

## Глава 4

# Центральная предельная теорема и нормальное распределение

---

## Содержание

---

|      |                                                           |    |
|------|-----------------------------------------------------------|----|
| 4.1. | Физическая задача: сигнал фотодетектора . . . . .         | 48 |
| 4.2. | Псевдослучайные числа . . . . .                           | 49 |
| 4.3. | Среднее и дисперсия суммы . . . . .                       | 49 |
| 4.4. | Центральная предельная теорема . . . . .                  | 51 |
| 4.5. | Нормальное распределение . . . . .                        | 52 |
| 4.6. | Квантили стандартного нормального распределения . . . . . | 53 |
| 4.7. | Центральные интервалы . . . . .                           | 54 |
| 4.8. | Совместимость двух измерений . . . . .                    | 55 |
| 4.9. | Задачи . . . . .                                          | 58 |
|      | Литература . . . . .                                      | 59 |

**Аннотация**

Центральная предельная теорема - одна из важнейших теорем в статистическом анализе. Она показывает фундаментально важную роль нормального распределения. Мы увидим как нормальное или гауссовое распределение возникает почти волшебным образом для описания суммы бесконечного количества случайных величин. Причём каждая из них может следовать почти любому распределению. Мы выведем среднее и дисперсию суммы, разберём точную формулировку теоремы и проследим гауссов предел в численном эксперименте. Нормальное распределение затем станет рабочим инструментом: через его квантили мы построим центральные области и сравним два независимых измерения.

**4.1. Физическая задача: сигнал фотодетектора**

Пусть короткая вспышка света выбивает в фотодетекторе  $N$  фотоэлектронов. Каждый фотоэлектрон создаёт на выходе электрический импульс со случайным зарядом  $Q_i$ . Электроника измеряет их сумму

$$Q_N = Q_1 + Q_2 + \dots + Q_N. \quad (4.1)$$

Даже при одинаковых вспышках заряды отдельных импульсов немного различаются. Обозначим средний заряд одного импульса через  $q_0$ , а его дисперсию — через  $s_q^2$ :

$$\mathbb{E}[Q_i] = q_0, \quad \text{Var}(Q_i) = s_q^2. \quad (4.2)$$

В простейшей модели импульсы независимы и подчиняются одному распределению. Тогда

$$\mathbb{E}[Q_N] = Nq_0, \quad \text{Var}(Q_N) = Ns_q^2. \quad (4.3)$$

Средний сигнал растёт как  $N$ , его стандартное отклонение — как  $\sqrt{N}$ . Относительная ширина пика равна

$$\frac{\sigma_{Q_N}}{\mathbb{E}[Q_N]} = \frac{s_q}{q_0\sqrt{N}}. \quad (4.4)$$

Если среднее число фотоэлектронов пропорционально энергии, этот результат даёт знакомый стохастический вклад в энергетическое разрешение, пропорциональный  $1/\sqrt{E}$ . Остаётся понять форму пика. Центральная предельная теорема даёт ответ:

$$\frac{Q_N - Nq_0}{s_q\sqrt{N}} \xrightarrow[N \rightarrow \infty]{d} \mathcal{N}(0, 1). \quad (4.5)$$

Вся задача уложилась в среднее суммы, дисперсию суммы и одну предельную теорему. В реальном детекторе флуктуирует также число фотоэлектронов, бывают корреляции, неоднородности и шум электроники. Они меняют коэффициент и добавляют другие слагаемые к разрешению. Упрощённая модель показывает сам механизм: средние независимых вкладов складываются, и их дисперсии тоже складываются.

Теперь разберём этот расчёт по частям.

## 4.2. Псевдослучайные числа

Для проверки центральной предельной теоремы устроим численный эксперимент. В первой главе мы уже генерировали псевдоданные и подбирали ширину распределения. Откуда компьютер берёт случайные числа?

Он этого не умеет делать! Любой генератор случайных чисел генерирует последовательность псевдослучайных чисел, потому что его программа представляет собой детерминированный алгоритм. Внутреннее состояние генератора полностью определяет следующее число. При удачно выбранном алгоритме последовательность проходит статистические тесты и имеет период, намного превышающий длину обычного расчёта. Для моделирования этого достаточно, хотя физического источника случайности внутри алгоритма нет.

