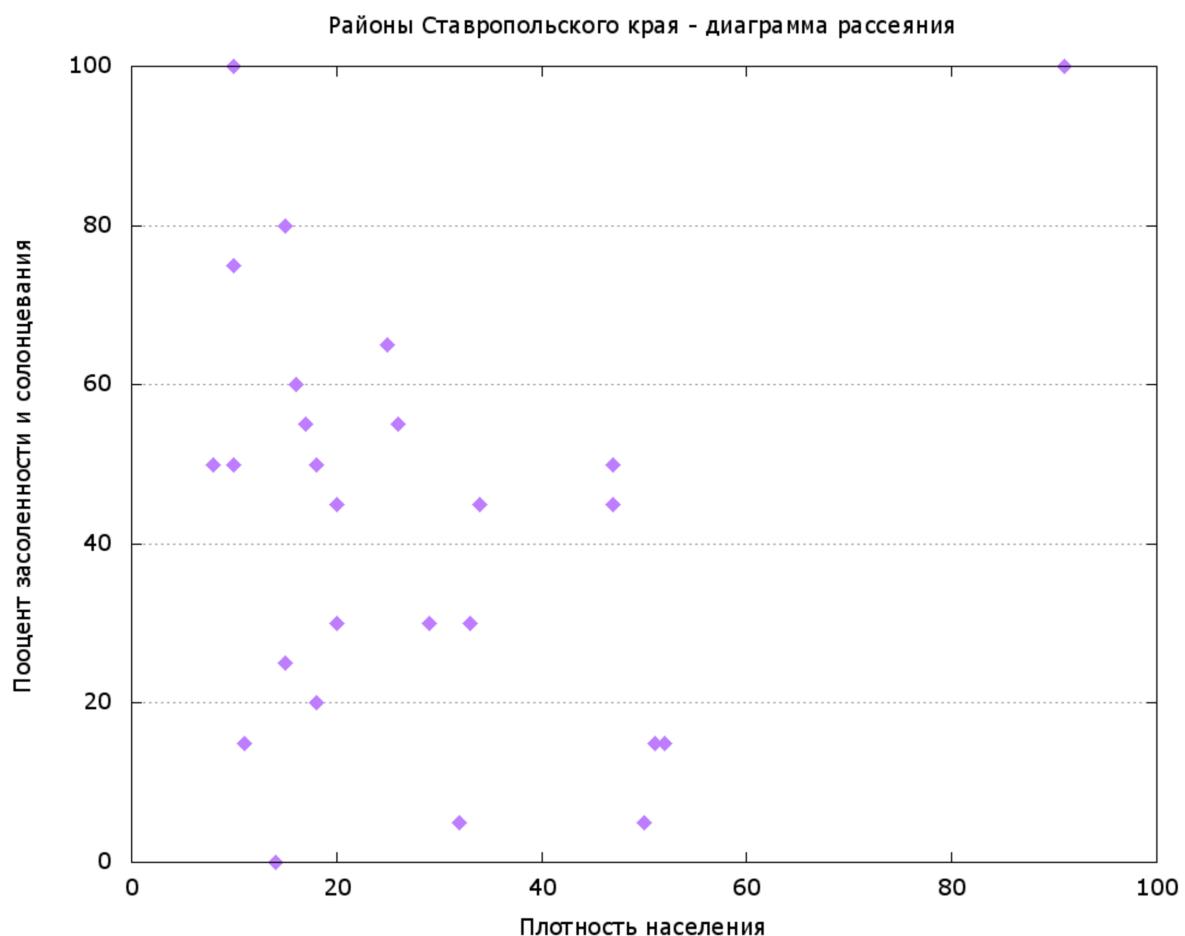


Рис. 19. Диаграмма рассеяния — связь засоленности почв и плотности населения районов Ставропольского края

Если каждый элемент выборки характеризуется двумя числами, то нашу выборку мы можем представить в виде двумерной 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.

Эти определения вводились для случайных величин, то есть для генеральных совокупностей. 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) , где $i = 1, \dots, n$, то коэффициент выборочной корреляции определяется по формуле

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y}.$$

В этой формуле использованы стандартные обозначения:

\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 тоже 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.

