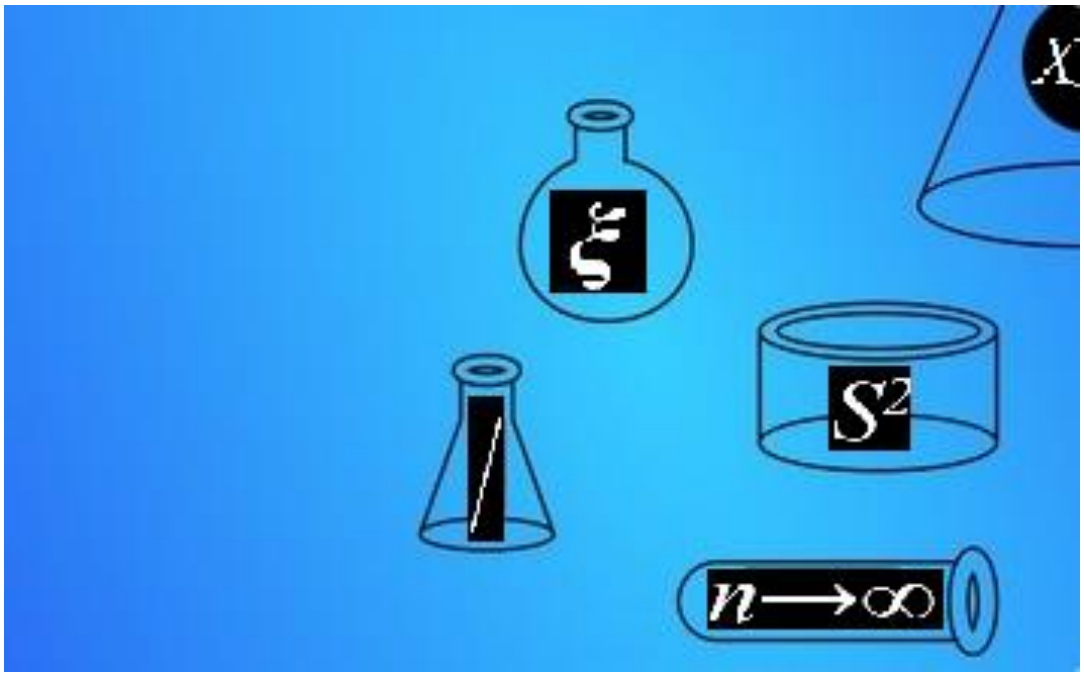
В. В. КОНКИНА, А. Д. ОБУХОВ, Ю. В. ЛИТОВКА

МЕТОДЫ ОБРАБОТКИ СТАТИСТИЧЕСКИХ ДАННЫХ



Тамбов
Издательский центр ФГБОУ ВО «ТГТУ»
2026

Министерство науки и высшего образования Российской Федерации

**Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Тамбовский государственный технический университет»**

В. В. КОНКИНА, А. Д. ОБУХОВ, Ю. В. ЛИТОВКА

МЕТОДЫ ОБРАБОТКИ СТАТИСТИЧЕСКИХ ДАННЫХ

Утверждено Ученым советом
ФГБОУ ВО «Тамбовский государственный технический университет»
в качестве учебного пособия для студентов направления подготовки
09.03.03 «Прикладная информатика», изучающих дисциплину
«Технологии обработки и хранения статистических данных»,
а также студентов направления подготовки 09.04.01 «Информатика
и вычислительная техника», изучающих дисциплину
«Введение в большие данные и анализ информации»,
очной и заочной форм обучения

Учебное электронное издание

**Тамбов
Издательский центр ФГБОУ ВО «ТГТУ»
2026**

УДК 004.04
ББК 16.23
К64

Рецензенты:

Доктор технических наук, доцент, доцент кафедры
математического моделирования и информационных технологий
ФГБОУ ВО «ТГУ им. Г. Р. Державина»
Д. С. Соловьев

Доктор технических наук, профессор, профессор кафедры
«Информационные процессы и управление» ФГБОУ ВО «ТГТУ»
В. А. Погонин

Конкина, В. В.

К64 Методы обработки статистических данных [Электронный ресурс] : учебное пособие / В. В. Конкина, А. Д. Обухов, Ю. В. Литовка. – Тамбов : Издательский центр ФГБОУ ВО «ТГТУ», 2026. – 1 электрон. опт. диск (CD-ROM). – ПК не ниже класса Pentium IV ; RAM 512 Mb ; необходимое место на HDD 2,18 Mb ; Windows 7/8/10/11 ; дисковод CD-ROM ; мышь. – Загл. с экрана.

ISBN 978-5-8265-2997-3

Рассмотрены принятие решений на основе статистических данных, выбор способов оценивания данных, которые минимизируют ошибку. Описаны доверительный интервал и бутстреп, математическая база A/B-тестирования и метод главных компонент. Показано исследование реальных задач, когда распределение случайной величины неизвестно, а также оценка параметров распределения и выбор лучшей оценки, построение доверительных интервалов, подтверждение и опровержение гипотезы, корректное сжатие данных. Показано, как проводить множественную проверку гипотез, находить матрицу ковариации. Приведен разбор метода главных компонент.

Предназначено для студентов направления подготовки 09.03.03 «Прикладная информатика», изучающих дисциплину «Технологии обработки и хранения статистических данных», а также студентов направления подготовки 09.04.01 «Информатика и вычислительная техника», изучающих дисциплину «Введение в большие данные и анализ информации», очной и заочной форм обучения.

УДК 004.04
ББК 16.23

ISBN 978-5-8265-2997-3

«В большинстве случаев информация более ценна,
чем аппаратное обеспечение, которое ее обрабатывает»

Грейс Хоппер[*]

ВВЕДЕНИЕ

Статистику все видят как-то по-своему. Для многих все сводится к A/B-тестам и среднему, для других – к длинным формулам и теоретическим рассуждениям. Статистика помогает людям с разным набором знаний обмениваться информацией, а также извлекать из данных выводы о реальном мире и проверять гипотезы. Большинство методов машинного обучения и анализа представляют собой статистические модели и описываются статистическим языком.

С помощью статистики выявляют скрытые взаимосвязи между данными. Статистика не отвечает на вопрос, какая модель в точности описывает реальность. Например, даже если вы миллион раз подбросите монетку, то не сможете сделать стопроцентный вывод, что орел и решка выпадают с равной вероятностью. Но вот что можно сделать с помощью статистики: научиться выбирать такие способы оценивания данных, чтобы ошибка была минимальной.

Статистические методы дают выводы в виде числовых оценок и доверительных интервалов. Но иногда не подходит ни числовая оценка, ни доверительный интервал. Вывод может быть связан с выбором одной

[*] Грейс Хоппер (9 декабря 1906 – 1 января 1992) – американская ученая и коммодор флота США. Будучи первооткрывательницей в своей области, она была одной из первых, кто писал программы для гарвардского компьютера Марк I. Она разработала первый компилятор для компьютерного языка программирования, развила концепцию машинно-независимых языков программирования, что привело к созданию COBOL, одного из первых высокоуровневых языков программирования. Ей приписывается популяризация термина debugging для устранения сбоев в работе компьютера.

из двух конфликтующих гипотез. Базовые статистические методы применяют в ситуациях, когда работа сводится к нормальным распределениям. Но так бывает не всегда. В этом случае применяют множественную проверку гипотез и метод оценки параметров – бутстреп. Если данных очень много и связи между ними неочевидны, используется метод главных компонент, сочетающий в себе статистику, объединенную с линейной алгеброй. Суть подхода в том, чтобы выделить самые значимые данные и исследовать их.

Случайные величины могут обозначаться заглавными латинскими буквами либо прописными греческими буквами. Оба способа равнозначны, оба встречаются в специальной литературе и привычны всем, кто погружен в изучение статистики. В таблице приведены названия некоторых букв (современные названия даны в скобках).

Буква	Название	Буква	Название	Буква	Название
α	альфа	ι	йота	ρ	ро
β	бета	κ	каппа	σ	сигма
γ	гамма	λ	лямбда	τ	тау
δ	дельта	μ	мю (ми)	υ	ипсилон
ϵ	эпсилон	ν	ню (ни)	ϕ	фи
ζ	дзета	ξ	кси	χ	хи
η	ита (эта)	o	омикрон	ψ	пси
θ	тета	π	пи	ω	омега

1. СТАТИСТИЧЕСКАЯ ОЦЕНКА ПАРАМЕТРОВ

1.1. ТОЧЕЧНЫЕ ОЦЕНКИ

Статистическая оценка $\tilde{\theta}$ параметра θ – это статистика, которая используется для оценивания неизвестного параметра распределения случайной величины.

Почти любую задачу оценки параметров можно представить в виде трех связанных задач:

1. Определить вид распределения случайных величин ξ_i.

2. Определить надежный алгоритм оценки параметров распределения.

3. Определить минимальный размер выборки n, необходимой для надежного определения параметров распределения.

Чтобы подобрать к задаче одно из известных распределений с конкретными параметрами, необходимо выполнить следующие шаги:

1. По описанию случайной величины предположить тип наиболее подходящего распределения.

2. Собрать выборку.

3. Вычислить статистику, которая лучшим образом описывает параметры распределения.

4. Проверить результаты. Для этого можно сравнить график полученного теоретического распределения с гистограммой, построенной по выборке. Если они похожи, то проверка пройдена. Если же нет, то нужно попробовать другое распределение.

Выборочным средним $\overline{X}$ выборки $X = (x_1, \ldots, x_n)$ – называют оценку математического ожидания распределения ξ, породившего выборку X:

$$\overline{X} = \tilde{E}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Выборочной дисперсией S^2 выборки $X = (x_1, \ldots, x_n)$ называют оценку дисперсии распределения ξ, породившего выборку X:

$$S^2 = \tilde{V}ar[X] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{X}^2.$$

Выборочной ковариацией $S_{X,Y}$ выборок $X = (x_1, \ldots, x_n)$ и $Y = (y_1, \ldots, y_n)$ называется оценка ковариации распределений ξ и η, породивших выборки X и Y:

$$S_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}x_i y_i - \overline{X}\overline{Y}.$$

Оценка параметра с помощью одного числа называется **точечной**.

В некоторых случаях одного параметра недостаточно, чтобы описать поведение случайной величины.

Интервальная оценка – это диапазон значений, в пределах которого с определенной степенью уверенности можно ожидать нахождение истинного значения параметра. Обычно она формируется в виде двух чисел, которые называются нижней и верхней границей интервала. Ключевое преимущество интервальной оценки в том, что она позволяет учесть неопределенность, связанную с оценкой.

Итак, в статистике оценки параметров бывают точечными и интервальными. Выбор между точечными и интервальными оценками зависит от конкретной ситуации и целей исследования.

Интервальные оценки позволяют учесть погрешность измерений, а точечные оценки обычно используются, когда необходима конкретность и простота.

Точечные оценки просто использовать, поэтому их применяют в случаях, когда нужна простота. Также они полезны, когда есть большой объем данных и уверенность в надежности оценок.

Получить более полное понимание данных, учесть возможную неопределенность и оценить точность результатов помогают интервальные оценки. Они особенно полезны при работе с небольшими выборками данных или при проведении научных исследований, где важно учесть все возможные факторы.

1.2. СМЕЩЕННЫЕ И НЕСМЕЩЕННЫЕ ОЦЕНКИ

При увеличении размера выборки логично ожидать, что ее статистики при этом становятся все больше похожи на соответствующие характеристики распределения случайной величины, которая выборку породила.

Например,

$$\tilde{E}[X] \to E[\xi] \text{ при } n \to \infty;$$

$$\tilde{V}ar[X] \to \tilde{V}ar[\xi] \text{при } n \to \infty$$

и т.д.

Это предположение. Проведем эксперимент в Python, чтобы его проверить.

Определим случайную величину ξ с известным распределением – например, $U(0, 1)$ и сгенерируем с ее помощью несколько выборок. Будем пошагово увеличивать размеры выборок и посмотрим, к чему стремится выборочное среднее.

```
1  import numpy as np
2
3  def my_mean(a, b, size):
4      # Генерируем случайную выборку
5      X = np.random.uniform(a, b, size)
6      # Выборочное среднее
7      set_mean = np.mean(X)
8      # Настоящее среднее
9      real_mean = (a + b)/2
10
11     print(size, set_mean, abs(real_mean - set_mean), sep = '\t|\t')
12
13 a = 0; b = 1; size = 10; size_factor = 2
14
15 print('Размер','Выборочное среднее','Отклонение от E[xi]', sep = '\t|\t')
16 print('-----------------------------------------------------------------')
17
18 i = 0
19 repeat_num = 15
20 while (i < repeat_num):
21     my_mean(a, b, size)
22     i = i + 1
23     size = size * size_factor
```

Результат:

Размер	Выборочное среднее	Отклонение от E[xi]
10	0.4975172837307367	0.0024827162692632743
20	0.5483768050346505	0.04837680503465047
40	0.5239472062930891	0.023947206293089107
80	0.4910969131674286 5	0.008903086832571 35
160	0.5336801696517407	0.03368016965174 07
320	0.5275006196140091	0.027500619614009 07
640	0.49858407118295245	0.0014159288170475515
1280	0.4965208869113939 5	0.0034791130888606045
2560	0.4981676404323229	0.0018323595676770776
5120	0.5025164195540995	0.002516419554099536
10240	0.5020302847235338	0.0020302847235338373
20480	0.5011013741034853	0.0011013741034853197
40960	0.4996754006910858 5	0.0003245993089141 53
81920	0.5008378082880462	0.0008378082880462 49
163840	0.4992828278409691	0.0007171721590308877

Ожидания оправдались: с ростом размера выборки выборочное среднее $\overline{X}$ стремится к математическому ожиданию $E[\xi]$.

Будем считать, что и другие выборочные статистики будут стремиться к своим аналогам для случайной величины.

На практике, как правило, невозможно постоянно увеличивать размеры выборки. Аналитики чаще всего работают с конечной выборкой, например, с результатами уже проведенного эксперимента. Размеры такой выборки не увеличить. И значит, нельзя опираться на то, что $n \to \infty$.

Вернемся к результатам эксперимента в Python. Выборка росла – и выборочное среднее $\overline{X}$ все ближе подходило к $E[\xi]$. Значит можно предположить, что в случае достаточно большого n с достаточно хорошей точностью выполнится серия приближенных равенств между характеристиками распределения и статистиками выборки:

$$\widetilde{E}[X] \approx E[\xi]; \quad \widetilde{Var}[X] \approx \widetilde{Var}[\xi]$$

и т.д.

Определить, является ли выборка достаточно большой – задача сложная и трудоемкая. Вместо этого можно строить оценки, которые хорошо показывают себя и на малых выборках.

Нет правильных и неправильных оценок, все они верные. При этом, можно оценивать одну и ту же характеристику одного и того же распределения на основе одной и той же выборки и получать разные значения.

Значит точечные оценки надо как-то сравнивать между собой. И критерий для такого сравнения есть – это информация о том, как именно оценки параметра стремятся к исходному значению самого параметра.

Определить, как оценка параметра или характеристики стремится к истинному значению параметра или характеристики поможет смещение.

Смещение оценки – разность между истинным значением θ параметра (характеристики) случайной величины и математическим ожиданием $E[\tilde{\theta}_n]$ оценки этого параметра (характеристики).

$$b(\theta) \approx E[\tilde{\theta}_n].$$

Следовательно, смещение оценки показывает, насколько далеки друг от друга математическое ожидание оценки параметра и истинное значение параметра.

– Если смещение оценки равно нулю или стремится к нему, то среднее полученных оценок при увеличении выборки будет приближаться к истинному значению параметра.

– Если смещение не равно нулю и не стремится к нему, то среднее оценок не попадет в точное значение параметра, как бы ни увеличивался размер выборки.

Получается, прежде чем вычислять оценку, надо проверить ее смещенность.

Оценки со смещением, стремящимся к нулю, применяются только для достаточно больших выборок.

Несмещенной оценкой $\tilde{\theta}$ параметра θ называют оценку θ, матожидание которой равно θ.

$$b(\theta) \approx E[\tilde{\theta}] - \theta = 0.$$

Другими словами, несмещенной будет оценка, смещенность которой равна 0, а остальные все смещенные.

Точечная оценка выборочной дисперсии не обладает нулевым смещением – значит она смещенная. Это неудобно, но есть решение: несмещенная оценка дисперсии.

Несмещенная оценка параметра дисперсии случайной выборки $X = (x_1, x_2, \ldots, x_n)$ распределения случайной величины ξ обозначается $\hat{s}^2$.

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2.$$

Чтобы получить такую оценку, взяли смещенную и умножили ее на $\frac{n}{n-1}$.

