Математическая статистика

Конспект лекций

Лектор: Лимар Иван Александрович

Автор: Никита Глейм (notakeith)

Весенний семестр 2026

Содержание

Лекция 1. Введение в математическую статистику

Что изучает математическая статистика

В этом семестре изучается **математическая статистика**. Чтобы понять её предмет, полезно сравнить с теорией вероятностей.

Теория вероятностей (грубо говоря): есть некоторая модель, задаётся распределение базовых параметров, и на основе этого находятся числовые характеристики или распределения функций этих параметров.

Математическая статистика имеет другую глобальную задачу. Допустим, есть совокупность всех студентов, и по тем или иным причинам нельзя обследовать каждый объект из этого набора. В контексте статистики такой полный набор называется **генеральной совокупностью**. Чтобы сделать вывод о генеральной совокупности, берут конечный поднабор из неё, который называется **выборкой**, и на основе выборки делают более-менее содержательные выводы о всей генеральной совокупности.

Иллюстрирующие примеры

Пример 1. Подбрасывание монеты роботом

Робот кидает монетку, мы видим результат: последовательность нулей и единиц. Какие задачи можно решать?

1. **Точечная оценка вероятности орла** — оценить вероятность орла в виде

числа. Резонной оценкой является:

$$\hat{p} = \frac{\text{количество орлов}}{\text{общее количество бросков}}$$

В статистике под словом «резонная» понимается «разумная». Можно оценивать в виде числа.

2. **Интервальная оценка** — указать интервал, в котором, скорее всего, лежит реальная вероятность орла. Иногда интервал нагляднее одного числа.
3. **Проверка гипотезы о честности монеты** — честно ли робот кидает монетку? Интуитивно: если доля орлов больше некоторого числа δ, то монетка нечестная. Но как выбрать порог δ? Кто-то скажет 0.51, кто-то 0.01, кто-то 0.1, 0.5 — всё это берётся «с потолка». **Методы статистики позволяют формально и обоснованно выбрать этот порог**.

   Также в статистике есть показатель **p-value (пэ-вэлю)**, который мы будем обсуждать. В терминах p-value тоже можно переписать критерий «if».

Пример 2. Клинические испытания вакцины/лекарства

Проводятся клинические испытания нового лекарства. Есть набор людей, для каждого из которых известны два атрибута: - болен / здоров - получил вакцину / не получил вакцину

Методы математической статистики позволяют **привести аргумент в пользу того, что вакцина действительно положительно влияет на здоровье человека**.

Пример 3. Цена квартиры

От чего может зависеть цена квартиры? - размер (площадь) - ситуация на рынке - регион (region) - расстояние до центра (distance) - возможно, какие-то другие факторы

Методы статистики позволяют сказать, действительно ли указанные факторы влияют на цену, или это ошибочные предположения.

Важное замечание: статистика — не серебряная пуля

Статистика **не доказывает** формальные утверждения так, как это делается в чистой фундаментальной математике.

Пример из медицины. Допустим, фармацевт провёл клиническое испытание, получил хорошие результаты, провёл статистический тест, p-value получилось меньше нужного значения. Он отправляет статью в журнал — и редактор её отклоняет (reject). Почему? Потому что был грубо нарушен **протокол**, принятый в медицинском сообществе. Нельзя в медицине «просто провести стат-тест и сказать, что вакцина эффективна».

Вывод: в каждой предметной области есть свои протоколы применения статистических методов. Курс посвящён именно самим статистическим методам, но нужно понимать, что в каждой предметной области есть своя специфика.

Репрезентативность выборки

Вопрос: какое свойство выборки позволяет сделать содержательные выводы о всей генеральной совокупности?

Ответ: репрезентативность.

> **Важно:** к утверждению «чем больше объём выборки, тем она репрезентативнее» нужно подходить очень осторожно.

Пример нерепрезентативной выборки

Социологический опрос жителей России. Анкета размещена только в интернете, заполнили сотни тысяч человек. Является ли выборка репрезентативной?

Нет, потому что есть люди, которые по тем или иным причинам не любят заполнять анкеты в интернете. Тем самым проигнорирован существенный кластер людей.

Виды выборок (на примере социологии)

- **Чисто случайная выборка** — объекты берутся случайно из генеральной совокупности.
- **Стратифицированная выборка** — есть кластеры (страты), которые заранее известны. В рамках каждого класса случайно выбираются люди, потом

всё объединяется в одну выборку. Например, делим людей на группы по возрасту, профессии и т.п., и из каждой группы случайно выбираем некоторое количество людей.

Это сильно зависит от **дизайна исследования** и **предметной области**.

> **Главный вывод:** утверждение «чем больше объём выборки, тем она репрезентативнее» **далеко не всегда правда**.

Простейшая выборка

Определение

Пусть имеется n вещественных чисел. Будем воспринимать их как n **независимых одинаково распределённых случайных величин**:

$$X_1, X_2, \dots, X_n$$

Распределение задаётся теоретической функцией распределения $F(t)$.

Почему случайные величины? Потому что в разных экспериментах могут получаться разные результаты. Например, выбираем 5 случайных студентов и измеряем рост — получаем один результат; выбираем других 5 — получаем другой результат.

Почему «простейшая»? Потому что величины независимы и одинаково распределены.

Цель

Оценить **неизвестную функцию распределения** F на основе выборки.

Обозначения

- Большие буквы $X_1, \dots, X_n$ — элементы выборки как случайные величины.
- Маленькие буквы $x_1, \dots, x_n$ — конкретная реализация выборки (когда важно подчеркнуть, что речь идёт о конкретных числах).
- Краткая запись: $X_1, \dots, X_n \sim F$ — выборка из распределения F.
- Иногда указывается конкретный класс распределений:

- $\mathcal{N}(\mu, \sigma^2)$ — нормальное распределение с матожиданием μ и дисперсией σ^2
- $\mathrm{Exp}(\lambda)$ — экспоненциальное распределение с параметром λ

Пример: запись $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ означает выборку из нормального закона с неизвестными μ и σ^2.

В первой части семестра почти всегда «выборка» = «простейшая выборка».

Эмпирическая функция распределения

Определение

Введём вспомогательную величину:

$$\nu_n(t) = \sum_{i=1}^{n} \mathbb{I}\{X_i \le t\}$$

— это **количество элементов выборки, не превосходящих** t.

Эмпирическая функция распределения:

$$\hat{F}_n(t) = \frac{\nu_n(t)}{n}$$

— это **доля (отношение) количества элементов выборки, не превосходящих** t**, к общему объёму выборки** n.

> **Терминология:** n называется **объём выборки** (или размер выборки).

График

График $\hat{F}_n(t)$ — это ступенчатая функция. Высота скачка зависит от количества элементов выборки в данной точке.

> **Замечание.** В разных учебниках функция распределения может быть непрерывна слева либо непрерывна справа. Здесь рассматривается непрерывная справа версия (знак $\le$).

Эмпирическая ФР как случайная величина

Сама выборка — случайные величины, поэтому $\widehat{F}_n(t)$ — это **функция от выборки**, то есть **тоже случайная величина**. В разных экспериментах будет получаться своя $\widehat{F}_n$. Значит, можно говорить о её распределении и числовых характеристиках.

Свойства эмпирической ФР

Распределение индикатора

Индикатор $\mathbb{I}\{X_i \leq t\}$ принимает значения 0 или 1. Это **распределение Бернулли** с параметром:

$$p = P(X_i \leq t) = F(t)$$

Это и есть теоретическая функция распределения.

Распределение $\nu_n(t)$

$\nu_n(t)$ — сумма n бернуллиевских величин, значит:

$$\nu_n(t) \sim \text{Bin}(n, F(t))$$

Отсюда:

$$\mathbb{E}\,\nu_n(t) = nF(t)$$

$$\text{Var}\,\nu_n(t) = nF(t)(1 - F(t))$$

Несмещённость

$$\mathbb{E}\,\widehat{F}_n(t) = \frac{\mathbb{E}\,\nu_n(t)}{n} = \frac{nF(t)}{n} = F(t)$$

В среднем эмпирическая функция распределения совпадает с теоретической.

> **Определение.** Свойство **несмещённости**: математическое ожидание оценки равняется самому оцениваемому параметру.
>
> **Практический смысл:** отсутствует систематическая ошибка. Можно получить оценку больше или меньше истинного значения, но в среднем — попадаешь в цель.

Состоятельность (закон больших чисел)

$\hat{F}_n(t)$ — усреднённая сумма независимых одинаково распределённых случайных величин. По **закону больших чисел**:

$$\hat{F}_n(t) \xrightarrow{P} F(t)$$

> **Определение.** Свойство **состоятельности**: оценка сходится по вероятности к параметру при $n \to \infty$.
>
> **Практический смысл:** оценка вообще «разумная» — при увеличении объёма выборки она становится всё ближе к реальному значению оцениваемого параметра.

Асимптотическая нормальность (ЦПТ)

По **центральной предельной теореме**:

$$\frac{\nu_n(t) - nF(t)}{\sqrt{nF(t)(1-F(t))}} \xrightarrow{d} \mathcal{N}(0,1)$$

Или, после преобразований (вынося n из числителя):

$$\sqrt{n} \cdot \frac{\hat{F}_n(t) - F(t)}{\sqrt{F(t)(1-F(t))}} \xrightarrow{d} \mathcal{N}(0,1)$$

> **Определение.** Свойство **асимптотической нормальности**: домноженная на $\sqrt{n}$ и нормированная разность между оценкой и оцениваемым параметром сходится по распределению к стандартной гауссовской величине.

Применение: построение доверительного интервала

Задача

Дано вещественное число $\gamma \in (0,1)$. Найти такое δ, чтобы:

$$P\left(\left|\hat{F}_n(t) - F(t)\right| < \delta\right) \geq \gamma$$

(можем считать n достаточно большим).

Решение через ЦПТ

Преобразуем выражение под вероятностью, домножив на $\sqrt{n}/\sqrt{F(t)(1-F(t))}$:

$$P\left(\frac{\sqrt{n}\,|\hat{F}_n(t)-F(t)|}{\sqrt{F(t)(1-F(t))}} < \frac{\sqrt{n}\,\delta}{\sqrt{F(t)(1-F(t))}}\right)$$

При больших n распределение этой величины близко к стандартному нормальному:

$$\approx \Phi\left(\frac{\sqrt{n}\,\delta}{\sqrt{F(t)(1-F(t))}}\right) - \Phi\left(-\frac{\sqrt{n}\,\delta}{\sqrt{F(t)(1-F(t))}}\right)$$

Используя свойства функции распределения стандартного нормального закона:

$$= 2\Phi\left(\frac{\sqrt{n}\,\delta}{\sqrt{F(t)(1-F(t))}}\right) - 1$$

Проблема и её решение

В аргументе содержится **неизвестное** выражение $F(t)(1-F(t))$ — а ведь F как раз и хотим оценить.

Идея: оценить $F(t)(1-F(t))$ **сверху**.

Функция $f(x) = x(1-x)$ — парабола ветвями вниз с точкой максимума в $x = 1/2$. Её максимум на $[0,1]$ равен $1/4$.

Значит:

$$F(t)(1-F(t)) \le \frac{1}{4} \implies \sqrt{F(t)(1-F(t))} \le \frac{1}{2}$$

В силу строгой монотонности Φ:

$$2\Phi\left(\frac{\sqrt{n}\,\delta}{\sqrt{F(t)(1-F(t))}}\right) - 1 \ge 2\Phi(2\sqrt{n}\,\delta) - 1$$

Требуем:

$$2\Phi(2\sqrt{n}\,\delta) - 1 \geq \gamma$$

$$\Phi(2\sqrt{n}\,\delta) \geq \frac{1+\gamma}{2}$$

Использование квантилей

Определение. Квантиль порядка α — такое число u_α, что:

$P(\xi \geq u_\alpha) \geq 1-\alpha$ (или эквивалентно) $F(u_\alpha) = \alpha$ (в непрерывном случае)

Геометрически: под плотностью вещественная прямая делится на две части: слева от u_α — масса α, справа — масса $1-\alpha$.

Обозначение. Для квантилей будет использоваться буква u (хотя для нормального закона иногда используется z).

Из неравенства $\Phi(2\sqrt{n}\,\delta) \geq (1+\gamma)/2$ получаем:

$$2\sqrt{n}\,\delta \geq u_{(1+\gamma)/2}$$

$$\boxed{\delta \geq \frac{u_{(1+\gamma)/2}}{2\sqrt{n}}}$$

Спойлер: это и есть **доверительный интервал** для теоретической функции распределения. К этой теме мы вернёмся в своё время.

Дополнительные теоремы для эмпирической ФР

Теорема Гливенко–Кантелли

Условия: простейшая выборка, F — теоретическая ФР, $\hat{F}_n$ — эмпирическая ФР.

Утверждение:

$$P\left(\sup_{t\in\mathbb{R}} \left|\hat{F}_n(t) - F(t)\right| \xrightarrow[n\to\infty]{} 0\right) = 1$$

То есть **с вероятностью 1** супремум модуля разности между эмпирической и теоретической ФР стремится к нулю.

Теорема Колмогорова

Условия: простейшая выборка, теоретическая ФР F должна быть **непрерывной**.

> **Важно:** теорема Колмогорова работает только для непрерывных распределений; для дискретных она не имеет места быть.

Обозначение:

$$D_n = \sqrt{n} \cdot \sup_{t \in \mathbb{R}} \left| \hat{F}_n(t) - F(t) \right|$$

Утверждение: D_n имеет предельное распределение, задаваемое функцией распределения Колмогорова:

$$P(D_n \leq x) \xrightarrow[n \to \infty]{} K(x)$$

где аналитический вид:

$$K(x) = \sum_{j=-\infty}^{\infty} (-1)^j \, e^{-2j^2 x^2}$$

Эта функция табулирована и реализована во многих статистических пакетах и библиотеках. Она используется для построения **критерия Колмогорова**.

Теорема Смирнова

Условия: есть две независимые выборки: - $X_1, \dots, X_n \sim F$ (непрерывное распределение) - $Y_1, \dots, Y_m$ — выборка того же распределения

$\hat{F}_n^X$, $\hat{F}_m^Y$ — соответствующие эмпирические ФР.

> **Замечание о независимости выборок:** если объединить обе выборки в одну, всё должно быть независимо. Например, функции от X_i и от Y_j независимы.

Обозначение:

$$D_{m,n} = \sqrt{\frac{mn}{m+n}} \cdot \sup_{t \in \mathbb{R}} \left| \hat{F}_n^X(t) - \hat{F}_m^Y(t) \right|$$

Утверждение: при $m, n \to \infty$:

$$P(D_{m,n} \leq x) \to K(x)$$

— **то же самое распределение Колмогорова** $K(x)$.

Поэтому в статистических пакетах два соответствующих критерия (Колмогорова и Смирнова) часто **объединены в одну функцию**.

Визуализация выборки

Способы визуализировать выборку:

1. **График эмпирической функции распределения** (ступенчатая функция).
2. **Гистограмма.** При построении в любой системе важно понимать: можно нормировать, можно не нормировать.

   Зачем нормировать? Пример: гистограмма оценок мальчиков и девочек. Если девочек меньше, чем мальчиков, ненормированная гистограмма для мальчиков будет явно выше. При сравнении двух гистограмм на одной картинке **лучше нормировать**.

   Существуют эмпирические (эвристические) принципы выбора оптимального количества интервалов в зависимости от объёма выборки.

   При увеличении объёма выборки гистограмма для непрерывного распределения стремится к **теоретической плотности**.
3. **Полигон частот.** Используется для дискретных величин: откладываются частоты, точки соединяются. Особой мудрости здесь нет; обычно гистограмма нагляднее.
4. **Кумулята** (упомянута вскользь). Если эмпирическую ФР «сгладить» прямыми через границы интервалов, получим кумуляту.

Эмпирическое распределение

Если зафиксировать конкретную реализацию $x_1, \dots, x_n$, то $\hat{F}_n$ — это функция, которая: - монотонно возрастает, - непрерывна справа, - стремится к 1 на $+\infty$, к 0 на $-\infty$,

то есть удовлетворяет всем свойствам функции распределения. Значит, **она задаёт некоторое распределение** — называемое **эмпирическим распределением**.

Описание эмпирического распределения

Случайная величина Y принимает значения $x_1, \dots, x_n$ со следующими вероятностями:

$$P(Y = x) = \frac{\#\{i : x_i = x\}}{n}$$

— количество элементов выборки, равных x, делённое на n.

Скачки эмпирической ФР

Если упорядочить элементы выборки и обозначить точку x_k (x_k, скажем, лежит между x_{k-1} и x_{k+1}), то **величина скачка** в точке x_k равна:

$$p_k = \frac{\#\{i : x_i = x_k\}}{n}$$

- Для **дискретного** распределения: при росте n величины скачков сходятся к реальным вероятностям p_k.
- Для **непрерывного** распределения: скачки уходят в 0.

Выборочные моменты

Теоретические моменты

Модель простейшей выборки $X_1, \dots, X_n \sim F$.

Теоретические моменты: - **Начальный момент порядка k:** $\alpha_k = \mathbb{E}\, X_1^k$ - **Центральный момент порядка k:** $\beta_k = \mathbb{E}(X_1 - \mathbb{E}\, X_1)^k$

Поскольку величины одинаково распределены: $\mathbb{E}\, X_1^k = \mathbb{E}\, X_2^k = \dots$

> **Предполагаем**, что нужные моменты существуют. Напоминание: есть распределения, где у матожидания и дисперсии большие проблемы. На-

пример, **распределение Коши** — у него **не существует матожидания и дисперсии**.

Выборочные моменты

Начальный выборочный момент порядка k:

$$\hat{\alpha}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Один из важнейших начальных выборочных моментов — **выборочное среднее**:

$$\hat{\alpha}_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Важное предостережение! Когда люди (либо не до конца разобравшись, либо специально желая ввести в заблуждение) приводят только средние — этого **далеко не достаточно**. Нужны и другие характеристики.

Центральный выборочный момент порядка k:

$$\hat{\beta}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$$

Выборочная дисперсия — наиболее важный центральный выборочный момент:

$$S_*^2 = \hat{\beta}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Выборочное стандартное отклонение (выборочное среднее квадратическое отклонение):

$$S_* = \sqrt{S_*^2}$$

Выборочные моменты как моменты эмпирического распределения

Предположим для простоты, что все $x_1, \dots, x_n$ различны. Тогда формула выборочного среднего:

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

— это среднее арифметическое, или, иными словами, сумма $X_i \cdot \frac{1}{n}$, где $\frac{1}{n}$ — вероятности значений в эмпирическом распределении.

> **Ключевое наблюдение:** выборочный момент — не что иное, как **математическое ожидание относительно эмпирического распределения**.

Это позволяет переносить «теоретические» формулы на «выборочные»:

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}\, X)^2 \implies S_*^2 = \hat{\alpha}_2 - \overline{X}^2$$

Почему? Потому что выборочные моменты — моменты относительно эмпирического распределения, и формула дисперсии переносится непосредственно.

Удобная нотация

Пусть $g : \mathbb{R} \to \mathbb{R}$ — некоторая функция. Введём обозначение:

$$\overline{g(X)} = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

Тогда: - $\hat{\alpha}_k = \overline{X^k}$ (взять $g(x) = x^k$) - $\hat{\beta}_k = \overline{(X - \overline{X})^k}$ (взять $g(x) = (x - \overline{X})^k$)

> Это **just notation** — просто обозначение, которое будет активно эксплуатироваться.

Свойства начальных выборочных моментов

Выборка — случайная, поэтому моменты — тоже случайные величины. Можно говорить об их распределении.

Несмещённость

$$\mathbb{E}\,\hat{\alpha}_k = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\,X_i^k = \frac{n\cdot\alpha_k}{n} = \alpha_k$$

Начальный выборочный момент — **несмещённая оценка** теоретического начального момента.

Состоятельность (закон больших чисел)

$\hat{\alpha}_k$ — усреднённая сумма независимых одинаково распределённых случайных величин X_i^k. По закону больших чисел:

$$\hat{\alpha}_k \xrightarrow{P} \alpha_k$$

Оценка **состоятельная**.

Асимптотическая нормальность (ЦПТ)

По центральной предельной теореме:

$$\sqrt{n}\cdot\frac{\hat{\alpha}_k - \alpha_k}{\sqrt{\mathrm{Var}(X_1^k)}} \xrightarrow{d} \mathcal{N}(0,1)$$

или, расписав дисперсию:

$$\sqrt{n}\cdot\frac{\hat{\alpha}_k - \alpha_k}{\sqrt{\mathbb{E}\,X_1^{2k} - (\mathbb{E}\,X_1^k)^2}} \xrightarrow{d} \mathcal{N}(0,1)$$

То есть:

$$\sqrt{n}\cdot\frac{\hat{\alpha}_k - \alpha_k}{\sqrt{\alpha_{2k} - \alpha_k^2}} \xrightarrow{d} \mathcal{N}(0,1)$$

Замена теоретических моментов выборочными в знаменателе

Проблема

В знаменателе ЦПТ-формулы стоят **неизвестные** характеристики α_{2k}, α_k. Это будет мешать решать статистические задачи (например, оценивать α_k).

> **Замечание про знак:** выражение $\alpha_{2k} - \alpha_k^2$ — это дисперсия, она **не отрицательна** (но может быть равна 0).

Решение: подставить выборочные аналоги

Утверждение: если в знаменатель подставить **выборочные моменты** вместо теоретических, сходимость к нормальному закону сохранится.

Рассматриваем модифицированную **статистику**:

> **Определение. Статистика** — это функция от выборки (для математически строгих — измеримая функция). Внимание: «статистика» — не только название предмета, но и термин!

$$T_n = \sqrt{n} \cdot \frac{\hat{\alpha}_k - \alpha_k}{\sqrt{\hat{\alpha}_{2k} - \hat{\alpha}_k^2}}$$

Доказательство сходимости (схема)

Преобразуем (умножаем и делим на теоретический корень):

$$T_n = \underbrace{\sqrt{n} \cdot \frac{\hat{\alpha}_k - \alpha_k}{\sqrt{\alpha_{2k} - \alpha_k^2}}}_{\xrightarrow{d} \mathcal{N}(0,1) \text{ по ЦПТ}} \cdot \underbrace{\sqrt{\frac{\alpha_{2k} - \alpha_k^2}{\hat{\alpha}_{2k} - \hat{\alpha}_k^2}}}_{\xrightarrow{P} 1}$$

Почему второй множитель сходится по вероятности к 1?

- $\hat{\alpha}_{2k} \xrightarrow{P} \alpha_{2k}$ (состоятельность)
- $\hat{\alpha}_k \xrightarrow{P} \alpha_k$ (состоятельность)
- Возведение в квадрат, разность, деление, извлечение корня — всё это **непрерывные функции** на области определения.

- Непрерывные функции **сохраняют сходимость по вероятности**.

Поэтому $\sqrt{\frac{\alpha_{2k}-\alpha_k^2}{\hat{\alpha}_{2k}-\hat{\alpha}_k^2}} \xrightarrow{P} 1$.

Используя свойство сходимости (произведение случайной величины, сходящейся по распределению, на величину, сходящуюся по вероятности к константе):

$$T_n \xrightarrow{d} \mathcal{N}(0,1)$$

Зачем нужна эта замена

В исходной записи в знаменателе стоят **неизвестные характеристики** — это потенциально мешает оценивать α_k. Чтобы упростить себе жизнь в оценивании α_k, теоретические характеристики заменяются выборочными, и оказывается, что данная величина также сходится к стандартной гауссовской величине.

> **Этот результат зафиксируем.** Он понадобится в дальнейшем — в частности, для построения **асимптотических доверительных интервалов** для математического ожидания и для проверки различных **статистических гипотез** о матожидании.

Что будет на следующей лекции

В следующий раз: - Разговор о **центральных выборочных моментах**. - В частности, поймём, **зачем нужны две выборочные дисперсии** (есть две выборочные дисперсии, и обе важны).

Лекция 2: Описательная статистика. Выборочные моменты, квантили и асимптотические свойства

Повторение материала прошлой лекции

Модель простейшей выборки

С математической точки зрения **простейшая выборка** — это набор случайных величин, которые: - независимы - одинаково распределены (i.i.d.)

Распределение этих величин описывается **теоретической функцией распределения** $F(x)$.

Эмпирическая функция распределения

В прошлый раз научились оценивать $F(x)$ с помощью **эмпирической функции распределения**:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \le x\}$$

Словами: это **доля элементов выборки, которые не превосходят заданного аргумента** x.

Свойства эмпирической функции распределения

1. **Состоятельность**: при $n \to \infty$ эмпирическая функция $\hat{F}_n(x)$ всё лучше оценивает теоретическую — $\hat{F}_n(x) \xrightarrow{P} F(x)$ по вероятности (в силу закона больших чисел).
   - Состоятельность означает: оценка стремится к оцениваемому параметру.
2. **Несмещённость**: $\mathbb{E}[\hat{F}_n(x)] = F(x)$ — в среднем эмпирическая функция распределения равна теоретической.
   - С практической точки зрения это означает **отсутствие систематической ошибки**.
3. **Асимптотическая нормальность** (по ЦПТ):

$$\sqrt{n} \cdot \frac{\hat{F}_n(x) - F(x)}{\sqrt{F(x)(1-F(x))}} \xrightarrow{d} \mathcal{N}(0,1)$$

Также упоминались: - Теорема Гливенко–Кантелли - Теорема Колмогорова–Смирнова - **Гистограмма** как графическая оценка теоретической плотности — при увеличении объёма выборки в силу ЗБЧ её график становится всё более похожим на график реальной плотности.

Выборочные начальные моменты

Обозначения: - $\alpha_k = \mathbb{E}[X_1^k]$ — теоретический k-й начальный момент - $\hat{\alpha}_k = \overline{X^k} = \frac{1}{n}\sum_{j=1}^{n} X_j^k$ — выборочный k-й начальный момент

Свойства: $\widehat{\alpha}_k$ является: - состоятельной оценкой α_k - несмещённой - асимптотически нормальной

Центральные выборочные моменты

Определения

Теоретический центральный k-й момент:

$$\beta_k = \mathbb{E}[(X_1 - \mathbb{E}X_1)^k]$$

Выборочный центральный k-й момент:

$$\widehat{\beta}_k = \overline{(X - \overline{X})^k} = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^k$$

В частности, для $k = 2$:

$$\widehat{\beta}_2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2$$

Ключевое наблюдение

Выборочный момент — это **не просто формулка, а момент относительно эмпирического распределения**. Поэтому свойства и соотношения, справедливые для теоретических моментов, справедливы и для выборочных.

Например, аналог формулы дисперсии:

$$\widehat{\beta}_2 = \overline{X^2} - \overline{X}^2$$

(среднее от квадрата минус квадрат среднего)

Состоятельность центральных выборочных моментов

Идея: произвольный k-й выборочный центральный момент — это некоторый **полином** от выборочных начальных моментов $\widehat{\alpha}_1, \widehat{\alpha}_2, \dots, \widehat{\alpha}_k$.

Поскольку: - начальные выборочные моменты сходятся к теоретическим по вероятности, - полином — непрерывная функция,

по теореме о сходимости непрерывной функции от сходящихся величин:

$$\widehat{\beta}_k \xrightarrow{P} \beta_k$$

Таким образом, **состоятельность есть**.

Несмещённость — есть проблемы (на примере дисперсии)

Введём обозначение **выборочной дисперсии** со звёздочкой:

$$S^{*2} = \widehat{\beta}_2 = \overline{X^2} - \overline{X}^2$$

Вычислим математическое ожидание:

$$\mathbb{E}[S^{*2}] = \mathbb{E}[\overline{X^2}] - \mathbb{E}[\overline{X}^2]$$

- Первое слагаемое: $\mathbb{E}[\overline{X^2}] = \alpha_2$ (поскольку начальный выборочный момент несмещён).
- Второе слагаемое (используем $\mathbb{E}[Y^2] = \mathbb{D}Y + (\mathbb{E}Y)^2$):

$$\mathbb{E}[\overline{X}^2] = \mathbb{D}[\overline{X}] + (\mathbb{E}\overline{X})^2 = \frac{\beta_2}{n} + \alpha_1^2$$

Подставляя:

$$\mathbb{E}[S^{*2}] = \alpha_2 - \frac{\beta_2}{n} - \alpha_1^2 = \beta_2 - \frac{\beta_2}{n} = \frac{n-1}{n}\beta_2$$

Вывод: $\mathbb{E}[S^{*2}] \neq \beta_2$ — выборочная дисперсия является **смещённой**.

Исправленная (несмещённая) выборочная дисперсия

Чтобы убрать смещение, вводят:

$$S^2 = \frac{n}{n-1}S^{*2} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$$

Эта величина называется **исправленной выборочной дисперсией**. Её свойство:

$$\mathbb{E}[S^2] = \beta_2$$

Таким образом, выборочных дисперсий **две штуки**: S^{*2} и S^2.

Выборочное стандартное отклонение: $S = \sqrt{S^2}$.

Промежуточный итог

У центральных выборочных моментов: - состоятельность - **несмещённость нарушается** (требуется поправка)

Дельта-метод

Зачем нужен

Для асимптотической нормальности **начальных** выборочных моментов и эмпирической функции распределения мы напрямую применяли ЦПТ. Но для **центральных** моментов это не работает: слагаемые вида $(X_j - \overline{X})^k$ **не являются независимыми**, потому что везде присутствует $\overline{X}$. Поэтому ЦПТ напрямую не применима — нужен **дельта-метод**.

Одномерная версия дельта-метода

Постановка. Пусть случайные величины ξ_n удовлетворяют:

$$\sqrt{n}(\xi_n - a) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Пусть $\varphi : \mathbb{R} \to \mathbb{R}$ — достаточно гладкая функция (столько раз дифференцируемая, сколько потребуется).

Вопрос: к чему сходится $\sqrt{n}(\varphi(\xi_n) - \varphi(a))$?

Обоснование одномерного дельта-метода

Шаг 0: $\xi_n - a \xrightarrow{P} 0$.

Действительно, рассмотрим:

$$\mathbb{P}(|\xi_n - a| < \varepsilon) = \mathbb{P}(\sqrt{n}|\xi_n - a| < \sqrt{n}\varepsilon)$$

Поскольку $\sqrt{n}(\xi_n - a) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, имеем:

$$\mathbb{P}(\sqrt{n}|\xi_n - a| < \sqrt{n}\varepsilon) \to \Phi_{0,\sigma^2}(+\infty) - \Phi_{0,\sigma^2}(-\infty) = 1 - 0 = 1$$

Значит, $\xi_n \xrightarrow{P} a$.

Шаг 1. Раскладываем по формуле Тейлора с остатком в форме Лагранжа:

$$\varphi(\xi_n) - \varphi(a) = \varphi'(a)(\xi_n - a) + \frac{\varphi''(\tilde{\xi}_n)}{2}(\xi_n - a)^2$$

где $\tilde{\xi}_n$ — между a и ξ_n.

Шаг 2. Домножим на $\sqrt{n}$:

$$\sqrt{n}(\varphi(\xi_n) - \varphi(a)) = \varphi'(a) \cdot \sqrt{n}(\xi_n - a) + \frac{\varphi''(\tilde{\xi}_n)}{2}\sqrt{n}(\xi_n - a)^2$$

- Первое слагаемое: $\varphi'(a) \cdot \sqrt{n}(\xi_n - a) \xrightarrow{d} \mathcal{N}(0, (\varphi'(a))^2\sigma^2)$
- Второе слагаемое: $\sqrt{n}(\xi_n - a)^2 = \underbrace{\sqrt{n}(\xi_n - a)}_{\to \mathcal{N}(0,\sigma^2)} \cdot \underbrace{(\xi_n - a)}_{\xrightarrow{P} 0} \xrightarrow{P} 0$, причём $\tilde{\xi}_n \xrightarrow{P} a$, $\varphi''(\tilde{\xi}_n)$ — ограничено.

В итоге всё второе слагаемое сходится к нулю по вероятности.

Результат (одномерный дельта-метод)

$$\boxed{\sqrt{n}(\varphi(\xi_n) - \varphi(a)) \xrightarrow{d} \mathcal{N}(0,\ (\varphi'(a))^2\sigma^2)}$$

Многомерная версия дельта-метода

Постановка: ξ_n — теперь случайный вектор, и

$$\sqrt{n}(\xi_n - a) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

где Σ — матрица ковариаций. Пусть $\varphi : \mathbb{R}^d \to \mathbb{R}$ — гладкая (непрерывно дифференцируемая) функция d переменных.

Утверждение:

$$\sqrt{n}(\varphi(\xi_n) - \varphi(a)) \xrightarrow{d} \mathcal{N}(0,\, \nabla\varphi(a)^\top \Sigma\, \nabla\varphi(a))$$

где $\nabla\varphi(a)$ — градиент (строчка из частных производных). Размерность согласуется: строчка × матрица × столбец = число.

Замечание: матрица ковариаций — это аналог дисперсии в многомерном случае.

Многомерная ЦПТ (для удобства использования)

Пусть $X_1, \dots, X_n$ — независимые одинаково распределённые случайные **векторы**, $\mathbb{E}X_1 = a$, $\mathbb{D}X_1 = \Sigma$. Пусть $S_n = \sum_{k=1}^{n} X_k$. Тогда:

$$\frac{S_n - na}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

В удобной для статистики форме:

$$\sqrt{n}\left(\frac{S_n}{n} - a\right) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

Теорема об асимптотической нормальности функций от начальных выборочных моментов

Постановка

Пусть $X_1, \dots, X_n$ — простейшая выборка (одномерные величины). Обозначим:

$$a = (\mathbb{E}X_1,\, \mathbb{E}X_1^2,\, \dots,\, \mathbb{E}X_1^k)$$

— вектор математических ожиданий случайного вектора $(X_1, X_1^2, \dots, X_1^k)$.

$$\Sigma = \mathbb{D}(X_1, X_1^2, \dots, X_1^k)$$

— матрица ковариаций этого случайного вектора.

$$\hat{a} = (\overline{X},\, \overline{X^2},\, \ldots,\, \overline{X^k})$$

— выборочный аналог a (это в точности S_n/n).

К $\hat{a}$ применима многомерная ЦПТ.

Пусть $\varphi : \mathbb{R}^k \to \mathbb{R}$ — гладкая функция.