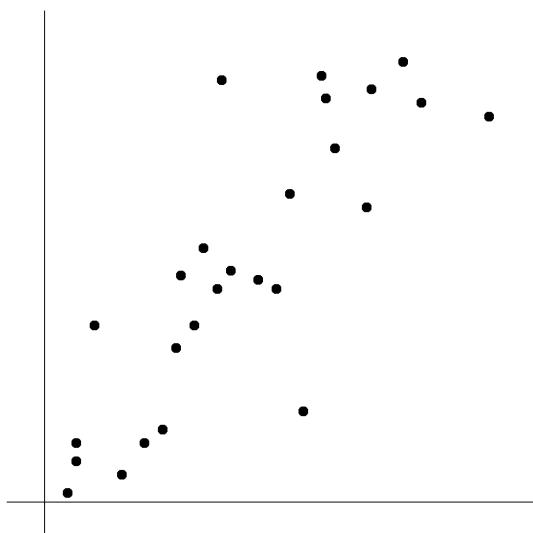


Рис. 20. Сильная положительная корреляция.
Коэффициент корреляции равен 0,82

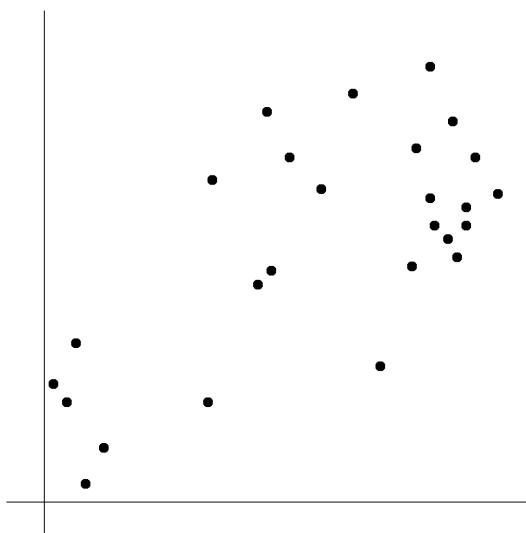


Рис. 21. Средняя положительная корреляция.
Коэффициент корреляции равен 0,68

Графические иллюстрации, показывающие, что означает то или иное значение коэффициента корреляции, приведены на рисунках 20 — 26. Важно также отметить, что коэффициент корреляции показывает наличие именно линейной связи между переменными, а не какой-либо другой функциональной зависимости. Это иллюстрирует рисунок 27.

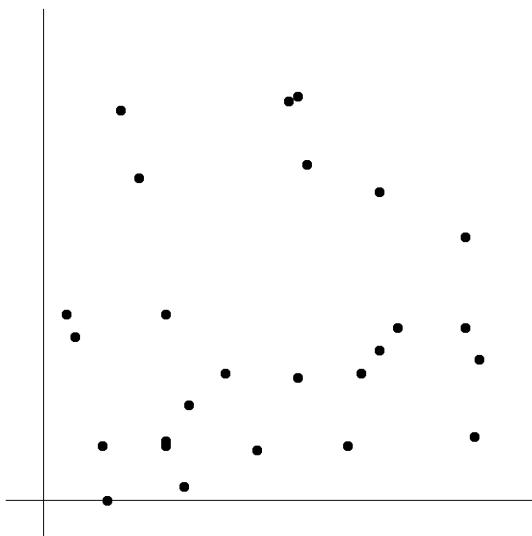
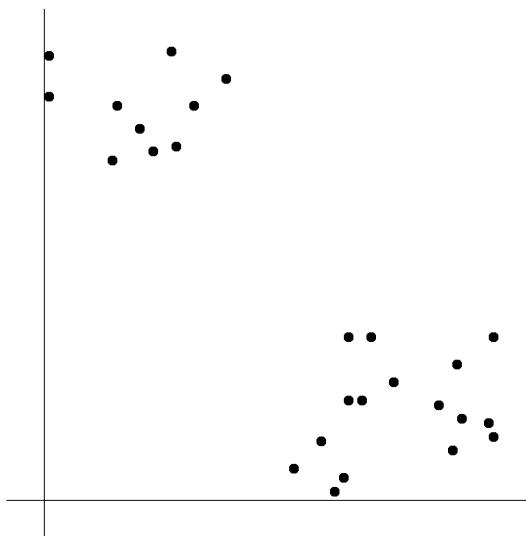
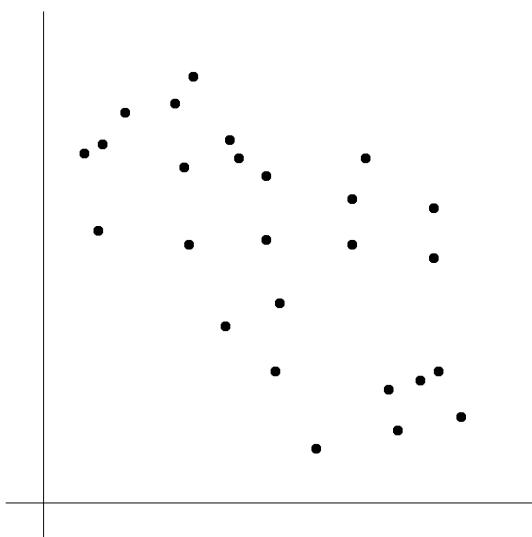