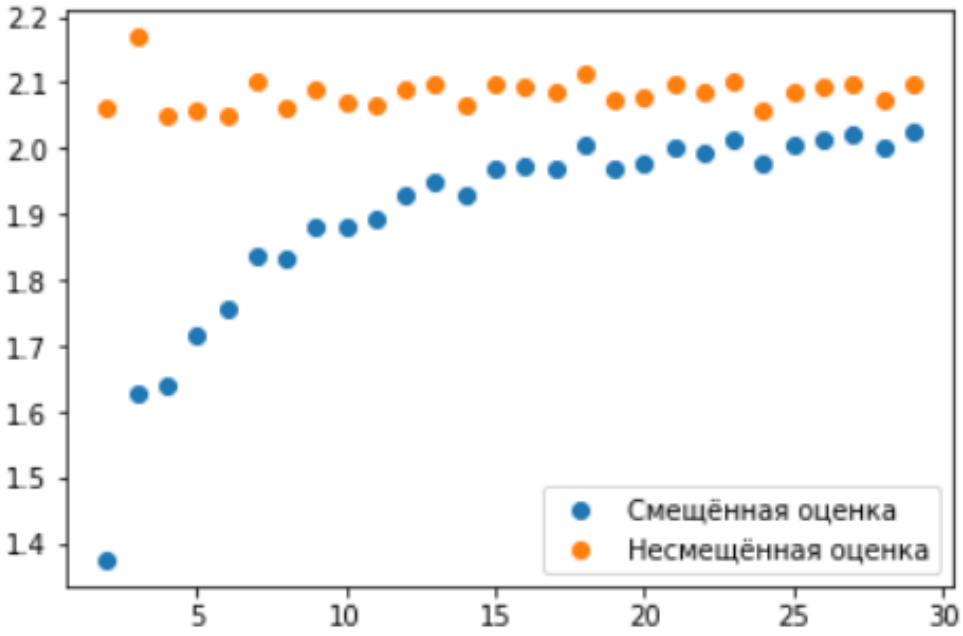
Посмотрим, как ведет себя несмещенная оценка на графике по сравнению со смещенной. Построим графики среднего значения набора несмещенной и смещенной оценок параметров.

```
import numpy as np
import matplotlib.pyplot as plt
a = 5
b = 10
i = 2
repeat_num = 30
theta1 = []
theta2 = []
while i < repeat_num:
    diff1 = []
    diff2 = []
    i = i + 1
    for j in range(1000):
        X = np.random.uniform(a, b, i)
        E1 = np.var(X)
        E2 = np.var(X, ddof=1)
        diff1.append(E1)
        diff2.append(E2)
    theta1.append(np.mean(diff1))
    theta2.append(np.mean(diff2))
    #diff.append(np.mean(theta) - (a + b) / 2)
plt.scatter(range(2, i), theta1)
plt.scatter(range(2, i), theta2)
plt.legend(["Смещённая оценка", "Несмещённая оценка"])
plt.show()
plt.show()
```

В коде несмещенная оценка дисперсии вычисляется иначе, чем смещенная. Для несмещенной добавился параметр `ddof=1`. Он определяет значение переменной k из знаменателя формулы, которая применяется в функции `np.var` для вычисления дисперсии выборки:

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2.$$

На рисунке приведен результат выполнения программы.



При небольшом размере выборки несмещенная оценка показывает более точный результат, чем смещенная. Но с увеличением выборки эти две оценки становятся практически одинаковы, потому что смещение смещенной оценки стремится к нулю.

Выборочная оценка ковариации тоже является смещенной оценкой.

Несмещенной выборочной ковариацией $S_{X,Y}$ выборок $X = (x_1, \ldots, x_n)$ и $Y = (y_1, \ldots, y_n)$ называют оценку ковариации распределений ξ и η, породивших выборки X и Y соответственно:

$$S_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})\,(y_i - \bar{Y}).$$

Краткая инструкция по анализу смещенности оценки:

Тип смещения/ Размер выборки	Равно нулю	Стремится к нулю	Не равно 0 и не стремится к 0
Малый	Можно применять	Лучше не применять	Не применять
Большой	Можно применять	Можно применять	Не применять

Перед тем как применять оценку, надо проанализировать ее смещение.

А иначе, если смещение оценки не равно нулю и не стремится к нему, можно получить значения, совсем не совпадающие с истинным значением параметра.

Несмещенная оценка лучше всех других работает на выборках небольшого объема. Оценка со стремящимся к нулю смещением хорошо работает только на больших наборах данных, а другие оценки почти никогда не имеют достаточной точности.

1.3. АНАЛИЗ ЭФФЕКТИВНОСТИ И СОСТОЯТЕЛЬНОСТИ ОЦЕНКИ

Несмещенная оценка лучше смещенной. Но бывает, что есть две несмещенные оценки. Для такого случая введем понятие эффективности.

Пусть $\tilde{\theta}_n$ и $\hat{\theta}_n$ – две оценки параметра θ по выборке объема n, смещенность которых равна или стремится к нулю. (Считаем, что либо обе оценки являются несмещенными, либо у обеих смещения стремятся к нулю). Оценка $\tilde{\theta}_n$ называется **более эффективной**, чем оценка $\hat{\theta}_n$, если при любых n ее дисперсия меньше, т.е.

$$Var(\tilde{\theta}_n) < Var(\hat{\theta}_n).$$

Нижний индекс n здесь для указания зависимости от размера выборки.

Из двух несмещенных оценок лучше та, которая более эффективна.

У более эффективной оценки меньше дисперсия, а значит, меньше диапазон ее возможных значений. Получается, если использовать более эффективную оценку, то с большей вероятностью попадем в маленькую окрестность истинного значения параметра.

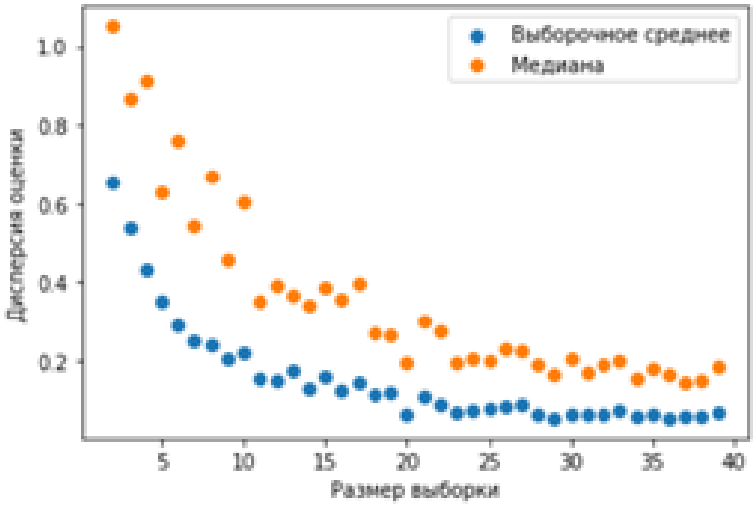
Рассмотрим две разные оценки математического ожидания: медиану и выборочное среднее. Обе они несмещенные. Выясним, какая из них более эффективная.

Оттолкнемся от определения более эффективной оценки.

Оценим дисперсию медианы и выборочного среднего с помощью выборочной дисперсии. Качество оценок проанализируем так же, как и раньше – с помощью выборки. Для этого возьмем 100 выборок каждого размера i и вычислим выборочную дисперсию для каждого размера, а затем – ее среднее на всех выборках заданного размера.

```python
import numpy as np
import matplotlib.pyplot as plt
a = 5
b = 10
i = 2
repeat_num = 40
theta1 = []
theta2 = []

while i < repeat_num:
    diff1 = []
    diff2 = []
    i = i + 1
    for j in range(100):
        X = np.random.uniform(a, b, i)
        E1 = np.mean(X)
        E2 = np.median(X)
        diff1.append(E1)
        diff2.append(E2)
    theta1.append(np.var(diff1, ddof = 1))
    theta2.append(np.var(diff2, ddof = 1))

plt.scatter(range(2, i), theta1)
plt.scatter(range(2, i), theta2)
plt.legend(["Выборочное среднее", "Медиана"])
plt.xlabel("Размер выборки")
plt.ylabel("Дисперсия оценки")
plt.show()
```

Результат:



Видно, что дисперсия выборочного среднего значительно меньше, чем дисперсия медианы. Значит в данном случае выборочное среднее – более эффективная оценка, чем медиана.

Для параметров распределения можно придумать много разных оценок.

Статистика $\tilde{\theta}_n$ называется **наиболее эффективной оценкой** параметра θ, если для любой другой оценки $\hat{\theta}_n$ выполнено условие

$$Var(\tilde{\theta}_n) < Var(\hat{\theta}_n).$$

Другими словами, наиболее эффективная оценка – это оценка с наименьшей дисперсией.

Разброс значений наиболее эффективной оценки меньше, чем у любой другой оценки. Если дисперсия одной оценки меньше, чем дисперсия другой, то первая оценка более эффективная.

Получается, чтобы выбрать оценку из двух несмещенных, надо взять наиболее эффективную из них.

Рассмотрим ситуацию, когда смещение хотя бы одной оценки не равно, а стремится к нулю. В таком случае эффективность не подходит, потому что надо одновременно оценить две характеристики: качество стремления математического ожидания оценки к истинному значению и малость дисперсии.

В этом случае надо использовать *MSE*.

MSE (среднеквадратичная ошибка) – это средняя сумма квадратов невязок. В случае, когда нужно определить точность оценки параметров, среднеквадратичная ошибка рассчитывается так:

$$MSE(\tilde{\theta}, \theta) = E\left[(\tilde{\theta} - \theta)^2\right].$$

Если размер выборки увеличивается, то *MSE* несмещенной выборочной дисперсии уменьшается. Поэтому у этой оценки есть важное свойство: при увеличении размера выборки несмещенная выборочная дисперсия становится точнее.

Значение *MSE* само по себе ничего не показывает, оно используется только для сравнения двух разных оценок. Поэтому

– формулировка «*MSE* оценки равна 30» – неинформативна;

– фраза «*MSE* одной оценки равна 30, а другой 150 – поэтому решили использовать первую оценку» – корректна.

MSE несмещенной выборочной дисперсии больше, чем *MSE* смещенной выборочной дисперсии. Получается, для нормального распределения смещенная оценка дисперсии точнее (больше похожа на истинное значение параметра), чем несмещенная.

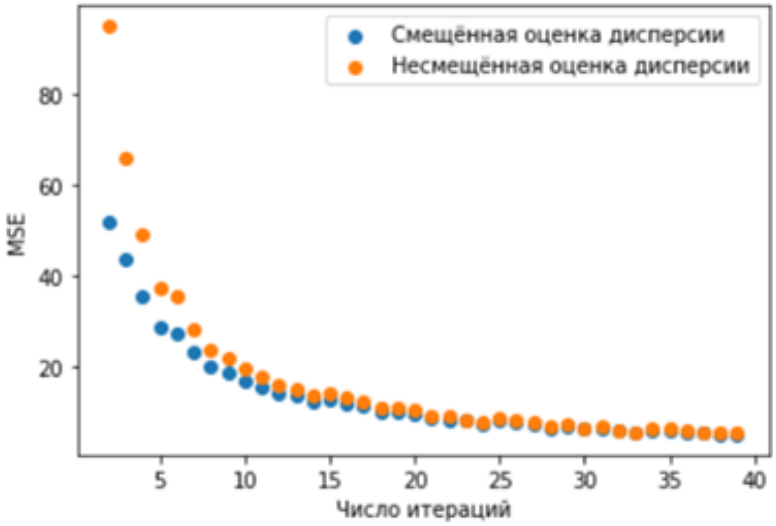
Подтвердим этот результат с помощью Python. Для этого проведем множество статистических экспериментов на выборках разного размера. Для выборки каждого размера рассчитаем несмещенную выборочную дисперсию и ее *MSE*, а также смещенную выборочную дисперсию и ее *MSE*. Чтобы сравнить полученные значения ошибок, выведем график.

На графике видно, что значение *MSE* для смещенной оценки действительно меньше значения *MSE* для несмещенной. А значит для нормального распределения смещенная выборочная дисперсия показывает более точный результат, чем несмещенная.

Получается, несмещенность оценки не гарантирует ее точности. Результат верен только для случая нормального распределения.

```python
import numpy as np
import matplotlib.pyplot as plt

mu = 5
sigma2 = 10
repeat_num = 40
i = 2

theta1 = []
theta2 = []

while i < repeat_num:
    diff1 = []
    diff2 = []
    i = i + 1
    for j in range(1000):
        X = np.random.normal(mu, np.sqrt(sigma2), i)
        E1 = np.var(X)
        E2 = np.var(X, ddof = 1)
        diff1.append((E1-sigma2)**2)
        diff2.append((E2-sigma2)**2)
    theta1.append(np.mean(diff1))
    theta2.append(np.mean(diff2))

plt.scatter(range(2, i), theta1)
plt.scatter(range(2, i), theta2)
plt.legend(["Смещённая оценка дисперсии", "Несмещённая оценка дисперсии"])
plt.xlabel("Число итераций")
plt.ylabel("MSE")
plt.show()
```

Результат:



Точечная оценка $\tilde{\theta}_n$ называется **состоятельной оценкой параметра** θ, если при $n \to \infty$ последовательность случайных величин $\tilde{\theta}_1, \tilde{\theta}_2, ..., \tilde{\theta}_n$ все ближе приближается к истинному значению параметра θ.

Пусть $\tilde{\theta}_n$ – статистическая оценка параметра θ распределения X. Если при $n \to +\infty$ $E(\tilde{\theta}_n) \to 0$ и $Var(\tilde{\theta}_n) \to 0$, то оценка $\tilde{\theta}_n$ состоятельная.

1.4. ОЦЕНКА МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

Метод максимального правдоподобия использует информацию о распределении, чтобы построить оценки параметров.

Для непрерывных случайных величин вероятность принятия значения в точке всегда равна нулю. Поэтому для непрерывной случайной величины оценивают не вероятность принятия конкретного значения, а вероятность попадания ее значения в конкретный отрезок. Поэтому вместо вероятности для непрерывных случайных величин рассматривают произведение функций плотности.

Пусть дана выборка $X = (x_1, x_2, \ldots, x_n)$, порожденная непрерывной случайной величиной ξ с плотностью вероятности $f_\xi(x; \theta)$. Тогда **функцией правдоподобия** называется функция

$$L(\theta) = \prod_{i=1}^{n} f_\xi(x_i; \theta).$$

Функция правдоподобия принимает значения от 0 до 1. Функции плотности у экспоненциального и нормального распределения различаются, а значит будут различны и функции правдоподобия.

Функция правдоподобия помогает определить, какая из выборок более вероятно порождена данным распределением. В этом случае известно распределение, а выборка – переменная величина. На практике чаще встречаются другие задачи. Обычно выборка задана и неизвестны ее пара-

метры. В таком случае используют ту же функцию, чтобы выбрать более подходящие параметры. Только теперь переменной становится параметр распределения.

Для выборки дискретной случайной величиной ξ функция правдоподобия записывается аналогично, под $f_\xi(x;\theta)$ понимается функция вероятности.

Здесь θ – параметр или набор параметров, который однозначно задает плотность вероятности (или функцию вероятности) $f_\xi(x;\theta)$.

В зависимости от распределения θ может быть одним параметром или набором параметров. Например, экспоненциальное распределение задается одним параметром, а равномерное и нормальное – двумя. Строго говоря, для распределений с несколькими параметрами θ – вектор, и в записи нужно это указывать. Когда рассматривают общий случай, для краткости пишут просто θ и понимают, что возможны разные варианты.

Будем рассматривать функцию правдоподобия именно с этой точки зрения – как функцию параметров распределения.

Метод максимального правдоподобия основан на поиске максимума функции правдоподобия.

Обозначим множество всех возможных параметров, которое пробегает θ как Θ. Если у распределения один параметр, то Θ – множество чисел. Если параметров у распределения несколько, то Θ – множество векторов.

Пусть $L(\theta)$ – функция правдоподобия случайной величины ξ, связанная с выборкой $X = (x_1, \ldots, x_n)$. Параметр $\tilde{\theta}$, найденный из условия

$$\tilde{\theta} = \arg\max_{\theta\in\Theta} L(\theta),$$

называется **оценкой максимального правдоподобия** параметра θ.

Если функция правдоподобия $L(\theta)$ дифференцируемая, то, чтобы найти ее точки максимума, можно приравнять производные к нулю или использовать метод градиентного спуска.

Получаем **алгоритм оценки параметров распределения**:

1. Предположить тип распределения исходной выборки.

2. Составить выражение функции правдоподобия для исходной выборки и выбранного распределения.

3. Найти значения оценок параметров, при которых функция правдоподобия достигает максимума. Это и будет оценка максимального правдоподобия.

У выборочных статистик есть недостаток: нужно каждый раз проверять их смещенность, эффективность, состоятельность. А оценки максимального правдоподобия лишены этого недостатка.

Если оценка $\tilde{\theta}$ параметра θ некоторого распределения $f_\xi(x; \theta)$ получена с помощью метода максимального правдоподобия, то она обязательно будет **состоятельной**.

Так как оценки максимального правдоподобия параметров получаются состоятельными, их не нужно дополнительно исследовать, а можно использовать сразу. В этом их преимущество по сравнению с выборочными оценками, и именно поэтому оценки максимального правдоподобия часто используются в практических задачах.

Если выборка $Y = (y_1, \ldots, y_n)$ порождена случайной величиной $\xi \sim \mathrm{Exp}(\lambda)$, то оценка $\tilde{\lambda}$ параметра λ по методу максимального правдоподобия имеет вид

$$\tilde{\lambda} = \frac{1}{\overline{Y}},$$

где $\overline{Y}$ – выборочное среднее.