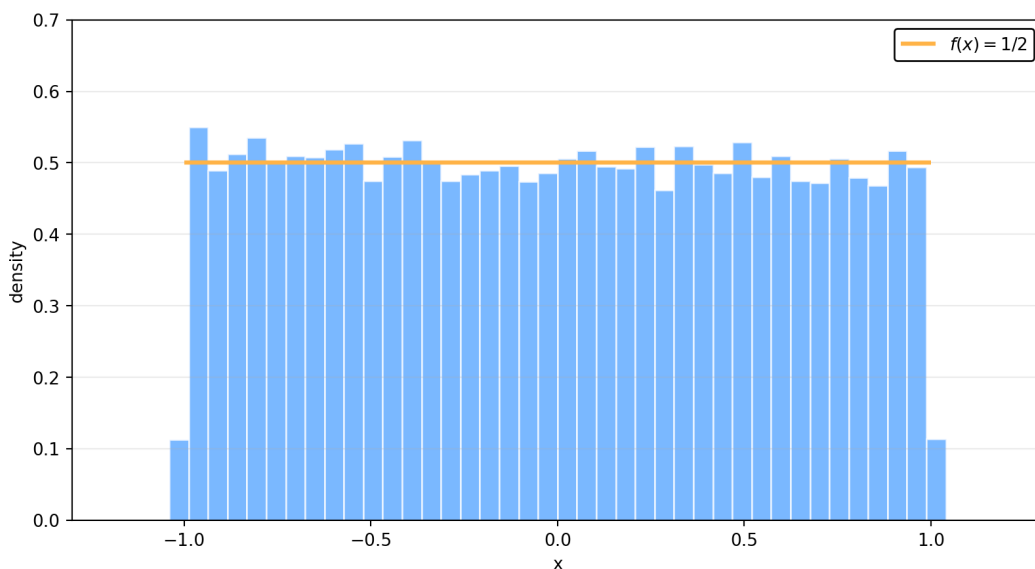
Начальное состояние генератора задаётся некоторым *начальным* числом, по английски *seed*. Одинаковое начальное состояние воспроизводит ту же последовательность. Это удобно при проверке программы: рисунок или численный результат можно получить ещё раз.

Генератор равномерного распределения на  $[-1, 1]$  в NumPy вызывается несколькими строками:

```
import numpy as np

rng = np.random.default_rng(7)
x = rng.uniform(-1, 1, size=20_000)
```

На рисунке 4.1 показана гистограмма такой выборки. Отдельные числа ничего не сообщают о форме распределения. На большой выборке становится видна постоянная плотность  $f(x) = 1/2$  внутри заданного интервала.



**Рис. 4.1.** Гистограмма 20 000 псевдослучайных чисел, равномерно распределённых на  $[-1, 1]$ ; горизонтальная линия показывает плотность  $f(x) = 1/2$

Гистограмма сама остаётся случайной: при другом начальном состоянии высоты столбцов изменятся. Распределение и его параметры при этом заданы в программе и остаются теми же.

## 4.3. Среднее и дисперсия суммы

Рассмотрим сумму  $t$  случайных величин

$$S_m = X_1 + X_2 + \dots + X_m. \quad (4.6)$$

Математическое ожидание линейно:

$$\mathbb{E}[S_m] = \sum_{i=1}^m \mathbb{E}[X_i]. \quad (4.7)$$

Для этого равенства независимость не требуется. Со средним всё складывается буквально. При вычислении дисперсии появляются произведения разных слагаемых. Их описывает ковариация.

**Ковариация** случайных величин  $X$  и  $Y$  определяется равенством

$$\begin{aligned} \text{cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \end{aligned} \quad (4.8)$$

При  $X = Y$  ковариация превращается в дисперсию:  $\text{cov}(X, X) = \text{Var}(X)$ .

#### Вывод

##### 4.3.1. Дисперсия суммы

Вычтем из каждого слагаемого его среднее и раскроем квадрат:

$$\begin{aligned} \text{Var}(S_m) &= \mathbb{E} \left[ \left( \sum_{i=1}^m (X_i - \mathbb{E}[X_i]) \right)^2 \right] \\ &= \sum_{i=1}^m \text{Var}(X_i) + 2 \sum_{i < j} \text{cov}(X_i, X_j). \end{aligned} \quad (4.9)$$

Для независимых величин  $\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\mathbb{E}[X_j]$  при  $i \neq j$ , поэтому все ковариации в двойной сумме равны нулю:

$$\text{Var}(S_m) = \sum_{i=1}^m \text{Var}(X_i). \quad (4.10)$$

Результат из уравнения 4.10 мы уже использовали для фотодетектора. Стандартные отклонения независимых вкладов складываются в квадратуре. Если между вкладами есть корреляции, двойная сумма в уравнении 4.9 сохраняется.