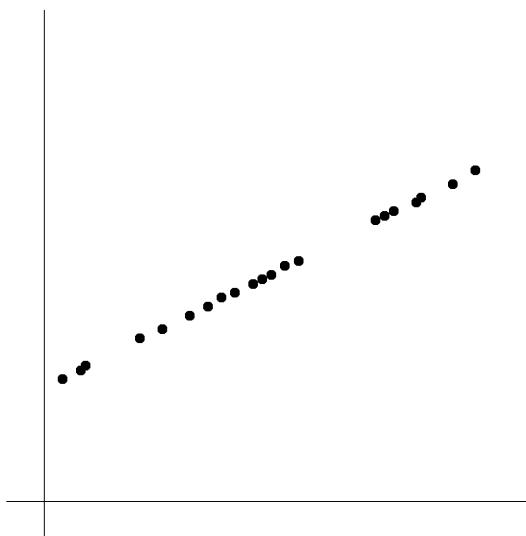
Рис. 22. Слабая корреляция.
Коэффициент корреляции
равен 0,11



**Рис. 24. Сильная
отрицательная корреляция.**
Коэффициент корреляции
равен -0,84



**Рис. 23. Средняя
отрицательная корреляция.**
Коэффициент корреляции
равен -0,62



**Рис. 25. 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 сгруппированности 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.

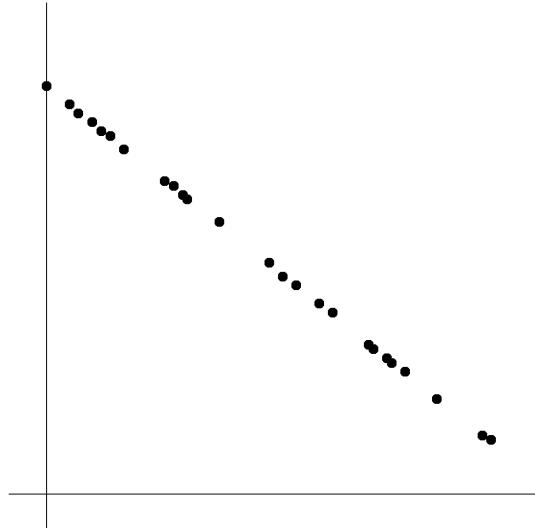


Рис. 26. Отрицательная линейная зависимость. Коэффициент корреляции равен -1

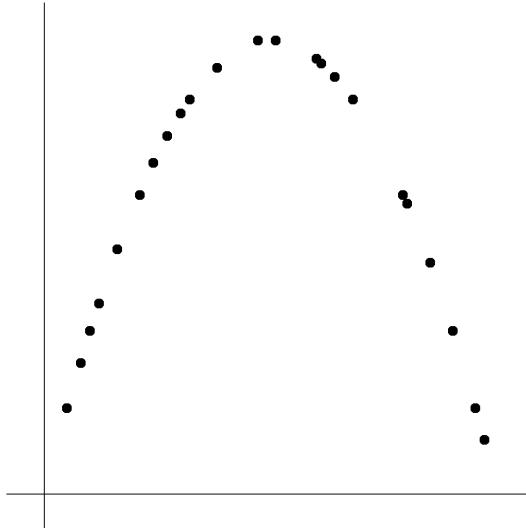


Рис. 27. Нелинейная функциональная зависимость. Коэффициент корреляции равен 0

Проверка гипотез о значении коэффициента корреляции

Рассмотрим задачу о проверке значимости коэффициента корреляции. Пусть на самом деле корреляции нет. Какие значения может принимать выборочный коэффициент 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