Утверждение 1

$$\sqrt{n}(\varphi(\hat{a}) - \varphi(a)) \xrightarrow{d} \mathcal{N}(0,\, \nabla\varphi(a)^\top \Sigma\, \nabla\varphi(a))$$

Это напрямую следует из дельта-метода.

Утверждение 2 (важно для практики)

Положим:

$$\sigma^2 = \nabla\varphi(a)^\top \Sigma\, \nabla\varphi(a)$$

Эта величина — функция от $\mathbb{E}X_1, \mathbb{E}X_1^2, \ldots, \mathbb{E}X_1^{2k}$ (поскольку на диагонали матрицы Σ стоит $\mathbb{D}X_1^k = \mathbb{E}X_1^{2k} - (\mathbb{E}X_1^k)^2$).

Будем считать, что σ^2 — **непрерывная** функция от своих аргументов. Тогда:

$$\frac{\sqrt{n}(\varphi(\hat{a}) - \varphi(a))}{\sigma(\hat{a}_{2k})} \xrightarrow{d} \mathcal{N}(0, 1)$$

где в знаменателе вместо теоретических моментов подставлены **выборочные аналоги**. Это работает, потому что знаменатель — непрерывная функция, и подстановка выборочных моментов сохраняет сходимость.

Асимптотическая нормальность выборочной дисперсии

Применение теоремы

Выборочная дисперсия:

$$S^{*2} = \overline{X^2} - \overline{X}^2 = \varphi(\overline{X}, \overline{X^2})$$

где $\varphi(x_1, x_2) = x_2 - x_1^2$.

Градиент:

$$\nabla\varphi(x_1, x_2) = (-2x_1,\ 1)$$

В точке $a = (\mathbb{E}X_1, \mathbb{E}X_1^2)$:

$$\nabla\varphi(a) = (-2\mathbb{E}X_1,\ 1)$$

Применяем дельта-метод

$$\sqrt{n}(S^{*2} - \mathbb{D}X_1) \xrightarrow{d} \mathcal{N}(0,\ \sigma^2)$$

где

$$\sigma^2 = (-2\mathbb{E}X_1,\ 1) \cdot \begin{pmatrix} \mathbb{D}X_1 & \operatorname{cov}(X_1, X_1^2) \\ \operatorname{cov}(X_1, X_1^2) & \mathbb{D}X_1^2 \end{pmatrix} \cdot \begin{pmatrix} -2\mathbb{E}X_1 \\ 1 \end{pmatrix}$$

Упрощение (упражнение)

После раскрытия:

$$\sigma^2 = \mathbb{E}(X - \mathbb{E}X)^4 - (\mathbb{D}X)^2 = \mu_4 - \beta_2^2$$

где μ_4 — четвёртый центральный момент.

Стандартизованный результат

$$\frac{\sqrt{n}(S^{*2} - \mathbb{D}X)}{\sqrt{\widehat{\beta}_4 - (S^{*2})^2}} \xrightarrow{d} \mathcal{N}(0, 1)$$

где $\widehat{\beta}_4$ — четвёртый выборочный центральный момент.

Вывод: выборочная дисперсия — **асимптотически нормальная** оценка.

Парные выборки. Выборочная ковариация и корреляция

Теоретические понятия (повторение)

Ковариация:

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}[XY] - \mathbb{E}X \cdot \mathbb{E}Y$$

Коэффициент корреляции:

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\mathbb{D}X \cdot \mathbb{D}Y}}$$

Парная выборка

В статистике часто возникает ситуация **парной выборки** — датафрейм длины n с двумя атрибутами:

$$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$$

Выборочная ковариация

$$\widehat{\text{cov}}(X, Y) = \overline{(X - \overline{X})(Y - \overline{Y})} = \overline{XY} - \overline{X} \cdot \overline{Y}$$

Выборочный коэффициент корреляции (Пирсона)

$$\hat{\rho}(X, Y) = \frac{\overline{(X - \overline{X})(Y - \overline{Y})}}{\sqrt{S_X^{*2} \cdot S_Y^{*2}}}$$

Это **нормированная** величина, принимающая значения от -1 до 1. Используется для оценки **меры линейной зависимости**.

Порядковые статистики и выборочные квантили

Вариационный ряд

Пусть $X_1, \dots, X_n$ — исходная выборка. **Сортируем по возрастанию** и получаем:

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

Это и есть **вариационный ряд**. Элементы $X_{(k)}$ называются **порядковыми статистиками**.

Замечание: некоторые авторы под вариационным рядом понимают **статистический ряд** — где сначала берётся `unique`, затем для каждого уникального значения считается количество вхождений ν_i:

$$(x_1, \nu_1), (x_2, \nu_2), \dots, (x_m, \nu_m)$$

после чего массив сортируется по x. Здесь ν_i — случайные величины (функции от выборки).

Теоретический квантиль (повторение)

Квантиль порядка α u_α — это число, такое что:

$$\mathbb{P}(X \geq u_\alpha) \geq 1 - \alpha \quad \text{и} \quad \mathbb{P}(X \leq u_\alpha) \geq \alpha$$

В **непрерывном случае** квантиль определяется однозначно:

$$F(u_\alpha) = \alpha$$

Геометрическая интерпретация: квантиль u_α делит вероятностную массу под графиком плотности на части α (слева) и $1 - \alpha$ (справа).

Выборочные квантили

Обозначение: $\hat{u}_\alpha$.

Граничные случаи: - $\hat{u}_0 = X_{(1)} = \min X_i$ — минимум - $\hat{u}_1 = X_{(n)} = \max X_i$ — максимум

Содержательный случай $\alpha \in (0, 1)$. Существует номер $k \in \{1, \dots, n\}$, такой что:

$$\frac{k-1}{n} < \alpha \leq \frac{k}{n}$$

Тогда:

$$\hat{u}_\alpha = X_{(k)} = X_{(\lceil n\alpha \rceil)}$$

(элемент вариационного ряда с номером $\lceil n\alpha \rceil$).

Связанные термины

Квартили (от лат. *quartus* — четвёртый): делят выборку на четыре равные (в смысле эмпирической вероятностной массы) части. - Нулевой квартиль = min - Первый квартиль (нижний) = $\widehat{u}_{1/4}$ - Второй квартиль = **медиана** = $\widehat{u}_{1/2}$ - Третий квартиль (верхний) = $\widehat{u}_{3/4}$ - Четвёртый квартиль = max

Перцентили: например, 74-й перцентиль = $\widehat{u}_{0.74}$.

Дециль: разбиение на десять частей.

Выборочная медиана

Часто определяется специальным образом в зависимости от чётности n:

- Если $n = 2m + 1$ (нечётно): $\widehat{\text{med}} = X_{(m+1)}$ — центральный элемент.
- Если $n = 2m$ (чётно): $\widehat{\text{med}} = \dfrac{X_{(m)} + X_{(m+1)}}{2}$ — среднее арифметическое двух центральных элементов вариационного ряда.

При программировании необходимо смотреть, какое именно определение используется в конкретной библиотеке.

Средства визуализации выборки

Box plot («ящик с усами»)

Также неформально: **«японские свечи»** (хотя это другой объект на самом деле).

Структура (вертикальная ориентация): - Прямоугольник (ящик): - **Нижняя граница** = первый (нижний) квартиль $\widehat{u}_{1/4}$ - **Средняя линия** = медиана $\widehat{u}_{1/2}$ - **Верхняя граница** = третий (верхний) квартиль $\widehat{u}_{3/4}$ - **Межквартильный размах** $\text{IQR} = \widehat{u}_{3/4} - \widehat{u}_{1/4}$ — аналог стандартного отклонения. - **«Усики»**: длиной обычно $1.5 \cdot \text{IQR}$ от границ ящика. - Точки за пределами «усов» отмечаются отдельно и трактуются как **выбросы**.

Внутри ящика сосредоточено **50% эмпирической вероятностной массы**.

Применение: на одной картинке можно нарисовать несколько box plot для разных категорий — это позволяет визуально сравнивать распределения.

Violin plot («скрипка»)

Неформально — это **box plot + гистограмма** на одной картинке. Точнее — аппроксимация плотности с двух сторон (KDE с очень узкими bin'ами), а внутри что-то вроде box plot.

Асимптотические результаты для порядковых статистик

Теорема об асимптотике среднего члена вариационного ряда

Условия: $X_1, \dots, X_n$ — выборка из непрерывного закона с теоретической плотностью f. Пусть $p \in (0,1)$ — фиксированное число.

Утверждение:

$$\sqrt{n} \cdot f(u_p) \cdot \frac{X_{(\lceil np \rceil)} - u_p}{\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}(0,1)$$

при $n \to \infty$.

Замечания: - u_p — теоретический квантиль порядка p. - Выборочный квантиль порядка p — **асимптотически нормальная** оценка теоретического квантиля. - Структура напоминает дисперсию распределения Бернулли $p(1-p)$. - При $p = 1/2$ получаем результат для **выборочной медианы**. - На русском языке эта теорема плохо гуглится — на английском лучше.

Теорема об асимптотике крайних членов вариационного ряда (более экзотическая)

Условия: те же — выборка из непрерывного закона.

Утверждение: Для фиксированных ℓ, s: - $n \cdot F(X_{(\ell)})$ сходится по распределению к Γ-распределению с параметрами $(\ell, 1)$. - $n \cdot (1 - F(X_{(n-s+1)}))$ сходится по распределению к Γ-распределению с параметрами $(s, 1)$.

При этом эти предельные распределения **независимы**.

Заключительные замечания

Что обсудили в курсе

- **Описательные статистики**: эмпирическая функция распределения, гистограмма, выборочные характеристики.
- **Хорошие свойства** выборочных характеристик: состоятельность, несмещённость (для исправленных версий), асимптотическая нормальность.

Важная оговорка: модель простейшей выборки

Все эти результаты получены в рамках модели **простейшей выборки** (i.i.d.), а это **сильное предположение**.

Проблема робастности

Если ослабить предположения модели (например, отказаться от полной независимости/одинаковой распределённости, допустить выбросы), оценки могут вести себя по-разному:

- **Выборочное среднее** — **неробастная** оценка: при наличии выбросов оно сильно искажается.
- **Медиана** — более **устойчивая** оценка к выбросам.

Это нетривиальная тема, на эту тему написано немалое количество **нетонких** книг. Конкретные подходы к борьбе с нарушением условий зависят от конкретной задачи и предметной области.

Где почитать про распределение порядковых статистик

Ивченко, Медведев — «Введение в математическую статистику» (упоминалась в списке литературы курса).

Лекция 3: Точечное оценивание параметров. Метод моментов

Общая постановка задачи

Пусть имеется **модель простейшей выборки**. С теоретической точки зрения это набор независимых одинаково распределённых случайных величин, распределение которых задаётся функцией распределения.

При этом будем предполагать, что функция распределения параметризуется неким параметром θ:

$$F(x;\theta), \quad \theta \in \Theta \subseteq \mathbb{R}^d$$

где Θ — множество допустимых значений параметра, а θ может быть d-мерным вектором.

Мотивация: часто есть основания предполагать, что выборка пришла из какого-то класса распределений. Например: - В биологии сами данные или их логарифмы аппроксимируются нормальным законом. - Для потоков событий нередко используется распределение Пуассона.

Цель: оценить неизвестный параметр θ в виде $\hat{\theta}$, где $\hat{\theta}$ — это какая-то функция от выборки.

> Напоминание: функция от выборки кратко называется **статистикой**.

Нам бы какая оценка не годится — хотелось, чтобы она удовлетворяла каким-то хорошим свойствам.

Свойства оценок

1. Состоятельность (consistency)

Неформально: при увеличении объёма выборки оценка становится ближе к истинному значению.

Определение: оценка $\hat{\theta}$ называется состоятельной, если она сходится по вероятности к θ:

$$\hat{\theta} \xrightarrow{P} \theta$$

> По-английски: **consistency**. Это базовое свойство, говорящее о том, что оценка вообще разумна.

2. Смещённость / несмещённость (bias / unbiasedness)

Смещение оценки определяется как:

$$\text{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

- Оценка **несмещённая** (unbiased) $\Longleftrightarrow \text{bias}(\hat{\theta}) = 0$
- Оценка **смещённая** $\Longleftrightarrow \text{bias}(\hat{\theta}) \neq 0$
- Оценка **асимптотически несмещённая** $\Longleftrightarrow \text{bias}(\hat{\theta}) \to 0$ при $n \to \infty$

Пример: обычная выборочная дисперсия — смещённая оценка теоретической дисперсии, но асимптотически несмещённая.

3. Асимптотическая нормальность

Оценка $\hat{\theta}$ называется **асимптотически нормальной**, если:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

то есть в пределе по распределению получается гауссовская величина с нулевым математическим ожиданием и какой-то матрицей ковариации Σ.

4. Оптимальность и эффективность

Гипотетический вопрос: если у нас есть несколько оценок, то как сравнить, какая оценка лучше? Нужна метрика.

Среднеквадратическая ошибка (Mean Squared Error)

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\|\hat{\theta} - \theta\|^2$$

где $\|x\|^2 = x^T x = \sum_i x_i^2$ — норма вектора $x = (x_1, \dots, x_n)$.

> **Важная ремарка:** среднеквадратическая ошибка (MSE) и среднее квадратическое отклонение — разные вещи! Хоть в русском языке слова “ошибка” и “отклонение” синонимы, в статистике они означают совершенно разное: - MSE — это $\mathbb{E}\|\hat{\theta} - \theta\|^2$ - Среднее квадратическое отклонение — это $\sqrt{\text{Var}(\hat{\theta})}$ (корень из дисперсии)

Определение оптимальной (эффективной) оценки Оценка $\hat{\theta}$ называется **оптимальной (эффективной)** в классе $\mathcal{T}$, если:

$$\hat{\theta} = \arg\min_{\tilde{\theta} \in \mathcal{T}} \mathrm{MSE}(\tilde{\theta})$$

Например, $\mathcal{T}$ может быть классом несмещённых оценок: оценка называется эффективной, если у неё наименьшая среднеквадратическая ошибка среди всех несмещённых оценок.

Важное замечание про терминологию Понятия оптимальности и эффективности часто отождествляют. Однако в некоторых книгах эти понятия различают:

- **Оптимальная оценка** — минимизирует MSE.
- **Эффективная оценка** в другом определении: $\hat{\theta}$ — эффективная оценка, если

$$\hat{\theta} = \arg\min_{\tilde{\theta} \in \mathcal{T}} \mathrm{tr}(\Sigma_{\tilde{\theta}})$$

где tr — след матрицы (trace), а Σ — матрица ковариации оценки.

> **Уточнение:** arg min — это значение аргумента, при котором достигается минимум функции. Класс $\mathcal{T}$ — это произвольный класс оценок, в котором мы ищем оптимум (например, класс несмещённых, класс линейных оценок и т.д.). Например, при изучении линейных моделей будет теорема Гаусса–Маркова про эффективность в классе линейных несмещённых оценок.

Связь свойств: разложение MSE

Распишем MSE

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^T (\hat{\theta} - \theta)\right]$$

Применим приём “плюс-минус $\mathbb{E}\hat{\theta}$”:

$$\hat{\theta} - \theta = (\hat{\theta} - \mathbb{E}\hat{\theta}) + (\mathbb{E}\hat{\theta} - \theta)$$

Раскрывая скобки и используя линейность математического ожидания, получим **четыре слагаемых**:

1. $\mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta})^T(\hat{\theta} - \mathbb{E}\hat{\theta})\right]$
2. $\mathbb{E}\left[(\mathbb{E}\hat{\theta} - \theta)^T(\mathbb{E}\hat{\theta} - \theta)\right]$
3. $\mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta})^T(\mathbb{E}\hat{\theta} - \theta)\right]$
4. $\mathbb{E}\left[(\mathbb{E}\hat{\theta} - \theta)^T(\hat{\theta} - \mathbb{E}\hat{\theta})\right]$

Анализ слагаемых

Слагаемые 3 и 4 равны нулю. Рассуждение:

- θ — константа.
- $\mathbb{E}\hat{\theta}$ — это число (не случайная величина), значит тоже константа.
- Следовательно, $\mathbb{E}\hat{\theta} - \theta$ — константа, которую можно вынести из-под знака математического ожидания.
- Остаётся $\mathbb{E}[\hat{\theta} - \mathbb{E}\hat{\theta}] = \mathbb{E}\hat{\theta} - \mathbb{E}\hat{\theta} = 0$.

  Транспонирование константы — это тоже константа (транспонированная), не важно, строчка или вектор.

Слагаемое 2 — это уже константа, поэтому $\mathbb{E}$ снимается. Оно равно квадрату нормы смещения:

$$\|\text{bias}(\hat{\theta})\|^2 = \|\mathbb{E}\hat{\theta} - \theta\|^2$$

Слагаемое 1 — расписав покомпонентно:

$$\sum_{i=1}^{d} \mathbb{E}\left[(\hat{\theta}_i - \mathbb{E}\hat{\theta}_i)^2\right] = \sum_{i=1}^{d} \text{Var}(\hat{\theta}_i) = \text{tr}(\Sigma_{\hat{\theta}})$$

(на диагонали матрицы ковариации стоят как раз дисперсии).

Итоговая формула

$$\boxed{\text{MSE}(\hat{\theta}) = \text{tr}(\Sigma_{\hat{\theta}}) + \|\text{bias}(\hat{\theta})\|^2}$$

Соображение №1: эффективность через след матрицы ковариации

Если оценка **несмещённая**, то bias $= 0$, и тогда:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{tr}(\Sigma_{\hat{\theta}})$$

То есть для несмещённых оценок оптимизация MSE — это то же самое, что оптимизация следа матрицы ковариации. Это объясняет, почему в некоторых книгах эффективность определяется именно через минимизацию следа матрицы ковариации.

> **Спойлер:** если оценка несмещённая (и выполняются некоторые условия, о которых будет сказано позже), то можно указать **нетривиальную нижнюю границу** для дисперсии оценки. Тривиальная граница — это, понятно, 0.

Связь свойств: асимптотическая несмещённость + дисперсия → 0 $\Longrightarrow$ состоятельность

> Для простоты рассмотрим случай размерности $d = 1$. Для общего d рассуждения аналогичны.

Утверждение

Пусть: 1. $\mathrm{bias}(\hat{\theta}) \to 0$ (асимптотическая несмещённость), 2. $\mathrm{Var}(\hat{\theta}) \to 0$.

Тогда $\hat{\theta}$ — состоятельная оценка.

Доказательство

Хотим оценить $P(|\hat{\theta} - \theta| > \varepsilon)$.

Запишем цепочку неравенств. Сначала “плюс-минус” $\mathbb{E}\hat{\theta}$:

$$\varepsilon < |\hat{\theta} - \theta| = |(\hat{\theta} - \mathbb{E}\hat{\theta}) + (\mathbb{E}\hat{\theta} - \theta)|$$

По неравенству треугольника:

$$|\hat{\theta} - \theta| \leq |\hat{\theta} - \mathbb{E}\hat{\theta}| + |\mathbb{E}\hat{\theta} - \theta|$$

Второе слагаемое — это смещение, которое стремится к нулю. Поэтому, начиная с некоторого n, оно становится меньше $\varepsilon/2$:

$$\varepsilon < |\hat{\theta} - \mathbb{E}\hat{\theta}| + \frac{\varepsilon}{2}$$

Откуда:

$$|\hat{\theta} - \mathbb{E}\hat{\theta}| > \frac{\varepsilon}{2}$$

Если из A следует B, то $P(A) \leq P(B)$. Значит:

$$P(|\hat{\theta} - \theta| > \varepsilon) \leq P\left(|\hat{\theta} - \mathbb{E}\hat{\theta}| > \frac{\varepsilon}{2}\right)$$

Применяем **неравенство Чебышёва**:

$$P\left(|\hat{\theta} - \mathbb{E}\hat{\theta}| > \frac{\varepsilon}{2}\right) \leq \frac{4\,\mathrm{Var}(\hat{\theta})}{\varepsilon^2} \xrightarrow{n\to\infty} 0$$

Значит, $\hat{\theta} \xrightarrow{P} \theta$, что и требовалось показать. □

Связь свойств: асимптотическая нормальность ⟹ состоятельность

Утверждение

Если $\hat{\theta}$ — асимптотически нормальная оценка, то она состоятельна.

Формальное доказательство

По определению асимптотической нормальности:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Распишем вероятность:

$$P(|\hat{\theta} - \theta| < \varepsilon) = P(|\sqrt{n}(\hat{\theta} - \theta)| < \varepsilon\sqrt{n})$$

В силу асимптотической нормальности:

$$P(|\sqrt{n}(\hat{\theta} - \theta)| < \varepsilon\sqrt{n}) \to F_{\mathcal{N}(0,\sigma^2)}(\varepsilon\sqrt{n}) - F_{\mathcal{N}(0,\sigma^2)}(-\varepsilon\sqrt{n})$$

При $n \to \infty$: - $F_{\mathcal{N}(0,\sigma^2)}(+\infty) = 1$ - $F_{\mathcal{N}(0,\sigma^2)}(-\infty) = 0$

Значит, выражение стремится к $1 - 0 = 1$. То есть оценка действительно состоятельна. □

Неформально про асимптотическую несмещённость

Из

$$\sqrt{n}(\hat{\theta} - \theta) \approx \mathcal{N}(0, \sigma^2)$$

неформально получаем:

$$\hat{\theta} - \theta \approx \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$$

При $n \to \infty$ это распределение “сжимается” в точку 0. Это неформальное рассуждение.

> □ **Важное замечание:** обратное неверно! Из состоятельности **не следует** даже асимптотическая несмещённость. Существует экзотический контрпример (его рассмотрим в следующий раз).

Метод моментов

Это первый из методов точечного оценивания параметров.

Постановка

Пусть имеется модель простейшей выборки $X_1, \dots, X_n$, теоретическая функция распределения параметризуется параметром $\theta = (\theta_1, \dots, \theta_d)$ — d-мерный параметр.

Алгоритм метода моментов

Шаг 1. Вводим функции $g_1, \dots, g_d$ такие, что существуют математические ожидания:

$$\mathbb{E}[g_1(X_1)], \quad \mathbb{E}[g_2(X_1)], \quad \dots, \quad \mathbb{E}[g_d(X_1)]$$

Шаг 2. Поскольку распределение зависит от θ, эти моментные характеристики тоже зависят от θ:

$$\mathbb{E}[g_k(X_1)] = m_k(\theta_1, \dots, \theta_d), \quad k = 1, \dots, d$$

Шаг 3. Переходим к **эмпирическим аналогам**. Заменяем теоретические математические ожидания выборочными средними:

$$\overline{g_k(X)} = \frac{1}{n} \sum_{i=1}^{n} g_k(X_i)$$

Шаг 4. Получаем систему уравнений на оценки $\hat{\theta}_1, \dots, \hat{\theta}_d$:

$$\begin{cases} \overline{g_1(X)} = m_1(\hat{\theta}_1, \dots, \hat{\theta}_d) \\ \overline{g_2(X)} = m_2(\hat{\theta}_1, \dots, \hat{\theta}_d) \\ \vdots \\ \overline{g_d(X)} = m_d(\hat{\theta}_1, \dots, \hat{\theta}_d) \end{cases}$$

Это система d уравнений на d неизвестных.

Шаг 5. Предположим, что существует и притом единственное решение. Тогда:

$$\hat{\theta}_k = \alpha_k \left(\overline{g_1(X)}, \dots, \overline{g_d(X)} \right), \quad k = 1, \dots, d$$

Это решение и называется **оценкой метода моментов**.

Почему “метод моментов”

По умолчанию в качестве функций g_k берут степенные:

$$g_k(x) = x^k$$

Тогда $\mathbb{E}[g_k(X)] = \mathbb{E}[X^k]$ — это k-й момент. Отсюда и название.

Свойства оценок метода моментов

1. Состоятельность. Если: - $\overline{g_k(X)}$ — состоятельные оценки $\mathbb{E}[g_k(X)]$, - функции α_k непрерывны,

то оценка метода моментов состоятельна.

Это обычно выполняется.

2. Асимптотическая нормальность. Если: - $\overline{g_k(X)}$ — асимптотически нормальные оценки, - функции α_k гладкие,

то имеет место асимптотическая нормальность по **дельта-методу**.

3. Смещённость. В общем случае про смещённость и несмещённость сказать ничего нельзя.

Плюсы и минусы метода

Плюсы: - Идея метода достаточно проста.

Минусы: - Часто получаются не очень эффективные оценки (это будет видно даже на учебных примерах).

Примеры применения метода моментов

Пример 1. Распределение Бернулли

Робот много раз кидает монетку, на входе последовательность нулей и единичек. Оценить вероятность p выпадения единички.

Берём дефолтную функцию $g(x) = x$. Математическое ожидание распределения Бернулли:

$$\mathbb{E}[X] = p$$

Переходим к эмпирическому аналогу:

$$\overline{X} = \hat{p}$$

Здесь всё разрешилось тривиально. Получили:

$$\boxed{\hat{p} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i}$$

То есть оценка вероятности успеха — это просто выборочное среднее (количество успехов / общее количество экспериментов).

Свойства: про выборочное среднее знаем, что это состоятельная, несмещённая, асимптотически нормальная оценка. Забегая вперёд — даже эффективная.

Пример 2. Распределение Пуассона

Вариант 1: $g(x) = x$ Математическое ожидание распределения Пуассона:

$$\mathbb{E}[X] = \lambda$$

Эмпирический аналог:

$$\hat{\lambda}_1 = \overline{X}$$

Вариант 2: $g(x) = x^2$ Используем то, что:

$$\mathbb{E}[X^2] = \text{Var}(X) + (\mathbb{E}[X])^2 = \lambda + \lambda^2$$

Эмпирический аналог:

$$\hat{\lambda}^2 + \hat{\lambda} - \overline{X^2} = 0$$

Это квадратное уравнение. Решаем:

$$\hat{\lambda} = \frac{-1 \pm \sqrt{1 + 4\overline{X^2}}}{2}$$

Формально два корня, но по смыслу задачи $\lambda > 0$, поэтому выбираем **положительный корень** (с плюсом):

$$\boxed{\hat{\lambda}_2 = \frac{-1 + \sqrt{1 + 4\overline{X^2}}}{2}}$$

Свойства: получилась гладкая (на области определения) функция от второго выборочного момента. Поэтому оценка состоятельная и асимптотически нормальная. Про смещённость конкретно сказать сложно.

Какая оценка лучше? Та, у которой меньше MSE. Забегая вперёд — **первая оценка** $\hat{\lambda}_1 = \overline{X}$ будет эффективной.

> Это хорошо иллюстрирует минус метода: разные функции g дают разные оценки, и не все они одинаково хороши.

Пример 3. Нормальное распределение $\mathcal{N}(\mu, b)$

Здесь b — дисперсия. Два неизвестных параметра, поэтому нужны два уравнения.

Берём: - $g_1(x) = x$, - $g_2(x) = x^2$.

Теоретические соотношения:

$$\mathbb{E}[X] = \mu$$

$$\mathbb{E}[X^2] = \text{Var}(X) + (\mathbb{E}[X])^2 = b + \mu^2$$

Эмпирические аналоги:

$$\overline{X} = \hat{\mu}$$

$$\overline{X^2} = \hat{b} + \hat{\mu}^2$$

Отсюда:

$$\boxed{\hat{\mu} = \overline{X}, \qquad \hat{b} = \overline{X^2} - (\overline{X})^2}$$

А $\overline{X^2} - (\overline{X})^2$ — это **выборочная дисперсия** S^{*2}.

Свойства: - $\hat{\mu} = \overline{X}$ — несмещённая оценка. - $\hat{b} = S^{*2}$ — смещённая оценка (но асимптотически несмещённая). - Обе оценки состоятельные и асимптотически нормальные.

> Это иллюстрирует, что в общем случае про смещённость метода моментов ничего конкретного сказать нельзя — здесь одна оценка несмещённая, другая смещённая.

Пример 4. Равномерное распределение $U[a, b]$

Берём те же функции $g_1(x) = x$, $g_2(x) = x^2$.

Математическое ожидание равномерного закона:

$$\mathbb{E}[X] = \frac{a+b}{2}$$

Дисперсия:

$$\text{Var}(X) = \frac{(b-a)^2}{12}$$

Тогда:

$$\mathbb{E}[X^2] = \frac{(b-a)^2}{12} + \left(\frac{a+b}{2}\right)^2$$

Эмпирические соотношения:

$$\overline{X} = \frac{\hat{a} + \hat{b}}{2}$$

$$\overline{X^2} = \frac{(\hat{b} - \hat{a})^2}{12} + \left(\frac{\hat{a} + \hat{b}}{2}\right)^2$$

Из второго уравнения, подставляя первое:

$$\frac{(\hat{b} - \hat{a})^2}{12} = \overline{X^2} - (\overline{X})^2 = S^{*2}$$

Снова получили выборочную дисперсию! Откуда:

$$(\hat{b} - \hat{a})^2 = 12 \cdot S^{*2}$$

$$\hat{b} - \hat{a} = \pm 2\sqrt{3} \cdot S^*$$

где $S^* = \sqrt{S^{*2}}$.

Выбираем знак “+”, так как $b - a > 0$.

Имеем систему:

$$\begin{cases} \hat{a} + \hat{b} = 2\overline{X} \\ \hat{b} - \hat{a} = 2\sqrt{3} \cdot S^* \end{cases}$$

Решая (сложение и вычитание):

$$\boxed{\hat{a} = \overline{X} - \sqrt{3} \cdot S^*, \qquad \hat{b} = \overline{X} + \sqrt{3} \cdot S^*}$$

Пример 5 (демонстрационный). Равномерное распределение $U[-\theta, \theta]$

Здесь интересный случай: функция $g(x) = x$ **не подходит**, потому что:

$$\mathbb{E}[X] = 0$$

— математическое ожидание не зависит от θ, поэтому уравнение бессмысленно.

Берём $g(x) = x^2$:

$$\mathbb{E}[X^2] = \frac{\theta^2}{3}$$

Эмпирический аналог даёт явное выражение для оценки $\hat{\theta}$.

> На демонстрации в Google Colab было показано: при объёме выборки 10 разброс оценки большой, а при объёме 10000 разброс существенно меньше, и распределение оценки концентрируется около реального параметра. Это иллюстрирует состоятельность и асимптотическую нормальность.

Что дальше

В следующий раз: 1. Будет приведён экзотический **контрпример**, показывающий, что из состоятельности не следует даже асимптотическая несмещённость. 2. Перейдём к следующему методу — **методу максимального правдоподобия** (maximum likelihood).

Лекция 4: Метод максимального правдоподобия и информация Фишера

1. Контрпример: асимптотическая нормальность ⇏ асимптотическая несмещённость

Напоминание из прошлой лекции

В прошлый раз были рассмотрены свойства оценок: - состоятельность, - эффективность, - асимптотическая нормальность, - несмещённость.

Было показано: **если оценка асимптотически нормальная, то она состоятельна**.

Сегодня покажем (обещанный контрпример), что **из асимптотической нормальности в общем случае НЕ следует асимптотическая несмещённость** (хотя обычно эта импликация имеет место). Пример экзотический, но формально корректный.

Построение контрпримера

Пусть выборка $X_1, \dots, X_n$ из нормального распределения $\mathcal{N}(0, \sigma^2)$.

Выборочное среднее $\bar{X}$ — состоятельная, несмещённая, асимптотически нормальная оценка для 0 (поскольку матожидание здесь равно 0).

Модифицируем оценку. Положим:

$$\hat{\theta} = \begin{cases} \bar{X}, & \text{с вероятностью } 1 - \frac{1}{n} \\ n, & \text{с вероятностью } \frac{1}{n} \end{cases}$$

Доказательство асимптотической нормальности $\hat{\theta}$

Рассмотрим функцию распределения $\sqrt{n}\,\hat{\theta}$ в точке t:

$P(\sqrt{n}\,\hat{\theta} \le t) = P\left(\sqrt{n}\,\hat{\theta} \le t \mid \text{случилось событие с вер. } 1 - \frac{1}{n}\right) \cdot \left(1 - \frac{1}{n}\right) + P\left(\sqrt{n}\,\hat{\theta} \le t \mid \text{случило}\right.$

Это эквивалентно:

$$P(\sqrt{n}\,\bar{X} \le t) \cdot \left(1 - \tfrac{1}{n}\right) + P\left(\sqrt{n} \cdot n \le t\right) \cdot \tfrac{1}{n}$$

При $n \to \infty$: - $\left(1 - \frac{1}{n}\right) \to 1$, - $\frac{1}{n} \to 0$, - $P(\sqrt{n}\,\bar{X} \le t) \to \Phi_{0,\sigma^2}(t)$ (т.к. выборочное среднее — асимптотически нормальная оценка), - второе слагаемое (ограниченная вероятность $\times \frac{1}{n}$) стремится к 0.

Итог: $\sqrt{n}\,\hat{\theta} \xrightarrow{d} \mathcal{N}(0, \sigma^2)$.

То есть $\hat{\theta}$ **— асимптотически нормальная оценка**, а значит и **состоятельная**.

Проверка асимптотической несмещённости

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[\bar{X}] \cdot \left(1 - \tfrac{1}{n}\right) + n \cdot \tfrac{1}{n} = 0 + 1 = 1$$

Таким образом, $\mathbb{E}[\hat{\theta}] = 1 \neq 0$ для любого n, **асимптотической несмещённости нет**.

> **NOTE: Вывод** Контрпример показывает: из асимптотической нормальности **не следует** асимптотическая несмещённость, хотя состоятельность из неё следует. Контринтуитивно, но формально верно.

---

2. Метод максимального правдоподобия

Нотация: дискретный и непрерывный случаи

На практике работают либо с дискретными, либо с непрерывными распределениями.

Случай	Функция	Обозначение
Дискретный	Функция вероятностей	PMF (probability mass function)
Непрерывный	Плотность вероятности	PDF (probability density function)

> **INFO: Унификация** В контексте метода максимального правдоподобия оба случая объединяются — будем использовать термин **«плотность»** и одну букву p для обоих случаев. Рассуждения в дискретном и непрерывном случаях идентичны.

Постановка задачи

Имеется простейшая выборка $X_1, X_2, \dots, X_n$ — независимые одинаково распределённые случайные величины с распределением, зависящим от параметра θ. Задача: оценить θ как функцию от выборки.

Функция правдоподобия

Поскольку элементы выборки независимы, **совместная плотность** есть произведение плотностей:

$$L(X, \theta) = \prod_{k=1}^{n} p(X_k, \theta)$$

Эта совместная плотность называется **функцией правдоподобия**.