##### 4.3.2. Равномерные слагаемые

Для численного эксперимента возьмём независимые одинаково распределённые величины

$$X_i \sim U(-1, 1). \quad (4.11)$$

Симметрия даёт  $\mathbb{E}[X_i] = 0$ . Плотность равна  $1/2$  на  $[-1, 1]$ , поэтому

$$\begin{aligned} \mathbb{E}[X_i^2] &= \frac{1}{2} \int_{-1}^1 x^2 dx = \frac{1}{3}, \\ \text{Var}(X_i) &= \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 = \frac{1}{3}. \end{aligned} \quad (4.12)$$

Для суммы  $S_m$  получаем

$$\mathbb{E}[S_m] = 0, \quad \text{Var}(S_m) = \frac{m}{3}. \quad (4.13)$$

Ширина самой суммы растёт с  $m$ . Чтобы сравнивать её форму при разных  $m$ , нужно каждый раз вернуть среднее к нулю, а дисперсию — к единице:

$$Z_m = \frac{S_m}{\sqrt{m/3}} = \frac{X_1 + \dots + X_m}{\sqrt{m/3}}. \quad (4.14)$$

#### 4.4. Центральная предельная теорема

Теперь можно сформулировать теорему без сокращений.

**Центральная предельная теорема.** Пусть  $X_1, \dots, X_m$  независимы, одинаково распределены и имеют конечные среднее и ненулевую дисперсию:

$$E[X_i] = \mu, \quad 0 < \text{Var}(X_i) = \sigma^2 < \infty.$$

Тогда

$$Z_m = \frac{\sum_{i=1}^m X_i - m\mu}{\sigma\sqrt{m}} \xrightarrow[m \rightarrow \infty]{d} \mathcal{N}(0, 1). \quad (4.15)$$

Стрелка  $\xrightarrow{d}$  означает сходимость по распределению. При каждом фиксированном  $z$  в точке непрерывности предельной функции распределения

$$P(Z_m \leq z) \longrightarrow \Phi(z),$$

где  $\Phi$  — CDF стандартного нормального распределения. В предыдущей главе мы увидели тот же предел через характеристические функции.

Для независимых, но неодинаково распределённых слагаемых одной конечности дисперсий мало. Один член суммы может сохранить заметную долю полного разброса и определять форму результата. Условия Линдеберга и Ляпунова исключают такое доминирование; соответствующие варианты теоремы приведены, например, в [1].



##### БИОГРАФИЯ

##### Александр Михайлович Ляпунов

1857–1918

Русский математик, известный работами по устойчивости движения и теории вероятностей. В начале XX века Ляпунов доказал вариант центральной предельной теоремы методом характеристических функций. Условие, названное его именем, ограничивает вклад далёких значений отдельных слагаемых.

Источник: MacTutor; фото MacTutor

В формуле 4.15 существенны центрирование и нормировка. Среднее ненормированной суммы равно  $m\mu$ , а стандартное отклонение —  $\sigma\sqrt{m}$ . И центр, и ширина уходят с ростом  $m$ . Для выборочного среднего  $\bar{X}_m = S_m/m$  та же нормированная величина имеет вид

$$Z_m = \frac{\bar{X}_m - \mu}{\sigma/\sqrt{m}}. \quad (4.16)$$

Отсюда появляется масштаб статистической неопределённости среднего  $\sigma/\sqrt{m}$ .

Теорема описывает распределение результатов повторяемого опыта. Она ничего не обещает отдельной реализации  $Z_m$ : значения порядка двух или трёх стандартных отклонений сохраняют ненулевую вероятность и при сколь угодно большом  $m$ .