$$r = \frac{w_2}{\sqrt{v_2^2 + \cdots + v_n^2}} = \frac{w_2}{\sqrt{w_2^2 + \cdots + w_n^2}}.$$

Обозначим для удобства

$$w = w_2, \quad \zeta^2 = w_3^2 + \cdots + w_n^2,$$

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 24. 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 аппроксирующую наши данные?

Решение этой задачи, то есть, 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.

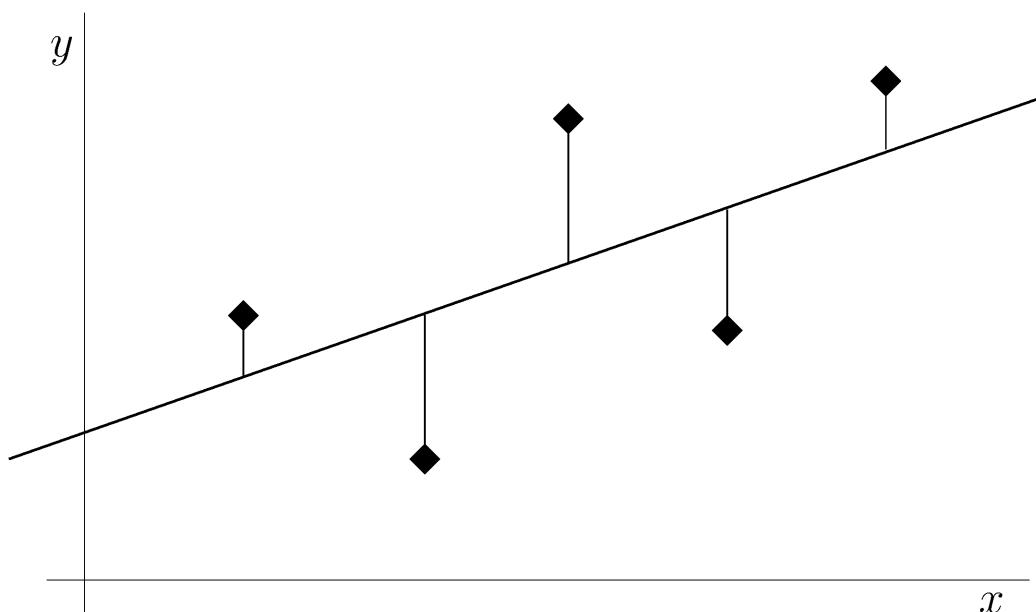


Рис. 28. Линейная регрессия

Назовем разность $(y_i - kx_i - b)$ между измеренным и предсказанным значениями *невязкой*. Потребуем теперь, следуя Гауссу, чтобы сумма квадратов невязок была минимальной.

На рисунке 28 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 играют и эстетические соображения, например, простота и изящество получаемых формул.

Во вторых, мы минимизируем квадраты разностей между искажениями, а не квадраты расстояний до прямой. Это означает, по сути, что иксы нам известны точно.

Вывод нормальных уравнений

Как 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 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 25. Для данных примера 24 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 уравнение 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 формулой $\text{cov}(x, y) = r\sigma_x\sigma_y$, которая сразу следует из определения коэффициента корреляции. 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 — определять, обуславливать).

Еще более важную роль этот коэффициент играет в множественной регрессии.

Множественная регрессия

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 регрессионный анализ”.

В общественных и естественных науках процедуры множественной регрессии чрезвычайно широко используются в исследованиях. В общем, множественная регрессия позволяет 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 пред-

сказывают 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, соответствующего 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:

$$y = b_0 + b_1x_1 + b_2x_2.$$

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

$$R^2 = \frac{SS_{\text{рег}}}{SS}.$$

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.