Идея метода

На интуитивном уровне $L(X, \theta)$ — это «вероятность выборки». Метод максимального правдоподобия предлагает подобрать θ так, чтобы эта вероятность была наибольшей.

Определение оценки максимального правдоподобия

> **IMPORTANT: Оценка максимального правдоподобия** $\hat{\theta}$ — это значение θ, при котором достигается максимум функции правдоподобия:
>
> $$\hat{\theta} = \arg\max_{\theta} L(X, \theta)$$

---

3. Алгоритм поиска оценки максимального правдоподобия

Пункт 0. Посмотреть и подумать

Возможно, удастся найти ответ, внимательно посмотрев на функцию правдоподобия — без вычислений (см. примеры с равномерным распределением и распределением Лапласа ниже).

Пункт 1. Логарифмирование

Используется свойство: производная логарифма функции

$$(\ln f(x))' = \frac{f'(x)}{f(x)}$$

Логарифм — строго монотонная функция, поэтому **точка максимума не меняется**. Удобно работать с $\ln L$ потому, что произведение превращается в сумму.

Пункт 2. Исследование на максимум

1. Рассмотреть $\ln L(X, \theta)$.
2. Вычислить производную $\dfrac{\partial \ln L}{\partial \theta}$.
3. Приравнять к нулю.
4. Проверить достаточные условия максимума.

---

4. Примеры применения метода максимального правдоподобия

Пример 1. Равномерное распределение $U[\theta_1, \theta_2]$

Плотность равномерного распределения:

$$p(x, \theta_1, \theta_2) = \frac{1}{\theta_2 - \theta_1} \cdot \mathbb{1}\{x \in [\theta_1, \theta_2]\}$$

Функция правдоподобия:

$$L(X, \theta_1, \theta_2) = \frac{1}{(\theta_2 - \theta_1)^n} \cdot \mathbb{1}\{X_1 \in [\theta_1, \theta_2], X_2 \in [\theta_1, \theta_2], \dots, X_n \in [\theta_1, \theta_2]\}$$

> **TIP: Замечание** Произведение индикаторов $\prod_k \mathbb{1}\{X_k \in [\theta_1, \theta_2]\}$ равно 1 тогда и только тогда, когда **все** X_k попадают в отрезок. Поэтому оно равно одному индикатору пересечения событий.

Анализ. Чтобы максимизировать L: - **Мысль А.** Индикатор должен быть равен 1, т.е. $\theta_1 \le \min_k X_k$ и $\theta_2 \ge \max_k X_k$. - **Мысль Б.** Знаменатель $(\theta_2 - \theta_1)^n$ должен быть минимальным, т.е. $\theta_2 - \theta_1$ — минимально.

Совмещая:

$$\hat{\theta}_1 = \min_k X_k, \qquad \hat{\theta}_2 = \max_k X_k$$

Пример 2. Распределение Лапласа

Плотность:

$$p(x, \theta) = \frac{1}{2}\, e^{-|x-\theta|}$$

(Распределение Лапласа с масштабным параметром 1.)

Функция правдоподобия:

$$L(X, \theta) = \frac{1}{2^n} \cdot \prod_{k=1}^{n} e^{-|X_k-\theta|} = \frac{1}{2^n} \cdot e^{-\sum_{k=1}^{n} |X_k-\theta|}$$

Множитель $\frac{1}{2^n}$ — константа. Чтобы максимизировать L, нужно **максимизировать аргумент экспоненты**, то есть **минимизировать**:

$$\sum_{k=1}^{n} |X_k - \theta| \;\longrightarrow\; \min_{\theta}$$

Поделив на n, получим $\mathbb{E}_{\hat{F}_n}[|X-\theta|]$ — матожидание относительно эмпирического распределения.

> **IMPORTANT: Известный факт из теории вероятностей** Медиана минимизирует среднее абсолютное отклонение.

Поэтому:

$$\hat{\theta} = \text{медиана выборки}$$

Пример 3. Биномиальное распределение $\mathrm{Bin}(m, p)$**,** m **известно**

Функция вероятностей:

$$p(x, p) = C_m^x \cdot p^x \cdot (1-p)^{m-x}$$

Функция правдоподобия:

$$L(X, p) = \prod_{k=1}^{n} C_m^{X_k} \cdot p^{X_k} \cdot (1-p)^{m-X_k}$$

Логарифм:

$$\ln L(X,p) = \sum_{k=1}^{n} \left[\ln C_m^{X_k} + X_k \ln p + (m - X_k) \ln(1-p) \right]$$

Дифференцируем по p (член с $C_m^{X_k}$ не зависит от p):

$$\frac{\partial \ln L}{\partial p} = \sum_{k=1}^{n} \left[\frac{X_k}{p} - \frac{m - X_k}{1-p} \right]$$

Приведём к общему знаменателю $p(1-p)$:

$$\frac{1}{p(1-p)} \sum_{k=1}^{n} \left[X_k(1-p) - p(m - X_k) \right] = \frac{1}{p(1-p)} \sum_{k=1}^{n} \left[X_k - p \cdot m \right]$$

Вынесем n и m:

$$= \frac{n \cdot m}{p(1-p)} \left[\frac{\bar{X}}{m} - p \right]$$

Приравнивая к нулю:

$$\hat{p} = \frac{\bar{X}}{m}$$

Проверка максимума: при $p < \bar{X}/m$ производная положительна (функция возрастает), при $p > \bar{X}/m$ — отрицательна (функция убывает). Значит, это точка максимума.

> **NOTE: Частный случай** Если $m = 1$ — распределение Бернулли. Тогда $\hat{p} = \bar{X}$ — доля единиц в выборке.

Пример 4. Нормальное распределение $\mathcal{N}(\mu, b)$

Здесь b — дисперсия (а не σ^2). Плотность:

$$p(x, \mu, b) = \frac{1}{\sqrt{2\pi b}} e^{-\frac{(x-\mu)^2}{2b}}$$

Логарифм функции правдоподобия:

$$\ln L(X, \mu, b) = \sum_{k=1}^{n} \left[-\tfrac{1}{2} \ln(2\pi) - \tfrac{1}{2} \ln b - \frac{(X_k - \mu)^2}{2b} \right]$$

Шаг 1. Зафиксируем b, найдём $\widehat{\mu}$ (ленивый вариант) При фиксированном b максимизация $\ln L$ эквивалентна минимизации:

$$\sum_{k=1}^{n} (X_k - \mu)^2 \longrightarrow \min_{\mu}$$

Поделив на n, получаем $\mathbb{E}_{\hat{F}_n}[(X - \mu)^2]$ — матожидание квадрата отклонения.

> **IMPORTANT: Известный факт** Эта величина минимизируется при $\mu = \mathbb{E}[X]$ (для эмпирического распределения это выборочное среднее).
>
> *(Точкой минимума является матожидание, а само минимальное значение — дисперсия.)*

Поэтому:

$$\widehat{\mu} = \bar{X}$$

Шаг 2. Найдём $\widehat{b}$ Дифференцируем $\ln L$ по b:

$$\frac{\partial \ln L}{\partial b} = -\sum_{k=1}^{n} \frac{1}{2b} + \sum_{k=1}^{n} \frac{(X_k - \widehat{\mu})^2}{2b^2} = 0$$

Откуда:

$$\widehat{b} = \frac{1}{n} \sum_{k=1}^{n} (X_k - \widehat{\mu})^2 = S^{*2}$$

Это **выборочная дисперсия** (со звёздочкой).

> **SUMMARY: Итог для нормального закона**
>
> $$\widehat{\mu} = \bar{X}, \qquad \widehat{b} = S^{*2}$$
>
> Оценками максимального правдоподобия для параметров нормального закона являются выборочное среднее и выборочная дисперсия.

(Это «ленивый» вариант: строго следовало бы исследовать функцию двух переменных через гессиан.)

Пример 5. Дискретное распределение на m значениях

Пусть выборка из дискретного распределения, принимающего значения $1, 2, \dots, m$ с вероятностями $p_1, p_2, \dots, p_m$.

Сколько неизвестных? $m - 1$ (так как $\sum p_k = 1$).

Группировка. Пусть ν_k — количество элементов выборки, равных k. Тогда:

$$\sum_{k=1}^{m} \nu_k = n$$

Функция правдоподобия:

$$L(X, p) = p_1^{\nu_1} \cdot p_2^{\nu_2} \cdot \dots \cdot p_m^{\nu_m}$$

с ограничением $p_1 + p_2 + \dots + p_m = 1$.

> **TIP: Замечание** Можно было бы использовать **множители Лагранжа**, но т.к. ограничение одно, проще выразить $p_m = 1 - p_1 - \dots - p_{m-1}$ и работать с функцией $m - 1$ переменной.

Подставляем:

$$\ln L(X, p) = \sum_{k=1}^{m-1} \nu_k \ln p_k + \nu_m \ln (1 - p_1 - \dots - p_{m-1})$$

Дифференцируем по p_j ($j \in \{1, \dots, m-1\}$):

$$\frac{\partial \ln L}{\partial p_j} = \frac{\nu_j}{p_j} - \frac{\nu_m}{p_m} = 0$$

Откуда:

$$\nu_j \cdot p_m = \nu_m \cdot p_j \quad \text{для всех } j = 1, \dots, m-1$$

Просуммируем все эти уравнения (по j от 1 до $m - 1$):

$$p_m \cdot (\nu_1 + ... + \nu_{m-1}) = \nu_m \cdot (p_1 + ... + p_{m-1})$$

Используя $\nu_1 + ... + \nu_{m-1} = n - \nu_m$ и $p_1 + ... + p_{m-1} = 1 - p_m$:

$$p_m \cdot (n - \nu_m) = \nu_m \cdot (1 - p_m)$$

$$p_m \cdot n - p_m \cdot \nu_m = \nu_m - \nu_m \cdot p_m$$

$$\boxed{\widehat{p}_m = \frac{\nu_m}{n}}$$

Подставляя обратно в $\nu_j p_m = \nu_m p_j$:

$$\widehat{p}_j = \frac{\nu_j}{n} \quad \text{для всех } j$$

SUMMARY: Интерпретация Чтобы оценить вероятность исхода типа j, нужно количество исходов типа j разделить на общее количество испытаний. Это пример, на который будем ссылаться в дальнейшем.

---

5. Информация Фишера

Условия регулярности

Информация Фишера определяется в рамках условий регулярности (для одномерного случая, $\theta \in \mathbb{R}$).

WARNING: Замечание о терминологии В разных книжках условия регулярности могут немного отличаться. То, что ниже — один из стандартных вариантов.

Условие 1. Если $\theta_1 \neq \theta_2$, то распределение при θ_1 не равно распределению при θ_2 (идентифицируемость).

Условие 2. Носитель распределения **не зависит от** θ. - Множество значений случайной величины не зависит от параметра. - *Пример: равномерное распределение $U[\theta_1, \theta_2]$ — НЕ регулярно, т.к. носитель зависит от параметров.*

Условие 3. Функция $p(x, \theta)$ дифференцируема по θ столько раз, сколько нужно.

Условие 4. Внесение дифференцирования по θ под знак интеграла — законная операция:

$$\frac{\partial}{\partial \theta} \int \dots \, dx = \int \frac{\partial}{\partial \theta} \dots \, dx$$

(не всегда верно в общем случае, но мы работаем там, где верно).

Условие 5. $\mathbb{E}[V^2(X, \theta)] < \infty$ (вводится далее).

Вклад выборки

DEFINITION: Вклад выборки

$$V(X, \theta) = \frac{\partial \ln L(X, \theta)}{\partial \theta}$$

— логарифмическая производная функции правдоподобия.

Интуиция термина «вклад выборки» Аналитически найти точку максимума L удаётся не всегда — иногда задача решается только численно. Один из простейших численных методов — **градиентный спуск**:

$$x_{k+1} = x_k - \alpha \cdot f'(x_k)$$

Здесь: - если мы правее минимума — $f'(x_k) > 0$ и сдвиг идёт влево (правильно); - если мы левее минимума — $f'(x_k) < 0$ и сдвиг идёт вправо (правильно); - чем больше $|f'|$, тем больше шаг — тем быстрее сходимость.

В многомерном случае вместо производной — **градиент** (вектор частных производных).

Применяя к функции правдоподобия: чем больше по модулю $V(X, \theta)$, тем быстрее численный метод сойдётся к оценке. Поэтому V называется «вкладом выборки» — чем больше вклад, тем лучше (быстрее находится оценка).

Проблема: $V(X, \theta)$ — случайная величина (зависит от X). Хотим унифицировать в виде числовой характеристики.

Матожидание вклада выборки

Рассмотрим тождество:

$$1 = \int L(X, \theta)\, dX$$

(плотность интегрируется в 1).

Дифференцируем по θ:

$$0 = \frac{\partial}{\partial \theta} \int L(X, \theta)\, dX = \int \frac{\partial L(X, \theta)}{\partial \theta}\, dX$$

Воспользуемся **трюком**: умножим и разделим на L:

$$0 = \int \frac{\partial L / \partial \theta}{L} \cdot L\, dX = \int \frac{\partial \ln L}{\partial \theta} \cdot L\, dX = \int V(X, \theta) \cdot L(X, \theta)\, dX$$

Это есть матожидание V:

$$\boxed{\mathbb{E}[V(X, \theta)] = 0}$$

В среднем вклад выборки равен нулю. Не очень информативно — рассмотрим другую характеристику.

Определение информации Фишера

Мера разброса относительно нуля — **дисперсия**.

DEFINITION: Информация Фишера

$$I(\theta) = \text{Var}[V(X, \theta)] = \text{Var}\left[\frac{\partial \ln L(X, \theta)}{\partial \theta}\right]$$

Свойство 1. Аддитивность по выборке

$$V(X, \theta) = \frac{\partial \ln L}{\partial \theta} = \sum_{k=1}^{n} \frac{\partial \ln p(X_k, \theta)}{\partial \theta}$$

(логарифм произведения = сумма логарифмов).

Так как X_k независимы, слагаемые независимы. Дисперсия суммы независимых = сумма дисперсий:

$$I(\theta) = \sum_{k=1}^{n} \operatorname{Var}\left[\frac{\partial \ln p(X_k, \theta)}{\partial \theta}\right]$$

Все слагаемые одинаково распределены, поэтому:

$$\boxed{I(\theta) = n \cdot i(\theta)}$$

где $i(\theta)$ — **информация Фишера для одного наблюдения**:

$$i(\theta) = \operatorname{Var}\left[\frac{\partial \ln p(X, \theta)}{\partial \theta}\right]$$

Свойство 2. Информация Фишера через матожидание квадрата

Поскольку $\mathbb{E}\left[\dfrac{\partial \ln p(X, \theta)}{\partial \theta}\right] = 0$, а дисперсия при нулевом матожидании совпадает с матожиданием квадрата:

$$i(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln p(X, \theta)}{\partial \theta}\right)^2\right]$$

Свойство 3. Альтернативная формула через вторую производную

Продифференцируем тождество $\int \dfrac{\partial \ln p}{\partial \theta} \cdot p\, dx = 0$ ещё раз по θ (для одного наблюдения, индекс k опускаем — все X_k одинаково распределены):

$$0 = \frac{\partial}{\partial \theta} \int \frac{\partial \ln p(X, \theta)}{\partial \theta} \cdot p(X, \theta)\, dX$$

Вносим производную под интеграл и применяем правило произведения:

$$0 = \int \frac{\partial^2 \ln p}{\partial \theta^2} \cdot p\, dX + \int \frac{\partial \ln p}{\partial \theta} \cdot \frac{\partial p}{\partial \theta}\, dX$$

Во втором интеграле умножим и разделим на p:

$$\int \frac{\partial \ln p}{\partial \theta} \cdot \frac{\partial p / \partial \theta}{p} \cdot p \, dX = \int \left(\frac{\partial \ln p}{\partial \theta} \right)^2 \cdot p \, dX = \mathbb{E} \left[\left(\frac{\partial \ln p}{\partial \theta} \right)^2 \right]$$

Получаем:

$$0 = \mathbb{E} \left[\frac{\partial^2 \ln p}{\partial \theta^2} \right] + \mathbb{E} \left[\left(\frac{\partial \ln p}{\partial \theta} \right)^2 \right]$$

Второе слагаемое равно $i(\theta)$, откуда:

IMPORTANT: Альтернативная формула для информации Фишера

$$\boxed{i(\theta) = -\mathbb{E} \left[\frac{\partial^2 \ln p(X, \theta)}{\partial \theta^2} \right]}$$

Часто удобнее для вычислений, чем определение через дисперсию.

Замечание о записи

В выкладках для одного наблюдения индекс k можно опустить — поскольку все X_k одинаково распределены, можно считать $k = 1$ или просто писать X без индекса.

Что будет в следующей лекции

1. Конкретные примеры вычисления информации Фишера для разных распределений.
2. Связь информации Фишера с **методом максимального правдоподобия**.
3. Связь информации Фишера с **оптимальностью оценок** (в частности, неравенство Крамера-Рао).

Лекция 5: Информация Фишера, неравенство Рао-Крамера и доверительные интервалы

1. Информация Фишера: напоминание определений

В прошлый раз была введена **информация Фишера**.

Информация Фишера для всей выборки определяется как дисперсия логарифмической функции правдоподобия:

$$I_n(\theta) = D\left(\frac{\partial \ln L(x,\theta)}{\partial \theta}\right)$$

Для одного наблюдения:

$$i(\theta) = D\left(\frac{\partial \ln p(x,\theta)}{\partial \theta}\right)$$

В силу того, что математическое ожидание этой величины равно нулю, дисперсия совпадает с математическим ожиданием квадрата:

$$i(\theta) = E\left[\left(\frac{\partial \ln p(x,\theta)}{\partial \theta}\right)^2\right]$$

Альтернативная формула (через вторую производную):

$$i(\theta) = -E\left[\frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2}\right]$$

---

2. Пример 1. Распределение Бернулли

Рассмотрим выборку из распределения Бернулли. Плотность (точнее, вероятность):

$$p(x,p) = p^x(1-p)^{1-x}$$

где $x \in \{0,1\}$: исход 1 с вероятностью p, исход 0 с вероятностью $1-p$.

Шаг 1. Логарифм:

$$\ln p(x,p) = x\ln p + (1-x)\ln(1-p)$$

Шаг 2. Первая производная по p:

$$\frac{\partial \ln p}{\partial p} = \frac{x}{p} - \frac{1-x}{1-p}$$

Шаг 3. Вторая производная по p:

$$\frac{\partial^2 \ln p}{\partial p^2} = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$$

Шаг 4. Информация Фишера для одного наблюдения (используем формулу через вторую производную, со знаком минус):

$$i(p) = -E\left[\frac{\partial^2 \ln p}{\partial p^2}\right] = E\left[\frac{x}{p^2} + \frac{1-x}{(1-p)^2}\right]$$

Пользуемся линейностью матожидания и тем, что $E[x] = p$:

$$i(p) = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$

Итог: для распределения Бернулли

$$\boxed{i(p) = \frac{1}{p(1-p)}}$$

Информация Фишера для всей выборки:

$$I_n(p) = \frac{n}{p(1-p)}$$

---

3. Пример 2. Равномерное распределение

Здесь нужно быть внимательным. **Информация Фишера не определена**, так как **не выполняются условия регулярности**.

> **Необходимое условие регулярности:** множество значений случайной величины не должно зависеть от параметра.

Для равномерного распределения множество значений зависит от параметра — поэтому модель **нерегулярна** и информация Фишера для неё не определяется.

---

4. Многомерная информация Фишера

Если параметр θ не одномерный, а многомерный, формулу можно обобщить. Информационная **матрица Фишера**:

$$I(\theta)_{ij} = -E\left[\frac{\partial^2 \ln p(x,\theta)}{\partial\theta_i\,\partial\theta_j}\right]$$

---

5. Пример 3. Нормальное распределение $N(\mu, b)$

Здесь $b = \sigma^2$ — дисперсия. Плотность:

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

Логарифм плотности:

$$\ln p(x,\mu,b) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln b - \frac{(x-\mu)^2}{2b}$$

Первые производные

По μ (первые два слагаемых обнуляются):

$$\frac{\partial \ln p}{\partial \mu} = \frac{x-\mu}{b}$$

По b:

$$\frac{\partial \ln p}{\partial b} = -\frac{1}{2b} + \frac{(x-\mu)^2}{2b^2}$$

Вторые производные

По μ дважды:

$$\frac{\partial^2 \ln p}{\partial \mu^2} = -\frac{1}{b}$$

Смешанная (по μ и b):

$$\frac{\partial^2 \ln p}{\partial \mu\, \partial b} = -\frac{x-\mu}{b^2}$$

По b дважды:

$$\frac{\partial^2 \ln p}{\partial b^2} = \frac{1}{2b^2} - \frac{(x-\mu)^2}{b^3}$$

Информационная матрица

Берём $-E[\cdot]$ от каждой второй производной.

- $-E\left[-\frac{1}{b}\right] = \frac{1}{b}$
- Смешанная: $-E\left[-\frac{x-\mu}{b^2}\right] = \frac{1}{b^2} \cdot E[x-\mu] = 0$ (т. к. $E[x] = \mu$)
- По b дважды: $-E\left[\frac{1}{2b^2} - \frac{(x-\mu)^2}{b^3}\right] = -\frac{1}{2b^2} + \frac{E[(x-\mu)^2]}{b^3} = -\frac{1}{2b^2} + \frac{b}{b^3} = \frac{1}{2b^2}$

  Здесь использовано, что $E[(x-\mu)^2] = D(x) = b$.

Итог — информационная матрица для нормального распределения:

$$\boxed{I(\mu, b) = \begin{pmatrix} \frac{1}{b} & 0 \\ 0 & \frac{1}{2b^2} \end{pmatrix}}$$

---

6. Неравенство Рао–Крамера

Эта теорема устанавливает нетривиальную нижнюю границу дисперсии несмещённой оценки.

Формулировка

Пусть выполнены условия: - модель **регулярна** (в смысле, обсуждавшемся ранее); - $\tau(\theta)$ — оцениваемая функция, $\tau \in C^1$ (непрерывно дифференцируема); - в частном случае $\tau(\theta) = \theta$ — оценивается сам параметр; - $T(x)$ — оценка функции $\tau(\theta)$; - $E[T(x)] = \tau(\theta)$ (оценка **несмещённая**).

Тогда:

$$D(T(x)) \geq \frac{(\tau'(\theta))^2}{n \cdot i(\theta)}$$

Доказательство

Стартуем из:

$$\tau(\theta) = E[T(x)] = \int T(x) \cdot L(x, \theta)\, dx$$

Модель регулярна — продифференцируем тождество по θ (регулярность позволяет вносить производную под интеграл):

$$\tau'(\theta) = \int T(x) \cdot \frac{\partial L(x, \theta)}{\partial \theta}\, dx$$

Трюк: домножим и разделим на функцию правдоподобия:

$$\tau'(\theta) = \int T(x) \cdot \frac{\partial \ln L(x, \theta)}{\partial \theta} \cdot L(x, \theta)\, dx$$

Здесь использовано: $\dfrac{\partial \ln L}{\partial \theta} = \dfrac{1}{L} \cdot \dfrac{\partial L}{\partial \theta}$ — логарифмическая производная.

То есть это **матожидание произведения**:

$$\tau'(\theta) = E[T(x) \cdot V(x, \theta)]$$

где $V(x,\theta) = \dfrac{\partial \ln L(x,\theta)}{\partial \theta}$ — **вклад выборки**.

В прошлый раз было показано, что $E[V(x,\theta)] = 0$. Значит,

$$\tau'(\theta) = E[T(x)\cdot V(x,\theta)] - E[T(x)]\cdot \underbrace{E[V(x,\theta)]}_{=0} = \text{Cov}(T(x), V(x,\theta))$$

Возведём в квадрат:

$$(\tau'(\theta))^2 = \text{Cov}^2(T(x), V(x,\theta))$$

Применяем **вероятностный аналог неравенства Коши–Буняковского** для ковариации:

$$\text{Cov}^2(T,V) \le D(T)\cdot D(V)$$

А $D(V(x,\theta))$ — это в точности $n\cdot i(\theta)$ (информация Фишера для всей выборки). Отсюда

$$(\tau'(\theta))^2 \le D(T(x))\cdot n\cdot i(\theta) \quad\Longrightarrow\quad D(T(x)) \ge \frac{(\tau'(\theta))^2}{n\cdot i(\theta)}$$

Что и требовалось доказать. ■

---

7. Замечания к неравенству Рао–Крамера

Замечание 1. Связь с MSE

Вспомним:

$$\text{MSE} = D(T) + (\text{смещение})^2$$

Если оценка **несмещённая**, то MSE $= D(T)$. Значит, при выполнении условий регулярности и несмещённости оценки можно дать нижнюю границу не только для дисперсии, но и для MSE.

> **Если в регулярной модели несмещённая оценка достигает нижней границы Рао-Крамера, то она оптимальная.**

То есть в несмещённой ситуации в регулярной модели оценка оптимальна тогда и только тогда, когда её дисперсия достигает нижней границы Рао–Крамера.

Замечание 2. Многомерная формулировка

Пусть $\tau(\theta)$ — функция из $\mathbb{R}^d \to \mathbb{R}$, и $T(x)$ — несмещённая оценка для $\tau(\theta)$. Тогда:

$$D(T(x)) \geq \frac{1}{n} \cdot \nabla\tau(\theta)^\top \cdot I^{-1}(\theta) \cdot \nabla\tau(\theta)$$

---

8. Возвращение к примерам — проверка оптимальности

Бернулли

Получили: $i(p) = \dfrac{1}{p(1-p)}$. Нижняя граница Рао–Крамера:

$$D(\hat{p}) \geq \frac{p(1-p)}{n}$$

Стандартная оценка — выборочное среднее $\hat{p} = \bar{x}$ (доля единиц). Это **несмещённая** оценка, и

$$D(\bar{x}) = \frac{1}{n} D(\text{Bern}) = \frac{p(1-p)}{n}$$

Дисперсия совпадает с нижней границей $\Rightarrow$ **выборочное среднее — оптимальная оценка для p в распределении Бернулли**.

> **Вопрос из аудитории:** а если бы дисперсия не совпала? **Ответ:** если бы дисперсия была больше, мы бы сказали, что оценка не оптимальна. В общем случае задача поиска оптимальной оценки не разрешима, но известно, что **оценка максимального правдоподобия в регулярном случае асимптотически эффективна** (см. ниже).

Нормальное распределение, оценка для μ

Пусть $\tau(\mu, b) = \mu$. Градиент: $\nabla\tau = (1, 0)^\top$.

Информационная матрица $I = \begin{pmatrix} 1/b & 0 \\ 0 & 1/(2b^2) \end{pmatrix}$. Обратная: $I^{-1} = \begin{pmatrix} b & 0 \\ 0 & 2b^2 \end{pmatrix}$.

Нижняя граница:

$$D(\hat{\mu}) \geq \frac{1}{n}(1,0) \begin{pmatrix} b & 0 \\ 0 & 2b^2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{b}{n}$$

Стандартная оценка матожидания — $\hat{\mu} = \bar{x}$ (выборочное среднее). Это несмещённая оценка с $D(\bar{x}) = \dfrac{b}{n}$.

Снова дисперсия совпала с нижней границей ⇒ **выборочное среднее — оптимальная оценка для матожидания нормального закона**.

---

9. Асимптотическая нормальность ОМП

Формулировка теоремы

Условия: - $\left|\dfrac{\partial^3 \ln p(x,\theta)}{\partial \theta^3}\right| \leq M(x)$, причём $E[M(x)] < \infty$; - модель регулярна.

Утверждение:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0,\, i^{-1}(\theta)\right)$$

Здесь $\hat{\theta}$ — **оценка максимального правдоподобия**.

Интерпретация

Неформально: при больших n

$$\hat{\theta} \approx N\left(\theta,\, \frac{1}{n \cdot i(\theta)}\right)$$

То есть **асимптотическая дисперсия ОМП совпадает с нижней границей Рао-Крамера**. Поэтому можно говорить, что оценки максимального правдоподобия **асимптотически эффективны**.

---

10. Переход к доверительным интервалам

Точечное оценивание даёт оценку в виде конкретного числа. Но число не всегда наглядно. Иногда удобнее давать оценку в виде **диапазона** — это и есть **доверительные интервалы** (confidence intervals).

Определение

Пусть выборка $x_1, \dots, x_n$ из распределения F с параметром θ (одномерным). Рассмотрим две статистики $L(x)$ и $R(x)$.

Будем говорить, что $(L(x), R(x))$ образуют **доверительный интервал уровня доверия** $1-\alpha$, если

$$P(\theta \in (L(x), R(x))) \geq 1-\alpha$$

Практическая интерпретация

Обычно $1-\alpha$ берут 90%, 95% или 99%.

Пример. Пусть уровень доверия 95%. Если провести 100 экспериментов и для каждой выборки построить свой доверительный интервал, то **хотя бы в 95 случаях из 100 реальное значение параметра попадёт в построенный интервал**. То есть «хорошими» в этом смысле будут не менее 95 из 100 интервалов.

> **Замечание об обозначениях.** Уровень доверия часто обозначают буквой $\gamma = 1-\alpha$. Тогда квантили будут порядка $\frac{1-\gamma}{2}$ слева и $\frac{1+\gamma}{2}$ справа. В лекции используется обозначение через α — это связано с другой задачей (проверка гипотез), которую рассмотрим позже.

---

11. Общая схема построения доверительного интервала

Шаг 1. Найти функцию $g(x, \theta)$ — статистику, аналитически зависящую от выборки и параметра, такую что **распределение** $g(x, \theta)$ **не зависит от** θ.

Шаг 2. Записать вероятность

$$P(L \leq g(x, \theta) \leq R) = 1-\alpha$$

Шаг 3. На графике плотности отсечь: - слева вероятностную массу $\alpha/2$; - справа вероятностную массу $\alpha/2$; - посередине останется $1-\alpha$.

Тогда:

$$L = q_{\alpha/2}, \quad R = q_{1-\alpha/2}$$

— квантили распределения статистики g.

Шаг 4. Разрешить неравенство относительно θ — получится доверительный интервал.

---

12. Доверительный интервал для матожидания при известной дисперсии

Условия: выборка $x_1, \dots, x_n$ из $N(\mu, \sigma^2)$, причём σ^2 известна. Строим доверительный интервал для μ.

Рецепт 1 (плохой). Использование одного элемента

Статистика:

$$g_1(x) = \frac{x_1 - \mu}{\sigma} \sim N(0,1)$$

Её распределение не зависит от μ.

Зажимаем квантилями стандартного нормального (распределение симметрично относительно 0):

$$-q_{1-\alpha/2} \le \frac{x_1 - \mu}{\sigma} \le q_{1-\alpha/2}$$

Разрешаем относительно μ:

$$\boxed{\mu \in \left(x_1 - \sigma \cdot q_{1-\alpha/2},\ x_1 + \sigma \cdot q_{1-\alpha/2}\right)}$$

Рецепт 2 (хороший). Использование всей выборки

Статистика:

$$g_2(x) = \sqrt{n} \cdot \frac{\bar{x} - \mu}{\sigma} \sim N(0,1)$$

(центрировали и нормировали — снова стандартное нормальное).

Зажимаем квантилями:

$$-q_{1-\alpha/2} \leq \sqrt{n} \cdot \frac{\bar{x} - \mu}{\sigma} \leq q_{1-\alpha/2}$$

Разрешаем относительно μ:

$$\boxed{\mu \in \bar{x} \pm \frac{\sigma \cdot q_{1-\alpha/2}}{\sqrt{n}}}$$

Сравнение

Какой интервал лучше? Второй, потому что: - в нём участвует **вся выборка** (а не только первый элемент); - **длина интервала уменьшается** с ростом n (за счёт деления на $\sqrt{n}$); - **середина интервала** ($\bar{x}$) при увеличении объёма выборки становится всё ближе к реальному значению параметра.

Оба интервала имеют один и тот же уровень доверия $1 - \alpha$, но второй — содержательно лучше.

Терминология

В контексте доверительных интервалов выражения вида **«оценка $\pm$ что-то»** возникают часто. Величина перед квантилью называется **стандартной ошибкой** (SE, Standard Error).

> Для квантилей нормального закона нередко используют букву z (вместо q). На лекции используется q, но из контекста всегда понятно, какое распределение имеется в виду.

---

13. Три важных вспомогательных распределения

Перед тем как переходить к следующим задачам, нужно ввести три распределения, играющих ключевую роль в статистике.

A. Распределение хи-квадрат χ^2_n

Пусть $x_1, x_2, \dots, x_n$ — независимые случайные величины, каждая со стандартным нормальным распределением $N(0,1)$. Тогда

$$\sum_{k=1}^{n} x_k^2 \sim \chi^2_n$$

Параметр n — **число степеней свободы** (это просто количество независимых слагаемых).

> **Связь с гамма-распределением:** χ^2_n — это гамма-распределение с параметрами $\left(\frac{n}{2}, \frac{1}{2}\right)$. То есть класс распределений χ^2 содержится в классе гамма-распределений.

B. Распределение Стьюдента t_n

Пусть $x_0, x_1, \dots, x_n$ — независимые $N(0,1)$. Рассмотрим:

$$T_n = \frac{x_0}{\sqrt{\frac{1}{n}\sum\limits_{k=1}^{n} x_k^2}}$$

Тогда T_n имеет **распределение Стьюдента** (t-распределение) с n степенями свободы: $T_n \sim t_n$.

Под корнем стоит **усреднённый** χ^2.

Свойства распределения Стьюдента

1. **Симметрично относительно нуля.** Числитель — стандартное гауссовское (симметрично), знаменатель — неотрицательная константа.
2. **При больших n близко к нормальному.** По закону больших чисел знаменатель $\to 1$, и остаётся гауссовская величина. То есть при больших n: $t_n \approx N(0,1)$.

C. Распределение Фишера $F_{n,m}$

Пусть χ^2_n имеет распределение хи-квадрат с n степенями свободы, а χ^2_m — независимая случайная величина с распределением хи-квадрат с m степенями свободы.

Тогда

$$F_{n,m} = \frac{\chi^2_n / n}{\chi^2_m / m}$$

имеет **распределение Фишера** с параметрами n и m.