Универсального числа слагаемых, после которого разрешается написать «распределение уже гауссово», тоже нет. Если существует третий абсолютный центральный момент

$$\rho_3 = \mathbb{E}[|X_i - \mu|^3],$$

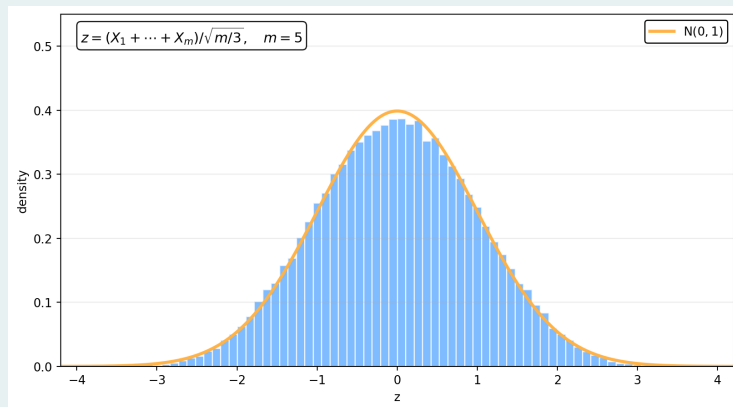
неравенство Берри—Эссеена даёт оценку

$$\sup_z |P(Z_m \leq z) - \Phi(z)| \leq C \frac{\rho_3}{\sigma^3 \sqrt{m}}, \quad (4.17)$$

где  $C$  — универсальная константа. Отношение  $\rho_3/\sigma^3$  зависит от исходного распределения. Кроме того, малая абсолютная ошибка CDF может быть заметной относительной ошибкой в далёком хвосте. При поиске редких сигналов проверка гауссова приближения в центральной области ещё не решает задачу для хвостовой вероятности.

#### 4.4.1. Как меняется форма суммы

Для равномерных слагаемых из уравнения 4.11 при  $m = 1$  распределение остаётся равномерным. При  $m = 2$  свёртка двух прямоугольников даёт треугольник. Дальше углы сглаживаются, и распределение нормированной суммы приближается к стандартному гауссу.



#### Интерактив. Форма нормированной суммы

Меняйте число слагаемых и объём выборки. Гистограмма показывает новую реализацию численного эксперимента, а гладкая кривая — плотность  $\mathcal{N}(0, 1)$ .



Вернёмся к фотодетектору. Формула 4.5 является прямым применением ЦПТ к зарядам однофотозлектронных импульсов. Гауссова форма пика и закон  $1/\sqrt{N}$  имеют одно происхождение. Первое относится к форме распределения нормированной суммы, второе — к её относительной ширине.

### 4.5. Нормальное распределение

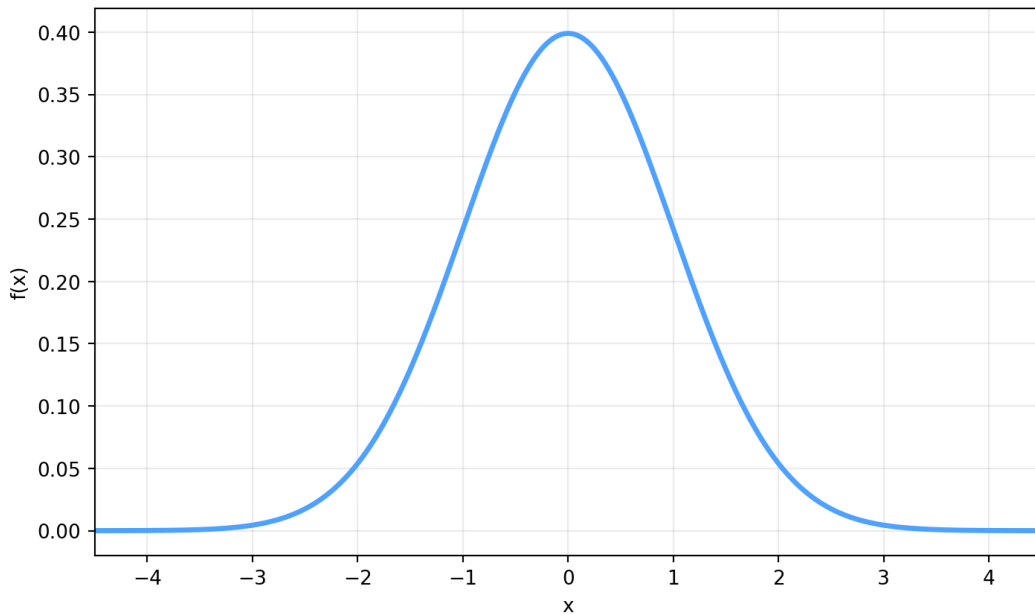
Нормальное, или гауссово, распределение уже появлялось в главах 2 и 3. Здесь оно нужно как числовая шкала отклонений.

