



Статистика

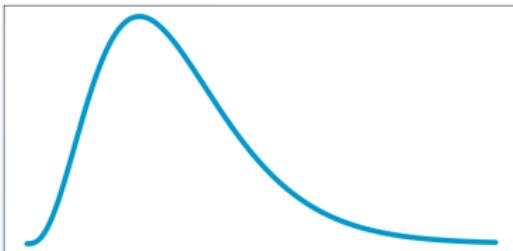
DS-поток

Лекция 9

Непараметрический подход

Продолжение

Что численно характеризует симметричность распределения?



Перебираем моменты:

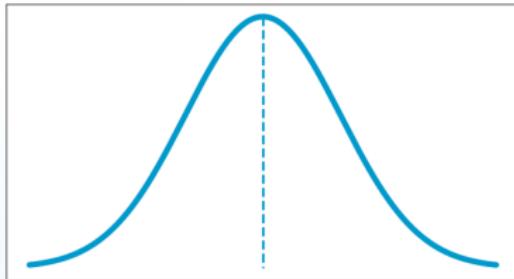
$a = EX$ — отвечает за среднее значение

$\sigma^2 = DX = E(X - a)^2$ — отвечает за разброс значений

Идем дальше...

$\kappa = \frac{1}{\sigma^3} E(X - a)^3$ — коэффициент асимметрии (skewness)
мера симметричности распределения

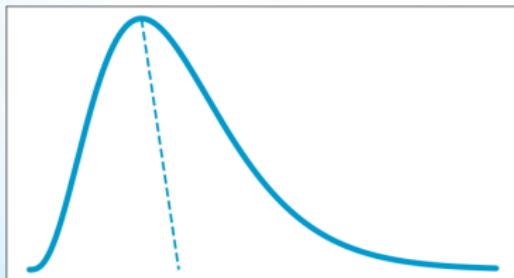
$\gamma = \frac{1}{\sigma^4} E(X - a)^4 - 3$ — коэффициент эксцесса (kurtosis)
мера остроты пика распределения



$$X \sim \mathcal{N}(0, 1)$$

$$\kappa = 0$$

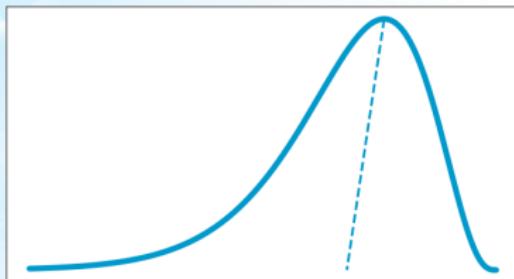
распределение симметрично



$$X \sim \Gamma(1/2, 4)$$

$$\kappa = 1$$

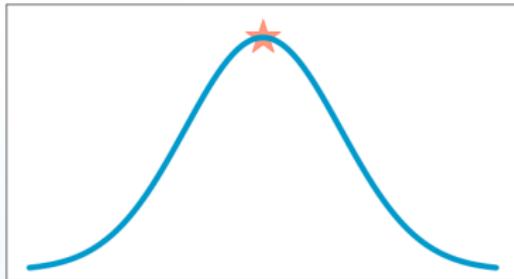
правый хвост тяжелее левого



$$-X \sim \Gamma(1/2, 4)$$

$$\kappa = -1$$

левый хвост тяжелее правого

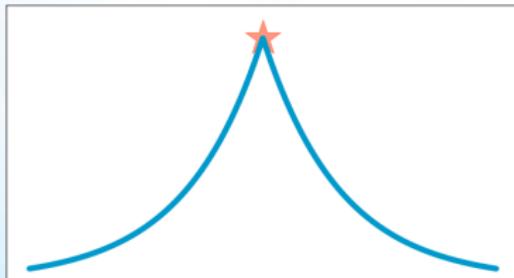


$$X \sim \mathcal{N}(0, 1)$$

$$\gamma = 0$$

сглаженный пик

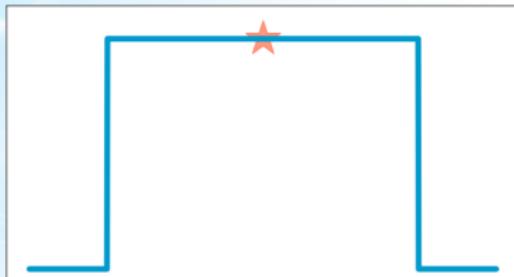
(тройка вычитается чтобы тут $\gamma = 0$)



$$X \sim \text{Laplace}$$

$$\gamma = 3$$

острый пик



$$-X \sim U[0, 1]$$

$$\gamma = -1.2$$

ровный пик

Пусть $X = (X_1, \dots, X_n)$ — выборка.

Посчитаем оценку методом подстановки

$$\hat{\kappa} = \frac{1}{\hat{\sigma}^3} \int_{\mathbb{R}} (x - \hat{a})^3 d\hat{F}_n(x) = \frac{1}{n\hat{\sigma}^3} \sum_{i=1}^n (X_i - \hat{a})^3.$$

Хотелось бы получить доверительный интервал для значения κ ...

Решение

Пусть $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ — параметрический случай.

Применяем многомерную ЦПТ:

$$\sqrt{n} \left(\begin{pmatrix} \bar{X} \\ \bar{X^2} \\ \bar{X^3} \end{pmatrix} - \begin{pmatrix} E_\theta X_1 \\ E_\theta X_1^2 \\ E_\theta X_1^3 \end{pmatrix} \right) \xrightarrow{d_\theta} \mathcal{N}(0, \Sigma(\theta)), \quad \text{где } \Sigma(\theta) = \dots$$

Применяем дельта метод с функцией $\tau(x, y, z) = \frac{z - 3xy^2 + 2x^3}{y - x^2}$.