Где используются эти распределения

- **Нормальное** — при построении доверительного интервала для матожидания, если дисперсия известна (или, забегая вперёд, при больших объёмах выборки — по ЦПТ).
- χ^2 — при построении доверительного интервала для дисперсии.
- **Стьюдент** — при построении доверительного интервала для матожидания, если дисперсия неизвестна.
- **Фишер** — при построении доверительного интервала для отношения дисперсий.

---

14. Доверительный интервал для дисперсии при известном матожидании

Условия: выборка из $N(\mu, \sigma^2)$, μ **известно**, строим интервал для σ^2.

Почему нельзя использовать прежнюю статистику

Если попробовать взять $g(x) = \sqrt{n} \cdot \dfrac{\bar{x} - \mu}{\sigma}$, то при разрешении неравенства относительно σ возникнут проблемы: $\bar{x} - \mu$ может быть как положительным, так и отрицательным, и при делении/умножении на эту величину знаки неравенств будут меняться по-разному. Это неудобно.

Правильная статистика

Заметим, что $\dfrac{x_k - \mu}{\sigma} \sim N(0, 1)$. Значит, по определению χ^2:

$$\sum_{k=1}^{n} \frac{(x_k - \mu)^2}{\sigma^2} \sim \chi^2_n$$

Это распределение не зависит от σ.

Зажимаем квантилями

Распределение χ^2_n **не симметрично относительно нуля** (плотность сосредоточена на $[0, \infty)$, асимметрична), поэтому **обе квантили нужно считать честно**:

$$q_{\alpha/2} \leq \sum_{k=1}^{n} \frac{(x_k - \mu)^2}{\sigma^2} \leq q_{1-\alpha/2}$$

где q — квантили распределения χ^2_n.

Разрешаем относительно σ^2

Из неравенства $\sum_k (x_k - \mu)^2 / \sigma^2 \leq q_{1-\alpha/2}$ получаем $\sigma^2 \geq \dfrac{\sum_k (x_k - \mu)^2}{q_{1-\alpha/2}}$.

Аналогично с другой стороны.

Итог:

$$\boxed{\sigma^2 \in \left(\frac{\sum\limits_{k=1}^{n} (x_k - \mu)^2}{q_{1-\alpha/2}}, \ \frac{\sum\limits_{k=1}^{n} (x_k - \mu)^2}{q_{\alpha/2}} \right)}$$

где квантили — распределения χ^2_n.

---

15. Теорема Фишера

Перед следующей задачей понадобится ключевая теорема. В разных источниках в неё включают разные пункты, приведём основные.

Условия: выборка $x_1, \ldots, x_n$ из гауссовского закона $N(\mu, \sigma^2)$.

Пункт 1

$$\frac{n \cdot S^2}{\sigma^2} = \frac{(n-1) \cdot S^{*2}}{\sigma^2} \sim \chi^2_{n-1}$$

где - $S^2 = \frac{1}{n}\sum\limits_{k=1}^{n}(x_k - \bar{x})^2$ — **смещённая** выборочная дисперсия; - $S^{*2} = \frac{1}{n-1}\sum\limits_{k=1}^{n}(x_k - \bar{x})^2$ — **несмещённая** выборочная дисперсия.

> **Неформально, почему** $n-1$**.** В каждом слагаемом $(x_k - \bar{x})^2$ участвует выборочное среднее $\bar{x}$, которое «портит» независимость слагаемых. За счёт этой связи число степеней свободы уменьшается на единицу.

Пункт 2

$\bar{x}$ и S^2 **независимы** (а также $\bar{x}$ и S^{*2} независимы).

> Это не очевидное наблюдение: в обеих статистиках на первый взгляд участвует $\bar{x}$ — казалось бы, они должны быть зависимы. Однако для выборок из нормального закона эти статистики **независимы**. Это нетривиальное свойство именно гауссовского распределения.

Эти два пункта потребуются для решения следующих задач.

---

16. Доверительный интервал для дисперсии при неизвестном математическом ожидании

Условия: выборка из $N(\mu, \sigma^2)$, μ **неизвестно**, строим интервал для σ^2.

Здесь нельзя использовать предыдущий рецепт, поскольку в нём фигурировало μ. На помощь приходит теорема Фишера.

По теореме Фишера:

$$\frac{n \cdot S^{*2}}{\sigma^2} \sim \chi^2_{n-1}$$

(Замечание Ивана Александровича: в записи может быть как S^2, так и S^{*2} — нужно следить, что именно записано: смещённая или несмещённая дисперсия. Здесь идёт зажатие именно этой статистики.)

Зажимаем квантилями

$$q_{\alpha/2} \leq \frac{n \cdot S^{*2}}{\sigma^2} \leq q_{1-\alpha/2}$$

где квантили — распределения χ^2_{n-1}.

Разрешаем относительно σ^2

$$\boxed{\sigma^2 \in \left(\frac{n \cdot S^{*2}}{q_{1-\alpha/2}},\ \frac{n \cdot S^{*2}}{q_{\alpha/2}} \right)}$$

где квантили распределения χ^2_{n-1}.

---

17. Доверительный интервал для матожидания при неизвестной дисперсии

Условия: выборка из $N(\mu, \sigma^2)$, дисперсия **неизвестна**, строим интервал для μ.

Здесь нельзя использовать рецепт со стандартным нормальным, поскольку в нём фигурирует σ.

Подбираем статистику

Рассмотрим:

$$T = \sqrt{n-1} \cdot \frac{\bar{x} - \mu}{S} = \sqrt{n} \cdot \frac{\bar{x} - \mu}{S^*}$$

> **Почему S, а не S^2?** Физически $\bar{x} - \mu$ — это «метры», а S^2 — это «метры в квадрате». Математически: при нормировании мы делим на **стандартное отклонение**, а не на дисперсию.

Распределение этой статистики

Перепишем:

$$T = \frac{\sqrt{n} \cdot (\bar{x} - \mu)/\sigma}{\sqrt{S^{*2}/\sigma^2}}$$

- В числителе: $\sqrt{n} \cdot \dfrac{\bar{x} - \mu}{\sigma} \sim N(0,1)$ — стандартная гауссовская величина.
- В знаменателе под корнем: $\dfrac{S^{*2}}{\sigma^2}$ связано с χ^2_{n-1} по теореме Фишера, причём поделенным на число степеней свободы.
- По теореме Фишера числитель и знаменатель **независимы**.

По определению распределения Стьюдента (отношение нормального к корню из «усреднённого χ^2») получаем:

$$T \sim t_{n-1}$$

Доверительный интервал

Распределение Стьюдента **симметрично относительно нуля**, поэтому:

$$-q_{1-\alpha/2} \leq \sqrt{n} \cdot \frac{\bar{x} - \mu}{S^*} \leq q_{1-\alpha/2}$$

где q — квантили распределения t_{n-1}.

Разрешая относительно μ:

$$\boxed{\mu \in \bar{x} \pm \frac{S^* \cdot q_{1-\alpha/2}}{\sqrt{n}}}$$

где $q_{1-\alpha/2}$ — квантиль распределения Стьюдента t_{n-1}.

> Это **доверительный интервал для матожидания нормального закона при неизвестной дисперсии** — в нём как раз и используется распределение Стьюдента.

---

18. Итоговая таблица: сводка доверительных интервалов для $N(\mu, \sigma^2)$

Параметр	Что известно	Используемое распределение	Доверительный интервал
μ	σ^2 известно	$N(0,1)$	$\bar{x} \pm \dfrac{\sigma \cdot q_{1-\alpha/2}}{\sqrt{n}}$
μ	σ^2 неизвестно	t_{n-1}	$\bar{x} \pm \dfrac{S^* \cdot q_{1-\alpha/2}}{\sqrt{n}}$
σ^2	μ известно	χ^2_n	$\left(\dfrac{\sum(x_k-\mu)^2}{q_{1-\alpha/2}}, \dfrac{\sum(x_k-\mu)^2}{q_{\alpha/2}}\right)$
σ^2	μ неизвестно	χ^2_{n-1}	$\left(\dfrac{nS^{*2}}{q_{1-\alpha/2}}, \dfrac{nS^{*2}}{q_{\alpha/2}}\right)$

---

19. Что будет в следующий раз

- Доверительный интервал для **разности матожиданий** двух выборок.
- Доверительный интервал для **отношения дисперсий** (здесь будет работать распределение Фишера).
- Ивана Александровича обещал прислать листинг с конкретными числовыми примерами.

Лекция 6: Доверительные интервалы и введение в проверку статистических гипотез

Повторение: определение доверительного интервала

Формальное определение. Доверительный интервал $[L(x), R(x)]$ задаётся условием:

$$P(\theta \in [L(x), R(x)]) \geq 1 - \alpha$$

где $1-\alpha$ — **уровень доверия**.

Содержательная интерпретация. Если уровень доверия 95% и мы рассматриваем 100 выборок, для каждой считаем доверительный интервал, то хотя бы в 95 случаях из 100 реальное значение параметра окажется в построенном доверительном интервале.

Что было раньше: на прошлой лекции рассматривались доверительные интервалы для параметров нормального закона: - для мат. ожидания при известной дисперсии - для мат. ожидания при неизвестной дисперсии - для дисперсии при известном мат. ожидании - для дисперсии при неизвестном мат. ожидании

Задача 5: Доверительный интервал для разности мат. ожиданий (известные дисперсии)

Постановка

Даны две **независимые** выборки: - $x_1, \dots, x_n$ из $N(\mu_x, \sigma_x^2)$ - $y_1, \dots, y_m$ из $N(\mu_y, \sigma_y^2)$

Дисперсии σ_x^2 и σ_y^2 **известны**. Нужно построить доверительный интервал для $\tau = \mu_y - \mu_x$.

Построение

Шаг 1. Распределения выборочных средних:

$$\bar{x} \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right), \qquad \bar{y} \sim N\left(\mu_y, \frac{\sigma_y^2}{m}\right)$$

Шаг 2. Из независимости выборок:

$$\bar{y} - \bar{x} \sim N\left(\mu_y - \mu_x,\ \frac{\sigma_y^2}{m} + \frac{\sigma_x^2}{n}\right)$$

> **Важное замечание.** Дисперсия суммы (или разности) независимых случайных величин — это **сумма** дисперсий, независимо от того, плюс это или минус. Если бы здесь стоял минус, то могло бы получиться отрицательное значение, что невозможно для дисперсии.

Шаг 3. Центрируем и нормируем:

$$\frac{\bar{y} - \bar{x} - \tau}{\sqrt{\dfrac{\sigma_y^2}{m} + \dfrac{\sigma_x^2}{n}}} \sim N(0, 1)$$

Шаг 4. Зажимаем статистику между квантилями (используем симметрию стандартного нормального закона) и разрешаем неравенство относительно τ.

Ответ

$$\boxed{\tau \in \bar{y} - \bar{x} \pm u_{1-\alpha/2}\sqrt{\frac{\sigma_y^2}{m} + \frac{\sigma_x^2}{n}}}$$

Задача 6: Доверительный интервал для разности мат. ожиданий (равные неизвестные дисперсии)

Постановка

Те же две независимые гауссовские выборки, но теперь: - дисперсии **неизвестны** - известно, что $\sigma_x^2 = \sigma_y^2 = \sigma^2$

Цель та же: построить доверительный интервал для $\tau = \mu_y - \mu_x$.

Идея

В предыдущей задаче мы получили стандартную гауссовскую величину. Сейчас её знаменатель содержит неизвестное σ^2. Идея — построить статистику с распределением Стьюдента.

Напоминание определения t-распределения: в числителе — стандартная гауссовская величина, в знаменателе — корень квадратный из χ^2, делённого на число степеней свободы; числитель и знаменатель независимы.

Применение теоремы Фишера

По теореме Фишера:

$$\frac{ns_x^{*2}}{\sigma^2} \sim \chi^2_{n-1}, \qquad \frac{ms_y^{*2}}{\sigma^2} \sim \chi^2_{m-1}$$

где s^{*2} — смещённая выборочная дисперсия.

Поскольку x и y независимы, при сложении степени свободы складываются:

$$\frac{ns_x^{*2} + ms_y^{*2}}{\sigma^2} \sim \chi^2_{n+m-2}$$

При этом числитель (выборочные средние) и знаменатель (выборочные дисперсии) **независимы** — также по теореме Фишера.

Построение статистики

$$T = \frac{\dfrac{\bar{y} - \bar{x} - \tau}{\sqrt{\sigma^2/m + \sigma^2/n}}}{\sqrt{\dfrac{1}{n+m-2}\left(\dfrac{ns_x^{*2}}{\sigma^2} + \dfrac{ms_y^{*2}}{\sigma^2}\right)}} \sim t_{n+m-2}$$

Ключевой момент: σ^2 в числителе и знаменателе **сокращаются**.

После упрощения:

$$T = \frac{(\bar{y} - \bar{x} - \tau)\sqrt{(n+m-2)\,mn}}{\sqrt{(m+n)(ns_x^{*2} + ms_y^{*2})}} \sim t_{n+m-2}$$

Зажатие между квантилями

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

Это работает потому, что распределение Стьюдента **симметрично** относительно нуля.

Ответ

$$\boxed{\tau \in \bar{y} - \bar{x} \pm t_{1-\alpha/2}\sqrt{\frac{(ns_x^{*2} + ms_y^{*2})(m+n)}{mn(n+m-2)}}}$$

 Замечание Ивана Александровича. Это самая громоздкая задача на сегодня — дальше будет проще.

 Если дисперсии неравны и неизвестны — задача формально неразрешима в таком виде (в общем случае точного решения нет — это

так называемая проблема Беренса–Фишера).

Задача 7: Доверительный интервал для отношения дисперсий (мат. ожидания неизвестны)

Постановка

Две независимые гауссовские выборки $x_1, \dots, x_n$ и $y_1, \dots, y_m$. Мат. ожидания μ_x, μ_y **неизвестны**. Построить доверительный интервал для σ_y^2/σ_x^2.

Применение теоремы Фишера

$$\frac{(n-1)s_x^2}{\sigma_x^2} \sim \chi^2_{n-1}, \qquad \frac{(m-1)s_y^2}{\sigma_y^2} \sim \chi^2_{m-1}$$

(здесь s^2 — несмещённая выборочная дисперсия)

Построение F-статистики

По определению F-распределения (отношение двух χ^2, делённых на свои степени свободы):

$$F_{n,m} = \frac{\dfrac{(n-1)s_x^2}{\sigma_x^2} \cdot \dfrac{1}{n-1}}{\dfrac{(m-1)s_y^2}{\sigma_y^2} \cdot \dfrac{1}{m-1}} = \frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2} = \frac{s_x^2}{s_y^2} \cdot \frac{\sigma_y^2}{\sigma_x^2} \sim F_{n-1,\, m-1}$$

Зажатие между квантилями

$$P\left(F_{\alpha/2} \le \frac{s_x^2}{s_y^2} \cdot \frac{\sigma_y^2}{\sigma_x^2} \le F_{1-\alpha/2} \right) = 1 - \alpha$$

 Распределение Фишера **не симметрично** относительно нуля, поэтому используются обе квантили: $F_{\alpha/2}$ и $F_{1-\alpha/2}$.

Ответ

После разрешения относительно σ_y^2/σ_x^2 (важно: при делении на дробь неравенство переворачивается):

$$\boxed{\frac{\sigma_y^2}{\sigma_x^2} \in \left[F_{\alpha/2}\,\frac{s_y^2}{s_x^2},\ F_{1-\alpha/2}\,\frac{s_y^2}{s_x^2} \right]}$$

Задача 8: Доверительный интервал для отношения дисперсий (мат. ожидания известны)

Постановка

То же, но μ_x и μ_y **известны**.

Идея

Формально можно использовать прежнюю статистику, но при малом объёме выборки лучше иметь больше степеней свободы.

Используем тот факт, что:

$$\sum_{k=1}^{n} \frac{(x_k - \mu_x)^2}{\sigma_x^2} \sim \chi_n^2, \qquad \sum_{k=1}^{m} \frac{(y_k - \mu_y)^2}{\sigma_y^2} \sim \chi_m^2$$

(степеней свободы на одну больше, чем в задаче 7)

Построение F-статистики

$$F = \frac{\dfrac{1}{n}\sum_{k=1}^{n} \dfrac{(x_k - \mu_x)^2}{\sigma_x^2}}{\dfrac{1}{m}\sum_{k=1}^{m} \dfrac{(y_k - \mu_y)^2}{\sigma_y^2}} \sim F_{n,m}$$

Дальше — стандартная процедура: зажатие между квантилями и разрешение относительно σ_y^2/σ_x^2.

«Универсальный» рецепт (в кавычках)

Постановка

Пусть $x_1, \dots, x_n$ — выборка из непрерывного распределения с функцией распределения F_θ.

Утверждение (а)

Случайная величина $u_i = F_\theta(x_i)$ имеет **равномерное распределение** на $[0, 1]$.

Доказательство. Для строго возрастающей F_θ:

$$P(u_i \le t) = P(F_\theta(x_i) \le t) = P(x_i \le F_\theta^{-1}(t)) = F_\theta(F_\theta^{-1}(t)) = t$$

(для $t \in [0, 1]$). Это функция распределения равномерного закона на $[0, 1]$.

Утверждение (б)

$v_i = -\ln u_i$ распределено по **экспоненциальному закону** с параметром 1.

Доказательство. Для $t > 0$:

$$P(v_i \le t) = P(-\ln u_i \le t) = P(\ln u_i \ge -t) = P(u_i \ge e^{-t}) = 1 - e^{-t}$$

Это функция распределения экспоненциального закона с параметром 1.

Утверждение (в)

$$\sum_{i=1}^{n} v_i = -\sum_{i=1}^{n} \ln F_\theta(x_i) \sim \Gamma(n, 1)$$

(сумма независимых экспоненциальных случайных величин с одним параметром даёт гамма-распределение).

Почему «в кавычках»?

Формально мы получили статистику с хорошим (известным) распределением для очень широкого класса задач. Но дальше нужно зажимать между квантилями и разрешать неравенство относительно θ. На статистику навешан и логарифм,

и функция распределения — **получается неравенство, которое часто либо очень трудно разрешимо, либо в принципе неразрешимо** относительно θ.

Пример: $F_\theta(t) = \frac{1}{2} + \frac{1}{\pi}\arctan(t-\theta)$ (распределение Коши со сдвигом). Статистика:

$$-\sum_{i=1}^{n} \ln\left(\frac{1}{2} + \frac{1}{\pi}\arctan(x_i - \theta)\right)$$

Зажать это между квантилями и разрешить относительно θ — крайне неприятная задача.

Асимптотические доверительные интервалы

Определение

$[L(x), R(x)]$ — **асимптотический доверительный интервал**, если:

$$\lim_{n\to\infty} P(\theta \in [L(x), R(x)]) \geq 1 - \alpha$$

Общая схема построения

1. Находим статистику $g(x, \theta)$, у которой существует **предельное распределение**, не зависящее от θ.
2. Зажимаем статистику между квантилями **предельного** распределения:

$$P(q_{\alpha/2} \leq g(x,\theta) \leq q_{1-\alpha/2}) \approx 1 - \alpha$$

3. Разрешаем неравенство относительно θ.

Применение А: Асимптотический ДИ для мат. ожидания

Условие

Существует дисперсия.

Использование

Выборочная средняя — асимптотически нормальная оценка:

$$\frac{\sqrt{n}(\bar{x}-\mu)}{s} \xrightarrow{d} N(0,1)$$

Зажатие между квантилями

$$-u_{1-\alpha/2} \leq \frac{\sqrt{n}(\bar{x}-\mu)}{s} \leq u_{1-\alpha/2}$$

Ответ

$$\boxed{\mu \in \bar{x} \pm u_{1-\alpha/2}\frac{s}{\sqrt{n}}}$$

Стандартная ошибка

> **Определение.** В контексте доверительных интервалов **стандартной ошибкой** называется величина, на которую умножается квантиль, — то есть $\frac{s}{\sqrt{n}}$.

Частный случай: ДИ для параметра распределения Бернулли

Постановка

Выборка из распределения Бернулли с параметром p. Мат. ожидание = p, дисперсия = $p(1-p)$.

Сходимость

$$\frac{\sqrt{n}(\bar{x}-p)}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1)$$

Проблема

В знаменателе p — неизвестно. Если оставить как есть, при разрешении неравенства p окажется и в числителе, и в знаменателе, да ещё под корнем.

Решение — подстановка состоятельной оценки

Подставляем выборочную оценку $\hat{p} = \bar{x}$ (она же — оценка методом моментов и оценка максимального правдоподобия). Сходимость к стандартной гауссовской величине сохраняется.

Ответ

$$\boxed{p \in \bar{x} \pm u_{1-\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}}$$

Когда такая подстановка допустима? Только если оценка **состоятельна**. В асимптотическом ДИ оценки близки к реальному значению, и сходимость сохраняется.

Применение Б: Асимптотический ДИ для медианы

Условие

Выборка из **непрерывного** распределения.

Использование

Выборочная медиана (порядковая статистика с номером $\lfloor n/2 \rfloor$) — асимптотически нормальная оценка теоретической медианы:

$$\sqrt{n} \cdot f(m) \cdot \frac{x_{(\lfloor n/2 \rfloor)} - m}{\sqrt{1/2 \cdot 1/2}} \xrightarrow{d} N(0,1)$$

где m — теоретическая медиана, f — плотность.

Ответ

$$\boxed{m \in x_{(\lfloor n/2 \rfloor)} \pm \frac{u_{1-\alpha/2}}{2\sqrt{n}\, f(m)}}$$

Проблема и решение

В формуле присутствует $f(m)$ — неизвестная плотность в неизвестной точке. Решение — подставить состоятельные оценки: вместо m использовать выборочную медиану $x_{(\lfloor n/2 \rfloor)}$.

 Это типичный приём: подстановка состоятельной оценки на место неизвестной величины.

Применение B: Асимптотический ДИ для дисперсии

Использование

Выборочная дисперсия — асимптотически нормальная оценка:

$$\frac{\sqrt{n}(s^{*2} - \sigma^2)}{\sqrt{\hat{\beta}_4 - s^{*4}}} \xrightarrow{d} N(0,1)$$

где $\hat{\beta}_4 = \overline{(x - \bar{x})^4}$ — четвёртый выборочный центральный момент.

Ответ

$$\boxed{\sigma^2 \in s^{*2} \pm \frac{u_{1-\alpha/2}}{\sqrt{n}} \sqrt{\hat{\beta}_4 - s^{*4}}}$$

Тонкость

 Левая граница может оказаться отрицательной, что для дисперсии бессмысленно.

Для мат. ожидания это нормально, для дисперсии — нет. Поэтому такой подход работает только при **очень большом объёме выборки**: при $n \to \infty$ дробь $\dfrac{1}{\sqrt{n}} \to 0$, и левая граница перестаёт быть отрицательной.

Применение Г: ДИ через оценку максимального правдоподобия

Утверждение

Если $\hat{\theta}$ — оценка максимального правдоподобия для θ, и модель регулярна, то:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{1}{i(\theta)}\right)$$

где $i(\theta)$ — **информация Фишера**.

Применение

Подставив состоятельную оценку $\hat{\theta}$ в информацию Фишера:

$$\sqrt{i(\hat{\theta})} \cdot \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, 1)$$

Отсюда стандартным образом извлекается доверительный интервал для θ.

Применение Д: ДИ через порядковые статистики (экзотический рецепт)

Утверждения

Для выборки из непрерывного распределения:

$$n \cdot F(x_{(\ell)}) \xrightarrow{d} \Gamma(\ell, 1)$$

$$n \cdot (1 - F(x_{(n+1-s)})) \xrightarrow{d} \Gamma(s, 1)$$

где ℓ и s — фиксированные.

> Эти соотношения встречались при изучении порядковых статистик. Чисто гипотетически из них можно извлекать асимптотические доверительные интервалы.

Упражнения для самостоятельного решения

- Для равномерного распределения $U[0, \theta]$ построить ДИ для θ через порядковые статистики (применение Д).
- Для распределения Пуассона $\text{Pois}(\lambda)$ построить асимптотический ДИ для λ через ОМП (применение Г).

Часть 2. Введение в проверку статистических гипотез

 Это пока **предварительные мысли** о формулировке задачи. Строгая постановка будет на следующей лекции.

Установка для размышления

Иван Александрович просит рассуждать: 1. **Максимально рационально.** 2. С точки зрения человека, у которого **нет опыта** в данной предметной области (как «рациональный инопланетянин»).

В каждой ситуации нужно выделить: - **Дефолтное предположение** (по умолчанию). - **Альтернативное предположение.**

Ситуация 1. Уголовный суд

Контекст: сферическая страна в вакууме с **континентальной** системой права (суд опирается на законы; в отличие от **прецедентной** системы, как в Великобритании или США, где суд опирается на предыдущие решения по похожим делам).

Происходит уголовное дело, подсудимого обвиняют в убийстве. Вы — судья.

- H_0 (по умолчанию): человек **не виновен**.
- H_1 (альтернатива): человек **виновен в убийстве** (в конкретном преступлении!).

Тонкий момент. Альтернатива конкретна. Если по ходу дела выяснится, что подсудимый занимался мошенничеством, но обвинение в убийстве не доказано, — судья скажет «не виновен» **относительно данной альтернативы**. Это другая задача.

Ситуация 2. Робот кидает монетку

- H_0**:** монетка честная.
- H_1**:** монетка нечестная (например, выпадает слишком много орлов или слишком много решек).

Ситуация 3. Измерение температуры

Хотим понять, здоров человек или болен, измеряя температуру.

- H_0**:** человек не болен (средняя температура $= 36{,}6$).
- H_1**:** человек болен (средняя температура $\neq 36{,}6$).

Уточнения альтернативы в зависимости от контекста

Контекст	Альтернатива
Общий случай	средняя $\neq 36{,}6$
Инфекционная больница (для инфекций характерна повышенная температура)	средняя $> 36{,}6$
Заболевания с пониженной температурой	средняя $< 36{,}6$

Ситуация 4. Влияет ли вещество на здоровье

- H_0**:** вещество не влияет на здоровье.
- H_1**:** возможны разные варианты:
  - Просто влияет (любым образом).
  - Влияет положительно (если мы фармацевты).
  - Влияет отрицательно (если разрабатываем биооружие).

 Вывод. Альтернатива формулируется в зависимости от того, **что именно мы хотим проверить**.

Общая схема: H_0 и H_1

Нулевая гипотеза H_0

Это **предположение по умолчанию**. Конкретные проявления: - Если изучаем связь явлений: H_0 = явления **не связаны**. - Если замеряем показатель: H_0 = показатель принимает **типичное** значение. - Если сравниваем две совокупности: H_0 = они **одинаковые**.

Альтернативная гипотеза H_1

Это то, что мы **хотим «доказать»** (в кавычках, потому что стат-тесты — это не строгий метод доказательства, а статистический метод валидации данных).

- Подозреваем некую **аномалию** — отклонение от нормы.
- Подозреваем, что **связь есть**.
- Подозреваем, что показатель принимает **аномальные значения**.

Важное замечание

> ❗ H_0 и H_1 **не всегда дополняют друг друга** до полного пространства возможностей.
>
> Пример из суда: H_0 = «не виновен», H_1 = «виновен в убийстве». Но возможны и другие сценарии (например, мошенничество), которые не покрываются ни H_0, ни H_1.
>
> В курсе будут рассматриваться ситуации, где H_1 — это отрицание H_0, но это далеко не всегда так.

Зачем нужны эти содержательные рассуждения

Стат-тесты нужно правильно **применять**. Чтобы их применять, надо понимать, **из каких соображений** формулируются H_0 и H_1 для каждой конкретной ситуации.

> **План на следующую лекцию:** строгая математическая постановка задачи проверки статистических гипотез и описание общей схемы процедуры проверки, которая выдаёт ответ «H_0» или «H_1».

Лекция 7: Проверка статистических гипотез

1. Постановка задачи проверки гипотез

1.1. Гипотезы H_0 и H_1

Для каждой ситуации формулируются два предположения:

- **Нулевая гипотеза (H_0)** — предположение «по умолчанию». Если рассматриваются какие-то явления, то по умолчанию они никак не связаны; если рассматривается некоторый показатель — он принимает типичное значение.
- **Альтернативная гипотеза (H_1)** — наше «подозрение», то, что мы хотим доказать.

  NOTE: Важно Сумма H_0 и H_1 **не всегда даёт всё пространство возможностей** — то есть не обязательно $H_0 \cup H_1 = \Omega$.

1.2. Определение статистического критерия

Статистический критерий (statistical test) — это функция, возвращающая одно из двух решений: принять H_0 или отвергнуть H_0.

Формально объявим декларацию функции:

$$\delta(X, H_0, H_1, \alpha) \to \{\text{accept } H_0,\ \text{reject } H_0\}$$

где: - X — **выборка в широком смысле**. Это не обязательно простейшая выборка из независимых одинаково распределённых случайных величин; в общем случае это произвольный датафрейм (например, аргументация прокурора и защиты в суде). - H_0 — нулевая гипотеза. - H_1 — альтернативная гипотеза. - α — **уровень значимости** (significance level). Типичные значения: $0.1,\ 0.05,\ 0.01,\ 0.001$ (хотя можно задавать любые).

> **WARNING: Критическая ремарка о смысле решения** - **«Принять H_0»** не означает доказательства истинности H_0. Это означает лишь, что **данные не противоречат нулевой гипотезе относительно заданной альтернативы**. - **«Отвергнуть H_0»** автоматически не доказывает истинности H_1. Это лишь говорит, что **данные скорее противоречат нулевой гипотезе и свидетельствуют в пользу альтернативы**.

Стат-тест — не «серебряная пуля», а скорее **средство аргументации**.

Пример (аналогия с уголовным судом) Если подсудимый подозревается в убийстве, и в ходе разбирательства приводятся факты о мошенничестве, судья скажет «невиновен» — потому что рассматривается именно дело об убийстве, а не о мошенничестве. То есть «принять H_0» = «H_0 не опровергнуто **относительно данной альтернативы**».

---

2. Принцип работы статистического критерия

2.1. Статистика критерия

Под капотом критерия работает функция $T(X)$ — **статистика критерия**.

Принцип выбора статистики:

> Статистику критерия выбирают так, чтобы её распределение **в точности имело** или **стремилось** (при $n \to \infty$) к некоторому «хорошему» распределению — **при условии истинности нулевой гипотезы**.
>
> **IMPORTANT: > Распределение $T(X)$ рассматривается** именно при условии истинности H_0**.**

2.2. Область принятия и критическая область

Множество всех значений случайной величины T разбивается на две **непересекающиеся** области:

$$\text{supp}(T) = T_0(\alpha) \sqcup T_1(\alpha)$$

- $T_0(\alpha)$ — **область принятия нулевой гипотезы**.
- $T_1(\alpha)$ — **критическая область**.

Распределение вероятностной массы:

$$P(T(X) \in T_0(\alpha) \mid H_0) = 1 - \alpha$$
$$P(T(X) \in T_1(\alpha) \mid H_0) = \alpha$$

Для **асимптотических критериев** равенство выполняется в пределе.

2.3. Основной if-statement критерия

$$\text{if } T(x) \in T_0(\alpha) \Rightarrow \text{принять } H_0, \quad \text{иначе} \Rightarrow \text{отвергнуть } H_0$$

Здесь x (маленькое) — **конкретная реализация** выборки.

Пример с монеткой Если кинуть монетку 100 раз и получить 95 решек — мы скорее скажем, что монетка нечестная. Но **гипотетически** для честной монетки такой исход возможен (хоть и с очень малой вероятностью). Поэтому выводы носят **нестрогий** характер.

2.4. Недостатки прямого if-statement

- У разных тестов **разные распределения** T.
- Сама статистика устроена по-разному (где-то суммируем, где-то усредняем).
- Критические области бывают **трёх типов**.

Хотелось бы **унифицированный показатель** — это p-value (см. ниже).

---

3. Три типа критических областей

Критические области выбираются не произвольно — почти во всех тестах встречается одна из трёх ситуаций.

3.1. Правосторонний тест

- Справа выделяется вероятностная масса α.
- Слева выделяется $1 - \alpha$.

$$T_0(\alpha) = (-\infty,\ q_{1-\alpha}]$$

где $q_{1-\alpha}$ — квантиль порядка $1 - \alpha$.

Называется правосторонним, потому что **критическая область находится справа**.

3.2. Левосторонний тест

- Слева выделяется вероятностная масса α.
- Справа $1 - \alpha$.

$$T_0(\alpha) = [q_\alpha,\ +\infty)$$

(или до супремума носителя случайной величины — в общем случае). Критическая область слева.

3.3. Двусторонний тест

- И слева, и справа выделяется по $\alpha/2$.

$$T_0(\alpha) = [q_{\alpha/2},\ q_{1-\alpha/2}]$$

Критическая область — с обеих сторон.

> **INFO: Замечание** Гипотетически возможны и более экзотические ситуации (например, разбиение на 3 куска), но в практически интересных тестах встречаются только эти три типа.

---

4. p-value

p-value — унифицированный показатель, позволяющий заменить громоздкий if-statement на простое сравнение с α.

4.1. Определения p-value по типам тестов

Правосторонний случай:

$$p\text{-value}_{\text{right}} = P\big(T(X) > T_{\text{набл}} \mid H_0\big)$$

Левосторонний случай:

$$p\text{-value}_{\text{left}} = P\big(T(X) < T_{\text{набл}} \mid H_0\big)$$

Двусторонний случай:

$$p\text{-value} = 2 \cdot \min\left(p\text{-value}_{\text{left}},\ p\text{-value}_{\text{right}}\right)$$

> **TIP: Упрощение** Если распределение T симметрично относительно нуля, формулу для двустороннего случая можно упростить (это часто встречается в литературе).

4.2. Геометрический смысл (правосторонний случай)

Допустим, наблюдаемое значение T_1 попало в область принятия $T_0(\alpha)$. Тогда:

$$p\text{-value}(T_1) > \alpha$$

Если T_2 попало в правый хвост (критическую область):

$$p\text{-value}(T_2) < \alpha$$

4.3. Геометрический смысл (двусторонний случай)

Допустим: - T_1 — в области принятия (центр): - $p\text{-value}_{\text{right}} > \alpha/2$ - $p\text{-value}_{\text{left}} > \alpha/2$ - $\Rightarrow$ $p\text{-value}(T_1) > \alpha$ - T_2 — в правом хвосте: - $p\text{-value}_{\text{left}} > \alpha/2$, $p\text{-value}_{\text{right}} < \alpha/2$ - min $< \alpha/2 \Rightarrow p\text{-value}(T_2) < \alpha$ - T_3 — в левом хвосте: - $p\text{-value}_{\text{right}} > \alpha/2$, $p\text{-value}_{\text{left}} < \alpha/2$ - $\Rightarrow p\text{-value}(T_3) < \alpha$

4.4. Унифицированный if-statement через p-value

$$\text{if } p\text{-value} > \alpha \Rightarrow \text{принять } H_0, \quad \text{иначе} \Rightarrow \text{отвергнуть } H_0$$

p-value всегда лежит в $[0, 1]$ (это вероятность), что делает критерий универсальным.