Если выборка $Y = (y_1, \ldots, y_n)$ порождена случайной величиной $\xi \sim U(a, b)$, то в качестве оценок $\tilde{a}$ и $\tilde{b}$ параметров a и b по методу максимального правдоподобия принимают

$$\tilde{a} = \min(Y), \quad \tilde{b} = \max(Y).$$

Пусть $X = (x_1, x_2, \ldots, x_n)$, – выборка, порожденная случайной величиной ξ с функцией плотности вероятности (или функцией вероятности в дискретном случае) $f_\xi(x; \theta)$ с набором параметров θ. Функцию

$$\ln L(\theta) = \ln\left(\prod_{i=1}^{n} f_\xi(x_i; \theta)\right) = \sum_{i=1}^{n} \ln(\, f_\xi(x_i; \theta))$$

называют **логарифмом функции правдоподобия**.

Когда вычисления сложные, исследование функции правдоподобия заменяют исследованием ее логарифма. Логарифм произведения равен сумме логарифмов, поэтому вычисление производной произведения заменяется на вычисление производной суммы величин. Часто это значительно упрощает решение задачи.

Когда применяем логарифм к функции правдоподобия, получаем композицию функций $\ln L(p)$. Натуральный логарифм – возрастающая функция, поэтому ее применение к функции правдоподобия сохранит точки экстремума исходной функции и их тип. Значит точки максимума обеих функций совпадают. Получается, можно заменить исследование функции правдоподобия исследованием ее логарифма.

Если выборка $Y = (y_1, \ldots, y_n)$ порождена случайной величиной $\xi \sim N(\mu, \sigma^2)$, то оценки максимального правдоподобия $\tilde{\mu}$ и $\tilde{\sigma}^2$ параметров μ и $\tilde{\sigma}$ имеют вид

$$\begin{cases} \tilde{\mu} = \bar{Y}; \\ \tilde{\sigma} = \sqrt{\hat{S}^2}, \end{cases}$$

где $\bar{Y}$ – выборочное среднее; $\hat{S}^2$ – смещенная выборочная дисперсия.

Получается, для нормального распределения оценками максимального правдоподобия являются выборочное среднее и смещенная выборочная дисперсия.

Итог по методу максимального правдоподобия:

– основа подхода проста – поиск точки максимума функции;

– метод помогает получить оценки параметров для распределений;

– оценки получаются состоятельными, а значит их можно использовать на практике;

– метод универсальный, подходит для анализа любых распределений;

– он помогает определить, к какому типу распределения наиболее вероятно принадлежит выборка.

Именно поэтому метод максимального правдоподобия часто используют в реальных задачах.

1.5. МЕТОД МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ В ДИСКРЕТНОМ СЛУЧАЕ

Рассмотрим оценку максимального правдоподобия параметров **геометрического распределения**.

Если выборка $Y = (y_1, \ldots, y_n)$ порождена случайной величиной $\xi \sim \text{Geom}(p)$, то оценка $\tilde{p}$ параметра p по методу максимального правдоподобия имеет вид

$$\tilde{p} = \frac{1}{\overline{Y}},$$

где $\overline{Y}$ – выборочное среднее.

Рассмотрим оценку максимального правдоподобия параметров **распределения Пуассона**.

Если выборка $Y = (y_1, \ldots, y_n)$ порождена случайной величиной $\xi \sim \text{Poisson}(\lambda)$, то оценка $\tilde{\lambda}$ параметра λ по методу максимального правдоподобия имеет вид

$$\tilde{\lambda} = \overline{Y},$$

где $\overline{Y}$ – выборочное среднее.

Рассмотрим оценку максимального правдоподобия параметров **биномиального распределения**.

Если выборка $Y = (y_1, \ldots, y_n)$ порождена случайной величиной $\xi \sim \text{Binomial}(m, p)$ с произвольным параметром m, то оценка $\tilde{p}$ параметра p по методу максимального правдоподобия имеет вид

$$\tilde{p} = \frac{\bar{Y}}{m},$$

где $\bar{Y}$ – выборочное среднее.

Надо найти максимум функции

$$g(m) = \ln L\left(m, \frac{Y}{m}\right),$$

когда m меняется от включительно до бесконечности.

Выберем достаточно большое число N и будем считать, что значение m, в котором достигается максимум функции $g(m)$, находится в отрезке $[\max(Y), N]$. Тогда задача сводится к перебору конечного набора значений.

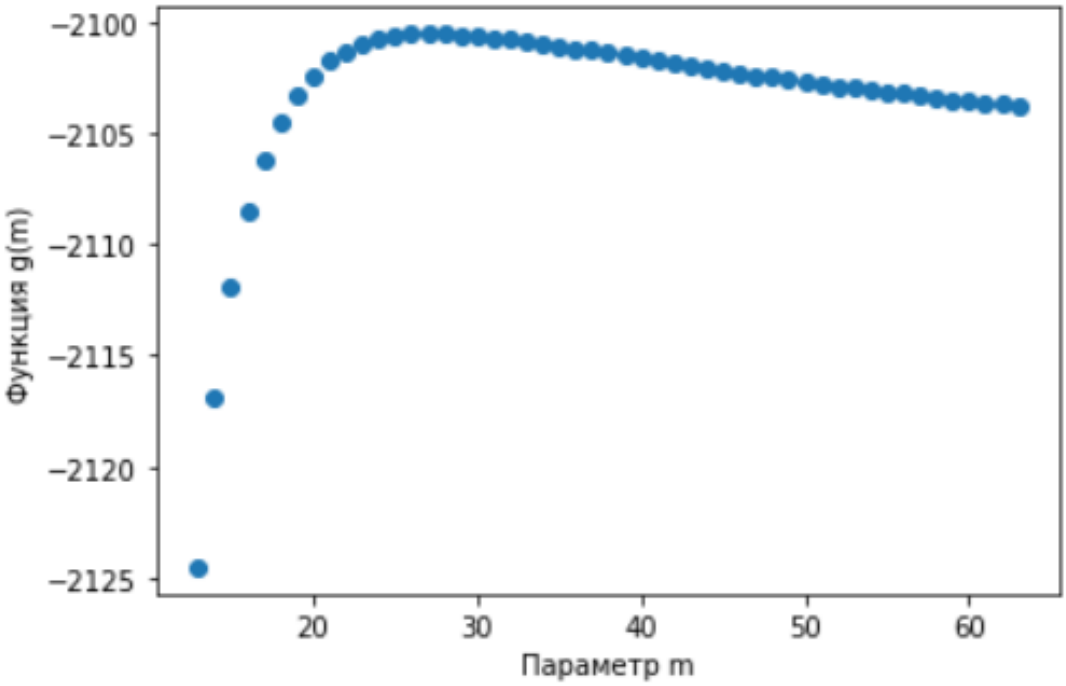
Параметр m можно оценить численно перебором значений функции $g(m)$. В таком случае есть два варианта.

1. На графике $g(m)$ не видно максимума. Тогда с большой вероятностью выборка Y не распределена биномиально. Не стоит пытаться оценить ее параметры.

2. На графике $g(m)$ видна точка максимума. Тогда с большой вероятностью выборка Y распределена биномиально. Полученные по алгоритму оценки состоятельны.

Рассмотрим выборку, порожденную биномиальным распределением с известными параметрами. Настоящие значения параметров: $p = 0{,}2$, $m = 24$. Оценки по методу максимального правдоподобия:

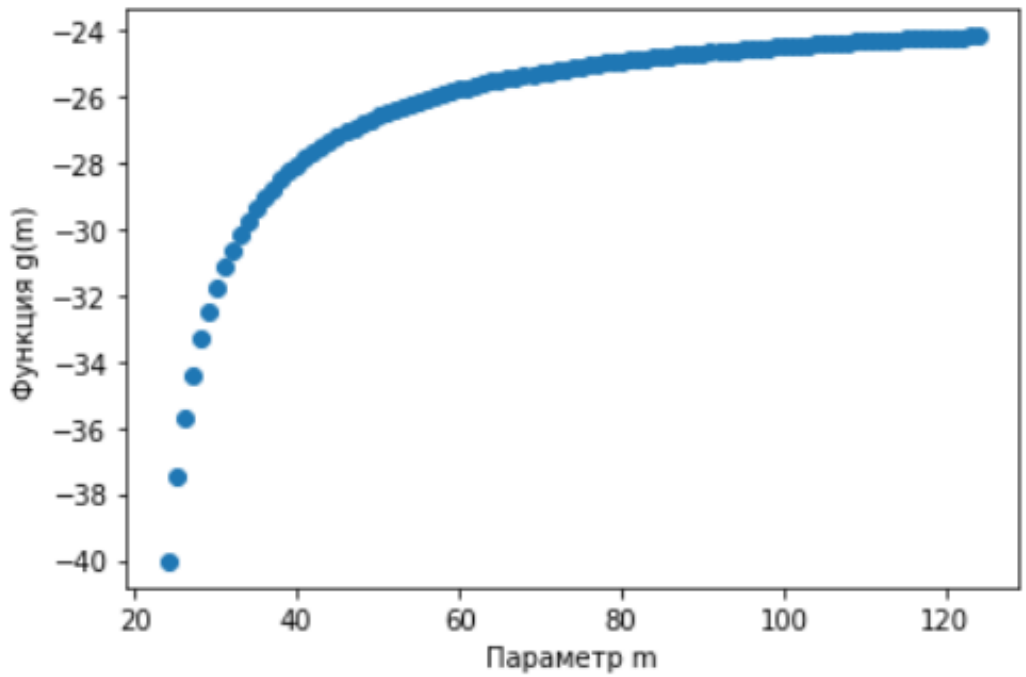
$$p = 0.17877777777777779,\ m = 27.$$



На графике видно, как функция $g(m)$ меняется с ростом параметра m. В окрестности истинного значения m у функции пик.

Применим созданный инструмент к проверке выборки X = (2, 16, 24, 10, 23). Будем перебирать значения из диапазона [max(X), max(X) + 100]. Оценки по методу максимального правдоподобия:

$$p = 0{,}12096774193548387,\ m = 124.$$



У этой функции нет максимума. Дело в том, что выборка не распределена биномиально, поэтому и оценка ведет себя некорректно. Оценка параметра здесь очень маловероятная.

1.6. ЛИНЕЙНАЯ РЕГРЕССИЯ С ВЕРОЯТНОСТНОЙ ТОЧКИ ЗРЕНИЯ

Дан набор векторов X_i и вектор Y. Формула линейной регрессии с учетом случайных ошибок:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \gamma,$$

где X_i – данные наборы векторов; γ – ошибка.

Если $\gamma \sim N(\mu, \sigma^2)$, то оценки для коэффициентов совпадают с оценками, полученными с помощью метода наименьших квадратов с метрикой *MSE*, и вычисляются так:

$$\beta^T = (X^T X)^{-1} X^T Y^T.$$

Общая дисперсия – это дисперсия исходного набора точек:

$$SS_{tot} = \frac{1}{n}\sum_{i=0}^{n}(y_i - \bar{y})^2.$$

Объясненная дисперсия – дисперсия набора точек, лежащих на линии регрессии:

$$SS_{res} = \frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y})^2.$$

Пусть дана выборка Y, на основе которой построили линейную регрессию. Тогда **выборочным коэффициентом детерминации** (R^2) называется величина, равная

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}},$$

где SS_{tot} – общая дисперсия; SS_{reg} – объясненная дисперсия; SS_{res} – необъясненная дисперсия.

2. СТАТИСТИЧЕСКИЕ ЭКСПЕРИМЕНТЫ И ПРОВЕРКА ГИПОТЕЗ

2.1. ВВЕДЕНИЕ В ИНТЕРВАЛЬНУЮ ОЦЕНКУ ПАРАМЕТРОВ

В разделе 1 была представлена точечная оценка параметра распределения. Например, если аналитик изучает возраст пользователей, точечной оценкой будет среднее арифметическое по собранным наблюдениям. Точечная оценка вычисляется на основе выборки. Предполагается, что полученное число близко к истинному параметру случайной величины.

Точечные оценки часто используют в статистике и прикладном анализе данных, поскольку это просто, а результат хороший.

Но у точечных оценок есть недостаток – иногда они существенно отклоняются от реального значения параметра. Особенно часто такое происходит, если данные имеют большую дисперсию или их немного. Например, в медицинских исследованиях бывает сложно собрать большое количество данных, при этом погрешность в оценке может иметь критические последствия.

Поэтому существует другой подход к оценке параметров – он учитывает погрешность.

Когда формируем выборку один раз и рассчитываем оценку по ней, погрешность может быть высока. Если же повторить эксперимент много раз, то получим много оценок, сможем проанализировать их и таким образом учесть погрешность. Такая концепция называется в статистике **многократным повторением эксперимента**. В реальной жизни часто нет возможности так делать, но есть статистические методы, которые позволят хорошо оценить такой процесс.

Напомним, что для непрерывной случайной величины вероятность попадания в точку равна нулю.

Использовать среднее по выборке, чтобы оценить среднее в генеральной совокупности, – это естественный шаг. Но нужно учитывать, что такая оценка связана со случайностью.

Пусть X – это случайная величина, которую исследуем. А X_i – это случайная величина, отражающая распределение для отдельного наблюдения. Предположим, что все X_i имеют такое же распределение, как X.

Тогда случайная величина, отражающая распределение выборочных средних:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

При $n \to \infty$ плотность распределения случайной величины $\overline{X}_n$ будет все больше похожа на функцию плотности нормального распределения с параметрами

$$N\left(\mu X, \frac{\sigma^2 X}{n}\right).$$

Следовательно, $E[\overline{X}_n] = \mu X$, а $Var[\overline{X}_n] = \frac{\sigma^2 X}{n}$.

Стандартная ошибка (от англ. Standard Error, *SE*) – это стандартное отклонение выборочных средних, которое зависит от количества наблюдений в выборке и стандартного отклонения случайной величины X.

$$SE = \sigma\overline{X}_n = \frac{\sigma X}{\sqrt{n}}.$$

Распределение выборочных средних не всегда подчиняется нормальному закону распределения. Например, если выборка малого размера или исходное распределение сильно отличается от нормального, то распределение выборочных средних может существенно отличаться от нормального.

Согласно центральной предельной теореме, математическое ожидание распределения выборочных средних равно математическому ожиданию исходного распределения.

Чем больше размер выборки, тем меньше дисперсия распределения выборочных средних. Дисперсия распределения выборочных средних рассчитывается по формуле

$$Var[\overline{X}_n] = \frac{\sigma^2 X}{n}.$$

Размер выборки – в знаменателе. Чем больше знаменатель, тем меньше дробь. Таким образом, с увеличением размера выборки дисперсия распределения выборочных средних уменьшается.

Чтобы получать более надежные выводы, практические эксперименты не обязательно повторять много раз. Многократное повторение эксперимента – теоретическая концепция. На практике повторять эксперименты часто невозможно или слишком дорого. Чтобы повысить точность и надежность результатов, в реальных задачах обычно стараются увеличить размер выборки в одном эксперименте.

2.2. ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ

Учесть погрешность, возникающую при оценке параметров распределения, помогают интервальные оценки – они представляют диапазон, в пределах которого с определенной степенью уверенности ожидают нахождение истинного значения параметра.

Концепция многократного повторения эксперимента помогает оценить среднее значение. Если провести много экспериментов с большой выборкой, то получится распределение выборочных средних. По центральной предельной теореме оно подчиняется нормальному закону распределения.

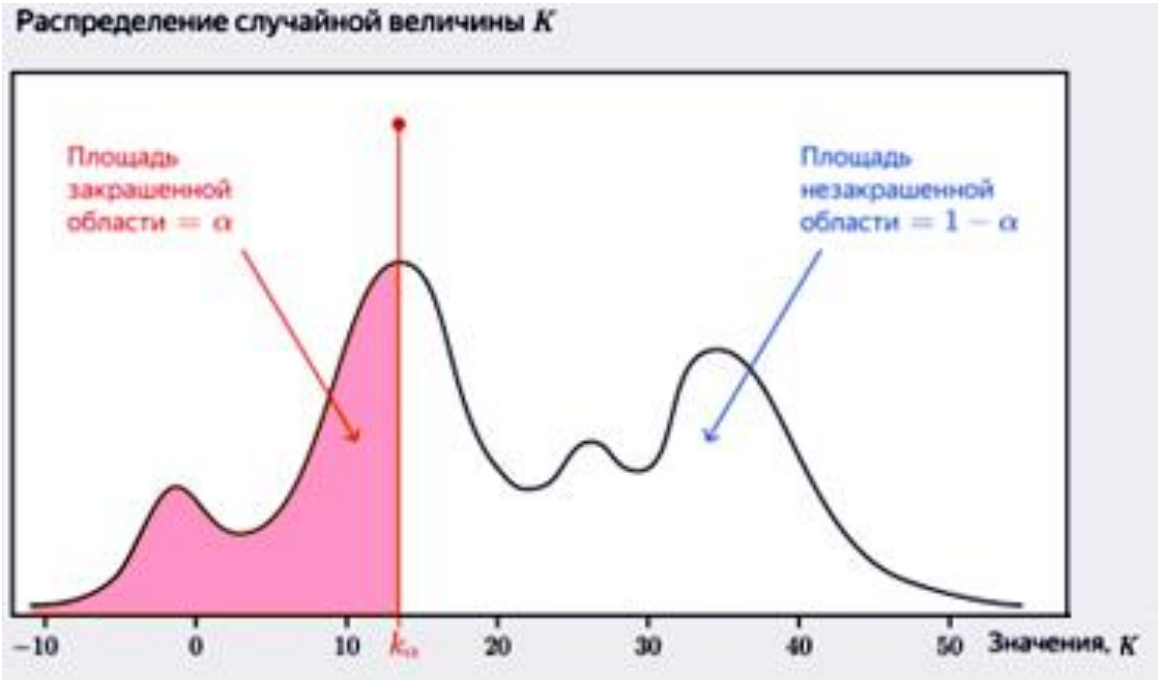
Значение, ниже которого находится определенный процент наблюдений из распределения, называется **квантилем** распределения.