Проверим эту гипотезу. В качестве альтернативной гипотезы рассмотрим гипотезу о том, что хотя бы один из регрессионных коэффициентов отличен от нуля.

В случае справедливости нулевой гипотезы можно доказать, что случайная величина $SS_{\text{пер}}/\sigma^2$ 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

$$F = \frac{SS_{\text{пер}}/k}{SS_{\text{ост}}/(n-k-1)}$$

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 =FPACПОБР(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 признакомстве 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 “себлазн” 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. В книге Ю. Н. Благовещенского “Тайны корреляционных связей в статистике” [1] приведен такой пример. Оказывается, если взять статистические данные по числу студентов и количеству экономических преступлений во всех 89 регионах Российской Федерации, то между этими переменными можно обнаружить сильную положительную связь — коэффициент корреляции равен 0,9.

Объяснение этой корреляции, конечно, состоит в том, что обе переменные зависят от населения региона, в более населенном регионе больше как число студентов, так и количество экономических преступлений.

И еще один пример. Оказывается, 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.

Задачи

Задача 84. 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 совокупности.

Задача 85. Написать систему нормальных уравнений для определения коэффициентов квадратичной зависимости вида

$$y = ax^2 + bx + c.$$

Задача 86. 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 .

В задачах 87 — 90 дана двумерная выборка $(x_i, y_i), i = 1, \dots, n$.
а). Найти выборочный коэффициент корреляции между x и y .
б). Построить линию регрессии y на x .
в). При уровне значимости α проверить гипотезу о значимости коэффициента корреляции.

Задача 87. $\alpha = 0,01$.

x_i	1,37	2,00	3,01	3,85	5,07	6,16	6,87	8,39
y_i	27,28	28,13	29,18	29,87	30,91	32,50	33,21	33,57

Задача 88. $\alpha = 0,02$.

x_i	0,78	1,85	2,57	3,97	4,63	5,85	7,05	7,71
y_i	23,60	20,82	17,53	15,10	12,40	8,64	5,83	3,49

Задача 89. $\alpha = 0,02$.

x_i	1,04	2,35	3,44	4,47	5,25	5,56	6,88	7,54
y_i	26,68	26,46	24,95	23,96	22,70	21,62	20,97	20,01

Задача 90. $\alpha = 0,05$.

x_i	0,78	1,96	2,73	3,86	4,79	5,89	6,52	7,59
y_i	28,59	27,78	27,05	26,08	24,61	24,44	23,08	21,66

Таблицы

В данном приложении приведены некоторые таблицы значений квантилей различных распределений. Напомним, что α -квантиль распределения с плотностью вероятности $\rho(x)$ — это решение z уравнения

$$\int_{-\infty}^z \rho(x)dx = \alpha.$$

Таблицы далеко не полны. Тем не менее их можно использовать для решения расчетных задач. Однако их основное предназначение — быть своего рода ориентиром при использовании электронных таблиц.

Квантили нормального распределения

0,5	0,51	0,52	0,53	0,54	0,55	0,56	0,57	0,58	0,59
0	0,025	0,050	0,075	0,100	0,126	0,151	0,176	0,202	0,228
0,6	0,61	0,62	0,63	0,64	0,65	0,66	0,67	0,68	0,69
0,253	0,279	0,305	0,332	0,358	0,385	0,412	0,440	0,468	0,496
0,7	0,71	0,72	0,73	0,74	0,75	0,76	0,77	0,78	0,79
0,524	0,553	0,583	0,613	0,643	0,674	0,706	0,739	0,772	0,806
0,8	0,81	0,82	0,83	0,84	0,85	0,86	0,87	0,88	0,89
0,842	0,878	0,915	0,954	0,994	1,036	1,080	1,126	1,175	1,227
0,9	0,91	0,92	0,93	0,94	0,95	0,96	0,97	0,98	0,99
1,282	1,341	1,405	1,476	1,555	1,645	1,751	1,881	2,054	2,326
0,955	0,965	0,975	0,985	0,994	0,995	0,996	0,997	0,998	0,999
1,695	1,812	1,960	2,170	2,512	2,576	2,652	2,748	2,878	3,090

Квантили распределения хи-квадрат

n — число степеней свободы.