4.5. Неформальная интерпретация p-value

> **DANGER: Распространённая ошибка** p-value **НЕ является** вероятностью того, что H_0 верна (или что мы её примем). С вероятностью принятия гипотезы p-value никак не связан.

Правильная интерпретация:

> p-value — это **вероятность того, что статистика критерия примет более экстремальное значение** относительно наблюдаемого (при условии истинности H_0).

«Экстремальные» значения — это значения, попадающие в хвост(ы): - В правостороннем тесте: значения **больше** наблюдаемого. - В двустороннем тесте — значения, **большие по модулю** наблюдаемого.

---

5. Терминология: статистическая значимость

Если мы отвергаем H_0, говорят, что **результат является статистически значимым**.

> Например: «доказана нечестность монетки с уровнем значимости $\alpha = 0.05$» = «отвергнута нулевая гипотеза о честности с $\alpha = 0.05$».

Отсюда и название α — **уровень значимости**.

---

6. Ошибки I и II рода

6.1. Таблица ошибок

	Реальность: H_0	**Реальность:** H_1
Тест: H_0	True Negative	False Negative (ошибка II рода)
Тест: H_1	False Positive (ошибка I рода)	True Positive

> **NOTE: Ремарка** К этой таблице нужно относиться **философски** — ведь мы говорили, что отвержение $H_0 \neq$ доказательство H_1. Здесь предполагается, что в реальности либо H_0, либо H_1 верно (при хорошей формулировке гипотез). Гипотетически возможна ситуация, когда **ни** H_0**, ни** H_1 **не верны**, но при разумной формулировке гипотез после предварительного анализа данных такого не возникает.

6.2. Ошибка I рода (False Positive)

Определение: тест отверг H_0, но в реальности H_0 верна.

$$P(\text{ошибка I рода}) = \alpha$$

То есть, **регулируя уровень значимости, мы регулируем ошибку I рода**.

6.3. Ошибка II рода (False Negative)

Определение: тест принял H_0, но в реальности верна H_1.

$$\beta = P(T(X) \in T_0(\alpha) \mid H_1)$$

β — вероятность ошибки II рода.

6.4. Состоятельность критерия

Критерий состоятелен, если:

$$\beta \to 0 \text{ при } n \to \infty$$

6.5. Мощность критерия

Мощность — это $1 - \beta$, то есть вероятность отвергнуть H_0, если в действительности верна H_1.

$$\text{Мощность} = 1 - \beta = P(T(X) \in T_1(\alpha) \mid H_1)$$

6.6. Терминология Positive/Negative (аналогия с медициной)

Аналогия: пациент пришёл к врачу. - H_0: пациент здоров. - H_1: пациент болен.

Положительный анализ → есть аномалия → выбираем H_1 → **Positive**. Отрицательный анализ → нет аномалии → выбираем H_0 → **Negative**.

Реальность Тест	Тест: H_0 (Negative)	Тест: H_1 (Positive)
H_0 верна	True Negative	**False Positive** (ошибка I рода)

Реальность Тест	Тест: H_0 (Negative)	Тест: H_1 (Positive)
H_1 верна	**False Negative** (ошибка II рода)	True Positive

---

7. Связь между α и β

«Сделаем α маленьким, и будет нам счастье» — **не работает!**

Пример: спам-классификатор

- H_0: письмо не является спамом.
- H_1: письмо является спамом.

Классификатор A — всё помещает во «Входящие» (всегда выбирает H_0): - $\alpha = 0$ (ошибка I рода исключена). - Но β велико — спам попадает во «Входящие».

Классификатор B — всё помещает в «Спам» (всегда выбирает H_1): - $\beta = 0$ (ошибка II рода исключена). - Но α велико — нормальные письма попадают в спам.

Мораль

> Как правило, **чем меньше α, тем больше β**. В общем случае аналитическую зависимость α от β написать нельзя, но в некоторых хороших ситуациях можно.

Стандартный подход

На практике: 1. Фиксируется **допустимый порог ошибки I рода** (α). 2. Среди тестов с заданным α выбирается тот, у которого β **минимален** (то есть мощность максимальна).

---

8. Связь доверительных интервалов и стат-тестов

8.1. Напоминание о доверительных интервалах

Чтобы построить доверительный интервал для параметра θ, рассматривается статистика $G(X, \theta)$, которая имеет (или стремится к) распределение случайной величины U, **не зависящее от** θ.

Затем зажимаем статистику между квантилями:

$$P(q_{\alpha/2} \leq G(X, \theta) \leq q_{1-\alpha/2}) = 1 - \alpha$$

Это очень похоже на область принятия нулевой гипотезы!

8.2. Преобразование в стат-тест

Рассмотрим гипотезы: - $H_0 : \ \theta = \theta_0$ - $H_1 : \ \theta \neq \theta_0$

При H_0 подставляем θ_0 вместо θ:

$$P(q_{\alpha/2} \leq G(X, \theta_0) \leq q_{1-\alpha/2}) = 1 - \alpha$$

8.3. Эквивалентность

> **IMPORTANT: Ключевое утверждение** Принять H_0 с уровнем значимости α **равносильно** тому, что значение θ_0 попадает в **доверительный интервал уровня доверия** $1 - \alpha$.

То есть: разрешая неравенство относительно θ_0, мы получаем доверительный интервал, и принятие H_0 означает попадание θ_0 в этот интервал.

---

9. Z-тест для одной выборки (тест о математическом ожидании)

9.1. Постановка

Пусть выборка достаточно большая. Хотим проверить:

- $H_0 : \ E[X] = \mu_0$

Альтернатива может быть трёх видов (в зависимости от наших подозрений): - $H_1 : \ E[X] > \mu_0$ — правосторонний тест - $H_1 : \ E[X] < \mu_0$ — левосторонний тест - $H_1 : \ E[X] \neq \mu_0$ — двусторонний тест

9.2. Статистика критерия

Используется та же статистика, что и для построения асимптотического доверительного интервала для матожидания:

$$T(X) = \frac{\sqrt{n}\,(\bar{X} - \mu_0)}{S} \xrightarrow{d} \mathcal{N}(0,1)$$

где S — выборочное стандартное отклонение.

9.3. Выбор типа критической области

При H_0 ($E[X] = \mu_0$) статистика принимает значения **около нуля** (правило 3-х сигм для $\mathcal{N}(0,1)$: на диапазон $[-3,3]$ приходится $\approx 99.73\%$ массы).

Куда попадает статистика при истинной альтернативе?

- $H_1:\ E[X] > \mu_0 \to \bar{X} \to E[X] > \mu_0 \to$ числитель $> 0 \to$ статистика смещена **вправо → критическая область справа** (правосторонний тест).
- $H_1:\ E[X] < \mu_0 \to$ статистика смещена влево → **левосторонний тест**.
- $H_1:\ E[X] \neq \mu_0 \to$ статистика может быть как слева, так и справа → **двусторонний тест**.

9.4. Доказательство состоятельности (правосторонний случай)

Покажем, что $\beta \to 0$ при $n \to \infty$.

Вероятность ошибки II рода:

$$\beta = P\left(\frac{\sqrt{n}\,(\bar{X} - \mu_0)}{S} \leq q_{1-\alpha} \,\middle|\, E[X] > \mu_0 \right)$$

Прибавим и вычтем истинное мат. ожидание $E[X]$ в числителе:

$$\beta = P\left(\frac{\sqrt{n}\,(\bar{X} - E[X])}{S} + \frac{\sqrt{n}\,(E[X] - \mu_0)}{S} \leq q_{1-\alpha} \,\middle|\, E[X] > \mu_0 \right)$$

Перенесём:

$$\beta = P\left(\frac{\sqrt{n}\,(\bar{X} - E[X])}{S} \leq q_{1-\alpha} - \frac{\sqrt{n}\,(E[X] - \mu_0)}{S} \,\middle|\, E[X] > \mu_0 \right)$$

По ЦПТ (и её следствиям) первая дробь сходится к $\mathcal{N}(0,1)$, поэтому при больших n:

$$\beta \approx \Phi\left(q_{1-\alpha} - \frac{\sqrt{n}\,(E[X]-\mu_0)}{S}\right)$$

Анализ аргумента: - $q_{1-\alpha}$ — константа. - $E[X]-\mu_0 > 0$ (по альтернативе). - $S \to \sigma$ (выборочное стандартное отклонение сходится к теоретическому). - $\sqrt{n} \to \infty$.

Значит, $\frac{\sqrt{n}(E[X]-\mu_0)}{S} \to +\infty$, и аргумент функции Φ уходит на $-\infty$.

$$\Phi(-\infty) = 0 \Rightarrow \beta \to 0$$

Критерий **состоятелен**. □

9.5. Терминология: Z-тест

Z-тест — это критерий, у которого статистика критерия **точно имеет** или **сходится к** нормальному распределению. Жаргон сложился исторически (от обозначения нормальной величины через Z).

Вышеописанный тест — **Z-тест для одной выборки**, проверяющий гипотезу о математическом ожидании.

---

10. Важный частный случай: распределение Бернулли

10.1. Постановка

Пусть $X_1, \dots, X_n$ — выборка из распределения Бернулли с параметром p.

> Параметр p распределения Бернулли = матожидание этого распределения.

Гипотезы: - $H_0: \ p = p_0$ - H_1: одна из трёх ($p > p_0$, $p < p_0$, $p \neq p_0$).

10.2. Статистика критерия

Поскольку для Бернулли дисперсия $D[X] = p(1-p)$, статистика принимает **более конкретный вид** (теоретическая дисперсия известна при H_0):

$$T(X) = \frac{\sqrt{n}\,(\bar{X} - p_0)}{\sqrt{p_0(1-p_0)}} \xrightarrow{d} \mathcal{N}(0,1)$$

10.3. Тип критической области

Аналогично общему случаю Z-теста: - $H_1 : \ p > p_0$ — правосторонний. - $H_1 : \ p < p_0$ — левосторонний. - $H_1 : \ p \neq p_0$ — двусторонний.

10.4. Применение: проверка честности монетки

Для честной монетки $p = 1/2$, поэтому $p_0 = 1/2$.

В зависимости от подозрений: - «Орёл выпадает чаще» → $H_1 : \ p > 1/2$. - «Решка выпадает чаще» → $H_1 : \ p < 1/2$. - «Монетка просто нечестная» → $H_1 : \ p \neq 1/2$.

---

11. Как выбирать тип альтернативы — пример с врачами

Тип теста (одно- или двусторонний) зависит от того, **что мы хотим доказать**.

Пример: измерения температуры пациента.

- **Участковый врач-терапевт**: пришёл пациент с жалобой «мне плохо». Терапевт хочет понять — **есть ли вообще отклонение** от нормы (36.6°C). Альтернатива: $E[T] \neq 36.6$ → **двусторонний тест**.
- **Врач-инфекционист**: ищет инфекцию, для которой характерна **повышенная** температура. Альтернатива: $E[T] > 36.6$ → **правосторонний тест**.
- **Врач, ищущий болезнь с пониженной температурой**: альтернатива $E[T] < 36.6$ → **левосторонний тест**.

  SUMMARY: Резюме Альтернатива зависит от того, **что именно мы хотим доказать**. Тип критической области определяется и сутью теста, и формулировкой альтернативы.

---

12. Итоговая схема работы стат-критерия

1. Сформулировать H_0 и H_1.
2. Зафиксировать уровень значимости α.
3. Выбрать статистику критерия $T(X)$ с известным распределением **при** H_0.
4. Определить тип критической области (право-/лево-/двусторонний — зависит от H_1).
5. Вычислить $T(x)$ на конкретной реализации.

6. Вычислить **p-value**.
7. Принять решение:
   - p-value $> \alpha \to$ принять H_0.
   - p-value $\leq \alpha \to$ отвергнуть H_0.
8. (При необходимости) проверить **состоятельность** критерия: $\beta \to 0$ при $n \to \infty$.

---

13. Что дальше

В следующих лекциях: - Стат-тесты, связанные с **доверительными интервалами и нормальным распределением**. - Конкретные ситуации, где можно явно выписать зависимость α от β и определить **оптимальный критерий**.

Лекция 8: Статистические критерии (продолжение)

1. Критерий о медиане (одна выборка)

Постановка

Пусть $x_1, x_2, \dots, x_n$ — выборка из некоторого **непрерывного** распределения. Проверяем гипотезу о теоретической медиане:

$$H_0 : \text{med} = c$$

Альтернатива H_1 настраивается: med $\neq c$, med $> c$, либо med $< c$.

Идея построения статистики

Что может оценивать теоретическую медиану? **Элемент вариационного ряда, стоящий в центре.**

Напоминание: вариационный ряд — это упорядоченная (отсортированная) выборка.

Средний член вариационного ряда $x_{(n/2)}$, если выборка из непрерывного закона, **асимптотически нормален**:

$$\sqrt{n} \cdot p(c) \cdot \frac{x_{(n/2)} - c}{\sqrt{\frac{1}{2} \cdot \frac{1}{2}}} \xrightarrow{d} \mathcal{N}(0,1)$$

где $p(c)$ — плотность распределения в точке c. В знаменателе стоит $\sqrt{p(1-p)}$, и поскольку для медианы $p = 1/2$, получаем $\sqrt{1/2 \cdot 1/2} = 1/2$.

Тип критической области

Логика та же, что и для гипотез о мат. ожидании: - $H_1 : \text{med} > c$ — **правосторонняя** - $H_1 : \text{med} < c$ — **левосторонняя** - $H_1 : \text{med} \neq c$ — **двусторонняя**

---

2. Z-тест для одной выборки (дисперсия известна)

Постановка

Выборка $x_1, \dots, x_n$ из нормального закона $\mathcal{N}(\mu_x, \sigma_x^2)$, причём σ_x^2 **известна**.

$$H_0 : \mu_x = \mu_0, \qquad H_1 : \mu_x \neq \mu_0 \ / \ > \mu_0 \ / \ < \mu_0$$

Статистика критерия

Мат. ожидание оценивается выборочным средним $\bar{x}$. Центрируем и нормируем, чтобы получить **стандартное нормальное распределение**:

$$Z = \frac{\bar{x} - \mu_0}{\sigma_x / \sqrt{n}} \sim \mathcal{N}(0,1) \quad \text{при } H_0$$

Это ещё одна вариация **Z-теста**.

Терминология

> Если статистика критерия имеет нормальное распределение при условии истинности H_0, такой тест называется **Z-тест**.

---

3. T-тест для одной выборки (дисперсия неизвестна)

Постановка

Та же гипотеза $H_0 : \mu_x = \mu_0$, но теперь σ_x^2 **неизвестна**.

Статистика критерия

Заменяем σ на её оценку S (выборочное стандартное отклонение):

$$T = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} \sim t(n-1) \quad \text{при } H_0$$

Распределение **Стьюдента** с $n-1$ степенями свободы — следствие **теоремы Фишера**. Этот же факт всплывал при построении доверительного интервала для мат. ожидания нормального закона при неизвестной дисперсии.

Терминология

> **T-тест** — стат. критерий, у которого статистика имеет распределение Стьюдента.

Здесь рассмотрен T-тест для одной выборки на мат. ожидание.

⚠ Важная ремарка о применимости

Существует рекомендация: если объём выборки маленький (порядка $n \in [10, 20]$), то для проверки гипотезы о мат. ожидании нужно использовать T-тест.

НО: есть важная посылка, о которой часто забывают: - В крайнем случае T-тест более-менее адекватно работает, **если исходное распределение более-менее симметрично**. - Если от нормальности совсем отказываемся — к результатам теста нужно относиться **очень аккуратно**.

---

4. Критерий χ^2 для дисперсии (одна выборка)

Постановка

Выборка $x_1, \dots, x_n$ из нормального закона $\mathcal{N}(\mu_x, \sigma_x^2)$.

$$H_0 : \sigma_x^2 = \sigma_0^2$$

Альтернативы: $\neq$, $>$, $<$.

Статистика критерия

Дисперсию оценивает выборочная дисперсия S^{*2}. По **теореме Фишера**:

$$\frac{nS^2}{\sigma_0^2} \sim \chi^2(n-1) \quad \text{при } H_0$$

Анализ типов критической области

Рассмотрим мат. ожидание статистики, домножив и поделив на реальную дисперсию:

$$E\left[\frac{nS^2}{\sigma_0^2}\right] = E\left[\frac{nS^2}{\sigma_x^2} \cdot \frac{\sigma_x^2}{\sigma_0^2}\right]$$

Здесь $\frac{nS^2}{\sigma_x^2} \sim \chi^2(n-1)$ независимо от истинности H_0, и $E[\chi^2(n-1)] = n-1$.

> **Вопрос из аудитории:** «А что у нас сверху было с σ_0?»
>
> **Ответ:** В контексте «при условии истинности H_0» мы заменили σ_x^2 на σ_0^2. Отсюда — единичка в отношении.

Логика выбора критической области:

Случай	Поведение σ_x^2/σ_0^2	В среднем статистика	Критическая область
H_0 верна	$= 1$	$\approx n-1$	—
$\sigma_x^2 > \sigma_0^2$	> 1	больше $n-1$	**правосторонняя**
$\sigma_x^2 < \sigma_0^2$	< 1	меньше $n-1$	**левосторонняя**
$\sigma_x^2 \neq \sigma_0^2$	—	—	**двусторонняя**

---

5. Парная выборка — сведение к одной выборке

Что такое парная выборка

Есть n наблюдений, для каждого замерены **два показателя**:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

При этом априори считаем, что эти два фактора **зависимы** (то есть мы НЕ можем считать их независимыми выборками).

Гипотеза

$$H_0 : \mu_x = \mu_y$$

объём выборки достаточно большой.

Классический приём

Рассматриваем новую выборку:

$$u_i = x_i - y_i, \quad i = 1, \dots, n$$

Тогда исходная гипотеза эквивалентна:

$$H_0 : E[U] = 0$$

Это уже задача о мат. ожидании для **одной выборки** — её и решаем разобранными ранее методами.

Замечание Ивана Александровича: «Это классический рецепт».

---

6. F-тест для отношения дисперсий (две выборки)

Постановка

Две **независимые** выборки из нормального закона: - $x_1, \dots, x_n \sim \mathcal{N}(\mu_x, \sigma_x^2)$ - $y_1, \dots, y_m \sim \mathcal{N}(\mu_y, \sigma_y^2)$

$$H_0 : \sigma_x^2 = \sigma_y^2$$

Альтернативы: $\neq$, $>$, $<$.

Статистика критерия

Идея взята из доверительного интервала для отношения дисперсий, где всплывали два χ^2, и их отношение давало распределение Фишера.

$$F = \frac{S_x^2}{S_y^2} \sim F(n-1, m-1) \quad \text{при } H_0$$

Анализ критических областей

Домножим и поделим на отношение реальных дисперсий:

$$E\left[\frac{S_x^2}{S_y^2}\right] = E\left[\frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2}\right] \cdot \frac{\sigma_x^2}{\sigma_y^2}$$

Первое отношение — мат. ожидание распределения Фишера (не зависит от H_0).

Альтернатива	Отношение σ_x^2/σ_y^2	Среднее значение статистики	Критическая область
$\sigma_x^2 > \sigma_y^2$	> 1	больше $E[F]$	**правосторонняя**
$\sigma_x^2 < \sigma_y^2$	< 1	меньше $E[F]$	**левосторонняя**
$\sigma_x^2 \neq \sigma_y^2$	—	—	**двусторонняя**

Терминология

> Критерии, у которых статистика имеет распределение Фишера, называются **F-тесты**. Этот F-тест сравнивает дисперсии двух выборок.

---

7. Z-тест для двух выборок (мат. ожидания, дисперсии известны)

Постановка

$x_1, \dots, x_n \sim \mathcal{N}(\mu_x, \sigma_x^2)$, $y_1, \dots, y_m \sim \mathcal{N}(\mu_y, \sigma_y^2)$, выборки **независимы**, дисперсии **известны**.

$$H_0 : \mu_x = \mu_y$$

Построение статистики

- $\bar{x} \sim \mathcal{N}(\mu_x, \sigma_x^2/n)$
- $\bar{y} \sim \mathcal{N}(\mu_y, \sigma_y^2/m)$
- В силу независимости: $\bar{x} - \bar{y} \sim \mathcal{N}\left(\mu_x - \mu_y,\ \dfrac{\sigma_x^2}{n} + \dfrac{\sigma_y^2}{m}\right)$ (дисперсии **складываются**)

Стандартизуем:

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{\sigma_x^2}{n} + \dfrac{\sigma_y^2}{m}}} \sim \mathcal{N}(0,1) \quad \text{при } H_0$$

(При H_0 числитель в среднем ноль.)

> Почему нормальное распределение → **Z-тест**? Исторически так сложилось. F-тест — от фамилии Fisher; «Z» же — историческая конвенция.

Модификация: ЦПТ-вариант (без нормальности)

Пусть теперь выборки x и y **независимы и достаточно большого объёма** (без предположения о нормальности). По ЦПТ:

$$\frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{\sigma_x^2}{n} + \dfrac{\sigma_y^2}{m}}} \xrightarrow{d} \mathcal{N}(0,1)$$

Если дисперсии **неизвестны** — подставляем их состоятельные оценки. В пределе по-прежнему получаем $\mathcal{N}(0,1)$.

Тип критической области

- $H_1 : \mu_x > \mu_y$: разность $\mu_x - \mu_y > 0$, статистика в среднем положительна → **правосторонняя**
- $H_1 : \mu_x < \mu_y$: → **левосторонняя**
- $H_1 : \mu_x \neq \mu_y$: → **двусторонняя**

---

8. T-тест для двух выборок (дисперсии равны и неизвестны)

Постановка

Две независимые выборки из нормального закона, **дисперсии равны**, но **неизвестны**:

$$\sigma_x^2 = \sigma_y^2 = \sigma^2 \text{ (неизвестна)}$$

$$H_0 : \mu_x = \mu_y$$

Построение статистики

Если бы дисперсия была известна:

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim \mathcal{N}(0, 1)$$

Дисперсию не знаем — **оцениваем**. Чтобы получить распределение Стьюдента, нужно поделить на корень из усреднённого χ^2.

Откуда взять χ^2? Из **теоремы Фишера**: - $\frac{nS_x^2}{\sigma^2} \sim \chi^2(n-1)$ - $\frac{mS_y^2}{\sigma^2} \sim \chi^2(m-1)$

Так как выборки независимы, **сумма** этих χ^2 — снова χ^2 с суммой степеней свободы:

$$\frac{nS_x^2 + mS_y^2}{\sigma^2} \sim \chi^2(n + m - 2)$$

(Это следует из формального определения χ^2 как суммы квадратов независимых стандартных гауссовских величин.)

Итоговая статистика

Делим стандартную нормальную величину на корень из (χ^2 / число степеней свободы):

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \cdot \frac{1}{\sqrt{\frac{nS_x^2 + mS_y^2}{\sigma^2(n+m-2)}}}$$

> **Прелесть в том, что σ^2 сокращаются** — в финальной статистике дисперсии нет.

После упрощения получаем:

$$T \sim t(n+m-2) \quad \text{при } H_0$$

> Окончательное «причёсывание» формулы оставлено как **упражнение**.

Это **T-тест для двух выборок** (сравнивает мат. ожидания).

---

9. Простой рецепт проверки однородности

Что такое однородность

Однородность двух выборок означает, что **распределения двух выборок одинаковы**.

Рецепт (для нормально распределённых выборок)

1. **F-тест** на равенство дисперсий ($H_0 : \sigma_x^2 = \sigma_y^2$).
2. **Если** H_0 принята → **T-тест** на равенство мат. ожиданий.

Почему «простой» в кавычках

- Простой только на бумаге — на практике вычислений много.

- **Важная посылка:** рецепт реально проверяет однородность, **только если выборки из нормального закона** (нормальный закон полностью задаётся μ и σ^2).

Устойчивость к нарушению посылок

- **F-тест** более-менее устойчив к нарушению предположения о нормальности.
- **T-тест** — менее устойчив.

Применение T-теста: A/B-тестирование

Пример приложения T-теста для двух выборок — **A/B-тестирование**: - Часть пользователей видит старую версию сайта (группа A). - Часть пользователей — новую версию (группа B). - Анализируем, как ведут себя пользователи (достигают ли целевого показателя). - T-тест помогает сравнить мат. ожидания целевой метрики между группами.

---

10. T-критерий Уэлча (упоминание)

T-критерий Уэлча — модификация T-теста для ситуации, когда **нельзя считать**, что $\sigma_x^2 = \sigma_y^2$. То есть отказываемся от предположения о равенстве дисперсий.

> Подробное обсуждение, какая там статистика и откуда она берётся — отложено.

---

11. Критерий согласия Колмогорова

Постановка

Простая выборка $x_1, \dots, x_n$. Проверяем гипотезу:

$$H_0 : F = F_0$$

где F_0 — **обязательно непрерывная** функция распределения.

Альтернатива (классическая): $H_1 : F \neq F_0$.

Статистика критерия

$$D_n = \sqrt{n} \cdot \sup_x |F_n(x) - F_0(x)|$$

где F_n — **эмпирическая функция распределения**.

Теорема Колмогорова

При условии истинности H_0:

$$P(D_n \leq t) \xrightarrow{n\to\infty} K(t)$$

где $K(t)$ — **функция распределения Колмогорова**:

$$K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2t^2}$$

«Тут меня лучше проверить — мог немного набрать.»

В стат-библиотеках есть численная реализация $K(t)$.

Тип критической области

- При H_0: $F_n \approx F_0$, супремум близок к 0 → статистика близка к 0.
- При нарушении H_0: статистика существенно больше 0 (модуль).
- → Критическая область **правосторонняя**.

Замечания и нюансы

1. Размер выборки. Если выборка объёма уже несколько десятков, асимптотика более-менее адекватная. В качестве критического значения берут квантиль порядка $1 - \alpha$ распределения Колмогорова.

2. Сложные гипотезы. Можно проверять не равенство конкретной F_0, а гипотезу о принадлежности параметрическому семейству F_θ. Но распределение статистики тогда будет более нетривиальным.

3. Проверка нормальности. Чисто гипотетически критерий Колмогорова можно использовать для проверки согласованности с нормальным законом. **НО лучше использовать специализированные критерии:** - **Тест Шапиро-Уилка**. -

Тест Жака-Бера (Jarque-Bera): статистика играется с **асимметрией** и **эксцессом** ($A + E$). У стандартного нормального распределения $A = 0$, $E = 0$.

Терминология: критерий согласия

> **Критерий согласия** — тест, проверяющий **согласованность данных с заданным вероятностным распределением**.

Критерий Колмогорова — пример критерия согласия.

---

12. Критерий однородности Смирнова

Постановка

Две **независимые** выборки. Проверяем:

$$H_0 : F_X = F_Y$$

где F_X, F_Y — **непрерывные** функции распределения.

$$H_1 : F_X \neq F_Y$$

Статистика критерия

$$D_{m,n} = \sqrt{\frac{mn}{m+n}} \cdot \sup_x |F_n(x) - F_m(x)|$$

где F_n и F_m — эмпирические функции распределения для выборок x и y соответственно.

Предельное распределение

$$P(D_{m,n} \leq t) \to K(t)$$

— то же самое распределение Колмогорова!

Тип критической области

- При H_0: $F_X \approx F_Y \to$ супремум близок к нулю.
- При H_1: существенное отличие, модуль $\to$ большое значение.
- $\to$ **Правосторонняя**.

Замечание

Поскольку **предельное распределение совпадает** у критериев Колмогорова и Смирнова, в некоторых стат-пакетах эти **два критерия объединены в одну функцию**.

Терминология: критерий однородности

> **Критерий однородности** — тест, проверяющий равенство распределений двух и более выборок.

Критерий Смирнова работает с **двумя** выборками.

---

13. Дискретизация распределений

Зачем

Критерии типа χ^2 работают с **дискретными распределениями с конечным множеством значений**. Иногда нужно применить их в других ситуациях.

Случай 1: Дискретное распределение с бесконечным (счётным) множеством значений

Например, пуассоновская случайная величина (значения $0, 1, 2, \ldots$).

Идея: оставить первые n значений, а **«хвост» объединить в одно значение**.

Было	Стало
$1, 2, 3, \ldots, n, n+1, \ldots$	$1, 2, 3, \ldots, n, \{> n\}$
$p_1, p_2, p_3, \cdots, p_n, p_{n+1}, \cdots$	$p_1, p_2, p_3, \cdots, p_n, \sum_{k>n} p_k$

> **Замечание из аудитории:** «А можно ли это делать оптимальным образом? Например, в хвост брать самые невероятные.»

Ответ: Да, идея совершенно верная и разумная.

Случай 2: Абсолютно непрерывное распределение

Идея: разбить вещественную ось на **конечное число интервалов** $\Delta_1, \dots, \Delta_k$.

Вероятность попадания в интервал:

$$p(\Delta_i) = \int_{\Delta_i} p(x)\,dx$$

Случайная величина теперь принимает k значений (номер интервала). Крайние интервалы могут быть бесконечными (от $-\infty$ или до $+\infty$).

Итог

Если у нас дискретное распределение со счётным множеством значений или **любое** непрерывное распределение — можем свести задачу к ситуации с дискретным распределением с конечным множеством значений.

---

14. Критерий согласия Пирсона χ^2

Постановка

- Дискретная случайная величина с конечным множеством значений (без потери общности — $1, 2, \dots, n$).
- Этим значениям сопоставлен вектор вероятностей $P = (p_1, \dots, p_n)$.
- ν_k — **количество элементов в выборке, равных** k (наблюдаемая частота).

Гипотеза

$$H_0 : P = P_0 = (p_{0,1}, p_{0,2}, \dots, p_{0,n})$$

$$H_1 : P \neq P_0$$

Пример (из обсуждения): многократно бросают кубик. Вопрос — честный ли? Для честного кубика $P_0 = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$.

Статистика критерия

$$\chi^2 = \sum_{k=1}^{n} \frac{(\nu_k - np_{0,k})^2}{np_{0,k}}$$

При условии истинности H_0:

$$\chi^2 \xrightarrow{d} \chi^2(n-1)$$

Логика типа критической области

- ν_k — наблюдаемые частоты, $np_{0,k}$ — ожидаемые частоты при H_0.
- При H_0: $\nu_k \approx np_{0,k}$ → статистика **малая**.
- При H_1: расхождение большое → статистика **большая**.
- → **Правосторонняя** критическая область (как в большинстве классических версий — там либо квадрат, либо модуль).

Рекомендации к применению

- Объём выборки n — желательно **хотя бы несколько десятков**.
- ν_k должны быть хотя бы **4 или 5** в каждой ячейке (точная цифра — порядка этого).

Демонстрация при $n = 2$

Покажем, что при $n = 2$ в пределе действительно получается $\chi^2(1)$.

Шаг 1. Распишем явно:

$$\chi^2 = \frac{(\nu_1 - np_{0,1})^2}{np_{0,1}} + \frac{(\nu_2 - np_{0,2})^2}{np_{0,2}}$$

Шаг 2. Связи между переменными при $n = 2$: - $p_{0,2} = 1 - p_{0,1}$ (вероятности в сумме дают 1). - $\nu_2 = n - \nu_1$ (количества в сумме дают n).

Тогда $\nu_2 - np_{0,2} = (n - \nu_1) - n(1 - p_{0,1}) = -\nu_1 + np_{0,1} = -(\nu_1 - np_{0,1})$.

Шаг 3. Квадрат не чувствует знака. Выносим $(\nu_1 - np_{0,1})^2/n$ за скобки:

$$\chi^2 = \frac{(\nu_1 - np_{0,1})^2}{n}\left(\frac{1}{p_{0,1}} + \frac{1}{1-p_{0,1}}\right)$$

Шаг 4. Приводим к общему знаменателю:

$$\frac{1}{p_{0,1}} + \frac{1}{1-p_{0,1}} = \frac{1}{p_{0,1}(1-p_{0,1})}$$

Шаг 5. Получаем:

$$\chi^2 = \frac{(\nu_1 - np_{0,1})^2}{np_{0,1}(1-p_{0,1})} = \left(\frac{\nu_1 - np_{0,1}}{\sqrt{np_{0,1}(1-p_{0,1})}}\right)^2$$

Шаг 6. Куда сходится выражение в скобках?

ν_1 — это количество исходов типа 1 в n испытаниях. Это **биномиальная** случайная величина $\text{Bin}(n, p_{0,1})$, у которой: - $E[\nu_1] = np_{0,1}$ - $\text{Var}(\nu_1) = np_{0,1}(1-p_{0,1})$

По **ЦПТ**:

$$\frac{\nu_1 - np_{0,1}}{\sqrt{np_{0,1}(1-p_{0,1})}} \xrightarrow{d} \mathcal{N}(0,1)$$

Шаг 7. Квадрат стандартной нормальной величины — это $\chi^2(1)$. Получили требуемое: $\chi^2(n-1) = \chi^2(1)$ при $n = 2$. □

> В общем случае логика рассуждений похожая, но выкладки более громоздкие.