Квантили существуют не только для стандартного нормального распределения, а вообще для любого распределения. Рассмотрим общий случай.

Пусть K – произвольная непрерывная случайная величина. Она может принимать разные значения k с разной вероятностью. Эта вероятность задается функцией плотности распределения.

Есть еще один способ задать поведение случайной величины, который тесно связан с функцией распределения $F(k)$.

Квантиль уровня α – это значение k_α, которое делит распределение на две части. Квантиль устанавливает границу так, что доля значений, которые меньше или равны k_α, составляет α, а доля значений, которые больше или равны k_α, составляет 1 – α.



Следовательно, число k_α – это одно из возможных значений случайной величины K, а α – это вероятность, которая может принимать значения от 0 до 1.

Математически это выражается так:

$$\begin{cases} P(K \le k_\alpha) = F(k_\alpha) = \alpha; \\ P(K > k_\alpha) = 1 - F(k_\alpha) = 1 - \alpha. \end{cases}$$

Например, медиана – это квантиль уровня 0,5, т.е. точка, ниже которой находится ровно 50% значений распределения.

На рисунках показано, как меняются квантили для стандартного нормального распределения в зависимости от уровня α.

Для заданного распределения квантиль можно найти с помощью **обратной функции распределения**, которую часто называют функцией квантилей. В Python она находится в библиотеке `scipy` и называется `ppf` (от англ. percent point function).

Допустим, надо найти квантиль уровня 0,99 для стандартного нормального распределения Z. Это точка, ниже которой находится 99% значений распределения. Ее обозначают как z_{99}. Вот как это можно сделать с помощью `scipy`:

```
from scipy.stats import norm

# Параметры стандартного нормального распределения
Z = norm(loc=0, scale=1)

# Находим квантиль уровня 0.99
z_99 = Z.ppf(0.99)
print(z_99)
```

Критические значения помогают определить «наиболее вероятный» диапазон значений случайной величины.

Пусть $z_{\alpha/2}$ и $z_{1-\alpha/2}$ – критические значения для стандартного нормального распределения. Тогда

$$P(z_{\alpha/2} \le Z \le z_{1-\alpha/2}) = 1-\alpha.$$

В некоторых источниках встречаются определения со знаками строгого неравенства:

$$P(z_{\alpha/2} < Z < z_{1-\alpha/2}) = 1-\alpha.$$

Для непрерывных случайных величин вероятность того, что величина примет конкретное значение, равна нулю. Получается, для непрерывных распределений вероятность попадания в интервал, заданный строгим неравенством, равна вероятности попадания в интервал, заданный нестрогим неравенством с такими же границами.

Это значит, что записи равноправны:

$$P(z_{\alpha/2} \le Z \le z_{1-\alpha/2}) = P(z_{\alpha/2} < Z < z_{1-\alpha/2}) = 1-\alpha.$$

α называют **уровнем значимости**. А вероятность 1 – α – **уровнем доверия**. Эти параметры задает исследователь. В качестве уровня значи-

мости обычно используют значения 0,01, 0,05, 0,1. Чем меньше α, тем больше уровень доверия, а значит, тем шире интервал между критическими значениями.

Пусть $x_1, \ldots, x_n$ – случайная выборка из произвольного распределения, где параметр σ_X^2 известен, а распределение выборочных средних является нормальным. Тогда для заданного уровня значимости $\alpha \in (0, 1)$ **доверительный интервал** для μ_X имеет вид:

$$\left(\overline{X}_n - z_{1-\alpha/2} \frac{\sigma X}{\sqrt{n}},\ \overline{X}_n - z_{\alpha/2} \frac{\sigma X}{\sqrt{n}} \right).$$

Именно границы доверительного интервала включают в себя случайную величину и меняются от выборки к выборке. Поэтому правильнее говорить, что вероятность связана с надежностью оценки границ интервала, а не с самим оцениваемым параметром θ.

Вместо математического ожидания можно оценивать и другие параметры. Искомый параметр называют θ. Границы доверительного интервала $\hat{\theta}_1$ и $\hat{\theta}_2$ обычно представляют собой статистики – численные характеристики выборки.

Пусть набор данных $x_1, \ldots, x_n$ – реализация выборки $X_1, \ldots, X_n$, где $X_i \sim F$. Пусть θ – интересующий нас параметр распределения F, а $\alpha \in (0, 1)$.

Если для любого возможного значения θ верно

$$P\left(\hat{\theta}_1(X_1, ..., X_n) < \theta < \hat{\theta}_2(X_1, ..., X_n)\right) = 1 - \alpha,$$

то интервал

$$\hat{\theta}_1(x_1, ..., x_n),\ \hat{\theta}_2(x_1, ..., x_n)$$

называют $100 \cdot (1 - \alpha)\%$ **доверительным интервалом** для параметра θ.

2.3. ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ И РАСПРЕДЕЛЕНИЕ СТЬЮДЕНТА

Стандартное отклонение генеральной совокупности не всегда известно. В некоторых случаях можно использовать данные на основе предыдущего опыта. Но когда впервые появляется новое явление, взять значение стандартного отклонения неоткуда. Это ограничение можно обходить.

Итак, когда σ неизвестна, тот факт, что $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$, не поможет при выводе доверительного интервала.

Лучшая несмещенная оценка для стандартного отклонения в рамках выборки – выборочное стандартное отклонение

$$S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}.$$

Заменим σ на S_n, получим случайную величину

$$T = \frac{\overline{X}_n - \mu}{S_n / \sqrt{n}}.$$

Если выборочные средние распределены нормально, то случайная величина T будет иметь особенное распределение. Его называют **распределением Стьюдента**, или t-распределением.

Когда используется выборочное стандартное отклонение S_n вместо истинного стандартного отклонения σ, в оценках появляется дополнительная неопределенность. Это связано с тем, что S_n само по себе является случайной величиной, которая изменяется от выборки к выборке. В результате распределение, которое используется для вывода доверительных интервалов, учитывает эту дополнительную неопределенность.

В контексте распределения Стьюдента используют термин **степени свободы**. Обозначение: $m = n - 1$, где n – размер выборки. Степени свободы описывают количество значений в наборе данных, которые могут варьироваться, когда известны некоторые статистические параметры выборки.

В контексте выборочного стандартного отклонения это означает, что если известно среднее значение выборки, то из n наблюдений только $n - 1$ могут свободно варьироваться. Последнее значение строго определяется выбранными $n - 1$ значениями и средним. Это ограничение и является причиной, по которой степень свободы для выборочного стандартного отклонения составляет $n - 1$.

Степени свободы определяются как количество значений в выборке, которые могут свободно варьироваться, когда известны некоторые статистические характеристики (например, среднее значение).

Форма распределения Стьюдента зависит от размера выборки n. А точнее – от количества степеней свободы.

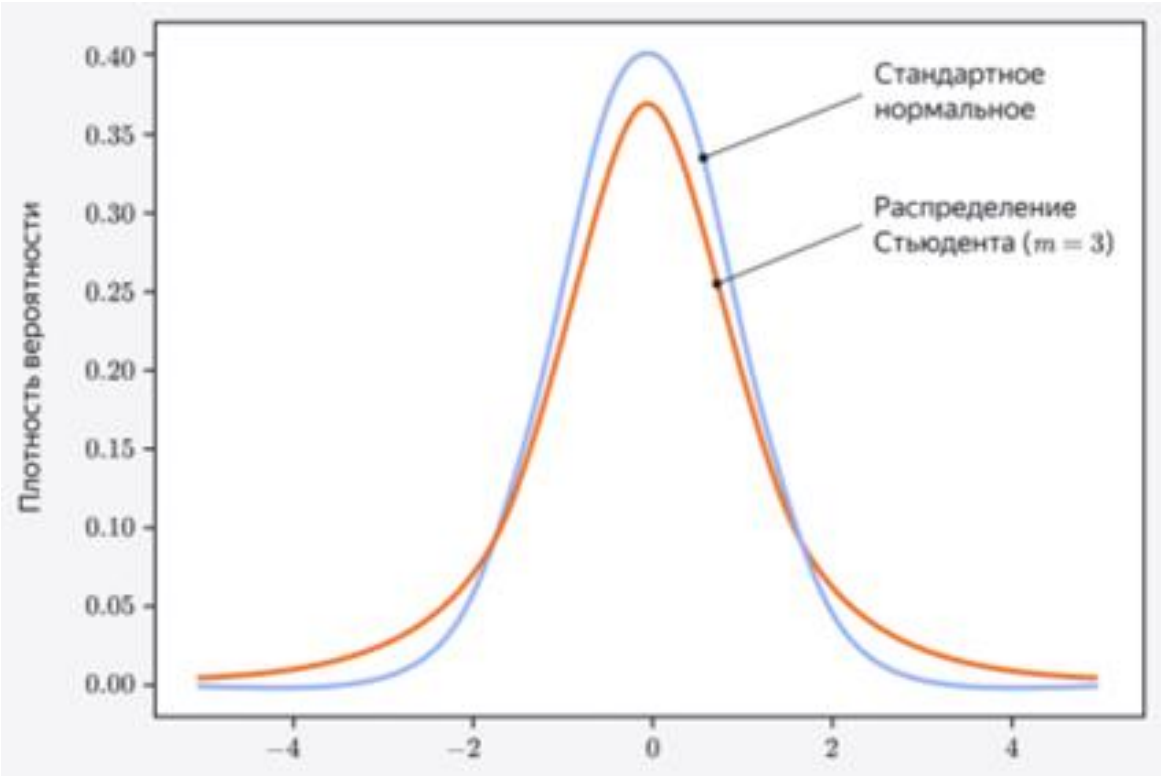
Количество степеней свободы рассчитывают так:

$$m = n - 1,$$

где n – размер выборки.

График распределения Стьюдента напоминает график нормального распределения. Он тоже имеет колокол и симметричен относительно нуля.

Сравним графики распределения Стьюдента, например, с тремя степенями свободы и график стандартного нормального распределения:



На иллюстрации заметно, например, что вероятность получить значение –4 для распределения Стьюдента выше, чем для стандартного нормального распределения. В отличие от нормального распределения, *t*-распределение обладает «тяжелыми» хвостами.

Распределение Стьюдента шире по сравнению со стандартным нормальным распределением. Это отражает увеличенную неопределенность из-за использования S_n вместо σ.

С увеличением размера выборки распределение Стьюдента сужается, и его хвосты «облегчаются». График становится все более похожим

на нормальное распределение. Это согласуется с интуитивным пониманием того, что с увеличением объема данных уменьшается неопределенность оценок.

Отличия в работе распределения Стьюдента и стандартного нормального заметны только при небольших размерах выборки.

Распределение Стьюдента – распределение колоколообразной формы, симметричное относительно нуля. Его плотность имеет более тяжелые и длинные хвосты, чем у стандартного нормального распределения: $f(x)$ стремится к нулю при x, стремящемся к $+\infty$ и $-\infty$, но медленнее, чем в стандартном нормальном распределении.

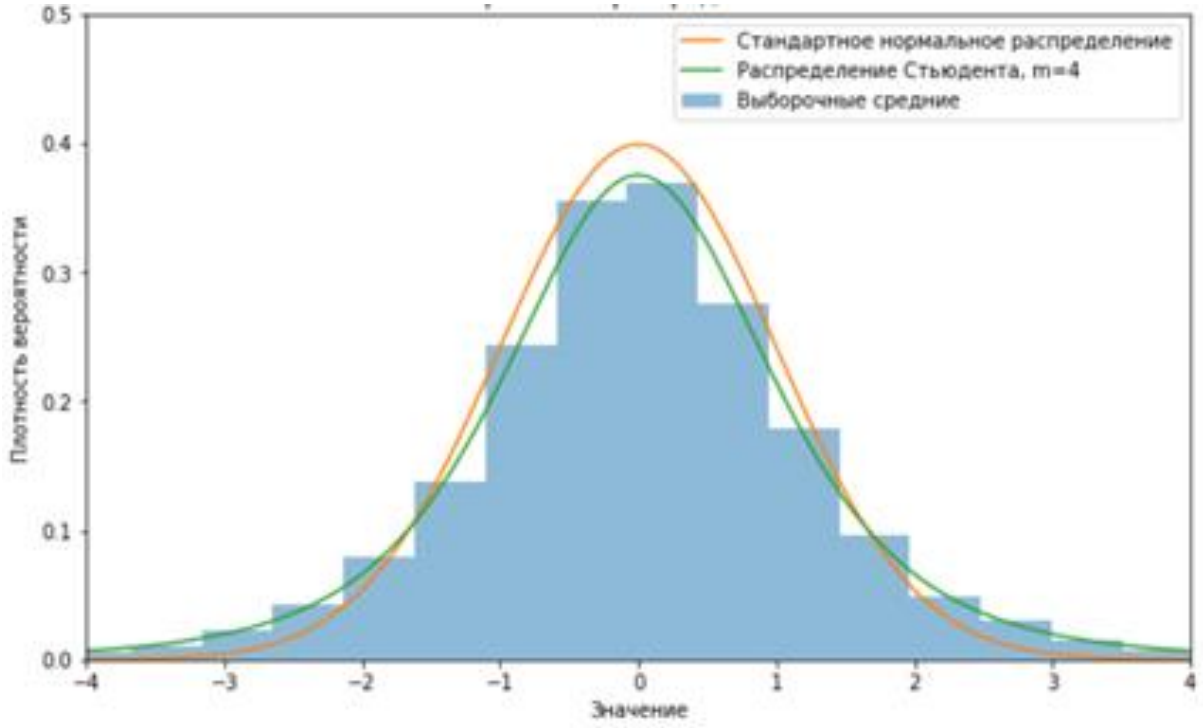
Распределение Стьюдента, или t-распределение, не является стандартным нормальным распределением. Распределение Стьюдента используют, когда стандартное отклонение генеральной совокупности неизвестно. Количество степеней свободы для t-распределения обычно определяется как размер выборки минус один $(n - 1)$. С увеличением размера выборки t-распределение становится ближе к стандартному нормальному распределению. «Тяжелые» хвосты t-распределения означают, что вероятности крайне больших или малых значений больше, чем у стандартного нормального распределения. Это и объясняет более медленное убывание его плотности к нулю на краях распределения. Как и стандартное нормальное распределение, распределение Стьюдента симметрично относительно нуля. Поэтому его математическое ожидание тоже равно нулю.

Распределение Стьюдента точнее оценивает распределение. Особенно ярко это видно в области хвостов. Малый объем выборки увеличивает вероятность получить значительное отклонение статистики от нуля, и распределение Стьюдента это хорошо отражает.

Сравнение распределений изображено далее на рисунке.

Таким образом, распределение Стьюдента позволяет адаптироваться к неизвестному стандартному отклонению генеральной совокупности и обеспечить точные и надежные оценки доверительных интервалов даже

при относительно малых размерах выборки. Поэтому распределение Стьюдента часто используют в статистическом анализе, когда сталкиваются с ограниченными данными и неизвестным стандартным отклонением.



Случайная величина $T = \dfrac{\overline{X}_n - \mu}{S_n / \sqrt{n}}$ будет иметь распределение Стьюдента, если распределение выборочных средних $\overline{X}_n$ имеет нормальное распределение.

Для построения доверительного интервала нужны критические значения.

Введем границы:

$$P(a \le T \le b) = P(T \le b) - P(T \le a),$$

где a и b – **критические значения**, $a = t_{(n-1),\,\alpha/2}$ – нижняя граница; $b = t_{(n-1),\,1-\alpha/2}$ – верхняя граница.

У t появился дополнительный индекс, который показывает степени свободы.

Границы – это квантили. Они показывают вероятность, что случайная величина, которая имеет распределение Стьюдента с $n-1$ степенями

свободы, примет значение меньше или равное заданному. Это значит, что

$$P(T \le t_{n-1,\alpha/2}) = \alpha/2,$$

$$P(T \le t_{n-1,1-\alpha/2}) = 1-\alpha/2.$$

Найдем критические значения для этого распределения с помощью Python. В модуле `stats` из библиотеки `SciPy` распределение Стьюдента обозначается лаконично: *t* – по второму названию распределения.

Для примера вычислим значение $t_{10,\,0.01}$.

Результат: –2.7637694574478893.

```
from scipy import stats

# Определим случайную величину.
# С помощью параметра df передадим количество степеней свобод
T = stats.t(df=10)

# Вычислим значение t_10,0.01.
t_001 = T.ppf(q=0.01)

# Выведем полученное значение.
print(t_001)
```

Пусть $x_1, \ldots, x_n$ – случайная выборка из произвольного распределения, где параметр σ_X^2 известен, а распределение выборочных средних является нормальным. Тогда для заданного уровня значимости $\alpha \in (0, 1)$ **доверительный интервал** для μ_X имеет вид

$$\left(\bar{x}_n - t_{(n-1),1-\alpha/2}\frac{S_n}{\sqrt{n}},\ \bar{x}_n - t_{(n-1),\alpha/2}\frac{S_n}{\sqrt{n}}\right).$$

Пусть $x_1, \ldots, x_n$ – случайная выборка из произвольного распределения, где параметр σ_X^2 известен, а распределение выборочных средних является нормальным. Тогда для заданного уровня значимости $\alpha \in (0, 1)$ **нижний односторонний доверительный интервал** для μ_X имеет вид:

$$\left(\bar{x}_n - t_{(n-1),1-\alpha} \frac{S_n}{\sqrt{n}},\ +\infty \right).$$

Верхний односторонний доверительный интервал для μ_X имеет вид:

$$\left(-\infty,\ \bar{x}_n - t_{(n-1),\alpha} \frac{S_n}{\sqrt{n}} \right).$$

2.4. ВВЕДЕНИЕ В ПРОВЕРКУ ГИПОТЕЗ

Интервальное оценивание исследует выборку и помогает сделать выводы о генеральной совокупности. Существует другой инструмент, который работает на эту же цель, но иначе – статистические тесты.