Случайная величина  $X \sim \mathcal{N}(\mu, \sigma^2)$  имеет плотность

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right], \quad \sigma > 0. \quad (4.18)$$

Плотность достигает максимума при  $x = \mu$ , математическое ожидание равно  $\mu$ , а дисперсия —  $\sigma^2$ . В записи  $\mathcal{N}(\mu, \sigma^2)$  вторым параметром стоит именно дисперсия.

На рисунке 4.2 показана плотность при  $\mu = 0$  и  $\sigma = 1$ .



**Рис. 4.2.** Плотность стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1

Площадь под всей кривой равна единице. Высота кривой является плотностью; вероятность задаётся площадью над выбранным интервалом оси  $x$ .

#### 4.5.1. Стандартизация

Если  $X \sim \mathcal{N}(\mu, \sigma^2)$ , то линейное преобразование

$$Z = \frac{X - \mu}{\sigma} \quad (4.19)$$

даёт  $Z \sim \mathcal{N}(0, 1)$ . Число  $z$  показывает отклонение от среднего в единицах стандартного отклонения. Результат  $x = \mu + 2\sigma$  соответствует  $z = 2$  при любых  $\mu$  и  $\sigma$ .

## 4.6. Квантили стандартного нормального распределения

Кумулятивная функция стандартной нормальной величины равна

$$\begin{aligned} \Phi(z) &= P(Z \leq z) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2/2} dt \\ &= \frac{1}{2} \left[ 1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right) \right]. \end{aligned} \quad (4.20)$$

Квантиль уровня  $\beta$  — это число  $z_\beta$ , для которого

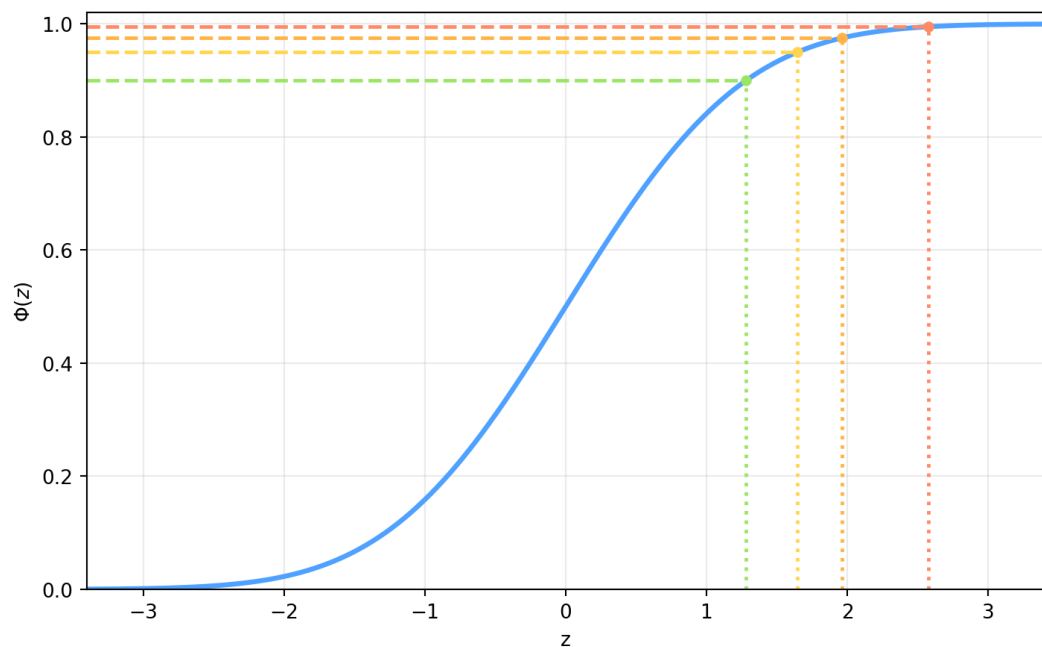
$$\Phi(z_\beta) = \beta. \quad (4.21)$$

Иными словами, слева от  $z_\beta$  лежит доля вероятности  $\beta$ . Несколько часто используемых значений собраны в таблице 4.1.