---

15. Сводная таблица всех критериев лекции

Критерий	Что проверяет	Распределение статистики	Крит. область
Критерий о медиане	$\text{med} = c$ (1 выборка)	$\mathcal{N}(0,1)$ асимпт.	по альтернативе
Z-тест (1 выборка)	$\mu = \mu_0$, σ^2 известна	$\mathcal{N}(0,1)$	по альтернативе
T-тест (1 выборка)	$\mu = \mu_0$, σ^2 неизвестна	$t(n-1)$	по альтернативе
χ^2-тест на дисперсию	$\sigma^2 = \sigma_0^2$	$\chi^2(n-1)$	по альтернативе
Парная выборка	$\mu_x = \mu_y$ (зависимые)	через разность u_i	по альтернативе
F-тест	$\sigma_x^2 = \sigma_y^2$	$F(n-1, m-1)$	по альтернативе
Z-тест (2 выборки)	$\mu_x = \mu_y$, σ^2 известны	$\mathcal{N}(0,1)$	по альтернативе
T-тест (2 выборки)	$\mu_x = \mu_y$, $\sigma_x^2 = \sigma_y^2$ неизв.	$t(n+m-2)$	по альтернативе
Уэлч	$\mu_x = \mu_y$, дисперсии не равны	(упоминание)	—
Колмогоров	$F = F_0$ (1 выборка)	Колмогорова $K(t)$	правосторонняя
Смирнов	$F_X = F_Y$ (2 выборки)	Колмогорова $K(t)$	правосторонняя
Пирсон χ^2	$P = P_0$ (дискретное)	$\chi^2(n-1)$	правосторонняя

---

16. Конвенции терминологии

Тест	Определение
Z-тест	Статистика имеет (асимпт.) нормальное распределение при H_0
T-тест	Статистика имеет распределение Стьюдента при H_0

Тест	Определение
F-тест	Статистика имеет распределение Фишера при H_0
χ^2**-тест**	Статистика имеет (асимпт.) распределение хи-квадрат при H_0
Критерий согласия	Проверяет согласованность данных с заданным распределением
Критерий однородности	Проверяет равенство распределений двух и более выборок

---

17. Упоминавшиеся, но не разобранные подробно тесты

- **Тест Романовского** — обсуждался у соседних групп; даёт более слабый вывод (выборочное и теоретическое распределения совпадают «случайно/неслучайно»), тогда как обычные критерии дают более сильный вывод (нет оснований отвергать H_0).
- **Тест Шапиро-Уилка** — для проверки нормальности.
- **Тест Жака-Бера** — для проверки нормальности (через асимметрию и эксцесс).
- **Тест Уэлча** — модификация T-теста при неравных неизвестных дисперсиях.

Лекция 9: Статистические критерии (продолжение)

1. Примеры применения базовых критериев

1.1. Проверка гипотезы о математическом ожидании (честная монета)

Постановка задачи: Монету подбросили 4096 раз, орёл выпал 2000 раз. Является ли монета честной?

Гипотезы: - H_0: $p = 0{,}5$ (монета честная, p — реальная вероятность успеха) - H_1: альтернатива может быть специфицирована тремя способами: - правосторонняя: $p > 0{,}5$ - левосторонняя: $p < 0{,}5$ - двусторонняя: $p \neq 0{,}5$

По сути проверяется, верно ли, что математическое ожидание равняется конкретному значению.

Статистика критерия:

$$Z = \frac{\bar{X} - \mu}{\sqrt{D}} \cdot \sqrt{n}$$

то есть (выборочное среднее минус мат. ожидание), делённое на корень квадратный из дисперсии, умноженное на корень из объёма выборки n.

Распределение статистики: при условии истинности H_0 статистика имеет **стандартное нормальное распределение**.

p-value: напоминание — это вероятность того, что мы получим более экстремальные значения относительно наблюдаемого. - Для правосторонней альтернативы — правосторонний p-value - Для левосторонней альтернативы — левосторонний p-value - Для двусторонней альтернативы — двусторонний p-value

Пример вывода: при каком уровне значимости мы опровергнем нулевую гипотезу? Иными словами, p-value должен быть меньше уровня значимости. Если уровень значимости больше чем $0{,}067$ — гипотеза будет отвергнута.

> **Технический момент:** в коде используется модуль `scipy.stats` (импортируется как подмодуль `stats` из `scipy`).

1.2. Проверка гипотезы о дисперсии (сеть магазинов)

Постановка задачи: Есть сеть магазинов, известно среднее время и стандартное отклонение. Открыли новый магазин, посмотрели на 25 случайных покупателей. На уровне значимости 1% проверить гипотезу о том, что стандартное отклонение времени в новом магазине **больше**, чем во всей сети.

Гипотезы: - H_0: $\sigma = 5{,}5$ - H_1: $\sigma > 5{,}5$ (подозреваем большее отклонение — это идёт в альтернативу)

Статистика: распределена по χ^2 с $n-1$ степенями свободы (по теореме Фишера).

Тип критерия: правосторонний (это было показано на теории).

Расчёт p-value:

$$\text{p-value} = 1 - \text{CDF}(\text{stat})$$

Результат: получили p-value $\approx 0{,}67$ — гипотеза H_0 **принимается**.

1.3. F-тест на равенство дисперсий двух выборок

Постановка задачи: Есть две выборки. Для каждой даны среднее и стандартное отклонение. Проверить равенство дисперсий.

Метод: F-тест.

Критическая область: двусторонняя.

Результат: p-value большой — нулевая гипотеза **принимается**.

1.4. T-тест для сравнения математических ожиданий двух выборок

Использовали T-тест для двух выборок (рассматривали в одной из прошлых лекций).

Результат: p-value $\approx 0{,}0004$, при уровне значимости $0{,}05$ — нулевая гипотеза **отвергается**. Тест показал статистически значимый результат: средние не равны.

1.5. T-тест для парных выборок

Постановка: есть парная выборка (условно «до» и «после»). Хотим проверить, верно ли, что математическое ожидание «после» больше, чем «до».

Метод: альтернативная гипотеза формулируется в терминах разности — фактически в терминах третьей, новой выборки. Используется t-test для парных выборок.

Результат: нулевая гипотеза **принимается**.

1.6. Простой критерий согласия Пирсона (число π)

Рассмотрен пример про распределение цифр в десятичной записи числа π.

Результат: статистика χ^2 дала p-value $\approx 0{,}4$ — это больше типичного уровня значимости, гипотеза принимается.

---

2. Критерий согласия Пирсона для сложной гипотезы

2.1. Отличие от простого критерия

В **простом** критерии согласия Пирсона у нас была простая гипотеза вида:

$$H_0 : p = p_0$$

где p_0 — конкретное значение (например, вектор вероятностей). В предыдущих примерах мы спрашивали: «верно ли, что вектор вероятностей равен вектору, состоящему из $\frac{1}{10}$» (для цифр π).

В **сложной** гипотезе p_0 зависит от параметра θ:

$$H_0 : p = p_0(\theta)$$
$$H_1 : \neg H_0$$

2.2. Статистика критерия

Рассматривается статистика χ^2, аналогичная простому случаю:

$$\chi^2 = \sum_{k=1}^{N} \frac{(\nu_k - n \cdot p_{0k}(\theta))^2}{n \cdot p_{0k}(\theta)}$$

где $p_0(\theta) = (p_{01}(\theta), p_{02}(\theta), \dots, p_{0N}(\theta))$ — вектор вероятностей, зависящий от θ.

2.3. Проблема и решение

Проблема: θ — неизвестная величина.

Что можно сделать? Заменить θ на выборочную оценку. Точнее — на **оценку максимального правдоподобия (ОМП)**.

При некоторых ограничениях предельное распределение остаётся «хорошим».

2.4. Утверждение (теорема о сложном критерии Пирсона)

Пусть: - θ — вектор размерности r параметров - $r < N - 1$ (строго меньше) - $\frac{\partial p_0}{\partial \theta}$ непрерывна - $\frac{\partial^2 p_0}{\partial \theta^2}$ дважды непрерывна - Матрица $\left(\frac{\partial p_{0i}}{\partial \theta_g} \right)$, где $i = 1, \dots, N$ (большое N), $g = 1, \dots, r$ (маленькое r), имеет ранг r

Тогда статистика χ^2 при подстановке ОМП $\hat{\theta}$ сходится к распределению χ^2 с числом степеней свободы:

$$\mathrm{df} = N - 1 - r$$

Здесь $N-1$ — это то же, что было в простом критерии согласия Пирсона, а r — размерность параметра, который мы дополнительно оценили.

2.5. Пример: семьи с двумя детьми

Данные: 2027 семей с двумя детьми. Среди них: - 527 пар — два мальчика - 476 пар — две девочки (в восстановленных данных ≈ 476, в записи прозвучало ≈ 400 с уточнением «по две девочки») - 1017 пар — один мальчик и одна девочка (≈ 1017)

Вопрос: верно ли, что количество мальчиков в таких семьях можно считать случайной величиной с **биномиальным распределением** с соответствующими параметрами?

Тип гипотезы: сложная (надо оценить параметр p).

Размерность параметра: $r = 1$.

ОМП для биномиального распределения: выборочное среднее, делённое пополам:

$$\widehat{p} = \frac{0 \cdot \nu_0 + 1 \cdot \nu_1 + 2 \cdot \nu_2}{2n}$$

то есть нули умножаем на количество нулей, единицы на количество единичек, двойки на количество двоек, и делим на $2n$.

Степени свободы для χ^2: - Простой критерий дал бы $N - 1 = 3 - 1 = 2$ - Учитывая оценку $\widehat{p}$: $\mathrm{df} = 2 - 1 = 1$

Результат: p-value $\approx 0{,}734$ — существенно больше типичных уровней значимости, нулевая гипотеза **принимается**.

> **Замечание о коде:** иногда библиотечный код эволюционирует, и в новых версиях нужно писать иначе, чем раньше. Этот пример будет более детально разобран в следующий раз.

---

3. Критерий однородности χ^2

3.1. Постановка задачи

Имеется K **независимых выборок**. Чтобы задача об однородности была содержательной, предполагаем, что величины в каждой из выборок принимают **одинаковые значения**.

> **Пример некорректной постановки:** если выборка 1 — это «мальчик/девочка», а выборка 2 — это «средний балл», то задача о проверке однородности вряд ли будет содержательной.

Обозначения: - Значения, которые могут принимать величины: от 1 до N - p_i — вектор вероятностей для i-й выборки - n_i — объём i-й выборки - ν_{ig} — количество значений типа g в i-й выборке

3.2. Гипотезы

$$H_0 : p_1 = p_2 = ... = p_K$$

(назовём это общее значение p_0 — это просто обозначение, удобное для формулы)

$$H_1 : \neg H_0$$

3.3. Статистика критерия

Критерий однородности χ^2 — это **модификация** критерия согласия Пирсона.

$$\chi^2_{n_1,...,n_K} = \sum_{i=1}^{K} \chi^2_{n_i}$$

где локальный χ^2:

$$\chi^2_{n_i} = \sum_{g=1}^{N} \frac{(\nu_{ig} - n_i \cdot p_{0g})^2}{n_i \cdot p_{0g}}$$

> Поскольку в нулевой гипотезе все вероятности равны, в формуле стоит общее p_0.

3.4. Оценка p_0

Проблема: конкретное значение p_0 нам неизвестно.

Решение: оцениваем методом максимального правдоподобия:

$$\hat{p}_{0g} = \frac{\nu_{1g} + \nu_{2g} + ... + \nu_{Kg}}{n}$$

То есть берём суммарное количество элементов типа g по всем выборкам и делим на общий объём $n = n_1 + n_2 + ... + n_K$.

3.5. Распределение статистики и степени свободы

Шаг 1. Предположим на секунду, что p_0 известно и фиксировано. Тогда: - Локальный $\chi^2_{n_i}$ имеет $N-1$ степень свободы - Поскольку выборки независимы, степени свободы складываются (по формальному определению χ^2 как суммы квадратов гауссовских случайных величин) - Получаем: $K(N-1)$ степеней свободы

Шаг 2. Вспоминаем, что p_0 на самом деле неизвестно, и мы его оценили. От количества степеней свободы нужно отнять размерность вектора неизвестных параметров.

Сколько неизвестных в векторе p? Не N, а $N-1$ (есть уравнение связи: сумма вероятностей равна 1).

Итоговое количество степеней свободы:

$$\text{df} = K(N-1) - (N-1) = (K-1)(N-1)$$

Формула, которую несложно запомнить:

$$\boxed{\text{df} = (N-1)(K-1)}$$

В пределе:

$$\chi^2_{n_1,...,n_K} \xrightarrow{d} \chi^2_{(N-1)(K-1)}$$

Критическая область: правосторонняя.

3.6. Пример: два потока абитуриентов

Постановка: два потока абитуриентов получили какие-то результаты вступительных экзаменов. Можно ли считать эти потоки одинаковыми?

Метод: критерий однородности χ^2.

Степени свободы: 4 значения, 2 выборки $\Rightarrow$ df $= (4-1)(2-1) = 3$.

Результат: p-value $\approx 0{,}5$ — нулевая гипотеза **принимается** (потоки можно считать однородными).

---

4. Критерий независимости χ^2

4.1. Постановка задачи

Хотим проверить, что две случайные величины **независимы**.

Имеется парная выборка: $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$.

Предположения: - X принимает значения от 1 до N - Y принимает значения от 1 до M

Обозначения: - ν_{ig} — количество пар, где $X = i$, $Y = g$ - p_{Xi} — вероятность того, что $X = i$ - p_{Yg} — вероятность того, что $Y = g$ - p_{ig} — вероятность того, что $X = i$ и $Y = g$

4.2. Гипотезы

В терминах вероятностей условие независимости:

$$p_{ig} = p_{Xi} \cdot p_{Yg}$$

Нулевая гипотеза:

$$H_0 : \forall i, g \quad p_{ig} = p_{Xi} \cdot p_{Yg}$$

Альтернатива:

$$H_1 : \neg H_0$$

4.3. Таблица сопряжённости (Contingency Table)

Для визуализации критерия строим таблицу:

	$Y=1$	$Y=2$	...	$Y=M$	Σ
$X=1$	ν_{11}	ν_{12}	...	ν_{1M}	ν_{1*}
$X=2$	ν_{21}	ν_{22}	...	ν_{2M}	ν_{2*}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X=N$	ν_{N1}	ν_{N2}	...	ν_{NM}	ν_{N*}
Σ	ν_{*1}	ν_{*2}	...	ν_{*M}	n

В ячейках — количество пар с соответствующими значениями. В дополнительном столбце — суммы по строкам (ν_{i*}), в дополнительной строке — суммы по столбцам (ν_{*g}).

Контроль: сумма по последнему столбцу = сумма по последней строке = объём выборки n.

4.4. Статистика критерия

Записываем χ^2 в общем виде:

$$\chi^2 = \sum_{i,g} \frac{(\nu_{ig} - n \cdot p_{ig})^2}{n \cdot p_{ig}}$$

Подставляем H_0 ($p_{ig} = p_{Xi} \cdot p_{Yg}$):

$$\chi^2 = \sum_{i,g} \frac{(\nu_{ig} - n \cdot p_{Xi} \cdot p_{Yg})^2}{n \cdot p_{Xi} \cdot p_{Yg}}$$

4.5. Оценки вероятностей

Проблема: p_{Xi} и p_{Yg} нам не даны.

Оценки (по аналогии с предыдущим критерием):

$$\widehat{p}_{Xi} = \frac{\nu_{i*}}{n}$$

$$\widehat{p}_{Yg} = \frac{\nu_{*g}}{n}$$

То есть берём соответствующие маргинальные суммы из таблицы сопряжённости и делим на n.

4.6. Степени свободы

Шаг 1. Если p_{Xi} и p_{Yg} известны: - Количество значений: $M \cdot N$ - Степеней свободы: $MN - 1$

Шаг 2. На самом деле p_{Xi} и p_{Yg} неизвестны: - Количество неизвестных в p_X: $N - 1$ (с учётом уравнения связи) - Количество неизвестных в p_Y: $M - 1$

Итог:

$$\text{df} = MN - 1 - (N-1) - (M-1) = MN - N - M + 1$$

Раскладываем (выносим N за скобку):

$$\text{df} = N(M-1) - (M-1) = (N-1)(M-1)$$

$$\boxed{\text{df} = (N-1)(M-1)}$$

Критическая область: правосторонняя (как и во всех модификациях критерия согласия Пирсона).

4.7. Зачем нам нужны степени свободы?

Степени свободы нужны для того, чтобы: 1. **Посчитать критическую область:** для правосторонней области рассматриваем квантиль уровня $1 - \alpha$ распределения χ^2, а это распределение задаётся именно числом степеней свободы. 2. **Посчитать p-value:** p-value $= 1 - \text{CDF}(\text{stat})$ — функция распределения тоже зависит от количества степеней свободы.

4.8. Пример: вакцина и здоровье (данные о болезни)

В коде использована готовая реализация, которая считает таблицу сопряжённости автоматически.

Результат: - Статистика $\chi^2 \approx 26{,}01$ - p-value очень маленький

Вывод: гипотеза независимости **отвергается** $\Rightarrow$ можно говорить о том, что **вакцина влияет на здоровье**.

> **Замечание (тизер):** в примере с вакцинами интереснее доказать не просто, что вакцина влияет, а что она влияет в **положительную сторону**. Критерий можно соответствующим образом модифицировать. Это будет рассмотрено в следующих лекциях.

Лекция 10: Статистические тесты

1. Критерий на коэффициент корреляции Пирсона

Постановка задачи

Пусть имеется парная выборка:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

При условии: выборка пришла из **двумерного нормального (гауссовского) распределения**.

Гипотезы

- **Нулевая гипотеза** H_0**:** $\operatorname{cov}(X, Y) = 0$, что эквивалентно $\rho_{X,Y} = 0$ (теоретический коэффициент корреляции равен нулю).
- **Альтернативная гипотеза** H_1**:** $\rho \neq 0$, $\rho > 0$ или $\rho < 0$ (альтернативу можно специфицировать).

Статистика критерия

$$t = \frac{\sqrt{n-2} \cdot \widehat{\rho}_{\text{Pearson}}}{\sqrt{1 - \widehat{\rho}^2_{\text{Pearson}}}}$$

где $\widehat{\rho}_{\text{Pearson}}$ — выборочный коэффициент корреляции Пирсона:

$$\hat{\rho}_{\text{Pearson}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Распределение статистики

При условии истинности H_0 статистика t имеет **распределение Стьюдента с $n-2$ степенями свободы**.

> **Замечание Ивана Александровича:** формальный вывод этого факта довольно громоздкий, поэтому на лекции даётся только формулировка.

Важное замечание о связи с независимостью

Вспомним общее соотношение: - Из **независимости** $\Rightarrow$ **некоррелированность** (всегда верно). - Обратная импликация в общем случае **неверна**.

Однако для **компонент гауссовского вектора** обратная стрелочка работает: некоррелированность $\Leftrightarrow$ независимость. Это один из специальных случаев.

Следствия: - Если (X, Y) — гауссовский вектор, то критерий проверяет **независимость** (по-честному). - Гипотетически тест Стьюдента можно использовать и для негауссовской выборки, но тогда проверяется только **некоррелированность** (более слабое условие). - Для негауссовских выборок критерий лучше использовать при **достаточно больших** n, поскольку при больших n распределение Стьюдента становится близким к стандартному нормальному (вспомните формальное определение распределения Стьюдента).

> **Уточнение:** «достаточно большое n» относится именно ко второму случаю (негауссовская выборка). Если выборка гауссовская — критерий работает при любом $n \geq 3$.

---

2. Критерий квантилей

Постановка задачи

Заданы: - Числа $q_1 < q_2 < \ldots < q_N$ - Числа $p_1 < p_2 < \ldots < p_N$

Выборка пришла из **непрерывного распределения**.

Гипотезы

Нулевая гипотеза H_0:

$$F(q_1) = p_1, \quad F(q_2) = p_2, \quad \dots, \quad F(q_N) = p_N$$

То есть проверяется, что q_k — квантиль порядка p_k для всех $k = 1, \dots, N$.

Альтернативная гипотеза H_1: отрицание H_0.

> **Замечание:** в общем случае H_1 не обязательно является отрицанием H_0, но в этой конкретной ситуации это так.

Дополнительные обозначения

Введём: - $q_0 = -\infty$, $q_{N+1} = +\infty$ - $p_0 = 0$, $p_{N+1} = 1$

Разобьём вещественную ось на полуинтервалы: - $\Delta_1 = [q_0, q_1) = (-\infty, q_1)$ - $\Delta_2 = [q_1, q_2)$ - ... - $\Delta_N = [q_{N-1}, q_N)$ - $\Delta_{N+1} = [q_N, q_{N+1}) = [q_N, +\infty)$

Введём приращения вероятностей: - $\Delta p_1 = p_1 - p_0$ - $\Delta p_2 = p_2 - p_1$ - ... - $\Delta p_N = p_N - p_{N-1}$ - $\Delta p_{N+1} = p_{N+1} - p_N$

Эквивалентная формулировка H_0

Система H_0 равносильна тому, что:

$$\mathbb{P}(X \in \Delta_k) = \Delta p_k, \quad k = 1, \dots, N+1$$

То есть вероятность попадания в каждый интервал Δ_k равна Δp_k.

Сведение к критерию согласия Пирсона χ^2

Пусть ν_k — количество элементов выборки, попавших в промежуток Δ_k. Тогда задача сводится к критерию согласия Пирсона со статистикой:

$$\chi^2 = \sum_{k=1}^{N+1} \frac{(\nu_k - n\Delta p_k)^2}{n\Delta p_k}$$

где n — объём выборки.

Распределение статистики и критическая область

При условии истинности H_0:

$$\chi^2 \xrightarrow{d} \chi^2_N$$

То есть распределение χ^2 с N степенями свободы (количество интервалов $N+1$ минус 1, как для простой гипотезы в критерии Пирсона).

Тип критической области: правосторонний (как и в критерии согласия Пирсона).

---

3. Критерий знаков (как частный случай критерия квантилей)

Постановка

Возьмём $N = 1$, $p_1 = \frac{1}{2}$. Тогда получаем критерий знаков.

Гипотеза

Проверяется:

$$F(c_0) = \frac{1}{2}$$

То есть верно ли, что **медиана** равна заданной константе c_0.

Статистика критерия

Поскольку у χ^2 только одна степень свободы, можно думать о статистике как о квадрате стандартной гауссовской величины. Поэтому используем без квадрата:

$$Z = \frac{\nu_1 - \frac{n}{2}}{\sqrt{n \cdot \frac{1}{2} \cdot \frac{1}{2}}} = \frac{\nu_1 - \frac{n}{2}}{\frac{\sqrt{n}}{2}}$$

где ν_1 — количество чисел в выборке, **меньших** потенциальной медианы c_0 (то есть попавших в Δ_1).

> Не возводим в квадрат, потому что хотим извлечь корень — рассматриваем сразу величину, сходящуюся к стандартному нормальному распределению по ЦПТ.

Распределение

При H_0 (по ЦПТ):

$$Z \xrightarrow{d} \mathcal{N}(0,1)$$

Альтернативы

Поскольку в пределе стандартное гауссовское распределение, альтернативы можно специфицировать: - $c \neq c_0$ (двусторонняя) - $c > c_0$ (правосторонняя) - $c < c_0$ (левосторонняя)

> **Историческое замечание:** ранее (в курсе) уже доказывался критерий согласия Пирсона для случая двух значений — и там получалось p и $1-p$, что как раз согласуется с настоящей формулировкой.

---

4. Применение критерия знаков к парной выборке

Постановка

Пусть имеется парная выборка:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

где обе компоненты имеют непрерывные распределения.

Гипотеза

$$H_0: \quad F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y), \quad F_X = F_Y$$

То есть X и Y: - **независимы** - **одинаково распределены**

Метод

Составляем новую выборку из разностей:

$$U_i = X_i - Y_i$$

Найдём $\mathbb{P}(U > 0)$ при условии H_0:

$$\mathbb{P}(U > 0) = \mathbb{P}(X - Y > 0) = \iint_{x-y>0} p(x,y)\,dx\,dy$$

При H_0 совместная плотность есть произведение одинаковых одномерных плотностей:

$$p(x,y) = p(x) \cdot p(y)$$

Тогда:

$$\mathbb{P}(X > Y) = \iint_{x>y} p(x)p(y)\,dx\,dy = \frac{1}{2}$$

Почему $\frac{1}{2}$? При перемене местами x и y подынтегральная функция не меняется. А плоскость разбивается на две равные части, и интеграл по всей плоскости равен 1. Значит, по симметрии каждый интеграл равен $\frac{1}{2}$.

Аналогично $\mathbb{P}(U < 0) = \frac{1}{2}$.

Вывод

Это означает, что **0 является медианой** распределения $U = X - Y$.

Поэтому к новой выборке $\{U_i\}$ применяем критерий знаков с $c_0 = 0$.

> **Важное замечание:** критерий знаков для такой задачи годится для **предварительного анализа** (по словам Ивана Александровича, неформально).

---

5. Ранговые критерии

Понятие ранга

Ранг элемента выборки — это его индекс в **вариационном ряде** (отсортированная по возрастанию выборка).

Проблема повторов

Если в выборке есть повторяющиеся значения, ранг можно определить разными способами:

Пример: выборка $(3, 1, 3, 1, 1, 2)$.

Вариационный ряд: $1, 1, 1, 2, 3, 3$.

Возможные подходы определения ранга: 1. **Минимальный ранг** — берётся минимальный из возможных рангов для группы повторов. - Ранг тройки = 5, ранг единицы = 1, ранг двойки = 4. 2. **Максимальный ранг.** - Ранг тройки = 6, ранг единицы = 3, ранг двойки = 4. 3. **Средний (среднеарифметический) ранг.** - Ранг тройки = 5.5, ранг единицы = 2, ранг двойки = 4. 4. **Различие между одинаковыми элементами** — присваиваем разным «единичкам» разные ранги в порядке появления: - Получится, например: $(5, 1, 6, 2, 3, 4)$ для исходной $(3, 1, 3, 1, 1, 2)$.

> **Практическое замечание:** при использовании рангового стат-теста нужно внимательно смотреть, как авторы/разработчики поступают с дублированными рангами. Дальнейшее изложение предполагает, что **все ранги уникальны**.

---

6. Критерий Уилкоксона / Манна–Уитни (Wilcoxon–Mann–Whitney)

> **Замечание:** формально это два разных теста, но они очень тесно связаны (аналогично критериям Колмогорова–Смирнова), поэтому их объединяют.

Постановка

Пусть есть две независимые выборки: - $X = (X_1, \dots, X_m)$ - $Y = (Y_1, \dots, Y_n)$

(возможно, разных объёмов).

Действие: объединяем их в одну выборку (union to one sample). Пусть R_i — ранг X_i в **объединённой выборке**.

Статистика Уилкоксона

$$T = R_1 + R_2 + \dots + R_m = \sum_{i=1}^{m} R_i$$

(сумма рангов первой выборки в объединённой).

Статистика Манна-Уитни

$$U_1 = \sum_{r=1}^{m} \sum_{s=1}^{n} \mathbb{I}\{X_r < Y_s\}$$

где $\mathbb{I}\{\cdot\}$ — индикаторная функция (1, если $X_r < Y_s$; 0 иначе).

Связь между статистиками

Эти две статистики связаны линейно:

$$T + U_1 = m \cdot n + \frac{n(n+1)}{2}$$

Это соотношение даётся без доказательства («просто так работает»).

Гипотезы (для критерия Манна-Уитни)

Нулевая гипотеза H_0: выборки **однородны** — распределение X совпадает с распределением Y.

Математическое ожидание U_1

$$\mathbb{E}[U_1] = m \cdot n \cdot \mathbb{P}(X_r < Y_s)$$

Если H_0 верна (то есть X и Y имеют одинаковое распределение, и обсуждавшимся выше способом), то:

$$\mathbb{P}(X_r < Y_s) = \frac{1}{2}$$

Отсюда:

$$\mathbb{E}[U_1]\big|_{H_0} = \frac{mn}{2}$$

Обозначим $a = \mathbb{P}(X_r < Y_s)$. При H_0: $a = \frac{1}{2}$.

Дисперсия

$$\mathrm{Var}(U_1)\big|_{H_0} = \frac{mn(n+m+1)}{12}$$

Альтернативы

Альтернативы формулируются через a или, эквивалентно, через медиану разности: - $a \neq \frac{1}{2}$ - $a > \frac{1}{2}$ - $a < \frac{1}{2}$

Это эквивалентно условию $X = Y + c$ (то есть **сдвиг распределения**).

> **Важная ремарка:** критерий Манна–Уитни хорошо ловит именно **сдвиги** распределений.

Распределение статистики

- При **малых** m и n — критическая область **табулирована**.
- При **больших** m и n:

$$U_1 \approx \mathcal{N}\left(\frac{mn}{2}, \frac{mn(n+m+1)}{12}\right)$$

---

7. Коэффициент корреляции Спирмена

Постановка

Парная выборка $(X_1, Y_1), \dots, (X_n, Y_n)$.

Сопоставим каждому элементу его ранг **в своей выборке**: - R_k — ранг X_k среди $X_1, \dots, X_n$ - S_k — ранг Y_k среди $Y_1, \dots, Y_n$

Определение

Коэффициент корреляции Спирмена — это выборочный коэффициент корреляции Пирсона между рангами:

$$\hat{\rho}_{\text{Spearman}} = \hat{\rho}_{\text{Pearson}}(R, S)$$

> **Замечание:** для уникальных рангов существуют упрощённые формулы для подсчёта (см. Ивченко–Медведев, Кобзарь).

Гипотезы

- H_0: корреляция равна 0
- H_1: корреляция $\neq 0$, > 0 или < 0

В зависимости от альтернативы выбирается тип критической области (двусторонняя, правосторонняя, левосторонняя).

Распределение статистики

- При **малых** n — табулировано.
- При **больших** n при условии H_0:

$$\text{статистика} \xrightarrow{d} \mathcal{N}(0,1)$$

---

8. Коэффициент корреляции Кендалла

Подготовка

Имеется последовательность пар рангов $(R_1, S_1), \dots, (R_n, S_n)$.

Шаг 1. Рассмотрим эти пары как датафрейм с двумя столбцами и **отсортируем** по первому атрибуту. После сортировки первый столбец становится $1, 2, 3, \dots, n$, а второй столбец — некоторая перестановка $T_1, T_2, \dots, T_n$.

Определение

Коэффициент корреляции Кендалла:

$$\tau = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sgn}(T_j - T_i)$$

где $\operatorname{sgn}(x)$ — функция знака:

$$\operatorname{sgn}(x) = \begin{cases} +1, & x > 0 \\ -1, & x < 0 \\ 0, & x = 0 \text{ (по соглашению)} \end{cases}$$

Интуиция

После сортировки по первому ключу мы смотрим на вторую строку и считаем **количество инверсий** (точнее, разность между конкордантными и дисконкордантными парами).

Предельные случаи

Ситуация после сортировки	Кол-во инверсий	τ
Полностью отсортировано **по возрастанию**	0	$+1$
Полностью отсортировано **по убыванию**	$\frac{n(n-1)}{2}$ (максимум)	-1
Перемешано случайно	около среднего	около 0

> **Смысл:** чем ближе τ к 0, тем более «случайной» (независимой) считается выборка. Количество инверсий — это мера случайности, мера отсутствия зависимости.

Распределение

- При **малых** n — табулировано.
- При **больших** n — нормальная аппроксимация:

$$\tau \approx \mathcal{N}\left(0, \frac{4}{9n}\right)$$

Гипотезы

- H_0: корреляции нет
- H_1: корреляция $\neq 0$, > 0 или < 0

Замечание о тестах Спирмена и Кендалла

Оба теста (Спирмена и Кендалла) **хорошо ловят монотонную зависимость**.

> **Предостережение Ивана Александровича:** будут показаны примеры в notebook, где тест принимает H_0 (отсутствие зависимости), но выборки на самом деле **зависимы** (немонотонная зависимость).

---

9. Критерий инверсий

Зачем нужен этот критерий

До сих пор все рассмотренные тесты предполагали работу с **моделью простейшей выборки** (одной или нескольких). Но это сильное предположение — далеко не все выборки простейшие. Есть тесты, которые проверяют **согласованность данных с моделью простейшей выборки**.

Что такое модель простейшей выборки

Случайные величины **независимы** и **одинаково распределены** (i.i.d.).

Постановка

Пусть имеются $X_1, X_2, \dots, X_n$ — **непрерывные** случайные величины (это важное предположение).

Гипотезы

H_0**:** величины $X_1, \dots, X_n$: - (а) независимы - (б) одинаково распределены

То есть совместная функция распределения есть произведение одномерных и при этом они одинаковые. Иными словами, числа соответствуют модели простейшей выборки.

H_1**:** $\neg H_0$.

Определение инверсии

Пара (X_i, X_j) при $i < j$ образует **инверсию**, если $X_j < X_i$.

Иными словами, в вариационном ряду X_j предшествует X_i.

Статистика

- t_1 — количество инверсий для X_1 (т.е. число пар X_1, X_j при $j > 1$, образующих инверсию).
- t_2 — количество инверсий для X_2.
- ...
- t_{n-1} — количество инверсий для X_{n-1}.

Статистика теста:

$$N = \sum_{i=1}^{n-1} t_i$$

— общее количество инверсий во всей выборке.

Идея критерия

Если H_0 верна, то:

$$\mathbb{P}(\text{любой перестановки}) = \frac{1}{n!}$$

То есть все возможные расстановки равновероятны.

Предельные случаи

Ситуация	Количество инверсий N
Выборка отсортирована по возрастанию	0
Выборка отсортирована по убыванию	$\frac{n(n-1)}{2}$
Перемешана случайно	около среднего

Если числа полностью отсортированы — индикатор того, что они вряд ли случайны. Если перемешаны — скорее случайны.

Распределение

- При **малых** n — распределение N табулировано.
- При **больших** n — нормальная аппроксимация:

$$N \approx \mathcal{N}\left(\frac{n(n-1)}{4}, \frac{n(n-1)(2n+5)}{72}\right)$$

> **Уточнение:** мат. ожидание $\frac{n(n-1)}{4}$ — это ровно половина максимального количества инверсий.
>
> **Откуда формулы:** во-первых, t_i независимы между собой; во-вторых, формулы можно получить через **производящие функции** (тесно связанные с характеристическими функциями). Подробный вывод см. в Ивченко–Медведев.

Заключительные замечания

Сводка рассмотренных тестов на лекции

Тест	Что проверяет	Ключевое предположение
Критерий Пирсона (на корреляцию)	$\rho_{XY} = 0$	гауссовский вектор
Критерий квантилей	$F(q_k) = p_k$	непрерывное распределение
Критерий знаков	медиана $= c_0$	непрерывное распределение
Манна–Уитни / Уилкоксона	однородность двух выборок	независимость выборок, ловит сдвиг
Спирмена	независимость в парной выборке	монотонная зависимость
Кендалла	независимость в парной выборке	монотонная зависимость
Критерий инверсий	модель простейшей выборки (i.i.d.)	непрерывные случайные величины

Главная мысль Ивана Александровича

Хотя тестов было рассмотрено много, на самом деле это лишь малая часть. Самое важное на этом этапе — **уловить общий принцип работы стат-теста**. Тогда при необходимости в конкретной задаче вы сможете самостоятельно подобрать подходящий тест и разобраться с ним.