Всегда формулируются две гипотезы:

- **нулевая гипотеза** обозначается H_0;
- **альтернативная гипотеза** обозначается H_1 или H_a.

Нулевая гипотеза формулируется как утверждение, которое необходимо проверить. Чаще всего в качестве нулевой гипотезы формулируют утверждение про отсутствие связи, отсутствие отличий, отсутствие какого-то явления. Альтернативная гипотеза обычно соответствует предположению, что связь, отличия или какое-то явление – есть.

Нулевая гипотеза считается истиной «по умолчанию». Альтернативную гипотезу принимают, если находятся факты, противоречащие нулевой.

Односторонняя гипотеза предполагает утверждение «больше» или «меньше». А в **двусторонней** гипотезе обычно утверждается «равно» или «не равно». Какую именно формулировать гипотезу – двустороннюю или одностороннюю, а если одностороннюю, то в какую сторону – зависит от контекста.

Чтобы определить, верна ли нулевая гипотеза, нужно проанализировать выборку. Обычно выбирают одно число – статистику выборки

$T = f(X_1, \ldots, X_n)$ и значения этой статистики сравнивают с некоторым числом. На основе этого сравнения делают вывод по гипотезам.

Обычно для статистики характерно такое поведение:

– нулевая гипотеза верна – статистика принимает умеренные значения;

– нулевая гипотеза неверна – статистика принимает большие или малые по модулю значения.

Правило, по которому гипотезу на основе реализации выборки отклоняют или не отклоняют, называется **статистическим критерием**.

Обычно статистический критерий формулируют в виде математического выражения, где статистику сравнивают с некоторым числом. Например, если выполнено условие $f(X_1, \ldots, X_n) > 0$, то нулевая гипотеза отклоняется, если не выполнено – не отклоняется. У этой статистики есть специальное название.

Пусть набор данных из элементов моделируется с помощью случайных величин $X_1, X_2, \ldots, X_n$. **Статистика критерия** – это любая выборочная статистика $T = f(X_1, \ldots, X_n)$, чье наблюдаемое значение t позволяет решить, стоит ли отвергнуть H_0.

T – это сумма независимых случайных величин, которые принимают значения 1 или 0. Выборочная статистика T подчиняется биномиальному распределению. Сумма случайных величин Бернулли всегда имеет распределение Binomial(n, p), где n – количество слагаемых, а p – вероятность успеха.

Критическая область – это множество $K \subset \boldsymbol{R}$, соответствующее всем значениям, для которых отвергаем H_0 в пользу H_1. Значения на границах критической области называют **критическими значениями**.

Иногда гипотезу отвергают или не отвергают на основе не критических значений, а некоторой вычисленной вероятности.

***P*-value**, или ***p*-значение** – вероятность получить такое или еще более экстремальное значение статистического критерия при условии, что нулевая гипотеза верна.

Другими словами, *p*-value – это вероятность того, что увидим изменения там, где их на самом деле нет.

Для ориентира при тестировании гипотез можно заранее установить отношение к ошибке. Таким отношением будет малое число α, например, 0,05, которое называется уровнем значимости.

Уровень значимости α – наибольшая допустимая вероятность, с которой можно позволить себе отвергнуть верную нулевую гипотезу.

Если *p*-value меньше уровня значимости, то принимаем решение, что надо отклонить нулевую гипотезу. А если *p*-value не меньше уровня значимости, то не отклоняем нулевую гипотезу.

Ситуация с некорректным отклонением нулевой гипотезы настолько важна в статистике, что имеет собственное название – ошибка первого рода.

Ошибка первого рода – ситуация, когда отвергнута верная нулевая гипотеза.

Вероятность такой ошибки обозначают как $P\,(T \in K \mid H_0)$:

– $T \in K$ – правило отклонения нулевой гипотезы;

– H_0 – условие, что нулевая гипотеза верна.

И эта вероятность – как раз то самое значение уровня значимости.

Вероятность ошибки первого рода и уровень значимости – это разные названия одной и той же вероятности.

Логично ожидать, что уровень значимости и размеры множества критических значений как-то связаны.

Представим, что уровень значимости маленький, т.е. готовность только на очень маленькую вероятность ошибки и желание быть уверенными в выводах. В таком случае множество критических значений будет

тоже маленьким. Ведь большую уверенность дадут только самые большие или самые маленькие наблюдаемые значения.

А если уровень значимости больше, т.е. готовность ошибаться с большей вероятностью, значит и множество критических значений будет больше.

Алгоритм тестирования гипотез.

1. Определить нулевую и альтернативную гипотезы:

– нулевая обозначается H_0;

– альтернативная обозначается H_1 или H_a.

2. Зафиксировать уровень значимости $\alpha \in (0, 1)$.

Уровень значимости α – наибольшая допустимая вероятность, с которой можно позволить себе отвергнуть верную нулевую гипотезу.

3. Выбрать статистику критерия T.

Статистика критерия $T = f(X_1, \ldots, X_n)$ – это любая выборочная статистика чье наблюдаемое значение t позволяет решить, стоит ли отвергнуть H_0.

4. На основе α определить критическую область для T.

Множество $K \subset \boldsymbol{R}$, соответствующее всем значениям, для которых отвергаем H_0 в пользу H_1, называют критической областью. Значения на границе критической области называют критическими значениями.

5. Сформулировать статистический критерий.

Статистика T и критическая область K задают вместе статистический критерий – правило, позволяющее принять или отвергнуть гипотезу на основе реализации выборки.

6. Найти значение T и сделать вывод:

– если $T = t \in K$ отвергаем H_0 в пользу H_1;

– иначе отвергнуть H_0 не можем.

Это общий сценарий, который применим для любого параметра генеральной совокупности. Но в нем не уточняется, например, как выбрать подходящую статистику T, для которой строится критическая область.

2.5. ТЕСТЫ ДЛЯ СРЕДНЕГО

На практике нередко бывает, что интересующая величина выражена в виде математического ожидания генеральной совокупности. И зачастую нужно выяснить, не отклоняется ли эта величина от некоторого конкретного значения.

Исследовать такое отклонение можно с помощью двух тестов: *Z*-тест и *T*-тест. Чтобы определить, какой тест лучше подходит в конкретном случае, нужно оттолкнуться от изначальных предположений о генеральной совокупности.

Предположим, есть выборка $X_1, X_2, \ldots, X_n$ из некоторого распределения. Математическое ожидание этого распределения неизвестно, поэтому проводим статистические тесты относительно его значения. Для этого можно рассмотреть одну из трех задач.

Первая заключается в том, чтобы протестировать следующий набор гипотез:

$H_0 : \mu = \mu_0$,

$H_1 : \mu \neq \mu_0$.

В этом случае нулевая гипотеза имеет простой вид, а альтернативная включает два варианта: $\mu > \mu_0$ **и** $\mu < \mu_0$. Такой тест называют **двусторонним**.

Двусторонние тесты проводят, чтобы выяснить, действительно ли среднее соответствует предположениям или оно отклоняется в бóльшую или в меньшую сторону.

Второй вариант задачи заключается в проверке вот таких гипотез:

$H_0 : \mu = \mu_0$,

$H_1 : \mu > \mu_0$.

Здесь альтернативная гипотеза дана в виде неравенства.

Третий вариант задачи похож на второй:

$H_0 : \mu = \mu_0$,

$H_1 : \mu < \mu_0$.

Во втором и третьем случае нулевая гипотеза тоже имеет простой вид, но при этом альтернативная включает только один из двух вариантов: $\mu > \mu_0$ **или** $\mu < \mu_0$. Такой тест называют **односторонним**.

Чтобы определить статистику критерия, в каждом из трех случаев применяют выборочное среднее:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Какой вид примет статистика критерия, зависит от того, известна ли в задаче **дисперсия генеральной совокупности**.

Когда известна дисперсия, применяется *Z*-тест. Если знаем, что $Var(X_i) = \sigma^2$, то, предполагая H_0, определяем статистику как

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}.$$

Когда неизвестна дисперсия, применяется *T*-тест. Тогда используют

$$Z = \frac{\overline{X} - \mu_0}{S / \sqrt{n}},$$

где *S* – выборочное стандартное отклонение несмещенной дисперсии.

В каждом из вариантов можно найти распределение статистики и таким образом построить критерий для тестов.

Одновыборочный *Z*-тест для среднего.

Предпосылки:

– $X_i \sim N(\mu, \sigma^2)$ и σ известна или

– $n > 30$ и σ известна.

Этапы:

1. Определить нулевую и альтернативную гипотезы для μ.
2. Зафиксировать уровень значимости $\alpha \in (0, 1)$.
3. Статистика: $Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,\ 1)$. Статистика критерия имеет **стандартное нормальное распределение**.
4. Определяем критическую область:

а) если $H_0 : \mu = \mu_0$, то $K = (-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, +\infty)$;

б) если $H_1 : \mu > \mu_0$, то $K = (z_\alpha, +\infty)$;

в) если $H_1 : \mu < \mu_0$, то $K = (-\infty, z_\alpha)$.

5. Если $Z = z \in K$, отвергаем H_0 в пользу H_1. Иначе не можем отвергнуть H_0.

Множество значений статистики критерия, которые не входят в критическую область K, называют **допустимой областью**.

Рассмотрим это множество подробнее.

Не отклоняем гипотезу в двустороннем тесте, если

$$\frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} < z_{\alpha/2}.$$

Можно переписать это условие как

$$\mu_0 \in \left(\overline{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right).$$

Это $(1 - \alpha)\cdot 100\%$ доверительный интервал для μ_0.

Между доверительными интервалами и тестированием гипотез действительно есть связь.

Для двустороннего теста отклоняют гипотезу H_0 тогда и только тогда, когда значение μ_0 не включено в $(1 - \alpha)\cdot 100\%$ двусторонний доверительный интервал для μ.

Для одностороннего теста отклоняют гипотезу H_0 тогда и только тогда, когда значение μ_0 не включено в $(1 - \alpha)\cdot 100\%$ односторонний доверительный интервал для μ.

Более того: на эти факты можно опираться, чтобы определить регион значений μ_0, для которых нулевая гипотеза H_0: $\mu = \mu_0$ не будет отклонена на уровне α.

Одновыборочный *T*-тест для среднего.

Предпосылки:

– $X_i \sim N(\mu, \sigma^2)$ и σ неизвестна или

– $n \geqslant 100$ и σ неизвестна.

Этапы:

1. Определить нулевую и альтернативную гипотезы для μ.
2. Зафиксировать уровень значимости $\alpha \in (0, 1)$.
3. Статистика: $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t(n - 1)$. Статистика критерия имеет **распределение Стьюдента с** $n - 1$ **степенями свободы.** S – выборочное стандартное отклонение.
4. Определяем критическую область:

а) если $H_0 : \mu = \mu_0$, то $K = (-\infty, -t_{\alpha/2}] \cup [t_{\alpha/2}, +\infty)$;

б) если $H_1 : \mu > \mu_0$, то $K = [t_\alpha, +\infty)$;

в) если $H_1 : \mu < \mu_0$, то $K = (-\infty, t_\alpha]$.

5. Если $T = t \in K$, отвергаем H_0 в пользу H_1. Иначе не можем отвергнуть H_0.

В обоих тестах результатом является отклонение или неотклонение нулевой гипотезы на основе критической области.

***p*-значение (*p*-уровнь значимости, фактический уровень значимости)** – наименьший уровень значимости α, на котором отклоняется нулевая гипотеза.

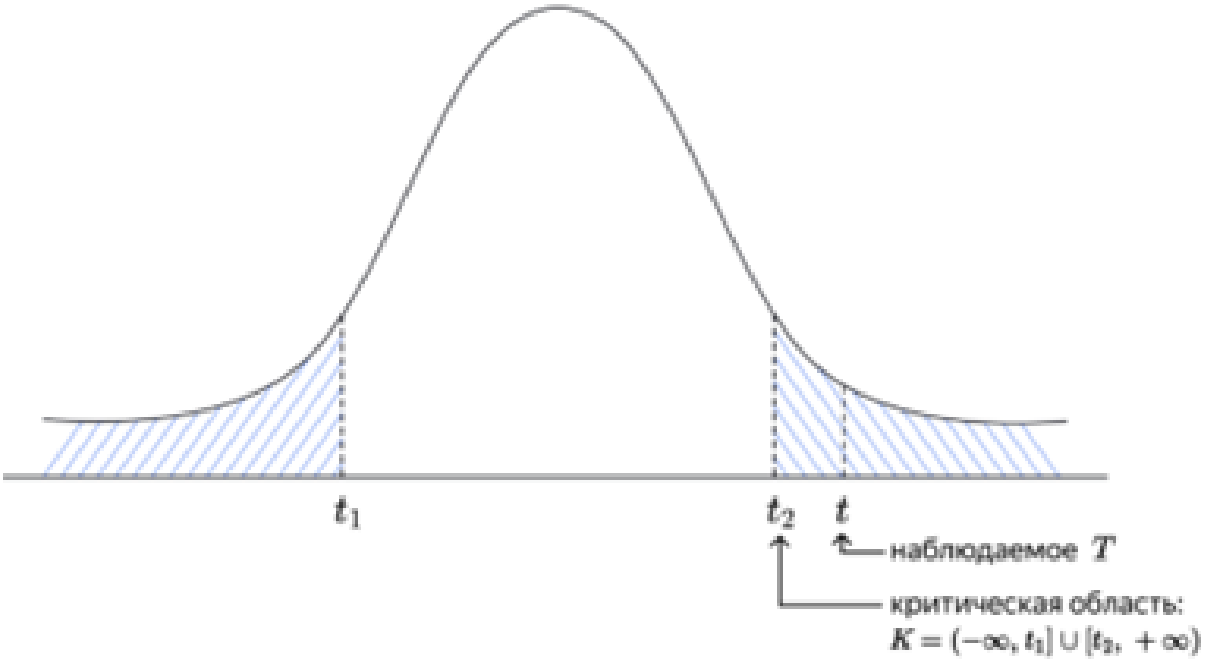
Другими словами, это вероятность получить значение статистики критерия как наблюдаемое или больше при условии, что нулевая гипотеза верна. Получается, при маленьком *p*-значении полученное значение статистики критерия очень маловероятно, если нулевая гипотеза верна. Поэтому готовы отклонить нулевую гипотезу.

Если проводим тест на уровне значимости α, используя некоторую статистику критерия T, то

$$T = t \in K \Leftrightarrow (p \text{ – значение для } t) \leqslant \alpha.$$

График ниже иллюстрирует тест, где значения статистики критерия T большие t_2 или меньшие t_1, определяют критическую область, где отвергаем нулевую гипотезу. В этом случае *p*-значение соответствует вероятности $P(|T| \geqslant t \mid H_0)$. А закрашенная область определяет α.

Распределение T при условии H_0:



Можно определить, отвергаем ли H_0 на уровне значимости α, сравнив t_2 и t (или, эквивалентно, α и p-значение). Поэтому p-значение иногда называют наблюдаемым уровнем значимости.

2.6. A/B-ТЕСТИРОВАНИЕ

A/B-тестирование – это исследовательский подход к изучению двух вариантов чего-либо в целях выбора более эффективного. Идея в том, что аудитория случайным образом делится на две группы, каждой из которых показывают разные варианты.

Для **контрольной группы A** все остается без изменений, а **исследуемая (экспериментальная) группа B** видит обновленную версию продукта. Сопоставив данные по двум группам, аналитик оценивает, значимо ли потенциальные изменения влияют на целевые метрики. В зависимости от результата принимается решение: внедрять ли новое или оставить все как есть.

A/B-тесты широко распространены в веб-дизайне и маркетинге, поскольку в этих сферах результаты легко поддаются измерению.

***Z*-тест для разницы средних из распределений с известными дисперсиями.**

Предположим, что есть независимые случайные выборки $X_1, \ldots, X_{nx}$ и $Y_1, \ldots, Y_{ny}$. Если

1) $X_i \sim N(\mu_x, \sigma_x^2)$ и $Y_i \sim N(\mu_y, \sigma_y^2)$, σ_x^2, σ_y^2 – известны или

2) $n_x > 30$ и $n_y > 30$, и σ_x^2, σ_y^2 – известны,

то на уровне значимости α можно провести тест

$H_0 : \mu_x = \mu_y$,

$H_1 : \mu_x \neq \mu_y$.

При расчете статистики для *Z*-теста применяют известные дисперсии. Когда тест проводят для двух выборок – учитывают две дисперсии, и в этом отличие от одновыборочного теста.

Тогда статистикой критерия будет

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}.$$

Нулевую гипотезу отклоняют, если $|T| > z_{\alpha/2}$.