**Таблица 4.1.** Квантили стандартного нормального распределения

| $\beta$ | $z_\beta$ |
|---------|-----------|
| 0.900   | 1.282     |
| 0.950   | 1.645     |
| 0.975   | 1.960     |
| 0.995   | 2.576     |

Рисунок 4.3 показывает, как горизонтальный уровень  $\beta$  на графике CDF определяет квантиль  $z_\beta$ .



**Рис. 4.3.** Кумулятивная функция стандартного нормального распределения; пунктирные линии отмечают уровни  $\beta$  и соответствующие квантили  $z_\beta$

## 4.7. Центральные интервалы

Квантиль  $z_\beta$  из уравнения 4.21 отсекает один левый хвост. Для симметричного распределения часто выбирают центральную область и оставляют в обоих хвостах одинаковую вероятность.

Пусть вероятность центральной области равна  $\gamma$ , а полная вероятность двух хвостов равна  $\alpha = 1 - \gamma$ . Тогда

$$P(-c \leq Z \leq c) = \gamma, \quad P(Z < -c) = P(Z > c) = \frac{\alpha}{2}, \quad (4.22)$$

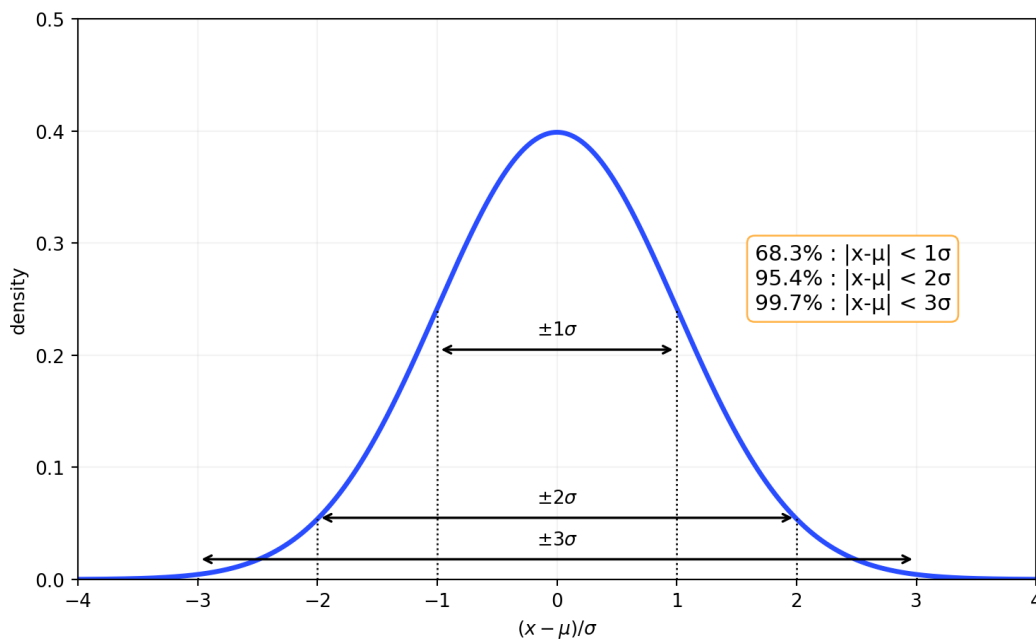
и граница области выражается через квантиль:

$$c = z_{(1+\gamma)/2} = z_{1-\alpha/2}. \quad (4.23)$$

Для границ  $c = 1, 2$  и  $3$  получаются значения из таблицы 4.2.

**Таблица 4.2.** Вероятности центральных гауссовых областей

| Центральная область | Вероятность внутри $\gamma$ | Вероятность каждого хвоста |
|---------------------|-----------------------------|----------------------------|
| $ Z  < 1$           | 0.6827                      | 0.1587                     |
| $ Z  < 2$           | 0.9545                      | 0.0228                     |
| $ Z  < 3$           | 0.9973                      | 0.00135                    |



**Рис. 4.4.** Плотность стандартного нормального распределения и центральные области в пределах одного, двух и трёх стандартных отклонений

На рисунке 4.4 границы  $\pm k\sigma$  отмечены на одной гауссовой кривой. Запись «интервал  $1\sigma$ » означает одну  $\sigma$  в каждую сторону от среднего; полная ширина такого интервала равна  $2\sigma$ .