Что дальше

Со следующего занятия начнётся новая большая тема — **линейные статистические модели**. Метод наименьших квадратов и линейная регрессия будут рассмотрены со статистической точки зрения.

Лекция 11: Линейная регрессия. Метод наименьших квадратов. Теорема Гаусса–Маркова

Введение

Сегодня начинается разговор про **линейные модели**, в частности — про **линейную регрессию**. Многие уже сталкивались с линейной регрессией и методом наименьших квадратов в других контекстах. Оказывается, эту, казалось бы, простую модель можно рассмотреть и со **статистической точки зрения**, чем мы и займёмся.

Постановка задачи

Рассмотрим модель в матричном виде:

$$y = Xc + \varepsilon$$

Раскроем смысл каждого объекта.

Матрица переменных X

X — это матрица $n \times m$ с вещественными коэффициентами. Это **матрица переменных**, где:

- n — количество наблюдений, доступных нам;
- m — количество переменных.

При этом X воспринимается как **не случайная** величина — это какой-то конкретный детерминированный набор.

Вектор коэффициентов c

$c \in \mathbb{R}^m$ — **неизвестный вектор коэффициентов** (вектор из m компонент).

Ошибка ε

ε — это **ошибка**, поскольку идеальная линейная зависимость встречается редко. Ошибка воспринимается как **случайная величина**.

ε — это вектор длины n:

$$\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$$

где ε_i соответствует i-му наблюдению.

Предположения на ошибку

На ошибку накладываются следующие предположения:

1. **Нулевое математическое ожидание:**

   $$\mathbb{E}\varepsilon_i = 0$$

   То есть в среднем ошибка равна нулю — это означает, что модель «более-менее адекватная».

2. **Некоррелированность** (но не независимость!):

   $$\text{cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$$

   На интуитивном уровне — мы независимо наблюдаем и i-ю, и j-ю строчку. Замечание: студент предложил предположение независимости и одинаковой распределённости, но Иван Александрович уточнил, что **независимость пока не предполагается** — только некоррелированность.

3. **Гомоскедастичность** — одинаковые дисперсии у ошибок:

   $$\mathbb{D}\varepsilon_i = \sigma^2$$

   Слово **гомоскедастичность** означает, что дисперсии у ошибок одинаковые. При этом σ^2 **неизвестна**.

Вектор наблюдений y

$y \in \mathbb{R}^n$ — **наблюдение зависимой переменной**.

Глобальная цель

Цель — «оценить» вектор коэффициентов c и величину σ^2 (которая называется **остаточная дисперсия**).

> Слово «оценить» написано в кавычках, потому что: - Можно дать **точечную** оценку; - Можно построить **доверительный интервал**; - Можно проверять **гипотезы**.
>
> То есть можно решать всякие разные статистические задачи касательно c и σ^2.

Пример: цена недвижимости

Допустим, рассматриваем цену недвижимости. Цена недвижимости (это y) может зависеть от разных факторов:

- расстояние от центра города;
- расстояние до ближайшего метро;
- и так далее.

Эти переменные образуют матрицу X:

$$X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{pmatrix}$$

- Первая строчка — значение переменных для первого наблюдения;
- Вторая строчка — для второго наблюдения; и т. д.

Например, столбец y — это `flat\_price` ($y_1, \dots, y_n$). Переменные: - `distance\_to\_center`: $x_{1,1}, \dots, x_{n,1}$; - `distance\_to\_nearest\_subway`: $x_{1,2}, \dots, x_{n,2}$; - и т. д.

Предполагаем, что цена линейно зависит от факторов:

$$y_1 = c_1 x_{1,1} + c_2 x_{1,2} + \dots + c_m x_{1,m} + \varepsilon_1$$

Грубо говоря, c_j — это **значимость** (коэффициент) при соответствующей переменной. Цель — оценить эти коэффициенты.

Замечание про свободный коэффициент c_0

Часто в линейных моделях фигурирует свободный коэффициент c_0. Однако его введение **не умаляет общности** записи. Если ввести c_0, то это частный случай рассмотренной ситуации:

$$c_0 \cdot 1$$

— то есть мы добавляем **фиктивную переменную**, равную единице для всех наблюдений. Поэтому общий вид $y = Xc + \varepsilon$ покрывает и случай со свободным членом.

Вспомогательная матрица A

Введём матрицу:

$$A = X^T X$$

На что она похожа? Это похоже на «**ковариацию**» между переменными (в кавычках!).

Действительно, строчка матрицы X^T — это столбец переменной. Если поделить на n, то получится почти выборочная ковариация. Формально это **не совсем** ковариация, но нечто, очень сильно напоминающее её. На интуитивном уровне про матрицу A можно думать как про вариацию между переменными.

Свойства матрицы A

- A — матрица $m \times m$ по построению.
- **Предполагаем:** $\text{rank}(A) = m$.

Это означает, что переменные **линейно независимы**. В контексте регрессионного анализа это называется **отсутствие мультиколлинеарности**.

> $\text{rank}(A) = m \Longleftrightarrow$ переменные линейно независимы $\Longleftrightarrow$ отсутствует мультиколлинеарность.

Также предполагаем, что **количество наблюдений существенно больше количества переменных**: $n \gg m$.

Оценка наименьших квадратов

Рассмотрим квадратическую ошибку:

$$S^2(c) = \sum_i \left(\sum_j x_{ij} c_j - y_i \right)^2$$

Или в матричном виде:

$$S^2(c) = (Xc - y)^T (Xc - y)$$

Оценка наименьших квадратов $\hat{c}$ — это оценка, которая **минимизирует** квадратическую ошибку:

$$\hat{c} = \arg\min_c S^2(c)$$

Утверждение: формула для $\hat{c}$

В рамках наших предположений можно написать **точную формулу**:

$$\boxed{\hat{c} = A^{-1} X^T y}$$

Доказательство

Обычно доказательство ведётся через дифференцирование $S^2(c)$ по c и приравнивание градиента к нулю. Однако докажем «в лоб» — по ходу доказательства получим **важное соотношение**, которое будет использовано в дальнейшем.

Рассмотрим $S^2(\hat{c} + h)$, где h — некоторое **приращение**. Распишем:

$$S^2(\hat{c} + h) = (X(\hat{c} + h) - y)^T (X(\hat{c} + h) - y)$$

Сгруппируем так:

$$= \big((X\hat{c} - y) + Xh\big)^T \big((X\hat{c} - y) + Xh\big)$$

Раскрываем скобки:

$$= \underbrace{(X\hat{c} - y)^T(X\hat{c} - y)}_{S^2(\hat{c})} + h^T X^T(X\hat{c} - y) + (X\hat{c} - y)^T Xh + h^T X^T Xh$$

Анализ перекрёстных членов Распишем $h^T X^T(X\hat{c} - y)$, подставляя $\hat{c} = A^{-1}X^T y$:

$$h^T X^T X \cdot A^{-1}X^T y - h^T X^T y$$

Поскольку $A = X^T X$ и A обратима (по предположению о ранге):

$$X^T X \cdot A^{-1} = A \cdot A^{-1} = I$$

Поэтому:

$$h^T X^T y - h^T X^T y = 0$$

Аналогично второй перекрёстный член:

$$(X\hat{c} - y)^T Xh = (XA^{-1}X^T y - y)^T Xh = y^T XA^{-1}X^T Xh - y^T Xh = y^T Xh - y^T Xh = 0$$

Итоговое соотношение Таким образом:

$$S^2(\hat{c} + h) = S^2(\hat{c}) + h^T X^T Xh = S^2(\hat{c}) + h^T Ah$$

Заметим, что $h^T Ah = h^T X^T Xh = (Xh)^T(Xh) \geq 0$ — скалярное произведение вектора на себя.

Поскольку $\text{rank}(A) = m$, матрица A **не вырождена**, значит, она **строго положительно определена**. Это означает: если $h \neq 0$, то $h^T Ah > 0$, то есть:

$$S^2(\hat{c} + h) > S^2(\hat{c})$$

Тем самым доказано, что $\hat{c} = A^{-1}X^T y$ действительно является минимумом. □

Важное соотношение, полученное по ходу доказательства

Если положить $c_1 = \hat{c} + h$, $c_2 = \hat{c}$, то $h = c_1 - c_2$, и мы получили:

$$\boxed{S^2(c_1) - S^2(c_2) = (c_1 - c_2)^T A(c_1 - c_2)}$$

Это соотношение будет использоваться в дальнейших выкладках.

Практическое замечание

С вычислительной точки зрения формула $\hat{c} = A^{-1}X^T y$ **не самая удобная**: нужно обращать матрицы, перемножать их. На практике обычно используются **численные методы**: - оптимизация исходной функции ошибок; - численное решение уравнения градиент = 0.

Теорема Гаусса–Маркова

Это **фундаментальная теорема** в рамках линейных моделей. Традиционно она формулируется для самой оценки наименьших квадратов, но здесь рассмотрим **более общее утверждение**.

Постановка

Рассмотрим линейную функцию от вектора коэффициентов:

$$\tau = Tc$$

где T — матрица $k \times m$, $k \leq m$, $\text{rank}(T) = k$.

> Если взять $T = I$ (единичная матрица), получим теорему Гаусса–Маркова для обычной оценки наименьших квадратов.

Введём оценку:

$$\hat{\tau} = T\hat{c}$$

Зачем нужно T?

В дальнейшем будут проверяться гипотезы о векторе c при **линейных ограничениях**. Соотношение $Tc = \tau$ как раз задаёт линейное ограничение. В качестве нулевой гипотезы стат-теста будет выступать предположение, что c удовлетворяет каким-то линейным ограничениям.

Формулировка

При выполнении всех предположений (некоррелированность ошибок, нулевое мат. ожидание, гомоскедастичность):

(а) $\hat{\tau}$ — **несмещённая** оценка для τ:

$$\mathbb{E}\hat{\tau} = \tau$$

(б) Матрица ковариаций $\text{cov}(\hat{\tau}) = \sigma^2 T A^{-1} T^T$, и $\hat{\tau}$ — **оптимальная** оценка для τ в классе **линейных по y несмещённых оценок**.

Доказательство (а): несмещённость

$$\mathbb{E}\hat{\tau} = \mathbb{E}[T\hat{c}] = \mathbb{E}[TA^{-1}X^T y]$$

T, A^{-1}, X^T — константы, выносим за знак мат. ожидания:

$$= TA^{-1}X^T\mathbb{E}y = TA^{-1}X^T\mathbb{E}[Xc + \varepsilon] = TA^{-1}X^T Xc$$

(поскольку $\mathbb{E}\varepsilon = 0$, а Xc — константа). Учитывая $X^T X = A$:

$$= TA^{-1}Ac = Tc = \tau \quad \blacksquare$$

Доказательство (б): матрица ковариаций

$$\text{cov}(\hat{\tau}) = \text{cov}(T\hat{c}) = T \cdot \text{cov}(\hat{c}) \cdot T^T$$

> **Замечание (вопрос студента):** в одномерном случае $\mathbb{D}(aX) = a^2\mathbb{D}X$, но в **многомерном** случае матрица ковариаций aX — это $A \cdot \text{cov}(X) \cdot A^T$.

Это **именно матрица ковариаций**, а не дисперсия в квадрате, потому что $\hat{\tau}$ — это случайный **вектор** (многомерная величина).

Считаем $\text{cov}(\hat{c})$:

$$\text{cov}(\hat{c}) = \text{cov}(A^{-1}X^T y) = A^{-1}X^T \cdot \text{cov}(y) \cdot XA^{-1}$$

Симметрия A^{-1}: $A = X^T X$ симметрична ($A^T = (X^T X)^T = X^T X = A$), значит, A^{-1} тоже симметрична. Поэтому $(A^{-1})^T = A^{-1}$.

Считаем $\text{cov}(y)$:

$$\text{cov}(y) = \text{cov}(Xc + \varepsilon)$$

Xc — константа, **сдвиг на матрицу ковариаций не влияет** (аналогично одномерному случаю, где $\mathbb{D}(X + a) = \mathbb{D}X$):

$$\text{cov}(y) = \text{cov}(\varepsilon)$$

Поскольку компоненты ε некоррелированы и имеют одинаковую дисперсию σ^2:

$$\text{cov}(\varepsilon) = \sigma^2 I$$

Подставляем:

$$\text{cov}(\hat{c}) = A^{-1}X^T \cdot \sigma^2 I \cdot XA^{-1} = \sigma^2 A^{-1} \underbrace{X^T X}_{A} A^{-1} = \sigma^2 A^{-1}$$

Итого:

$$\boxed{\text{cov}(\hat{\tau}) = \sigma^2 T A^{-1} T^T}$$

Введём обозначение:

$$D = TA^{-1}X^T$$

(к этому обозначению вернёмся позже).

Доказательство (б): оптимальность

Напоминание: критерий оптимальности Для несмещённых оценок: оценка оптимальна, если у неё **минимальная дисперсия**. В многомерном случае оптимизируется:

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta}-\theta)^T(\hat{\theta}-\theta)\big]$$

Можно показать, что:

$$\text{MSE}(\hat{\theta}) = \text{tr}(\text{cov}(\hat{\theta})) + \text{bias}^T\text{bias}$$

где tr — **след** матрицы (сумма диагональных элементов), а $\text{bias} = \mathbb{E}\hat{\theta} - \theta$.

Это обобщение одномерной формулы $\text{MSE} = \mathbb{D} + \text{bias}^2$.

В нашем случае оценка несмещённая, поэтому $\text{bias} = 0$ — нужно минимизировать $\text{tr}(\text{cov}(\hat{\tau}))$.

Шаг А: произвольная линейная несмещённая оценка Пусть $\hat{L} = Ly$ — произвольная линейная по y несмещённая оценка для τ:

$$\mathbb{E}[Ly] = \tau$$

С другой стороны:

$$\mathbb{E}[Ly] = L \cdot \mathbb{E}[Xc + \varepsilon] = LXc$$

Поскольку $\tau = Tc$, получаем $Tc = LXc$ для **любого** c. Отсюда:

$$\boxed{T = LX}$$

Шаг В: переобозначение Прибавим и вычтем $TA^{-1}X^T$:

$$L = \underbrace{(L - TA^{-1}X^T)}_{\hat{L}} + TA^{-1}X^T$$

Введём $\hat{L} = L - TA^{-1}X^T$. Тогда:

$$L = \hat{L} + TA^{-1}X^T$$

Дополнительное соотношение Из $T = LX$ домножим обе части на X справа... нет, у нас уже $T = LX$. Подставим $L = \hat{L} + TA^{-1}X^T$:

$$T = \hat{L}X + TA^{-1}\underbrace{X^TX}_{A} = \hat{L}X + T$$

Отсюда:

$$\boxed{\hat{L}X = 0}$$

Транспонируя: $X^T\hat{L}^T = 0$.

Шаг C: матрица ковариаций для Ly

$$\mathrm{cov}(Ly) = L \cdot \mathrm{cov}(y) \cdot L^T = \sigma^2 LL^T$$

Распишем $\sigma^2 LL^T$, подставляя $L = TA^{-1}X^T + \hat{L}$:

$$\sigma^2 LL^T = \sigma^2(TA^{-1}X^T + \hat{L})(TA^{-1}X^T + \hat{L})^T$$

Раскрываем:

$$= \sigma^2\big[TA^{-1}\underbrace{X^TX}_{A}A^{-1}T^T + TA^{-1}X^T\hat{L}^T + \hat{L}XA^{-1}T^T + \hat{L}\hat{L}^T\big]$$

Используем $\hat{L}X = 0$ и $X^T\hat{L}^T = 0$ — средние два слагаемых обнуляются:

$$\mathrm{cov}(Ly) = \sigma^2 TA^{-1}T^T + \sigma^2\hat{L}\hat{L}^T$$

Финальный шаг: оптимизация следа Получили:

$$\operatorname{cov}(Ly) = \underbrace{\sigma^2 T A^{-1} T^T}_{\operatorname{cov}(\hat{\tau}),\ \text{не зависит от выбора } L} + \underbrace{\sigma^2 \hat{L}\hat{L}^T}_{\text{зависит от } \hat{L}}$$

Считаем след:

$$\operatorname{tr}(\hat{L}\hat{L}^T) = \sum_i (\hat{L}\hat{L}^T)_{ii} = \sum_i \sum_j \hat{L}_{ij}^2$$

(диагональный элемент $(\hat{L}\hat{L}^T)_{ii}$ — это i-я строка, скалярно умноженная на саму себя, то есть сумма квадратов её элементов).

Минимум суммы квадратов достигается при $\hat{L}_{ij} = 0$ для всех i, j, то есть $\hat{L} = 0$. А это в точности означает, что $L = TA^{-1}X^T$ — то есть, что $Ly = \hat{\tau}$.

Таким образом, оценка наименьших квадратов **оптимальна** в классе линейных несмещённых оценок. □

Точечная оценка для σ^2

Найдём **несмещённую оценку** для остаточной дисперсии σ^2.

Шаг 1: вычислим $\mathbb{E}S^2(c)$

$$\mathbb{E}S^2(c) = \mathbb{E}\big[(Xc - y)^T (Xc - y)\big]$$

Поскольку $Xc - y = -\varepsilon$:

$$= \mathbb{E}[\varepsilon^T \varepsilon] = \mathbb{E} \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \mathbb{E}\varepsilon_i^2$$

Используем:

$$\mathbb{E}\varepsilon_i^2 = \mathbb{D}\varepsilon_i + (\mathbb{E}\varepsilon_i)^2 = \sigma^2 + 0 = \sigma^2$$

Итого:

$$\boxed{\mathbb{E}S^2(c) = n\sigma^2}$$

Шаг 2: вычислим $\mathbb{E}[S^2(c) - S^2(\hat{c})]$

Используем выведенное ранее соотношение:

$$S^2(c) - S^2(\hat{c}) = (\hat{c} - c)^T A(\hat{c} - c)$$

(подставили $c_1 = c$, $c_2 = \hat{c}$, $h = c - \hat{c}$, но из-за симметрии знак не важен).

Расписываем по компонентам:

$$\mathbb{E}[S^2(c) - S^2(\hat{c})] = \mathbb{E}\sum_{i,j}(\hat{c}_i - c_i)A_{ij}(\hat{c}_j - c_j)$$

По линейности мат. ожидания (и поскольку $c_i = \mathbb{E}\hat{c}_i$ — несмещённость):

$$= \sum_{i,j} A_{ij} \cdot \mathbb{E}[(\hat{c}_i - \mathbb{E}\hat{c}_i)(\hat{c}_j - \mathbb{E}\hat{c}_j)] = \sum_{i,j} A_{ij} \cdot \mathrm{cov}(\hat{c}_i, \hat{c}_j)$$

Замечание о матрице ковариаций $\hat{c}$ В теореме Гаусса–Маркова матрица ковариаций $\hat{\tau}$ равна $\sigma^2 T A^{-1} T^T$. Подставляя $T = I$, получаем:

$$\mathrm{cov}(\hat{c}) = \sigma^2 A^{-1}$$

Поэтому $\mathrm{cov}(\hat{c}_i, \hat{c}_j) = \sigma^2 (A^{-1})_{ij}$.

Продолжение вычислений

$$\mathbb{E}[S^2(c) - S^2(\hat{c})] = \sigma^2 \sum_{i,j} A_{ij}(A^{-1})_{ij}$$

Используя симметрию A: $A_{ij} = A_{ji}$, и заметим:

$$\sum_{i,j} A_{ji}(A^{-1})_{ij} = \sum_i (A \cdot A^{-1})_{ii} = \sum_i I_{ii} = m$$

(строка матрицы A умножается на столбец A^{-1} — это диагональный элемент произведения $AA^{-1} = I$).

Итого:

$$\mathbb{E}[S^2(c) - S^2(\hat{c})] = \sigma^2 \cdot m$$

Шаг 3: окончательная формула

Из шагов 1 и 2:

$$\mathbb{E}S^2(\hat{c}) = \mathbb{E}S^2(c) - \sigma^2 m = n\sigma^2 - m\sigma^2 = (n-m)\sigma^2$$

Откуда:

$$\boxed{\mathbb{E}\left[\frac{S^2(\hat{c})}{n-m}\right] = \sigma^2}$$

Таким образом, **несмещённая оценка остаточной дисперсии**:

$$\hat{\sigma}^2 = \frac{S^2(\hat{c})}{n-m}$$

Аналогия с выборочной дисперсией

Можно провести параллель с обычной выборочной дисперсией. Когда мы считали выборочную дисперсию, делённую на n, она оказывалась смещённой; чтобы сделать её несмещённой, мы делили на $n-1$.

Здесь аналогично: если бы мы делили квадратическую ошибку на n, оценка была бы **смещённой**. А деление на $n-m$ (разность между количеством наблюдений и количеством переменных) даёт **несмещённую** оценку σ^2.

Анонс следующей лекции

Сегодня были рассмотрены **точечные оценки** для $\hat{c}$ и σ^2. В следующий раз будут рассмотрены:

- **Интервальное оценивание** (доверительные интервалы);
- **Проверка различных статистических гипотез**.

Лекция 11: Линейная регрессия. Доверительные интервалы и проверка гипотез

Восстановление контекста

Рассматривается линейная модель:

$$y = Xc + \varepsilon$$

где: - c — вектор коэффициентов - X — матрица с элементами (матрица плана) - y — вектор значений - ε — вектор ошибок

Базовые предположения

1. $\mathbb{E}[\varepsilon] = 0$ — математическое ожидание ошибки равно нулю
2. Матрица ковариаций ошибок: $\text{Cov}(\varepsilon) = \sigma^2 \cdot E$, где E — единичная матрица

Это означает, что модель **гомоскедастичная**: матрица ковариаций диагональная, и на диагонали стоит одна и та же дисперсия.

Что было получено ранее

Найдена **оценка наименьших квадратов** (ОНК):

$$\hat{c} = A^{-1} X^T y$$

где $A = X^T X$.

Теорема Гаусса—Маркова (повторение): - $\hat{c}$ — несмещённая оценка - $\hat{c}$ — оптимальная (эффективная) в классе линейных несмещённых оценок

Также получена **несмещённая оценка остаточной дисперсии**:

$$\hat{\sigma}^2 = \frac{S^2(\hat{c})}{n - m}$$

где $S^2(\hat{c})$ — квадратическая ошибка для ОНК, n — число наблюдений, m — число переменных.

> До этого мы научились находить **точечные** оценки для c и для остаточной дисперсии. Теперь будем строить **доверительные интервалы**.

Усиление предположений: гауссовские ошибки

До сих пор: $\mathbb{E}[\varepsilon] = 0$ и диагональная матрица ковариаций.

Усиление: ε теперь — гауссовская величина:

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 E)$$

Так как y — линейное преобразование гауссовской величины, то y тоже гауссовская:

$$y \sim \mathcal{N}(Xc, \sigma^2 E)$$

Функция правдоподобия

Запишем плотность y при фиксированных c и σ^2 (многомерное нормальное распределение):

$$L(c, \sigma^2) = \frac{1}{(\sqrt{2\pi})^n \cdot \sigma^n} \exp\left(-\frac{1}{2\sigma^2}(y - Xc)^T(y - Xc)\right)$$

Здесь: - Определитель диагональной матрицы $\sigma^2 E$ равен σ^{2n}, корень даёт σ^n - Обратная матрица к $\sigma^2 E$ — это $\frac{1}{\sigma^2}E$

Связь МНК и метода максимального правдоподобия

Зафиксируем σ^2. Тогда максимизация L по c равносильна максимизации аргумента экспоненты, то есть **минимизации** выражения:

$$(y - Xc)^T(y - Xc) = S^2(c)$$

> Это и есть квадратическая ошибка! Минимум достигается на ОНК.

Вывод: При добавлении предположения о нормальности оценка наименьших квадратов **совпадает** с оценкой максимального правдоподобия:

$$\hat{c}_{\text{онк}} = \hat{c}_{\text{ммп}}$$

Следствие: Ранее было показано, что оценка максимального правдоподобия эффективна в классе **всех** несмещённых оценок (а не только линейных). То есть:

- В теореме Гаусса—Маркова (минимальные предположения): ОНК эффективна в классе **линейных** несмещённых оценок
- При добавлении нормальности: ОНК эффективна в классе **всех** несмещённых оценок

Теорема о нормальной регрессии

Условия: выполнены все предположения, плюс ошибка распределена нормально.

Утверждения:

1. $\hat{c} \sim \mathcal{N}(c, \sigma^2 A^{-1})$
2. $\dfrac{S^2(\hat{c})}{\sigma^2} \sim \chi^2_{n-m}$ (хи-квадрат с $n-m$ степенями свободы)
3. $\dfrac{S^2(c) - S^2(\hat{c})}{\sigma^2} \sim \chi^2_m$ (хи-квадрат с m степенями свободы)
4. Пары $\hat{c}$ и $S^2(\hat{c})$ — **независимы** (несмотря на то, что $S^2(\hat{c})$ зависит от $\hat{c}$)

   Эту теорему можно воспринимать как переформулировку **теоремы Фишера** (которая использовалась при построении доверительных интервалов для параметров нормального закона).

Доверительный интервал для дисперсии σ^2

Используем **результат 2** теоремы.

Запишем:

$$\mathbb{P}\left(q_{\alpha/2} \leq \frac{S^2(\hat{c})}{\sigma^2} \leq q_{1-\alpha/2}\right) = 1 - \alpha$$

где $q_{\alpha/2}$, $q_{1-\alpha/2}$ — квантили распределения χ^2_{n-m}.

Разрешая неравенство относительно σ^2:

$$\frac{S^2(\hat{c})}{q_{1-\alpha/2}} \leq \sigma^2 \leq \frac{S^2(\hat{c})}{q_{\alpha/2}}$$

Проверка гипотезы о дисперсии

Гипотеза: $H_0 : \sigma^2 = \sigma_0^2$

Статистика критерия:

$$T = \frac{S^2(\hat{c})}{\sigma_0^2}$$

При истинности H_0: $T \sim \chi^2_{n-m}$.

Виды альтернатив и критические области

Альтернатива H_1	Тип критерия	Критическая область
$\sigma^2 \neq \sigma_0^2$	Двусторонний	$[0, q_{\alpha/2}] \cup [q_{1-\alpha/2}, +\infty)$
$\sigma^2 > \sigma_0^2$	Правосторонний	$[q_{1-\alpha}, +\infty)$
$\sigma^2 < \sigma_0^2$	Левосторонний	$[0, q_{\alpha}]$

> **Замечание о терминологии.** Везде в записи используются **квантили**. В практических таблицах часто используются **критические значения**, которые могут обозначаться как Q_α (то, что в записи через квантили является $q_{1-\alpha}$). Важно понимать смысл и не путать.
>
> **Замечание о носителе.** Распределение χ^2 имеет носитель $[0, +\infty)$ (как сумма квадратов), поэтому отрицательных значений быть не может.

Доверительный интервал для коэффициента c_i

Из **результата 1** теоремы:

$$\frac{\hat{c}_i - c_i}{\sqrt{\sigma^2 (A^{-1})_{ii}}} \sim \mathcal{N}(0,1)$$

Но σ^2 неизвестна — оценим её через $\hat{\sigma}^2 = \dfrac{S^2(\hat{c})}{n-m}$.

Подставляя оценку, получаем:

$$\frac{\sqrt{n-m}(\hat{c}_i - c_i)}{\sqrt{S^2(\hat{c}) \cdot (A^{-1})_{ii}}} \sim t_{n-m}$$

Почему распределение Стьюдента?

По формальному определению: $t_k = \dfrac{\xi}{\sqrt{\chi^2_k / k}}$, где $\xi \sim \mathcal{N}(0,1)$.

В числителе у нас стандартная гауссовская величина, а в знаменателе — корень из хи-квадрата, делённого на число степеней свободы. Получается распределение Стьюдента с $n-m$ степенями свободы.

Доверительный интервал

$$c_i \in \hat{c}_i \pm t_{1-\alpha/2,\, n-m} \cdot \sqrt{\frac{S^2(\hat{c}) \cdot (A^{-1})_{ii}}{n-m}}$$

Величина $\sqrt{\dfrac{S^2(\hat{c}) \cdot (A^{-1})_{ii}}{n-m}}$ называется **стандартной ошибкой**.

Используется симметричность распределения Стьюдента относительно нуля.

---

t-тест значимости коэффициента линейной регрессии

Идея: проверить, действительно ли i-я переменная влияет на модель.

Нулевая гипотеза: $H_0 : c_i = 0$ (фактор не влияет)

Альтернативы (зависят от подозрений): - $c_i \neq 0$ (двусторонняя) - $c_i > 0$ (правосторонняя) - $c_i < 0$ (левосторонняя)

Статистика критерия:

$$T = \frac{\sqrt{n-m} \cdot \hat{c}_i}{\sqrt{S^2(\hat{c}) \cdot (A^{-1})_{ii}}}$$

При H_0: $T \sim t_{n-m}$.

Примеры выбора альтернативы

Пример 1. Цена недвижимости в зависимости от расстояния до центра. Подозрение: чем меньше расстояние, тем больше цена → **левосторонняя** альтернатива ($c_i < 0$).

Пример 2. Стоимость авто в зависимости от мощности. Подозрение: чем больше мощность, тем больше цена → **правосторонняя** альтернатива ($c_i > 0$).

Предсказание новых значений

До сих пор имели “тренировочный набор”: $y = Xc + \varepsilon$, по которому оценили c и σ^2.

Теперь — новое наблюдение:

$$y_\nu = x_\nu c + \varepsilon_\nu$$

где: - x_ν — новая строка наблюдений - $\varepsilon_\nu \sim \mathcal{N}(0, \sigma^2)$ - ε_ν и ε **независимы**

На уровне модели: $y_\nu \sim \mathcal{N}(x_\nu c, \sigma^2)$.

Оценка нового значения

$$\hat{y}_\nu = x_\nu \hat{c}$$

Распределение $\hat{y}_\nu$:

$$\hat{y}_\nu \sim \mathcal{N}(x_\nu c, \sigma^2 x_\nu A^{-1} x_\nu^T)$$

Здесь это **дисперсия** (число), а не матрица, потому что x_ν — строка.

Независимость $\hat{y}_\nu$ и y_ν

- $\hat{c}$ — функция от старого y, который функция от старого ε
- y_ν — функция от нового ε_ν
- Старый и новый ε независимы → $\hat{y}_\nu$ и y_ν независимы

Распределение разности

$$\hat{y}_\nu - y_\nu \sim \mathcal{N}\left(0,\, \sigma^2(1 + x_\nu A^{-1} x_\nu^T)\right)$$

(дисперсии складываются при независимости)

Стандартизация:

$$\frac{\hat{y}_\nu - y_\nu}{\sqrt{\sigma^2(1 + x_\nu A^{-1} x_\nu^T)}} \sim \mathcal{N}(0, 1)$$

Заменяя σ^2 на оценку:

$$\frac{\sqrt{n-m}(\hat{y}_\nu - y_\nu)}{\sqrt{S^2(\hat{c})(1 + x_\nu A^{-1} x_\nu^T)}} \sim t_{n-m}$$

Отсюда стандартным образом строится доверительный интервал для y_ν (зажимаем между квантилями и разрешаем неравенство).

Условные оценки наименьших квадратов

Понадобятся для описания F-критерия.

Постановка: вектор c удовлетворяет линейным ограничениям:

$$Tc = t_0$$

где: - T — матрица $k \times m$, $k \leq m$ - $\text{rank}(T) = k$ (ограничения линейно независимы)

Определение условной ОНК:

$$\hat{c}_T = \arg \min_{Tc = t_0} S^2(c)$$

Это задача оптимизации квадратичной функции при линейных ограничениях.

Аналитическая формула

$$\hat{c}_T = \hat{c} - A^{-1}T^T D^{-1}(T\hat{c} - t_0)$$

где:

$$D = TA^{-1}T^T$$

(матрица D возникала в теореме Гаусса—Маркова).

> Матрица D симметрична: $D^T = D$, поэтому $(D^{-1})^T = D^{-1}$. D^{-1} существует, потому что $\text{rank}(T) = k$.

Идея вывода

Аналогично доказательству обычной ОНК. Показывается:

$$S^2(\hat{c}_T + h) > S^2(\hat{c}_T)$$

для любого $h \neq 0$ такого, что $Th = 0$ (приращение в допустимом направлении).

Упражнение: показать, что $T\hat{c}_T = t_0$ (выполняется в одну строчку).

Ключевое наблюдение

Из результата прошлой лекции:

$$S^2(c) - S^2(\hat{c}) = (c - \hat{c})^T A(c - \hat{c})$$

Подставляя $c = \hat{c}_T$:

$$S^2(\hat{c}_T) - S^2(\hat{c}) = (\hat{c}_T - \hat{c})^T A (\hat{c}_T - \hat{c})$$

Используя формулу для $\hat{c}_T - \hat{c} = -A^{-1}T^T D^{-1}(T\hat{c} - t_0)$:

$$S^2(\hat{c}_T) - S^2(\hat{c}) = (T\hat{c} - t_0)^T D^{-1} \underbrace{T A^{-1} T^T}_{=D} D^{-1}(T\hat{c} - t_0)$$

$$= (T\hat{c} - t_0)^T D^{-1}(T\hat{c} - t_0)$$

Это **квадратичная форма** от $\hat{c}$. Так как $\hat{c}$ имеет нормальное распределение, и квадратичная форма построена с матрицей ранга k, эта величина связана с распределением χ^2_k **— число степеней свободы равно** k.

F-критерий для линейной модели

Общая формулировка

Гипотезы: - $H_0 : Tc = t_0$ - $H_1 : Tc \neq t_0$

Статистика критерия:

$$F = \frac{[S^2(\hat{c}_T) - S^2(\hat{c})]/k}{S^2(\hat{c})/(n-m)}$$

При истинности H_0: $F \sim F_{k,\, n-m}$ (распределение Фишера).

Обоснование правосторонней критической области

Знаменатель:

$$\mathbb{E}\left[\frac{S^2(\hat{c})}{n-m}\right] = \sigma^2$$

— всегда, независимо от истинности H_0.