Если $H_1 : \mu_x > \mu_y$, тогда критерий $T > z_\alpha$;

если $H_1 : \mu_x < \mu_y$, тогда критерий $T < z_\alpha$.

На практике дисперсии чаще неизвестны, и в таком случае *T*-тест подходит лучше.

Все зависит от того, можно ли предположить равные или неравные дисперсии в распределениях генеральной совокупности для каждой выборки. Для равных и неравных будут немного разные статистики критерия.

***T*-тест для разницы средних из распределений с равными дисперсиями.**

Предположим, что есть независимые случайные выборки $X_1, \ldots, X_{nx}$ и $Y_1, \ldots, Y_{ny}$. Если

1) $X_i \sim N(\mu_x, \sigma_x^2)$ и $Y_i \sim N(\mu_y, \sigma_y^2)$, $\sigma_x^2 = \sigma_y^2$ – неизвестны или

2) $n_x > 100$ и $n_y > 100$, и $\sigma_x^2 = \sigma_y^2$ – неизвестны,

то на уровне значимости α можно провести тест

$H_0 : \mu_x = \mu_y$,

$H_1 : \mu_x \neq \mu_y$.

Когда дисперсии распределений генеральных совокупностей для выборок неизвестны, учитывают две выборочные дисперсии в статистике. Правда, немного иначе, чем в *Z*-тесте.

Пусть $s = \sqrt{\dfrac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}}$.

Тогда статистикой критерия будет

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{s^2}{n_x} + \dfrac{s^2}{n_y}}}.$$

Нулевую гипотезу отклоняют, если $|T| > t_{\alpha/2,\ nx+\ ny-2}$.

Если $H_1 : \mu_x > \mu_y$, тогда критерий $T > t_{\alpha,\ nx+\ ny-2}$;

если $H_1 : \mu_x < \mu_y$, тогда критерий $T < t_{\alpha,\ nx+\ ny-2}$.

***T*-тест для разницы средних из распределений с неравными дисперсиями.**

Предположим, что есть независимые случайные выборки $X_1, \ldots, X_{nx}$ и $Y_1, \ldots, Y_{ny}$. Если

1) $X_i \sim N(\mu_x, \sigma_x^2)$ и $Y_i \sim N(\mu_y, \sigma_y^2)$, $\sigma_x^2 \neq \sigma_y^2$ – неизвестны или

2) $n_x > 100$ и $n_y > 100$, и $\sigma_x^2 \neq \sigma_y^2$ – неизвестны,

то на уровне значимости α можно провести тест

$H_0 : \mu_x = \mu_y$,

$H_1 : \mu_x \neq \mu_y$.

Несмотря на то, что здесь более слабые предположения, статистика критерия считается проще. Как в *Z*-тесте, используем только выборочные дисперсии.

Статистикой критерия будет

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_x^2}{n_y}}}.$$

Нулевую гипотезу отклоняют, если $|T| > t_{\alpha/2,\ \min\{nx,\ ny\}\ -1}$.

Если $H_1 : \mu_x > \mu_y$, тогда критерий $T > t_{\alpha,\ \min\{nx,\ ny\}\ -1}$;

если $H_1 : \mu_x < \mu_y$, тогда критерий $T < t_{\alpha,\ \min\{nx,\ ny\}\ -1}$.

При проведении теста есть вероятность столкнуться с ошибкой первого рода.

Ошибка второго рода возникает, если ошибочно **не** отклонить нулевую гипотезу. Это значит, что на самом деле тестируемая H_0 неверна, но на основе наших данных ее не отвергли.

Вероятность ошибки второго рода обозначается греческой буквой β.

Чтобы контролировать вероятность возникновения ошибки *первого* рода, нужно установить уровень значимости, который будет ограничивать ее сверху. А еще можно вычислить *p*-значение для теста: так узнаем наименьший уровень значимости, при котором можно отклонить нулевую гипотезу.

Нулевую гипотезу отвергают, когда есть веские доказательства в пользу альтернативной гипотезы. Но если отвергнуть нулевую не удалось, то тут одно из двух:

- либо нулевая гипотеза верна,
- либо допущена ошибка второго рода.

Мощность теста – это вероятность, что тест корректно отклонит нулевую гипотезу, когда альтернативная гипотеза верна.

Чтобы найти мощность теста, нужно вычесть из единицы вероятность ошибки второго рода: 1 – β.

Если мощность одного теста на одних и тех же данных равна 0,6, а другого 0,91, то вероятнее, будет совершена ошибка второго рода в первом тесте. Низкая мощность снижает чувствительность теста к обнаружению статистически значимых эффектов. Но, как следствие, первый эксперимент меньше подвержен ошибке первого рода – случайному отклонению верной нулевой гипотезы.

Чтобы определиться с размером выборки, нужно зафиксировать некоторые параметры. Прежде всего – минимальный размер замеряемого эффекта.

Минимальный размер эффекта (***MDE*** от англ. Minimum Detectable Effect) – это наименьшая разница в результатах *A* и *B*, которую может обнаружить статистический критерий.

Следующие показатели, которые необходимо зафиксировать, – это допустимые вероятности ошибок первого и второго рода. Первое контролируют с помощью уровня значимости α. Второе можно рассматривать напрямую, но чаще применяют смежный показатель – мощность. Чем больше выбранная мощность, тем меньше вероятность ошибки второго рода.

Эти показатели – вероятности ошибок первого и второго рода – напрямую связаны с размером выборки. Нельзя снизить выбранные вероятности, не меняя размер выборки, иначе тест попросту не будет улавливать нужных различий.

Найти минимальный размер выборки, который подойдет для конкретного теста, можно по формуле.

Опишем уравнением зависимость между всеми показателями для одностороннего теста:

$$\frac{MDE}{\sqrt{\frac{s_x^2 + s_y^2}{n}}} \geq z_\alpha + z_\beta.$$

Выражение слева похоже на статистику критерия. Разница в том, что в числителе не наблюдаемое значение различия, а желаемое. Это минимально допустимое значение альтернативной гипотезы.

Преодолев z_α, уже можно отклонить нулевую гипотезу. Но надо не просто преодолеть порог, а продвинуться чуть дальше – до нужной альтернативной гипотезы (что различие между группами не меньше *MDE*). За это и отвечает z_β.

Переорганизуем это выражение и получим универсальную формулу для нахождения размера выборки для *T*-теста.

Для **одностороннего *T*-теста** минимальный размер выборки определяется по формуле

$$n \geq \left(\frac{z_\alpha + z_\beta}{MDE} \right)^2 \left(s_x^2 + s_y^2 \right).$$

Для **двустороннего** –

$$n \geq \left(\frac{z_{\alpha/2} + z_\beta}{MDE} \right)^2 \left(s_x^2 + s_y^2 \right).$$

В случае с *Z*-тестом выборочные дисперсии просто заменяются на дисперсии распределений генеральных совокупностей.

3. МЕТОДЫ СТАТИСТИЧЕСКОЙ ПРОВЕРКИ ГИПОТЕЗ

3.1. ЛОГНОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ И НЕЛИНЕЙНОЕ ПРЕОБРАЗОВАНИЕ ДАННЫХ

Пусть распределение случайной величины задается плотностью вероятности, имеющей вид

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2} \right),$$

где $x > 0$, $\sigma > 0$.

Тогда величина X имеет логнормальное распределение с параметрами μ и σ: $X \sim LogN(\mu, \sigma^2)$.

В этой записи $LogN$ – название распределения, а не операция логарифмирования.

График логнормального распределения напоминает график нормального: у обоих графиков есть максимум, в обе стороны от которого функция стремится к нулю.

Но график нормального распределения стремится к нулю с одинаковой скоростью слева и справа. А у логнормального – по-разному: часть графика, которая находится левее максимума, очень быстро стремится к нулю, в то время как правая часть также стремится к нулю, но медленно.

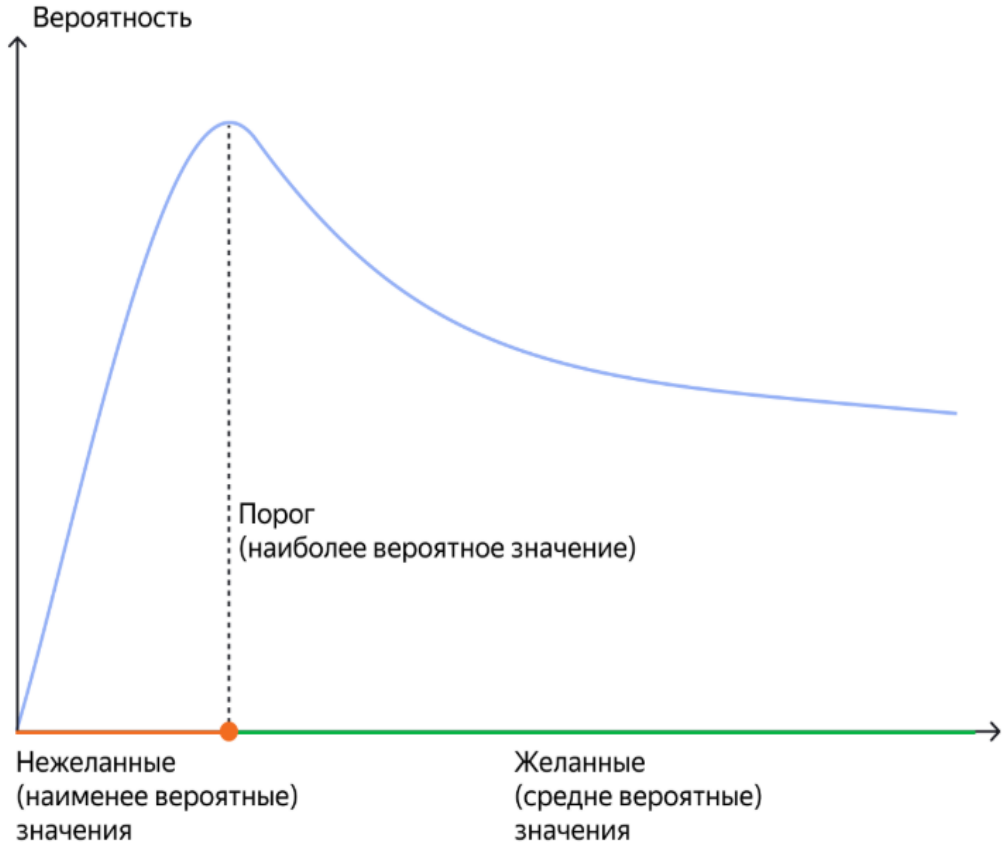
Поэтому такое распределение хорошо описывает процессы, когда:

– задано пороговое значение величины. Обычно есть план действий, чтобы его достичь, поэтому вероятность именно этого значения максимальна;

– отклонение в меньшую сторону очень нежелательно. Чем сильнее отклонение, тем меньше его вероятность;

– превышение порога требует сложных действий, поэтому эта ситуация менее вероятна, чем достижение порогового значения. Но эта

ситуация желанна, поэтому вероятность отклонения от порогового значения в большую сторону падает медленнее, чем отклонение в меньшую сторону.



Если величина $X \sim LogN(\mu, \sigma^2)$, то величина $Y = \ln(X)$ распределена $Y \sim N(\mu, \sigma^2)$. В обратную сторону это тоже работает: если $X \sim N(\mu, \sigma^2)$, то величина $Y = \exp(X)$ распределена $Y \sim LogN(\mu, \sigma^2)$.

Рассмотрим пример.

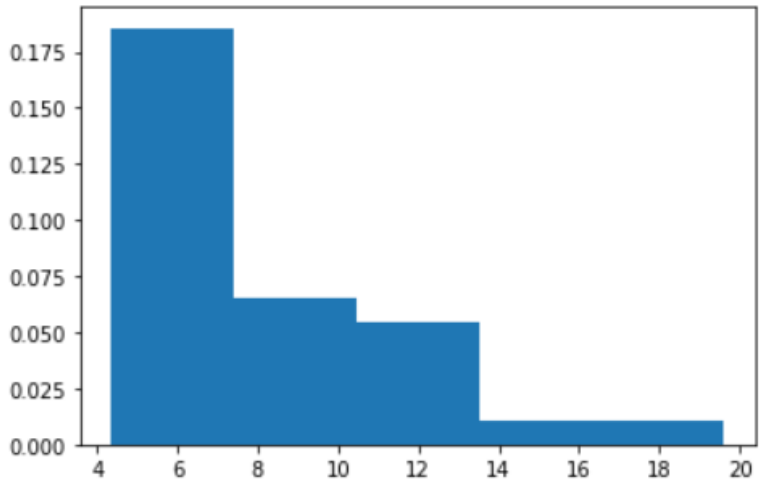
Есть набор данных:

X = [4.9406, 11.1565, 6.11626, 19.6103, 6.12839, 4.3416, 9.06663, 6.56027, 5.08844, 7.23608, 9.59747, 8.27959, 6.26848, 4.80749, 6.52615, 7.47267, 15.7362, 7.35348, 12.7505, 10.2956, 7.16994, 10.921, 12.022, 6.47511, 12.8345, 4.76712, 5.13015, 8.11151, 4.86316, 6.70681]

Построим гистограмму полученного набора данных.

В качестве первого подхода к анализу данного распределения определим его математическое ожидание. Заметим на гистограмме, что наибо-

лее вероятное значение этого распределения лежит в диапазоне [6, 8]. Найдем это значение.



Здесь объем выборки мал, при этом по гистограмме видно, что исходное распределение не нормальное, его дисперсия неизвестна. Значит тест покажет неточный результат и нет смысла его применять.

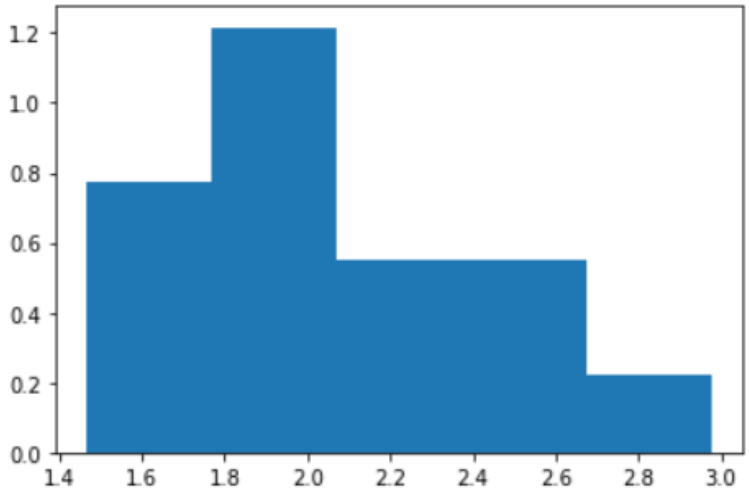
Прологарифмируем каждое значение исходной выборки и проанализируем полученные значения.

Метод преобразования данных, когда от исходной выборки переходят к выборке логарифмов этих данных, называется ***Log*-преобразованием**.

Чтобы совершить *Log*-преобразование с помощью Python, используем функцию `log` из библиотеки `NumPy`.

```
from scipy import stats
import matplotlib.pyplot as plt
import numpy as np

X = [4.9406, 11.1565, 6.11626, 19.6103, 6.12839, 4.3416, 9.06663, 6.56027,
     5.08844, 7.23608, 9.59747, 8.27959, 6.26848, 4.80749, 6.52615,
     7.47267, 15.7362, 7.35348, 12.7505, 10.2956, 7.16994, 10.921, 12.022,
     6.47511, 12.8345, 4.76712, 5.13015, 8.11151, 4.86316, 6.70681]

LogX = np.log(X)

plt.hist(LogX, density=True, bins=5)
plt.show()
```

Результат:



Видно, что полученное распределение похоже на нормальное.

Log-преобразование помогает в случаях, когда исходное распределение сильнее «скошено» влево. Он может сделать данные более похожими на нормально распределенные.

Математическое ожидание логарифма случайной величины не равно логарифму математического ожидания:

$$\ln E[\xi] \neq E[\ln \xi].$$

На графике видно, что математическое ожидание логарифмированной выборки отличается от логарифма математического ожидания исходной выборки.

Log-преобразование часто используют, чтобы проверить, равны ли математические ожидания двух совершенно не нормальных выборок.

По значению p_{value} можно сделать выводы:

– если $p_{value} < 0.05$, можно утверждать, что средние значения двух распределений различны;

– если $p_{value} > 0.05$, нельзя утверждать, что средние значения двух распределений различны, а значит, они могут быть равны.

3.2. НЕПАРАМЕТРИЧЕСКИЕ ТЕСТЫ

Тесты называют параметрическими, если они основаны на предположениях о параметрах генеральной совокупности и распределениях, из которых берутся данные:

– нормальность. Данные должны быть нормально распределены или иметь достаточный размер, чтобы использовать центральную предельную теорему;

– независимость. Данные должны быть реализацией случайной выборки;

– отсутствие выбросов. В данных не должно быть сильно больших или малых значений по сравнению с другими значениями выборки;

– равенство дисперсий. В большинстве тестов предполагают, что дисперсии анализируемых выборок примерно равны.

Существует и другой тип тестов – непараметрические. Они ничего не предполагают о параметрах генеральной совокупности. Эти тесты используют, если параметрические условия для тестирования параметров распределения не соблюдаются. Каждый параметрический тест имеет свой непараметрический эквивалент.