В физике частиц порог открытия обычно задают односторонней гауссовой значимостью  $5\sigma$ . Соответствующая хвостовая вероятность равна

$$p_0 = P(Z \geq 5) = 1 - \Phi(5) \approx 2.87 \cdot 10^{-7}. \quad (4.24)$$

Одностороннее и двустороннее соглашения дают разные вероятности при одном числе сигм. Кроме того, локальная хвостовая вероятность меняется после учёта поиска во многих точках параметра, а систематические неопределённости входят в самую статистическую модель. Надпись  $5\sigma$  не выполняет эти вычисления за исследователя [2].

## 4.8. Совместимость двух измерений

Пусть два измерения одной физической величины дали

$$x_1 \pm \sigma_1, \quad x_2 \pm \sigma_2. \quad (4.25)$$

Для начала сравним две пары:

$$1.0 \pm 0.5, \quad 2.0 \pm 0.5, \quad (4.26)$$

и

$$1.010 \pm 0.001, \quad 1.020 \pm 0.001. \quad (4.27)$$

В первой паре центральные значения отличаются на единицу, во второй — всего на 0.010. Этого сравнения недостаточно. Масштаб разности задают её собственные флуктуации.

#### 4.8.1. Распределение разности

Обозначим результаты ещё не выполненных измерений через  $X_1$  и  $X_2$  и введём

$$D = X_1 - X_2. \quad (4.28)$$

При гипотезе, что оба эксперимента измеряют одно истинное значение  $\mu$ ,

$$\mathbb{E}[D] = \mu - \mu = 0.$$

Из общей формулы для дисперсии суммы следует

$$\text{Var}(D) = \sigma_1^2 + \sigma_2^2 - 2 \text{cov}(X_1, X_2). \quad (4.29)$$

Если измерения независимы, ковариация равна нулю и

$$\sigma_D = \sqrt{\sigma_1^2 + \sigma_2^2}. \quad (4.30)$$

Для независимых гауссовых измерений разность также гауссова. Полученное значение разности переводится в стандартную нормальную шкалу:

$$z_{\text{obs}} = \frac{x_1 - x_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}. \quad (4.31)$$

Двусторонняя хвостовая вероятность для такой или ещё большей по модулю разности равна

$$p = 2 [1 - \Phi(|z_{\text{obs}}|)]. \quad (4.32)$$

Это  $p$ -значение относится к конкретной гипотезе: два измерения имеют одно среднее, их стандартные отклонения известны, а ошибки независимы и гауссовы.

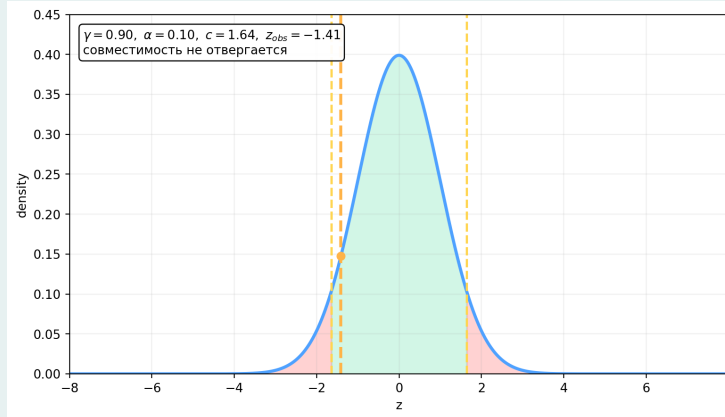
#### 4.8.2. Центральная область проверки

Выберем уровень значимости  $\alpha$ . Центральная область неотвержения имеет границу

$$c = z_{1-\alpha/2}. \quad (4.33)$$

Если  $|z_{\text{obs}}| \leq c$ , гипотеза совместимости не отвергается. При  $|z_{\text{obs}}| > c$  результат попадает в один из двух хвостов, и гипотеза отвергается на выбранном уровне  $\alpha$ . Та же проверка записывается короче:  $p \geq \alpha$  или  $p < \alpha$ .

Фраза «совместимость не отвергается» выбрана намеренно. Она не доказывает, что оба измерения правильны. Два эксперимента могут иметь общий неучтённый сдвиг и прекрасно соглашаться друг с другом.