Числитель: математическое ожидание разности квадратичных ошибок.

$$\mathbb{E}\left[\frac{1}{k}(S^2(\hat{c}_T) - S^2(\hat{c}))\right] = \frac{1}{k}\mathbb{E}\left[(T\hat{c} - t_0)^T D^{-1}(T\hat{c} - t_0)\right]$$

Расписываем как сумму:

$$= \frac{1}{k}\sum_{i,j}(D^{-1})_{ij}\,\mathbb{E}\left[(T\hat{c} - t_0)_i(T\hat{c} - t_0)_j\right]$$

Используя $\mathbb{E}[XY] = \mathrm{Cov}(X, Y) + \mathbb{E}[X]\mathbb{E}[Y]$, получаем две части:

Часть 1 (с ковариациями):

$$\frac{1}{k}\sum_{i,j}(D^{-1})_{ij}\,\mathrm{Cov}((T\hat{c} - t_0)_i, (T\hat{c} - t_0)_j)$$

Так как t_0 — константа, ковариация определяется только $T\hat{c}$. Матрица ковариации $T\hat{c}$:

$$\mathrm{Cov}(T\hat{c}) = T \cdot \mathrm{Cov}(\hat{c}) \cdot T^T = T \cdot \sigma^2 A^{-1} \cdot T^T = \sigma^2 D$$

Подставляя:

$$\frac{1}{k}\sum_{i,j}(D^{-1})_{ij} \cdot \sigma^2 D_{ij} = \frac{\sigma^2}{k} \cdot \mathrm{tr}(D^{-1}D) = \frac{\sigma^2}{k} \cdot k = \sigma^2$$

Здесь использовано: $\sum_{i,j}(D^{-1})_{ij}D_{ij} = \mathrm{tr}(D^{-1}D) = \mathrm{tr}(E_k) = k$.

Часть 2 (с произведениями матожиданий):

Используя $\mathbb{E}[\hat{c}] = c$:

$$\frac{1}{k}(Tc - t_0)^T D^{-1}(Tc - t_0)$$

Итого:

$$\mathbb{E}\left[\frac{1}{k}(S^2(\hat{c}_T) - S^2(\hat{c}))\right] = \sigma^2 + \frac{1}{k}(Tc - t_0)^T D^{-1}(Tc - t_0)$$

Анализ:

- Если H_0 верна ($Tc = t_0$): математическое ожидание числителя равно σ^2
- Если H_0 не верна: математическое ожидание числителя **строго больше** σ^2

Знаменатель всегда в среднем равен σ^2. Поэтому:

- При H_0: $\mathbb{E}[F] \approx 1$
- При H_1: $\mathbb{E}[F] > 1$

Критическая область — правосторонняя.

F-критерий “по умолчанию” (значимость модели в целом)

Стандартная модель

$$y_i = c_0 + x_{i1}c_1 + x_{i2}c_2 + ... + x_{im}c_m + \varepsilon_i$$

В библиотеках c_0 (свободный член) обычно выделяется отдельно.

Гипотеза по умолчанию

Нулевая гипотеза: все коэффициенты, кроме свободного, равны нулю:

$$H_0 : c_1 = c_2 = ... = c_m = 0$$

Альтернатива: H_1: хотя бы один $c_i \neq 0$ (то есть $\neg H_0$).

Это **проверка значимости модели в целом**.

Коэффициент детерминации R^2

Определение. Множественный коэффициент корреляции — это коэффициент корреляции между y и $\hat{y}$:

$$R = \text{corr}(y, \hat{y})$$

где $\hat{y} = X\hat{c}$.

Коэффициент детерминации:

$$R^2 = R^2(\text{множественный})$$

Связь с остаточной дисперсией

Имеет место соотношение:

$$S^2(\hat{c}) = (1 - R^2) \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Правая сумма — константа, зависящая от датасета.

Интерпретация

R^2	Остаточная дисперсия	Качество модели
Близко к 1	Маленькая	Модель адекватная
Близко к 0	Большая	Модель не очень адекватная

F-статистика через R^2

Для гипотезы по умолчанию F-статистика выражается через коэффициент детерминации:

$$F = \frac{R^2/m}{(1 - R^2)/(n - m - 1)}$$

Замечание Ивана Александровича: возможны небольшие неточности в коэффициентах — нужно перепроверить.

Что планируется на следующей лекции

- Модель **однофакторного дисперсионного анализа**
- Обобщения линейных моделей
- Как **проверять исходные предположения**
- Что делать, если матрица A **необратима** или плохо обратима

Лекция 13. Линейные модели. Однофакторный дисперсионный анализ, метод главных компонент, взвешенный МНК

Организационная часть

- На прошлой лекции была допущена опечатка в формуле статистики, выражающейся через коэффициент детерминации R^2.
- Для рассматриваемой модели верная формула:

$$F = \frac{n-m}{m-1} \cdot \frac{R^2}{1-R^2}$$

- Ранее коэффициент R^2 находили вручную для простого случая (двух переменных), затем переходили к более общим случаям.

---

1. Однофакторный дисперсионный анализ (One-way ANOVA)

1.1. Постановка задачи

В классической линейной регрессии переменные были **количественными**. Однако часто встречаются ситуации, когда переменная **факторная** (категориальная).

Категориальная переменная — переменная, которая может принимать конечное число различных значений (уровней фактора).

Примеры: - Буква в паспорте (М или Ж) - Номер курса (1, 2, 3, 4) - Специальность

Для каждого значения категориальной переменной у нас есть некий набор наблюдений.

1.2. Формальная модель

Рассмотрим модель вида:

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

где: - ε_{ij} — ошибки, $\varepsilon_{ij} \sim N(0, \sigma^2)$, независимы - $i = 1, \dots, I$ — уровни фактора

(значения категориальной переменной) - $j = 1, \dots, J_i$ — индекс наблюдения внутри группы i - J_i — количество наблюдений на уровне i - μ_i — среднее влияние фактора на уровне i

Объём выборки:

$$n = J_1 + J_2 + \dots + J_I$$

Замечание: для каждой группы количество наблюдений может быть разным.

1.3. Гипотезы

Хотим проверить, влияет ли фактор на показатель.

Нулевая гипотеза (фактор не влияет):

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_I$$

То есть среднее влияние не зависит от значения категориальной переменной.

Альтернативная гипотеза:

$$H_1 : \neg H_0$$

(существуют i, k такие, что $\mu_i \neq \mu_k$)

1.4. Кодирование категориальной переменной

Один из самых простых способов кодирования — сопоставление значениям категориальной переменной векторов из нулей и единиц (one-hot encoding):

$$1 \to (1, 0, 0, \dots, 0)$$

$$2 \to (0, 1, 0, \dots, 0)$$

$$\vdots$$

$$I \to (0, 0, \dots, 0, 1)$$

После такого кодирования модель сводится к общей модели линейной регрессии:

$$y = X\beta + \varepsilon$$

И гипотезы такого типа можно проверять с помощью **F-теста**.

1.5. F-статистика в общем виде

$$F = \frac{\hat{S}^2|_{H_0} - \hat{S}^2}{\hat{S}^2} \cdot \frac{n-m}{k}$$

где k — количество степеней свободы.

> Можно вывести из общего вида линейной регрессии, однако вывод получится громоздким. Решим задачу иначе — будет понятно, почему фигурирует слово «дисперсионный».

1.6. Вывод F-статистики через дисперсии

Разминка: безусловный минимум Рассмотрим:

$$S^2(\mu) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \mu_i)^2$$

Минимизируем эту функцию. Сумма разбивается на внешнюю (по i) и внутреннюю (по j). При различных i переменные μ_i независимы, поэтому каждую локальную сумму можно оптимизировать отдельно:

$$\sum_{j=1}^{J_1} (y_{1j} - \mu_1)^2, \quad \sum_{j=1}^{J_2} (y_{2j} - \mu_2)^2, \quad \ldots$$

Где достигается минимум?

Выражение $\sum_j (y_j - \mu)^2$ — это с точностью до множителя матожидание квадрата разности. Минимум достигается при **среднем**:

$$\hat{\mu}_i = \frac{1}{J_i} \sum_{j=1}^{J_i} y_{ij} = \bar{y}_{i*}$$

Если бы стоял модуль вместо квадрата, ответом была бы медиана.

Внутригрупповая дисперсия Подставив $\widehat{\mu}_i$ в S^2, получаем:

$$S^2_W = S^2(\widehat{\mu}) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{i*})^2$$

Это **внутригрупповая дисперсия** (within-group variance) — индекс W от *within*.

Степени свободы (безусловный случай) Общая формула:

$$df_W = n - I$$

(от количества наблюдений отнимаем количество групп)

Нередко все группы одинакового размера: $J_1 = J_2 = ... = J_I = J$. Тогда:

$$df_W = I \cdot J - I = I(J-1)$$

Минимум при условии H_0 При истинности H_0 все μ_i равны общему μ:

$$S^2(\mu) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \mu)^2$$

Минимум достигается при общем выборочном среднем:

$$\widehat{\mu}|_{H_0} = \bar{y}$$

Степени свободы при H_0 Количество линейно независимых ограничений в гипотезе

$$\mu_1 = \mu_2 = ... = \mu_I$$

равно $I - 1$ (а не общее число попарных уравнений, поскольку из $\mu_1 = \mu_2$ и $\mu_1 = \mu_3$ следует $\mu_2 = \mu_3$).

1.7. Разложение дисперсии

Введём обозначения: - S^2 — общая дисперсия (когда подставлено $\bar{y}$) - S_W^2 — внутригрупповая дисперсия - S_B^2 — межгрупповая дисперсия (between):

$$S_B^2 = \sum_{i=1}^{I} J_i(\bar{y}_{i*} - \bar{y})^2$$

Разложение:

$$S^2 = S_W^2 + S_B^2$$

Идейное обоснование (через формулу полного матожидания) Из теории вероятностей — формула разложения дисперсии:

$$D(Y) = E[D(Y|X)] + D[E(Y|X)]$$

- S_W^2 соответствует $E[D(Y|X)]$ — дисперсия внутри групп, усреднённая;
- S_B^2 соответствует $D[E(Y|X)]$ — разброс групповых средних относительно общего.

Действительно: - $D(Y|X)$ при фиксированной группе — внутренняя сумма в S_W^2; - внешняя сумма по i — это операция взятия матожидания; - $E(Y|X = i) = \bar{y}_{i*}$, а матожидание этой величины равно $\bar{y}$.

1.8. Итоговая F-статистика

$$F = \frac{S_B^2/(I-1)}{S_W^2/(n-I)}$$

При условии истинности H_0:

$$F \sim F(I-1,\, n-I)$$

(распределение Фишера со степенями свободы $I-1$ и $n-I$).

Тест правосторонний — это частный случай общего F-критерия, рассмотренного на прошлой лекции.

1.9. Обобщения

- **Two-way ANOVA** — две факторные переменные;
- **Многофакторный дисперсионный анализ** — больше двух факторов;
- **ANCOVA (ковариационный анализ)** — есть и числовые, и категориальные переменные.

  Соответствующие соотношения становятся гораздо более громоздкими; реализованы в стат-пакетах.

---

2. Метод главных компонент (PCA)

2.1. Мотивация

В модели линейной регрессии предполагалось, что $X^T X$ **не вырождена** (отсутствие мультиколлинеарности — линейная независимость переменных).

Две проблемы:

1. Чисто гипотетически матрица может оказаться вырожденной — переменные линейно зависимы. Одна выражается через другие.
2. При вычислении оценки наименьших квадратов нужно обращать $X^T X$. С численной точки зрения существуют показатели, от которых зависит скорость сходимости алгоритмов обращения. Может оказаться, что матрица обратима формально, но обращается очень медленно — это означает, что переменные **почти линейно зависимы**.

   Подробнее об этом — в курсе численных методов на 3-м курсе.

2.2. Идея метода

Обозначим $A = X^T X$. Эта матрица: - напоминает ковариационную матрицу; - **неотрицательно определена**.

Из неотрицательной определённости: - собственные числа неотрицательны: $\lambda_i \geq 0$; - собственные векторы можно выбрать **ортонормированными** (n линейно независимых).

2.3. Спектральное разложение

$$\Lambda = U^T A U$$

где: - $\Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$ — диагональная матрица собственных чисел, отсортированных по убыванию ($\lambda_1 \geq \lambda_2 \geq ... \geq 0$); - $U = [u_1\, u_2\, ...\, u_n]$ — ортонормированные собственные векторы.

Поскольку U ортогональная, $U^{-1} = U^T$, поэтому:

$$A = U \Lambda U^T$$

2.4. Введение новых переменных

Вводим новые переменные:

$$W = XU$$

Посчитаем $W^T W$:

$$W^T W = (XU)^T (XU) = U^T X^T X U = U^T A U = U^T (U \Lambda U^T) U = (U^T U) \Lambda (U^T U) = \Lambda$$

(использовали $U^T U = I$, поскольку U ортогональна).

Получили:

$$W^T W = \Lambda$$

— это **диагональная матрица**.

2.5. Свойства новых переменных

Поскольку $W^T W$ диагональна:

- **новые переменные линейно независимы (некоррелированы)** — недиагональные элементы нулевые;

- **новые переменные отсортированы по убыванию дисперсии** — на диагонали матрицы ковариаций стоят дисперсии.

В матричной форме:

$$\begin{pmatrix} w_1 & w_2 & ... & w_m \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & ... & x_n \end{pmatrix} \cdot U$$

Каждый w_k получается умножением старых переменных на k-й столбец матрицы U: - w_1 — самая большая дисперсия; - w_2 — поменьше; - и т.д.

2.6. Снижение размерности

Идея: дисперсия — мера разброса от матожидания. Если у переменной маленькая дисперсия, она почти не изменяется, фактически ведёт себя как константа — особой роли в модели не играет.

Алгоритм: оставляем только переменные с большой дисперсией; остальные отбрасываем.

Критерии остановки

1. **Порог по отдельной дисперсии:** оставляем w_i, для которых $\lambda_i > \tau$ (порог).
2. **Кумулятивный порог:** суммируем λ_i по убыванию, пока сумма не достигнет заданного порога.

2.7. Что получили

С помощью PCA борются с: - **некоррелированностью** (точнее, с проблемой мультиколлинеарности — делаем переменные некоррелированными); - **большим количеством переменных** — снижаем размерность.

---

3. Взвешенный метод наименьших квадратов (взвешенный МНК)

3.1. Мотивация

Ранее предполагали, что **дисперсии у ошибок одинаковые** (гомоскедастичность). Однако дисперсии могут различаться — модель **гетероскедастичная**.

Пусть теперь матрица ковариаций для ε — диагональная матрица с разными элементами:

$$\mathrm{Cov}(\varepsilon) = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)$$

Обычная оценка наименьших квадратов уже не будет оптимальной.

3.2. Новая функция ошибок

Введём взвешенную сумму квадратов:

$$S^2(c) = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(\sum_{j=1}^{m} x_{ij} c_j - y_i \right)^2$$

(каждое слагаемое умножается на $1/\sigma_i^2$ — учитываем веса).

3.3. Поиск оптимального c

Дифференцируем по c_k:

$$\frac{\partial S^2(c)}{\partial c_k} = \sum_{i=1}^{n} \frac{2}{\sigma_i^2} \left(\sum_{j=1}^{m} x_{ij} c_j - y_i \right) x_{ik}$$

Приравниваем к нулю:

$$\sum_{i=1}^{n} \frac{1}{\sigma_i^2} x_{ik} \sum_{j=1}^{m} x_{ij} c_j = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} x_{ik} y_i$$

3.4. Матричная запись

Введём диагональную матрицу весов:

$$W = \mathrm{diag} \left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \dots, \frac{1}{\sigma_n^2} \right)$$

Полученные уравнения переписываются в матричном виде:

$$X^T W X \cdot c = X^T W \cdot y$$

Это упражнение на вспоминание определения умножения матриц.

Отсюда:

$$\boxed{\hat{c} = (X^T W X)^{-1} X^T W y}$$

3.5. Свойства оценки

- Является **наилучшей несмещённой линейной оценкой** (доказывается аналогично теореме Гаусса—Маркова).
- При одинаковых дисперсиях ($\sigma_i^2 = \sigma^2$) формула превращается в обычную оценку МНК: $W = \frac{1}{\sigma^2} I$, множители σ^2 и σ^{-2} сокращаются:

$$\hat{c} = (X^T X)^{-1} X^T y$$

3.6. Что делать с неизвестными дисперсиями?

В формуле использованы дисперсии σ_i^2, но они **неизвестны**. Вместо них можно подставить их **оценки** — при некоторых условиях формулы будут корректно работать.

> Подробнее об этом сейчас не углубляемся.

3.7. Когда применять

Взвешенный МНК работает, когда модель **не является гомоскедастичной**.

---

4. Замечание о проверке предположений модели

В стандартной линейной регрессии предполагали: - ошибки распределены нормально; - ошибки не коррелированы; - гомоскедастичность (одинаковые дисперсии).

Для проверки этих предположений существуют **специально предназначенные стат-тесты**: - тесты на гомоскедастичность; - тесты на отсутствие корреляции ошибок; - тесты на нормальность распределения.

> В рамках курса подробно не разбираем; кому интересно — можно изучить самостоятельно.

---

5. Анонс следующих лекций

- **Тест отношения правдоподобия** (идейно новая вещь)
- **Различные модификации линейных моделей**

Лекция 14: Обобщённые линейные модели и критерий отношения правдоподобия

Введение

В предыдущих линейных моделях, которые рассматривались, выходная переменная была **количественной**. Однако выходная переменная не всегда количественная — она может быть и **категориальной**, в частности, бинарной.

Сегодня рассматриваются: - Один из простейших алгоритмов **бинарной классификации** — логистическая регрессия (с точки зрения статистики) - Регрессия Пуассона - Критерий отношения правдоподобия (простой и общий случай) - Лемма Неймана-Пирсона - Проверка значимости моделей

---

1. Логистическая регрессия

1.1. Постановка задачи

Вход: - X — матрица переменных (как и ранее) - y — вектор наблюдений зависимой переменной

Особенность: y_i принимает только два значения: 0 или 1.

То есть, если до сей поры y_i мог принимать любое количество значений, то теперь $y_i \in \{0, 1\}$ — вектор-столбец, в каждой компоненте которого записано 0 или 1.

1.2. Сигмоида

Для построения модели предлагается рассмотреть функцию:

$$f(t) = \frac{1}{1 + e^{-t}}$$

Эта функция называется **сигмоидой**.

Свойства графика сигмоиды: - При $t \to +\infty$: $f(t) \to 1$ - При $t \to -\infty$: $f(t) \to 0$ - В нуле: $f(0) = \frac{1}{2}$

1.3. Модель логистической регрессии

Будем воспринимать y_i как реализацию **бернуллиевской случайной величины**:

$$y_i \sim \text{Bern}\left(f(c_0 + c_1 x_{i1} + c_2 x_{i2} + ... + c_m x_{im})\right)$$

То есть y_i — это бернуллиевская случайная величина с параметром успеха (вероятностью единички), равным сигмоиде от линейной функции.

Цель: оценить коэффициенты $c_0, c_1, c_2, \dots, c_m$.

1.4. Оценка параметров методом максимального правдоподобия

Найдём точечную оценку параметров c_i с помощью **метода максимального правдоподобия (ММП)**.

Предполагаем, что наблюдения **независимы**. Тогда функция правдоподобия:

$$L = \prod_{i=1}^{n} \left(\frac{1}{1 + e^{-(c_0 + c_1 x_{i1} + ... + c_m x_{im})}} \right)^{y_i} \cdot \left(1 - \frac{1}{1 + e^{-(c_0 + c_1 x_{i1} + ... + c_m x_{im})}} \right)^{1 - y_i}$$

Берём **минус логарифм функции правдоподобия**:

$$-\ln L = -\sum_{i=1}^{n} \left[y_i \ln f(c_0 + c_1 x_{i1} + ... + c_m x_{im}) + (1 - y_i) \ln \left(1 - f(c_0 + c_1 x_{i1} + ... + c_m x_{im})\right) \right]$$

Для нахождения оценки максимального правдоподобия нужно **максимизировать** функцию правдоподобия, то есть **минимизировать** минус логарифм.

> **Важно:** аналитического решения здесь нет. Оптимизация проводится **численными методами** (например, градиентным спуском).

1.5. Связь с машинным обучением

Возможно, в других курсах при рассмотрении простейших алгоритмов бинарной классификации упомянут модель логистической регрессии и связанную с ней **функцию потерь (loss function)**.

> Эта стандартная функция потерь есть не что иное, как **минус логарифм функции правдоподобия** для модели логистической регрессии.

1.6. Свойства оценок

Поскольку оценки получены ММП, они являются **асимптотически нормальными**:

$$\hat{c}_i \text{ — асимптотически нормальные оценки}$$

Это позволяет: - Строить **доверительные интервалы** для c_i - **Проверять гипотезы** о значениях c_i

> Проверка значимости всей модели обсуждается позже.

---

2. Регрессия Пуассона

2.1. Постановка задачи

Вход: такой же, как раньше, но y_i — **категориальная переменная**, принимающая значения $0, 1, 2, \ldots$

2.2. Модель

Случайную величину y_i воспринимаем как **пуассоновскую** с параметром:

$$y_i \sim \text{Pois}\left(e^{c_0 + c_1 x_{i1} + \ldots + c_m x_{im}}\right)$$

2.3. Обобщённые линейные модели

> До этого рассматривались **линейные модели**: $y = Xc + \varepsilon$.

Здесь рассматриваются так называемые **обобщённые линейные модели**: - В предыдущем случае: бернуллиевская величина от сигмоиды от линейной функции - В пуассоновской регрессии: пуассоновская величина (от экспоненты) от линейной функции

2.4. Оценка параметров

Коэффициенты c_i оцениваются точно так же — методом максимального правдоподобия: 1. Записываем логарифм функции правдоподобия 2. Находим максимальное значение

Аналитического решения нет — только **численное решение**.

Оценки $\widehat{c}_i$ являются **асимптотически нормальными**, поэтому: - Можно строить доверительные интервалы - Можно проверять гипотезы

---

3. Критерий отношения правдоподобия (простой случай)

3.1. Простые гипотезы

Рассмотрим случай **простых гипотез**:

$$H_0 : f = f_0 \quad \text{vs} \quad H_1 : f = f_1$$

где f_0 и f_1 — непрерывны.

Что такое f? У нас есть простая выборка из какого-то распределения, и мы хотим проверить эти две гипотезы.

3.2. Статистика отношения правдоподобия

Введём функцию:

$$L(X) = \frac{L(X \mid H_1)}{L(X \mid H_0)}$$

— правдоподобие при первой гипотезе делим на правдоподобие при нулевой.

Решающее правило:

$$\text{если } L(X) \geq C, \text{ то принимаем } H_1, \text{ иначе } H_0$$

Логика: если $L(X) \geq C$, значит правдоподобие при H_1 больше, чем при H_0, поэтому принимаем H_1.

Вопрос: из каких соображений выбирать пороговую константу C?

3.3. Выбор константы C

Рассмотрим функцию:

$$\psi(C) = P(L(X) \geq C \mid H_0)$$

Какое событие она описывает?

Когда $L(X) \geq C$, мы принимаем H_1. При условии истинности H_0 это **ошибка первого рода**.

То есть $\psi(C)$ — это **вероятность ошибки первого рода**.

Свойства: - $\psi(0) = 1$

Рассмотрим ещё:

$$P(L(X) \geq C \mid H_1) \leq 1$$

С другой стороны, это интеграл от плотности:

$$P(L(X) \geq C \mid H_1) = \int_{\{x : L(x) \geq C\}} L(x \mid H_1)\, dx$$

По условию $L(x \mid H_1) \geq C \cdot L(x \mid H_0)$, поэтому:

$$\int_{\{L(x) \geq C\}} L(x \mid H_1)\, dx \geq \int_{\{L(x) \geq C\}} C \cdot L(x \mid H_0)\, dx = C \cdot \psi(C)$$

Откуда:

$$\psi(C) \leq \frac{1}{C}$$

В частности, $\psi(C) \xrightarrow{C\to\infty} 0$.

3.4. Подбор C под заданный уровень значимости

Предположение: функция $\psi(C)$ непрерывна.

Тогда для любого $\alpha \in (0,1)$ существует C_α такое, что:

$$\psi(C_\alpha) = \alpha$$

Таким образом, можно подобрать порог C так, чтобы вероятность ошибки первого рода в точности равнялась α.

---

4. Лемма Неймана-Пирсона

4.1. Формулировка

> **Лемма Неймана-Пирсона.** Пусть выполнено условие выше. Тогда критерий отношения правдоподобия является **оптимальным**, то есть имеет **минимальную вероятность ошибки второго рода** среди всех тестов, которые проверяют данные гипотезы и имеют вероятность ошибки первого рода α.
>
> Хотя это и лемма, на самом деле это **фундаментальное утверждение**.

4.2. Доказательство

Обозначения: рассмотрим другой тест с вероятностью ошибки первого рода α. Пусть: - g — статистика этого теста - $T_0(\alpha)$ — область принятия H_0 - $T_1(\alpha)$ — критическая область

Выкладка №1 Рассмотрим вероятность:

$$P\Big(\{L(X) \geq C_\alpha\} \setminus \{L(X) \geq C_\alpha \text{ и } g(X) \in T_1(\alpha)\} \mid H_0\Big)$$

Обозначим $B = \{L(X) \geq C_\alpha \text{ и } g(X) \in T_1(\alpha)\}$.

Тогда:

$$= P(L(X) \geq C_\alpha \mid H_0) - P(B \mid H_0) = \alpha - P(B \mid H_0)$$

(по построению C_α).

С другой стороны, у второго теста вероятность ошибки первого рода тоже α:

$$\alpha = P(g(X) \in T_1(\alpha) \mid H_0) - P(B \mid H_0)$$

Итого получаем:

$$P(\{L \geq C_\alpha\} \setminus B \mid H_0) = P(\{g(X) \in T_1(\alpha)\} \setminus B \mid H_0)$$

Выкладка №2 Теперь посчитаем при условии H_1:

$$P(\{L(X) \geq C_\alpha\} \setminus B \mid H_1) = \int_{\{L \geq C_\alpha\} \setminus B} L(x \mid H_1)\, dx$$

Для этого множества выполнено $L(x \mid H_1) \geq C_\alpha \cdot L(x \mid H_0)$, поэтому:

$$\geq C_\alpha \cdot \int_{\{L \geq C_\alpha\} \setminus B} L(x \mid H_0)\, dx = C_\alpha \cdot P(\{L \geq C_\alpha\} \setminus B \mid H_0)$$

Используя выкладку №1:

$$= C_\alpha \cdot P(\{g(X) \in T_1(\alpha)\} \setminus B \mid H_0)$$

Анализ события $\{g(X) \in T_1(\alpha)\} \setminus B$ Для x из этого множества $g(X) \in T_1(\alpha)$, но не выполнено $L(X) \geq C_\alpha$, т.е. $L(X) < C_\alpha$.

Это означает: $L(x \mid H_0) > \frac{1}{C_\alpha} L(x \mid H_1)$.

Воспользуемся этим неравенством:

$$C_\alpha \cdot \int_{\{g \in T_1(\alpha)\} \setminus B} L(x \mid H_0)\, dx > C_\alpha \cdot \frac{1}{C_\alpha} \int_{\{g \in T_1(\alpha)\} \setminus B} L(x \mid H_1)\, dx$$

$$= P(\{g(X) \in T_1(\alpha)\} \setminus B \mid H_1)$$

Завершение доказательства Прибавим к обеим частям $P(B \mid H_1)$:

$$P(L(X) \geq C_\alpha \mid H_1) > P(g(X) \in T_1(\alpha) \mid H_1)$$

То есть **мощность критерия отношения правдоподобия больше мощности любого другого теста**. А мощность — это $1-$ вероятность ошибки второго рода. ■

---

5. Пример: проверка простых гипотез о среднем нормального закона

5.1. Постановка

- H_0: выборка из $\mathcal{N}(0, 1)$
- H_1: выборка из $\mathcal{N}(1, 1)$

5.2. Вычисление отношения правдоподобия

$$L(X) = \frac{L(X \mid H_1)}{L(X \mid H_0)} = \frac{\frac{1}{(\sqrt{2\pi})^n} \prod_{i=1}^n e^{-\frac{1}{2}(x_i-1)^2}}{\frac{1}{(\sqrt{2\pi})^n} \prod_{i=1}^n e^{-\frac{1}{2}x_i^2}}$$

После сокращения:

$$L(X) = \prod_{i=1}^n e^{-\frac{1}{2}[(x_i-1)^2 - x_i^2]} = \prod_{i=1}^n e^{-\frac{1}{2}(-2x_i+1)}$$

5.3. Преобразование неравенства $L(X) \geq C$

Берём логарифм:

$$\sum_{i=1}^n \left(-\frac{1}{2}\right)(-2x_i + 1) \geq \ln C$$

Обозначим $\tilde{C} = \ln C$:

$$\sum_{i=1}^{n} x_i - \frac{n}{2} \geq \tilde{C}$$

$$\sum_{i=1}^{n} x_i \geq \tilde{C} + \frac{n}{2}$$

5.4. Распределение тестовой статистики

При условии H_0:

$$\sum_{i=1}^{n} x_i \sim \mathcal{N}(0, n)$$

При условии H_1:

$$\sum_{i=1}^{n} x_i \sim \mathcal{N}(n, n)$$

5.5. Условие на вероятность ошибки первого рода

$$P\left(\sum x_i \geq \tilde{C} + \frac{n}{2} \mid H_0\right) = 1 - \Phi\left(\frac{\tilde{C} + n/2}{\sqrt{n}}\right) = \alpha$$

где Φ — функция распределения стандартного нормального закона.

Дальше остаётся разрешить уравнение относительно $\tilde{C}$.

5.6. Геометрическая интерпретация

Имеем две гауссианы: 1. **Гауссиана №1** — плотность $\mathcal{N}(0, n)$ (при H_0) 2. **Гауссиана №2** — плотность $\mathcal{N}(n, n)$ (при H_1)

Отметим на оси константу (обозначим её через две волны $\tilde{\tilde{C}}$).

Вероятность ошибки первого рода (α): - Это ситуация: опровергаем H_0, но она верна - На графике: площадь под первой гауссианой **справа** от черты

Вероятность ошибки второго рода (β): - Это ситуация: принимаем H_0, но верна H_1 - $P(\sum x_i < \tilde{\tilde{C}} \mid H_1)$ - На графике: площадь под второй гауссианой **слева** от черты

5.7. Анализ trade-off

Если параметр $\tilde{\tilde{C}}$ варьировать: - **Двигаем вправо:** α уменьшается, β увеличивается - **Двигаем влево:** α увеличивается, β уменьшается

Лемма Неймана-Пирсона утверждает: если в критерии отношения правдоподобия для простых гипотез подобрать константу так, чтобы α в точности равнялась заданной величине, то этот критерий **оптимален** в плане минимизации β.

---

6. Общий критерий отношения правдоподобия

6.1. Постановка (сложные параметрические гипотезы)

Пусть имеется параметрическая гипотеза:

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta \setminus \Theta_0$$

То есть Θ_0 — некоторое подмножество параметров, а альтернатива — его дополнение.

6.2. Статистика

$$\Lambda_n = \frac{\sup_{\theta \in \Theta_0} L(X, \theta)}{\sup_{\theta \in \Theta} L(X, \theta)}$$

Здесь, в отличие от случая простых гипотез, нужна **оптимизация**: - **Числитель:** условная оптимизация ($\theta \in \Theta_0$) - **Знаменатель:** безусловная оптимизация по всему Θ

6.3. Асимптотическое распределение

Предположение: оценки максимального правдоподобия асимптотически нормальные (это выполняется в рамках условий регулярности).

Тогда:

$$-2 \ln \Lambda_n \xrightarrow{d} \chi^2_{m-r}$$

где: - m — размерность Θ (всего пространства параметров) - r — размерность Θ_0

6.4. Почему именно $-2\ln\Lambda_n$? (объяснение «на пальцах»)

- **Минус:** в классическом критерии отношения правдоподобия было «наоборот» — наверху правдоподобие при H_1, внизу при H_0. Минус условно «переворачивает» дробь.
- **Логарифм:** упрощает работу с произведениями
- **Двойка:** $2\ln x = \ln x^2$, а логарифм произведения это сумма. Получается похоже на сумму квадратов — отсюда и χ^2-распределение.

---

7. Применение: проверка значимости логистической регрессии

7.1. Постановка

Возвращаемся к модели логистической регрессии:

$$y_i \sim \text{Bern}\left(\frac{1}{1+e^{-(c_0+c_1x_{i1}+\ldots+c_mx_{im})}}\right)$$

Что значит проверить значимость модели? Хотим выяснить, действительно ли переменные x влияют на y.

7.2. Гипотезы

Нулевая гипотеза (по умолчанию: переменные не влияют):

$$H_0 : c_1 = c_2 = \ldots = c_m = 0$$

В этом случае остаётся только свободный коэффициент c_0, поэтому **размерность Θ_0 равна 1**.

Альтернативная гипотеза (формально):

$$H_1 : \exists k \text{ такое, что } c_k \neq 0$$

7.3. Размерности

- $\dim\Theta = m+1$ (модель описывается $m+1$ параметром: $c_0, c_1, \ldots, c_m$)
- $\dim\Theta_0 = 1$ (остался только c_0)

7.4. Отношение правдоподобия

$$\Lambda_n = \frac{\sup_{\theta \in \Theta_0} L(X, \theta)}{\sup_{\theta \in \Theta} L(X, \theta)}$$

По сути спрашиваем: действительно ли наша выборка — просто бернуллиевские величины, или же зависит от X?

> **Замечание про размерности.** Вся Θ — это все возможные значения $(c_0, c_1, \dots, c_m)$, их $m+1$ штука, поэтому размерность $m+1$. В Θ_0 все c_i при $i \geq 1$ занулены, остался только c_0 — размерность 1.

7.5. Аналогично для регрессии Пуассона

Нулевая гипотеза: все коэффициенты, кроме c_0, равны нулю. Альтернатива — отрицание H_0.

---

8. Построение критерия

Известно:

$$-2 \ln \Lambda_n \xrightarrow{d} \chi^2_{m-r}$$

Решающее правило:

Если $\Lambda_n > C$, то принимаем H_0, иначе H_1.

В терминах $-2 \ln \Lambda_n$ (знак неравенства меняется):

$$\text{если } -2 \ln \Lambda_n < \tilde{C}, \text{ то } H_0, \text{ иначе } H_1$$

В качестве пороговой константы $\tilde{C}$ берём **квантиль χ^2-распределения** с $m-r$ степенями свободы.