У непараметрических тестов есть недостаток: обычно непараметрические тесты имеют меньшую статистическую мощность, чем параметрические.

Это значит, что, когда нулевая гипотеза ложна, вероятность принять правильное решение – ниже. Большая статистическая мощность означает, что если существенные различия есть, то тест с большей вероятностью их обнаружит. Чтобы увеличить мощность непараметрического теста, берут выборку большего размера.

Результаты непараметрических тестов сложнее интерпретировать.

Многие непараметрические тесты используют не фактические данные, а значения, ранжированные по определенному правилу. При решении бизнес-задачи это бывает менее интуитивно понятно или полезно, чем работа с фактическими данными.

Если данные позволяют использовать параметрический тест, лучше применить его. Если данные не подходят для параметрического теста, то используют непараметрический аналог.

Параметрические тесты используют, когда:

– данные порождены ненормальным распределением, но выборка достаточно велика. *T*-тест с одной выборкой применяют, когда размер выборки больше 30;

– важна интерпретируемость результатов;

– важна большая статистическая мощность.

Непараметрические тесты используют, когда:

– объем выборки небольшой или данные не соответствуют нормальному распределению;

– размер выборки небольшой и нет уверенности в нормальности данных;

– область исследования лучше всего представлена медианой, а не средним;

– нужно проанализировать качественные данные, в которых может быть установлен порядок или когда выбросы нельзя удалить из данных.

Если в наборе данных есть выбросы, то *t*-тест применить не получится. Здесь можно использовать ***U*-критерий Манна–Уитни**.

***U*-критерий Манна–Уитни** – непараметрический статистический критерий, который используется для оценки различий между двумя независимыми выборками по уровню признака, измеренного количественно.

Статистика критерия: $U = \min\{U_1, U_2\}$, где

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1,$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2,$$

n_1, n_2 – количество значений в каждой из выборок; R_1, R_2 – ранги выборок.

При проведении теста Манна–Уитни

- на уровне значимости α;
- и с размерами выборок n_1 и n_2

нулевая гипотеза о том, что две выборки имеют одинаковое распределение и, как следствие, одинаковые средние, отвергается, если

$$U < U_a(n_1, n_2).$$

Иначе не можем утверждать, что две выборки различаются.

Краткий алгоритм расчета рангов в тесте Манна–Уитни:

1. Сформировать один общий список значений из обеих выборок.
2. Отсортировать значения по возрастанию.
3. Пронумеровать отсортированные значения.
4. Повторяющимся значениям назначить среднее арифметическое их рангов. Для остальных значений ранг совпадает с номером.

Чтобы перепроверить расчеты, сложите все получившиеся ранги и сравните их с суммой всех порядковых номеров. Если ошибок в расчетах нет, то суммы получатся одинаковыми.

Алгоритм применения теста с помощью *U*-критерия Манна–Уитни:

1. Принять в качестве нулевой гипотезы H_0, что две выборки, размерами n_1 и n_2, порождены одинаковым распределением.
2. Рассчитать ранги для обеих выборок.

3. Определить статистику критерия U, как минимальное из значений U_1 и U_2, где

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1,$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2.$$

4. Определить с помощью статистической таблицы критическое значение $U_\alpha(n_1, n_2)$.

5. Если $U < U_\alpha(n_1, n_2)$, то нулевая гипотеза об одинаковости распределений отвергается на уровне значимости α. Иначе сделать вывод о различиях в распределениях нельзя.

***T*-критерий Вилкóксона** – непараметрический статистический тест, который используют для проверки различий между двумя выборками парных измерений по уровню какого-либо количественного признака. Шкала измерений признака может быть непрерывной или порядковой.

Статистика критерия: $T = \min\{T_1, T_2\}$,

где T_1 – сумма рангов модулей отрицательных разностей соответствующих значений выборок, T_2 – сумма рангов положительных разностей соответствующих значений выборок.

При проведении теста Вилкоксона

– на уровне значимости;

– и с количеством ненулевых разностей

нулевая гипотеза о том, что две выборки имеют одинаковое распределение и, как следствие, одинаковые средние, отвергается, если

$$T < T_a(n).$$

Иначе не можем утверждать, что две выборки различаются.

Краткий алгоритм расчета рангов:

1. Вычислить разности между значениями до и после.
2. Найти модули получившихся разностей.
3. Отсортировать значения модулей по возрастанию.

4. Пронумеровать отсортированные модули.

5. Если модуль разности равен 0, то ранг ему не присваивается. Повторяющимся модулям назначить среднее арифметическое их рангов. Для остальных значений ранг совпадает с номером.

Чтобы перепроверить расчеты, сложите все получившиеся ранги и сравните их с суммой всех порядковых номеров для ненулевых разностей. Если ошибок в расчетах нет, то суммы получатся одинаковыми.

Алгоритм применения теста с помощью критерия Вилкоксона:

1. Принять в качестве нулевой гипотезы H_0, что две выборки одинакового размера порождены одним распределением.

2. Рассчитать ранги для обеих выборок. Определить n – количество элементов, имеющих ранг.

3. Определить статистику критерия T как минимальную из значений T_1 и T_2, где T_1 – сумма рангов для отрицательных разностей, а T_2 – сумма рангов для положительных разностей.

4. Определить с помощью статистической таблицы критическое значение $T_\alpha(n)$.

5. Если $T \leqslant T_\alpha(n)$, то нулевая гипотеза об одинаковости распределений отвергается на уровне значимости α. Иначе сделать вывод о различиях в распределениях нельзя.

3.3. БУТСТРЕП

Эмпирическая функция распределения, построенная по выборке $x_1, \ldots, x_n$ – это функция $F_n(x)$, определяемая формулой

$$F_n(x) = \frac{1}{n} \sum_{x_i \leq x} 1.$$

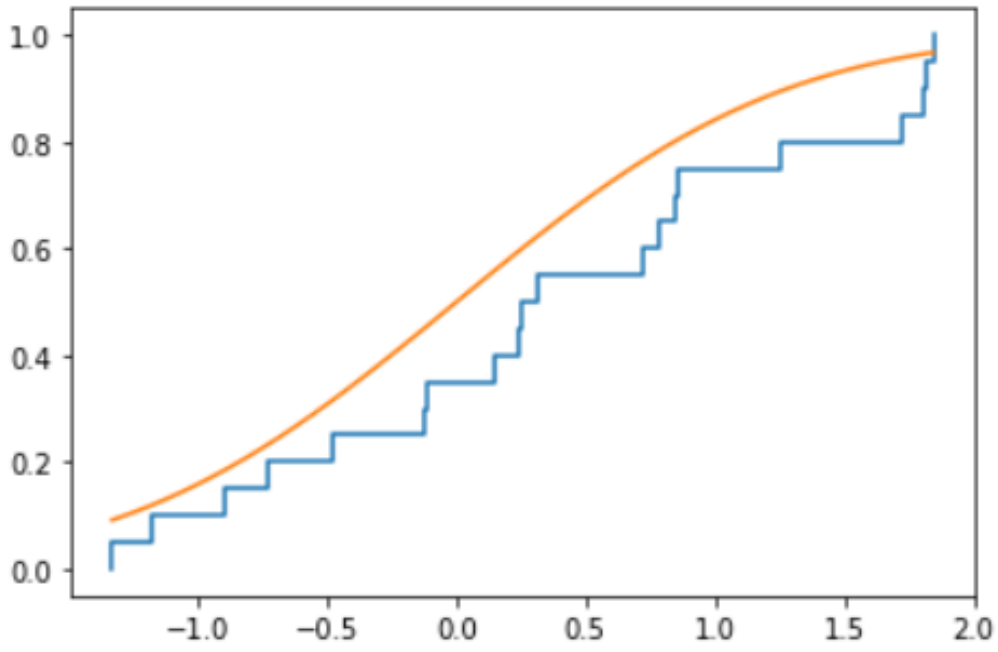
Эмпирическая функция распределения важна тем, что при увеличении размера выборки она стремится к функции распределения случайной величины, породившей эту выборку.

Запишем формально.

Если выборка $x_1, \ldots, x_n$ порождена случайной величиной X с функцией распределения $F_X(x)$, то при каждом $x \in \mathbf{R} : F_n(x) \to F_X(x)$, при $n \to \infty$.

Получается, что эмпирическая функция распределения позволяет приближенно восстановить искомую функцию распределения. И чем больше n, тем точнее это приближение.

Рассмотрим пример. Пусть дана выборка из значений. Посмотрим, как приближенно восстановить функцию распределения случайной величины, породившей эту выборку. Используем Python.



При маленькой исходной выборке эта функция несильно похожа на истинную функцию распределения, но все же она неплохо приближает исходное распределение.

Бутстреп – метод, который позволяет на основе исходной выборки получить множество новых наборов данных с помощью эмпирической функции распределения.

Чтобы оценка, вычисленная с помощью бутстрепа, была несмещенной, необходимо создавать выборки такого же объема, как и объем исход-

ной выборки. Количество итераций бутстрепа рекомендуют брать в диапазоне от 1000 до 10 000.

Чтобы провести полный анализ некоторой статистики случайной величины с помощью бутстрепа, необходимо:

1) создать набор выборок с помощью бутстрепа;

2) в каждой выборке определить значение интересующей нас характеристики;

3) собрать все полученные значения;

4) для полученного набора определить доверительный интервал.

Бутстреп плохо работает в задачах, когда набор данных большой или порожден сложным распределением.

3.4. МНОЖЕСТВЕННАЯ ПРОВЕРКА ГИПОТЕЗ

Проблема множественных сравнений – эффект, возникающий в результате тестирования нескольких статистических гипотез одновременно. При одновременной проверке большого количества гипотез вероятность совершить хотя бы одну ошибку первого рода, т.е. ошибочно отвергнуть верную нулевую гипотезу, сильно возрастает.

Количество ошибок первого рода в n экспериментах – это случайная величина, обозначим ее V. Ее значение надо взять под контроль: гарантировать, что оно не будет большим.

Но работать напрямую с V не очень удобно. Рассматривать все возможные варианты, где, как и сколько ошибок может возникнуть – это сложнее, чем контролировать вероятность хотя бы одной. Поэтому применяют специальные показатели, зависимые от V. Один из самых популярных вариантов – это FWER.

FWER (family-wise error rate, или групповая вероятность ошибки первого рода) – это вероятность совершить хотя бы одну ошибку первого рода:

$$FWER = P(V > 0).$$

Поправка Бонферрони – это самый простой метод противодействия проблеме множественных сравнений. Он позволяет определить уровень значимости для каждого эксперимента так, чтобы добиться желаемого условия: $FWER \leqslant \alpha$.

Для этого необходимо приравнять все уровни значимости к величине $\frac{\alpha}{n}$:

$$\alpha_1 = \alpha_2 = \cdots = \frac{\alpha}{n},$$

где $\frac{\alpha}{n}$ – заданная максимальная групповая вероятность ошибки первого рода, поделенная на количество экспериментов.

Тогда при тестировании каждой гипотезы надо сравнивать либо статистику критерия с критическим значением, зависящим от $\frac{\alpha}{n}$ вместо α_i, либо *p*-значение с $\frac{\alpha}{n}$ вместо α_i.

Есть еще один способ применить поправку Бонферрони при проверке гипотез с помощью *p*-значения: модифицировать не уровень значимости, а само *p*-значение. Формула

$$p_{new} = \min(1, n \cdot p_{old}).$$

Отклоним нулевую гипотезу, если $p \leqslant \alpha_i$. По изученной процедуре введения поправки Бонферрони новое правило отклонения будет выглядеть как $p \leq \frac{\alpha}{n}$.

Значит, чтобы добиться эквивалентного эффекта, изменяя *p*-значение, нужно умножить его на *n*. Тогда отклоним нулевую гипотезу, если $np \leqslant \alpha$.

4. МЕТОД ГЛАВНЫХ КОМПОНЕНТ

4.1. МАТРИЦА КОВАРИАЦИИ

Метод главных компонент представляет собой, по сути, сочетание линейной алгебры и статистики. Он применяется, когда данных очень много и связи между ними неочевидны. Суть подхода в том, чтобы выделить самые значимые данные и исследовать их. Тогда операции с матрицами совершаются быстрее, а результаты получаются нагляднее и проще для анализа.

Формулы ковариации:

$$Cov(x,\ y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y});$$

$$Cov(x,\ y) = Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Ковариация в Python. Чтобы реализовать математическую формулу в коде, бывает очень полезно описать порядок действий – как правило, в коде он сохраняется. Стоит учитывать особенности векторных операций в `NumPy`: одной командой можно выполнять операции сразу для всех чисел в векторе.

1) Считаем $\bar{x}, \bar{y}$.

```
x_mean, y_mean = x.mean(), y.mean()
```

2) Считаем $(x_i - \bar{x})$ для всех элементов x.

```
x_diff = x - x_mean
```

3) Считаем $(y_i - \bar{y})$ для всех элементов y.

```
y_diff = y - y_mean
```

4) Считаем произведения $(x_i - \bar{x})(y_i - \bar{y})$ для всех i.

```
prods = x_diff * y_diff
```

5) Считаем сумму полученных произведений и умножаем ее на $\frac{1}{n-1}$.

```
n = len(x)
cov = prods.sum() / (n - 1)
```

Итого:

```
x_mean, y_mean = x.mean(), y.mean()
x_diff = x - x_mean
y_diff = y - y_mean
prods = x_diff * y_diff
n = len(x)
cov = prods.sum() / (n - 1)
```

Получилось довольно объемно, можно это сжать. Большой разницы в читаемости между `x.mean()` и `x_mean` нет, уберем лишние переменные.

```
prods = (x - x.mean()) * (y - y.mean())
n = len(x)
cov = prods.sum() / (n - 1)
```

Обратим внимание, что сумма произведений компонент двух векторов, которую считаем, – это фактически скалярное произведение. Используем это наблюдение.

```
n = len(x)
cov = (x - x.mean()) @ (y - y.mean()) / (n - 1)
```

Таким образом, с помощью математики довольно сильно сократили код, при этом читаемость хуже не стала.

Матрица ковариации – матрица, содержащая попарные выборочные ковариации имеющихся наборов данных.

Матрица ковариации 2×2:

$$\begin{pmatrix} Var(x) & Cov(x,y) \\ Cov(x,y) & Var(y) \end{pmatrix}.$$

Матрица ковариации 2×2 в коде:

```
import numpy as np

X = np.array(
    [
        [1, 4],
        [3, 6],
        [2, 10],
        [3, 9],
        [2, 5],
        [0, 8],
        [1, 8],
        [1, 6],
        [2, 8],
        [0, 5],
    ]
)
x, y = X[:, 0], X[:, 1]
x_centered = x - x.mean()
y_centered = y - y.mean()
n = len(X)
cov = x_centered @ y_centered / (n - 1)
x_var = x_centered @ x_centered / (n - 1)
y_var = y_centered @ y_centered / (n - 1)
cov_mat = np.array([[x_var, cov], [cov, y_var]])
print(cov_mat)
```

Матрица ковариации 3×3 в коде (с использованием heatmap):

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# Получает массив из трех столбцов и создает
для него матрицу ковариации
def cov_matrix_3(X):
    x, y, z = X[:,  0], X[:, 1], X[:, 2]
    x_centered = x - x.mean()
    y_centered = y - y.mean()
    z_centered = z - z.mean()
    n = len(X)
    cov_xy = x_centered @ y_centered / (n - 1)
    cov_xz = x_centered @ z_centered / (n - 1)
```

```
        cov_yz = y_centered @ z_centered / (n - 1)
        x_var = x_centered @ x_centered / (n - 1)
        y_var = y_centered @ y_centered / (n - 1)
        z_var = z_centered @ z_centered / (n - 1)
        cov_mat = np.array([
            [x_var, cov_xy, cov_xz],
            [cov_xy, y_var, cov_yz],
            [cov_xz, cov_yz, z_var]
        ])
        return cov_mat

    meals = [1, 3, 2, 3, 2, 0, 1, 1, 2, 0]
    profile = [4, 6, 10, 9, 5, 8, 8, 6, 8, 5]
    meetings = [5, 2, 12, 6, 0, 3, 9, 2, 10, 7]

    data = np.array([meals, profile, meetings])   #
Создаем датасет.

    cov = cov_matrix_3(data.transpose())
    labs = ['meals', 'profile', 'meetings'] #
    Даем название строкам и столбцам.

    sns.heatmap(
        cov, # Матрица ковариации.
        annot=True, # Подписывать ли значения
    элементов матрицы.
        fmt='g', # Округление.
        xticklabels=labs, # Отсечки на оси X.
        yticklabels=labs  # Отсечки на оси Y.
    )
    plt.show()
```

4.2. МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Метод главных компонент – один из подходов к уменьшению размерности данных.

Часто размерность данных, с которыми работают аналитики, оказывается слишком большой, поэтому информацию сложно визуализировать. Одномерный массив можно представить гистограммой, двумерный – облаком точек, с трехмерным тоже можно что-то придумать. Наглядно изобразить пятимерный массив не получится. Аналитики решают эту задачу так: превращают массив большой размерности в массив меньшей размерности.

Например, для визуализации данных удобно свести многомерный массив к трехмерному или двумерному.

При таком преобразовании важно сохранить как можно больше информации. Один из подходов – построить из исходных признаков такое описание данных, которое сохранит наибольшее разнообразие объектов.

Для вычисления матрицы ковариации по формуле можно рассчитывать каждый элемент, а можно использовать более компактный способ – Python-метод `np.cov`:

```
cov_mat = np.cov(X)
```

Этот метод вычисляет дисперсии и ковариации строк матрицы *X*. Допустим, в задаче данные расположены по столбцам. Чтобы посчитать ковариацию столбцов, нужно добавить аргумент `rowvar=False`. Тогда метод сначала транспонирует массив, а затем вычислит дисперсии и ковариации транспонированной матрицы – ее строки как раз будут соответствовать столбцам исходной.

Другой способ вычислить ковариации столбцов – сразу передать в метод транспонированную матрицу.