n	0,005	0,01	0,95	0,98	0,99	0,995
1	0,000	0,000	3,841	5,412	6,635	7,879
2	0,010	0,020	5,991	7,824	9,210	10,597
3	0,072	0,115	7,815	9,837	11,345	12,838
4	0,207	0,297	9,488	11,668	13,277	14,860
5	0,412	0,554	11,070	13,388	15,086	16,750
6	0,676	0,872	12,592	15,033	16,812	18,548
7	0,989	1,239	14,067	16,622	18,475	20,278
8	1,344	1,646	15,507	18,168	20,090	21,955
9	1,735	2,088	16,919	19,679	21,666	23,589
10	2,156	2,558	18,307	21,161	23,209	25,188
11	2,603	3,053	19,675	22,618	24,725	26,757
12	3,074	3,571	21,026	24,054	26,217	28,300
13	3,565	4,107	22,362	25,472	27,688	29,819
14	4,075	4,660	23,685	26,873	29,141	31,319
15	4,601	5,229	24,996	28,259	30,578	32,801
16	5,142	5,812	26,296	29,633	32,000	34,267
17	5,697	6,408	27,587	30,995	33,409	35,718
18	6,265	7,015	28,869	32,346	34,805	37,156
19	6,844	7,633	30,144	33,687	36,191	38,582
20	7,434	8,260	31,410	35,020	37,566	39,997
21	8,034	8,897	32,671	36,343	38,932	41,401
22	8,643	9,542	33,924	37,659	40,289	42,796
23	9,260	10,196	35,172	38,968	41,638	44,181
24	9,886	10,856	36,415	40,270	42,980	45,559
25	10,520	11,524	37,652	41,566	44,314	46,928

Квантили распределения Стьюдента

n — число степеней свободы.

n	0,90	0,95	0,975	0,98	0,99	0,995
1	3,078	6,314	12,706	15,895	31,821	63,657
2	1,886	2,920	4,303	4,849	6,965	9,925
3	1,638	2,353	3,182	3,482	4,541	5,841
4	1,533	2,132	2,776	2,999	3,747	4,604
5	1,476	2,015	2,571	2,757	3,365	4,032
6	1,440	1,943	2,447	2,612	3,143	3,707
7	1,415	1,895	2,365	2,517	2,998	3,499
8	1,397	1,860	2,306	2,449	2,896	3,355
9	1,383	1,833	2,262	2,398	2,821	3,250
10	1,372	1,812	2,228	2,359	2,764	3,169
11	1,363	1,796	2,201	2,328	2,718	3,106
12	1,356	1,782	2,179	2,303	2,681	3,055
13	1,350	1,771	2,160	2,282	2,650	3,012
14	1,345	1,761	2,145	2,264	2,624	2,977
15	1,341	1,753	2,131	2,249	2,602	2,947
16	1,337	1,746	2,120	2,235	2,583	2,921
17	1,333	1,740	2,110	2,224	2,567	2,898
18	1,330	1,734	2,101	2,214	2,552	2,878
19	1,328	1,729	2,093	2,205	2,539	2,861
20	1,325	1,725	2,086	2,197	2,528	2,845
21	1,323	1,721	2,080	2,189	2,518	2,831
22	1,321	1,717	2,074	2,183	2,508	2,819
23	1,319	1,714	2,069	2,177	2,500	2,807
24	1,318	1,711	2,064	2,172	2,492	2,797
25	1,316	1,708	2,060	2,167	2,485	2,787

Квантили распределения Фишера с 8 степенями свободы знаменателя

n — число степеней свободы числителя.