#### Интерактив. Совместимость двух измерений

Изменяйте уровень  $\alpha$  и полученное значение  $z_{\text{obs}}$ . На графике показаны центральная область неотвержения, два хвоста и двустороннее  $p$ -значение.



#### 4.8.3. Два численных примера

Для пары из уравнения 4.26

$$z_{\text{obs}} = \frac{1.0 - 2.0}{\sqrt{0.5^2 + 0.5^2}} \approx -1.41, \quad p \approx 0.157. \quad (4.34)$$

При  $\alpha = 0.10$  критическая граница равна  $z_{0.95} \approx 1.645$ . Полученное значение лежит внутри центральной области, и совместимость не отвергается.

Для пары из уравнения 4.27

$$z_{\text{obs}} = \frac{1.010 - 1.020}{\sqrt{0.001^2 + 0.001^2}} \approx -7.07, \quad p \approx 1.5 \cdot 10^{-12}. \quad (4.35)$$

Разность здесь в сто раз меньше, чем в первом примере, и примерно в пять раз больше в единицах своей неопределённости. Физический смысл имеет второе сравнение.

#### 4.8.4. Общие неопределённости

Независимость двух измерений часто оказывается приближением. Эксперименты могут использовать одну внешнюю константу, общий расчёт потока или одну модель фона. Тогда в уравнении 4.29 остаётся ковариационный член.

Положительная ковариация уменьшает дисперсию разности: общий сдвиг сокращается при вычитании. Отрицательная ковариация её увеличивает. Для совместно гауссовых  $X_1$  и  $X_2$  разность остаётся гауссовой, и формула 4.31 обобщается заменой знаменателя на

$$\sqrt{\sigma_1^2 + \sigma_2^2 - 2 \text{cov}(X_1, X_2)}. \quad (4.36)$$

Если сами неопределённости оценены по небольшим выборкам или распределения имеют заметные негауссовы хвосты, стандартная нормальная шкала

требует отдельной проверки.

### Итоги главы

- Среднее суммы равно сумме средних. В её дисперсию входят дисперсии отдельных вкладов и все попарные ковариации.
- Для независимых одинаково распределённых величин с конечной ненулевой дисперсией центрированная и нормированная сумма сходится к стандартному нормальному распределению. Скорость сходимости зависит от исходного закона, особенно при работе с хвостами.
- Стандартизация переводит отклонение в единицы  $\sigma$ , а квантили задают односторонние и центральные области с выбранной вероятностью.
- Совместимость двух гауссовых измерений определяется распределением их разности. Общие источники неопределённости входят в её дисперсию через ковариацию.

## 4.9. Задачи

### Задача 1. Квантили и отклонения в сигмах

Пусть

$$X \sim N(12.0, 0.8^2).$$

Найдите:

1. центральный интервал 68%;
2. центральный интервал 95%;
3. односторонний квантиль уровня 0.95;
4. на сколько сигм отклоняется значение  $x = 13.4$  от среднего.

### Задача 2. Совместимость независимых измерений

Три независимых измерения дали:

$$x_1 = 125.12 \pm 0.30, \quad x_2 = 124.86 \pm 0.18, \quad x_3 = 125.05 \pm 0.24.$$

Для каждой из трёх пар вычислите  $z_{ij}$  и двустороннее  $p$ -значение. Какие пары совместимы на уровне значимости  $\alpha = 0.05$ ? Сравните порядок пар по абсолютной разности  $|x_i - x_j|$  и по  $|z_{ij}|$ .

### Задача 3. Фотоэлектронная статистика

Заряд одного фотоэлектронного импульса имеет среднее  $q_0 = 1.0$  пКл и стандартное отклонение  $s_q = 0.4$  пКл. Импульсы независимы.

1. Найдите средний полный заряд, его стандартное отклонение и относительную ширину для  $N = 25$  и  $N = 100$  фотоэлектронов.
2. Во сколько раз нужно увеличить  $N$ , чтобы относительная ширина уменьшилась вдвое?
3. Какое распределение полного заряда ожидается при большом фиксированном  $N$ ? Запишите нормированную величину, к которой применяется ЦПТ.

## Литература

- [1] Feller, W. *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons. 1971.
- [2] Cowan, G. *Statistical Data Analysis*. Oxford University Press. 1998.