```
cov_mat = np.cov(X, rowvar=False)
cov_mat = np.cov(X.T) # Равносильно
```

Чем больше собственное значение матрицы ковариации, тем больший разброс данных наблюдается вдоль оси, задаваемой соответствующим собственным вектором.

Единичные собственные векторы матрицы ковариации, задающие направления наибольшего распределения данных, называют главными компонентами.

Главные компоненты (Principal Components) – это ключевые элементы метода PCA (Principal Component Analysis), которые позволяют преобразовывать исходные данные в более удобную для анализа форму. Они представляют собой новые переменные, которые получаются в результате

линейного преобразования исходных данных. Главные компоненты устроены так, чтобы максимально упростить анализ и выделить наиболее важные закономерности.

PCA по шагам.

```
import numpy as np

data = np.array([
        [5, 12],
        [2, 5],
        [0, 2],
        [2, 10],
        [6, 25],
        [3, 14],
        [2, 12],
        [3, 23],
        [7, 26],
        [2, 16]
])
# Шаг 1. Вычесть среднее
data_meaned = data - np.mean(data, axis=0)
# Шаг 2. Вычислить матрицу ковариации
cov_mat = np.cov(data_meaned, rowvar = False)
# Шаг 3. Собственные значения и векторы
eigen_values , eigen_vectors =
np.linalg.eigh(cov_mat)
# Шаг 4. Сортировка
sorted_index = np.argsort(eigen_values)[::-1]
sorted_eigenvalue = eigen_values[sorted_index]
sorted_eigenvectors =
eigen_vectors[:,sorted_index]
# Шаг 5. Создать вектор-признак
n_components = 1 # Тут нужное количество
признаков
eigenvector_subset =
sorted_eigenvectors[:,0:n_components]
# Шаг 6. Новый датасет
data_reduced = data_meaned @ eigenvector_subset
print(data_reduced)
```

Сингулярное разложение матрицы (SVD) позволяет выделить главные компоненты, но без вычисления ковариационной матрицы.

PCA с помощью SVD.

```
from sklearn.decomposition import PCA
import numpy as np
import matplotlib.pyplot as plt

data = np.array([
        [5, 12],
        [2, 5],
        [0, 2],
        [2, 10],
        [6, 25],
        [3, 14],
        [2, 12],
        [3, 23],
        [7, 26],
        [2, 16]
])

data_centered = data - data.mean(0)

n_components = 2
U, S, Vt = np.linalg.svd(data_centered)
data_compressed = data_centered @ Vt.T[:,
0:n_components]

print(data_compressed)
```

4.3. ПРИМЕНЕНИЕ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ

PCA используют в практических задачах для повышения качества работы других алгоритмов.

PCA в `sklearn`.

Как и линейную регрессию, PCA не обязательно каждый раз воспроизводить своими руками по формулам. Можно использовать реализацию PCA из библиотеки машинного обучения `sklearn`.

Для примера, массив исходных данных запишем в `data`, количество компонент-столбцов, которое должно остаться после сжатия, обозначим как *k*.

Сначала импортируем класс PCA из подмодуля `decomposition`:

```
from sklearn.decomposition import PCA
```

Создадим специальный объект `pca` с помощью команды `PCA(n_components=k)`.

```
pca = PCA(n_components=3)
```

Объект pca готов к сжатию данных.

Чтобы сжать данные и уменьшить число столбцов, применим метод `pca.fit_transform(data)`. Он центрирует данные, найдет главные компоненты и спроецирует данные на оси. Метод вернет матрицу, количество строк у которой такое же, как у исходного массива `data`, а количество столбцов равно `n_components`.

```
from sklearn.decomposition import PCA
import numpy as np
import matplotlib.pyplot as plt

data = np.array([
        [5, 12],
        [2, 5],
        [0, 2],
        [2, 10],
        [6, 25],
        [3, 14],
        [2, 12],
        [3, 23],
        [7, 26],
        [2, 16]
])

pca = PCA(n_components=1)

data_compressed = pca.fit_transform(data)
print(data_compressed)
```

Результат:

```
[[  2.06608664]
 [  9.54010314]
 [ 12.89330406]
 [  4.65167289]
 [-10.85390341]
 [  0.5308573 ]
 [  2.69630079]
 [ -8.26831716]
 [-12.04166084]
 [ -1.21444342]]
```

Данные сжаты до одного столбца. Чтобы понять, что произошло, визуализируем исходные данные и их проекцию на главную компоненту. В этом помогут внутренние переменные объекта `pca`. Переменная `pca.components` содержит собственные векторы матрицы ковариации, `pca.explained_variance` хранит соответствующие собственные числа.

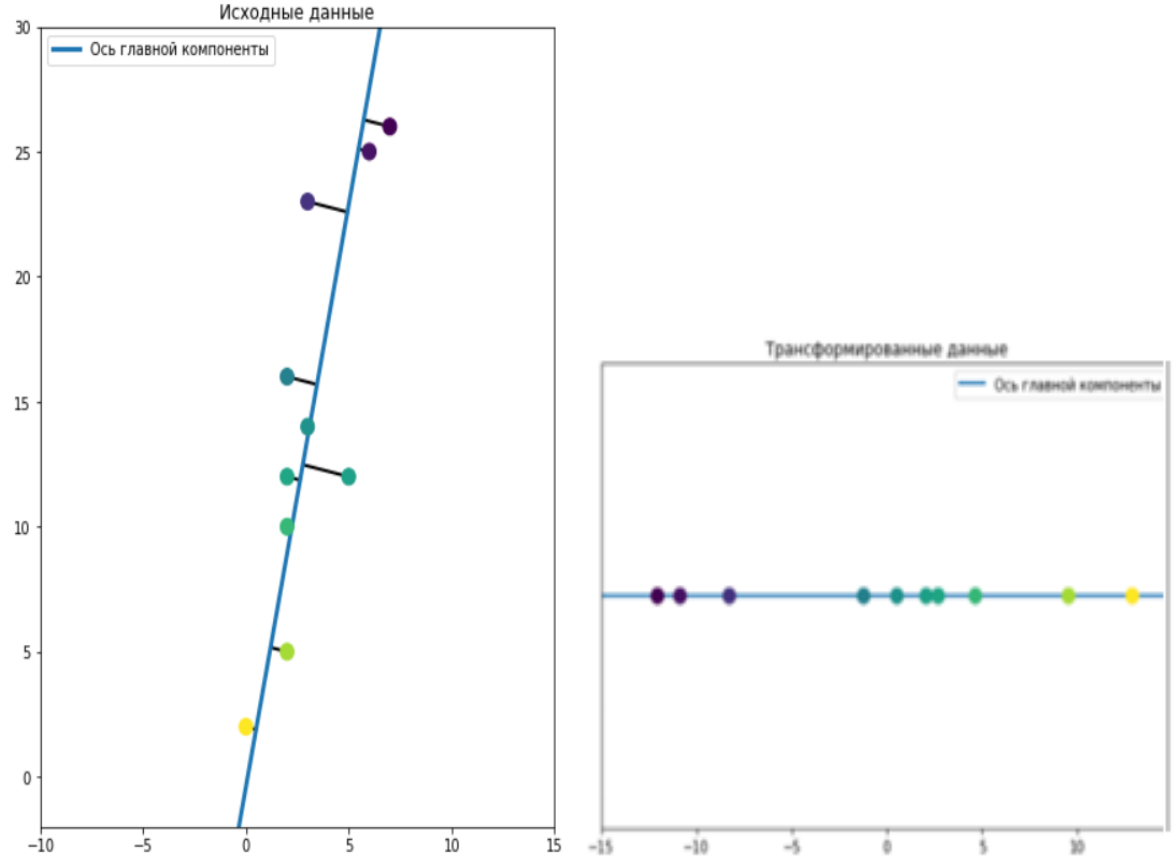
Код ниже строит графики.

```
from sklearn.decomposition import PCA
import numpy as np
import matplotlib.pyplot as plt

data = np.array([
        [5, 12],
        [2, 5],
        [0, 2],
        [2, 10],
        [6, 25],
        [3, 14],
        [2, 12],
        [3, 23],
        [7, 26],
        [2, 16]
])

# Создаем объект PCA
pca = PCA(n_components=1)

# Сжимаем данные.
```

```
    data_compressed = pca.fit_transform(data)

    # Строим собственный вектор, на который
    проецировались данные
    # Последний множитель помогает в красивой
    отрисовке
    magnificent_coef = 3
    principal_vector = pca.components_[0] *
    np.sqrt(pca.explained_variance_) *
    magnificent_coef

    # Рассчитываем координаты точек для визуализации
главной компоненты
    x_min = pca.mean_[0] - principal_vector[0]
    x_max = pca.mean_[0] + principal_vector[0]

    y_min = pca.mean_[1] - principal_vector[1]
    y_max = pca.mean_[1] + principal_vector[1]

    plt.figure(figsize=(18, 9))

    cmap = "viridis" # Задаем цветовую схему

    # Строим график с исходными данными
    plt.subplot(121)
    # Исходные данные.
    plt.scatter(data[:, 0], data[:, 1],
    c=data_compressed.flatten(), cmap=cmap, s=100)
    # Соответствующие точки на оси
    points_on_line =
    pca.inverse_transform(data_compressed)
    # Отрезки между точками и осью
    plt.plot(
        np.stack([data[:, 0], points_on_line[:, 0]],
    axis=1).T,
        np.stack([data[:, 1], points_on_line[:, 1]],
    axis=1).T,
        zorder=-2,
        linewidth=2,
        c="black",
    )
```

```
# Визуализируем главную ось
plt.plot(
    [x_min, x_max],
    [y_min, y_max],
    linewidth=3,
    zorder=-2,  # Размещаем линию под точками
    label="Ось главной компоненты",
)
# Задаем формат графика.
plt.xlim(-10, 15)
plt.ylim(-2, 30)
# Делаем отсечки на осях одинаковыми по масштабу
plt.gca().set_aspect("equal")
plt.legend()
plt.title("Исходные данные")

# Строим график со сжатыми данными
plt.subplot(122)
# Трансформированные данные
plt.scatter(
    data_compressed,
    np.zeros(data_compressed.shape),
    c=data_compressed,
    cmap=cmap,
    s=100,
)
# Главная ось.
plt.plot(
    [-20, 20],
    [0, 0],
    zorder=-2,  # Размещаем линию под точками
    linewidth=2,
    label="Ось главной компоненты",
)
# Задаем формат графика.
plt.xlim(-15, 15)
plt.ylim(-10, 10)
# Делаем отсечки на осях одинаковыми по масштабу
plt.gca().set_aspect("equal")
plt.yticks([])  # Отключаем засечки на оси Oy
plt.title("Трансформированные данные")
```

Результат:



Видно, что данные спроецировались на ось, вдоль которой они сильнее разбросаны. В результате исходные точки получили новые координаты на этой оси.

Выделение групп в данных с помощью PCA.

Перед тем как анализировать данные, часто нужно их предварительно обработать – найти в них зависимости или группы. Поэтому данные визуализируют. Для этого данные сжимают до двумерных с помощью PCA, затем отображают на графике и определяют, какие группы можно выделить в полученном наборе точек.

Цель нормализации данных стандартным отклонением – преобразовать столбцы так, чтобы масштаб их дисперсии стал одинаковым.

Чтобы определить количество главных компонент, собственные числа матрицы ковариации визуализируют в виде столбчатой диаграммы. Компоненты, на которых происходит резкое падение, отбрасывают.

Реализация t-SNE в Python:

```
from sklearn.manifold import TSNE
from sklearn.datasets import load_digits
import numpy as np
import matplotlib.pyplot as plt
X, y = load_digits(return_X_y=True)
points = TSNE(2).fit_transform(X)
plt.scatter(points[:, 0], points[:, 1],
cmap='rainbow', c=y)
```

ЗАКЛЮЧЕНИЕ

В данном пособии были рассмотрены следующие вопросы:

– исследование реальных задач, когда распределение случайной величины неизвестно;

– оценка параметров распределения и выбор лучших оценок;

– построение доверительных интервалов;

– грамотное подтверждение и опровержение гипотез;

– корректное сжатие данных.

Показан подход, который помогает подобрать к задаче одно из известных распределений с конкретными параметрами, а также оценка этих параметров.

СПИСОК ЛИТЕРАТУРЫ

1. Митина, О. А. Технологии организации, обработки и хранения статистических данных : учебное пособие / О. А. Митина, И. А. Юрченков // Лань : электронно-библиотечная система. – М. : РТУ МИРЭА, 2019. – 163 с. – URL : https://e.lanbook.com/book/171511 (дата обращения: 10.02.2025).

2. Теория вероятностей и математическая статистика : учебно-методическое пособие / О. М. Дмитриева, Т. Е. Рекина, Г. М. Полевая и др. // Лань : электронно-библиотечная система. – СПб. : СПбГУТ им. М. А. Бонч-Бруевича, 2020. – 63 с. – URL : https://e.lanbook.com /book/180163 (дата обращения: 10.02.2025).

3. Иванов, Б. Н. Теория вероятностей и математическая статистика : учебное пособие для вузов / Б. Н. Иванов. // Лань : электронно-библиотечная система. – 3-е изд., стер. – СПб. : Лань, 2024. – 224 с. – URL : https://e.lanbook.com/book/393053 (дата обращения: 10.02.2025).

4. Статистический анализ данных с использованием современных информационных технологий : учебное пособие / сост. : Р. А. Алборов и др. // Лань : электронно-библиотечная система. – Ижевск : УдГАУ, 2022. – 12 с. – URL : https://e.lanbook.com/book/422687 (дата обращения: 11.02.2025).

5. Тарасов, И. Е. Статистический анализ данных в информационных системах : учебно-методическое пособие / И. Е. Тарасов // Лань : электронно-библиотечная система. – М. : РТУ МИРЭА, 2020. – 96 с. – URL : https://e.lanbook.com/book/163854 (дата обращения: 11.02.2025).

6. Кочетыгов, А. А. Основы программирования на языке Python : учебное пособие / А. А. Кочетыгов // Лань : электронно-библиотечная

система. – Тула : ТулГУ, 2024. – 272 с. – URL: https://e.lanbook.com/book/427316 (дата обращения: 10.02.2025).

7. Системы баз данных: организация, инженерия, ведение : учебное пособие / О. В. Тараканов, Ю. А. Паршенкова, М. Ю. Конышев и др. // Лань : электронно-библиотечная система. – М. : РТУ МИРЭА, 2023. – 373 с. – URL : https://e.lanbook.com/book/368672 (дата обращения: 10.02.2025).

8. Курс «Основы математики для цифровых профессий» [Электронный ресурс]. – URL : https://start.practicum.yandex/math-foundations/

9. Курс «Математика для анализа данных» [Электронный ресурс]. – URL : https://practicum.yandex.ru/math-for-da-ds/

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ 3
1. Статистическая оценка параметров 5
1.1. Точечные оценки 5
1.2. Смещенные и несмещенные оценки 7
1.3. Анализ эффективности и состоятельности оценки 13
1.4. Оценка максимального правдоподобия 18
1.5. Метод максимального правдоподобия в дискретном случае 22
1.6. Линейная регрессия с вероятностной точки зрения 25
2. Статистические эксперименты и проверка гипотез 26
2.1. Введение в интервальную оценку параметров 26
2.2. Доверительные интервалы 28
2.3. Доверительные интервалы и распределение Стьюдента 33
2.4. Введение в проверку гипотез 39
2.5. Тесты для среднего 43
2.6. A/B-тестирование 47
3. Методы статистической проверки гипотез 53
3.1. Логнормальное распределение и нелинейное преобразование данных 53
3.2. Непараметрические тесты 57
3.3. Бутстреп 61
3.4. Множественная проверка гипотез 63
4. Метод главных компонент 65
4.1. Матрица ковариации 65
4.2. Метод главных компонент 68
4.3. Применение метода главных компонент 71
ЗАКЛЮЧЕНИЕ 78
СПИСОК ЛИТЕРАТУРЫ 79

Учебное электронное издание

КОНКИНА Виктория Викторовна
ОБУХОВ Артем Дмитриевич
ЛИТОВКА Юрий Владимирович

МЕТОДЫ ОБРАБОТКИ СТАТИСТИЧЕСКИХ ДАННЫХ

Учебное пособие

Редактирование И. В. Калистратовой
Графический и мультимедийный дизайнер Т. Ю. Зотова
Обложка, упаковка, тиражирование И. В. Калистратовой

ISBN 978-5-8265-2997-3

Подписано к использованию 27.02.2026.
Тираж 50 шт. Заказ № 22

Издательский центр ФГБОУ ВО «ТГТУ»
392000, г. Тамбов, ул. Советская,
д. 106/5, пом. 2, к. 14
Телефон 8(4752)63-81-08.
E-mail: izdatelstvo@tstu.ru