n	0,005	0,01	0,95	0,98	0,99	0,995
1	0,000	0,000	5,318	8,389	11,259	14,688
2	0,005	0,010	4,459	6,637	8,649	11,042
3	0,023	0,036	4,066	5,901	7,591	9,596
4	0,047	0,068	3,838	5,489	7,006	8,805
5	0,072	0,097	3,687	5,223	6,632	8,302
6	0,095	0,123	3,581	5,036	6,371	7,952
7	0,115	0,146	3,500	4,897	6,178	7,694
8	0,133	0,166	3,438	4,790	6,029	7,496
9	0,149	0,183	3,388	4,705	5,911	7,339
10	0,164	0,198	3,347	4,635	5,814	7,211
11	0,176	0,211	3,313	4,577	5,734	7,104
12	0,187	0,222	3,284	4,528	5,667	7,015
13	0,197	0,232	3,259	4,486	5,609	6,938
14	0,206	0,242	3,237	4,449	5,559	6,872
15	0,214	0,250	3,218	4,417	5,515	6,814
16	0,221	0,257	3,202	4,389	5,477	6,763
17	0,228	0,264	3,187	4,364	5,442	6,718
18	0,234	0,270	3,173	4,342	5,412	6,678
19	0,239	0,275	3,161	4,322	5,384	6,641
20	0,245	0,281	3,150	4,304	5,359	6,608
21	0,249	0,285	3,140	4,287	5,336	6,578
22	0,254	0,290	3,131	4,272	5,316	6,551
23	0,258	0,294	3,123	4,258	5,297	6,526
24	0,261	0,297	3,115	4,245	5,279	6,503
25	0,265	0,301	3,108	4,233	5,263	6,482

Квантили распределения Фишера с 28 степенями свободы знаменателя

n — число степеней свободы числителя.

n	0,005	0,01	0,95	0,98	0,99	0,995
1	0,000	0,000	4,196	6,087	7,636	9,284
2	0,005	0,010	3,340	4,513	5,453	6,440
3	0,024	0,038	2,947	3,851	4,568	5,317
4	0,050	0,072	2,714	3,475	4,074	4,698
5	0,079	0,106	2,558	3,228	3,754	4,300
6	0,106	0,138	2,445	3,052	3,528	4,020
7	0,132	0,166	2,359	2,920	3,358	3,811
8	0,156	0,192	2,291	2,817	3,226	3,649
9	0,177	0,214	2,236	2,733	3,120	3,519
10	0,196	0,234	2,190	2,664	3,032	3,412
11	0,214	0,252	2,151	2,606	2,959	3,322
12	0,229	0,269	2,118	2,556	2,896	3,246
13	0,244	0,283	2,089	2,513	2,842	3,180
14	0,257	0,297	2,064	2,475	2,795	3,123
15	0,269	0,309	2,041	2,442	2,753	3,073
16	0,280	0,320	2,021	2,413	2,716	3,028
17	0,291	0,330	2,003	2,386	2,683	2,988
18	0,300	0,340	1,987	2,363	2,653	2,952
19	0,309	0,349	1,972	2,341	2,626	2,919
20	0,317	0,357	1,959	2,321	2,602	2,890
21	0,325	0,365	1,946	2,303	2,579	2,863
22	0,332	0,372	1,935	2,287	2,559	2,838
23	0,339	0,378	1,924	2,272	2,540	2,815
24	0,345	0,384	1,915	2,258	2,522	2,794
25	0,351	0,390	1,906	2,244	2,506	2,775

Литература

1. Благовещенский Ю. Н. Тайны корреляционных связей в статистике. — М.: Научная книга, ИНФРА-М, 2009. — 158 с.
2. Ван дер Варден Б. Л. Математическая статистика. — М.: Издательство иностранной литературы, 1960. — 434 с.
3. Гамов Г., Стерн М. Занимательные задачи. — М.: Еditorial УРСС, 2003. — 144 с.
4. Гарднер М. А ну-ка, догадайся. — М.: Мир, 1984. — 213 с.
5. Гмурман В. Е. Руководство к решению задач по теории вероятностей и математической статистике. — М.: Высшая школа, 2004 — 407 с.
6. Гланц С. Медико-биологическая статистика. — М.: Практика, 1998 — 459 с.
7. Ивченко Г. И., Медведев Ю. И. Введение в математическую статистику. — М.: Издательство ЛКИ, 2010. — 600 с.
8. Лагутин М. Б. Наглядная математическая статистика. — М.: БИНОМ. Лаборатория знаний, 2007. — 472 с.
9. Монтгомери Д. К. Планирование эксперимента и анализ данных. — Л.: Судостроение, 1984. — 384 с.
10. Яу Н. Искусство визуализации в бизнесе. М.: Манн, Иванов и Фарбер, 2013. — 352 с.