Далее берем производные, перемножаем матрицы... Что за жесть...

СКОЛЬКО ЗАДАЧ НА ДЕЛЬТА МЕТОД МОЖНО РЕШИТЬ

**1**

ЗАДАЧУ

2-3

ЗАДАЧИ

4-5

ЗАДАЧ

6-8

ЗАДАЧ

11

ЗАДАЧ

○ появляется усталость, снижается внимание и память.

○ нарушается координация движений, ухудшается концентрация зрения, речь, появляется нервный тик, тошнота.

○ наступает чрезвычайная раздражительность, галлюцинации и бредовые идеи.

○ замедленная речь, дрожь конечностей, короткие периоды потери памяти, странности в поведении.

● (рекорд 17-летнего Р. Гарднера, * установленный в 1965 г.) фрагментированное мышление, безразличие ко всему, однозначение.



Посмотрим подробнее на жесть:

1. Посчитать первые 6 моментов, посчитать матрицу ковариаций;
2. Найти функцию для применения дельта-метода;
3. Взять производные;
4. Перемножить матрицы.

Три вопроса "А если ...":

1. А если лень?
2. А если это срочная бизнесовая задача,
а не домашка по статистике?
3. А если нет никакого параметрического семейства?

То есть рассматривается непараметрический случай. Ой...

5.3. Бутстреп

Постановка задачи

X_1, \dots, X_n — выборка из неизв. распр. P ;

$T(X_1, \dots, X_n)$ — некоторая статистика;

$v = V(T(X_1, \dots, X_n)) = G(P)$ — функционал, значение которого требуется оценить;

$\hat{v} = G(\hat{P}_n)$ — оценка методом подстановки.

В примере выше:

$T(X_1, \dots, X_n) = \hat{\kappa}$ — оценка коэффи. асимметрии

$v = D\hat{\kappa}$ — дисперсия оценки коэффи. асимметрии

$\hat{v} = D_{\hat{P}_n} \hat{\kappa}$ — оценка дисперсии оценки коэффи. асимметрии

Пример: оценка дисперсии

Дисперсия статистики

$$V(T(X_1, \dots, X_n)) = D T(X_1, \dots, X_n) = \\ = \int_{\mathcal{X}^n} T^2(x_1, \dots, x_n) dF(x_1) \dots dF(x_n) - \left(\int_{\mathcal{X}^n} T(x_1, \dots, x_n) dF(x_1) \dots dF(x_n) \right)^2.$$

Оценка методом подстановки имеет вид

$$\hat{v} = D_{\hat{P}_n} T(X_1, \dots, X_n) = \\ = \int_{\mathcal{X}^n} T^2(x_1, \dots, x_n) d\hat{F}_n(x_1) \dots d\hat{F}_n(x_n) - \left(\int_{\mathcal{X}^n} T(x_1, \dots, x_n) d\hat{F}_n(x_1) \dots d\hat{F}_n(x_n) \right)^2 = \\ = \frac{1}{n^n} \sum_{i_1=1}^n \dots \sum_{i_n=1}^n T^2(X_{i_1}, \dots, X_{i_n}) - \left(\frac{1}{n^n} \sum_{i_1=1}^n \dots \sum_{i_n=1}^n T(X_{i_1}, \dots, X_{i_n}) \right)^2,$$

Нужно совершить порядка n^n операций!!!



Решение проблемы



Монте-Карло!!!

Метод бутстрепа

Идея: приближенное вычисление \hat{v} методом Монте-Карло.

Этап 1. Генерация выборки из эмп. распределения \hat{P}_n .

Рассмотрим реализацию выборки $X_1(\omega) = x_1, \dots, X_n(\omega) = x_n$.

Тогда реализацией \hat{P}_n явл. распределение $U\{x_1, \dots, x_n\}$.

(с учетом повторений)

Генерация случ. величины из \hat{P}_n :

выбор случайного элемента из мн-ва $\{X_1, \dots, X_n\}$

Генерация выборки X_1^*, \dots, X_n^* из \hat{P}_n :

упоряд. выбор с возвращением n элементов из мн-ва $\{X_1, \dots, X_n\}$.

Другой вид записи:

1. $i_1, \dots, i_n \sim U\{1, \dots, n\}$.

2. $X^* = (X_1^*, \dots, X_n^*) = (X_{i_1}, \dots, X_{i_n})$ — бутстрепная выборка.

Метод бутстрепа

Этап 2.

Процедуру генерации выборок повторить B раз:

$$X_b^* = (X_{b1}^*, \dots, X_{bn}^*), \text{ где } 1 \leq b \leq B.$$

Далее по каждой выборке посчитаем значение статистики T , получив выборку значений $T_1^* = T(X_1^*), \dots, T_B^* = T(X_B^*)$.

Этап 3.

Полученную выборку использовать для аппроксимации значения оценки, которая называется *бутстрепной оценкой*.

Например, бутстрепная оценка дисперсии имеет вид

$$\hat{v}_{boot} = \frac{1}{B} \sum_{b=1}^B {T_b^*}^2 - \left(\frac{1}{B} \sum_{b=1}^B T_b^* \right)^2,$$

Схема метода бутстрепа

$X = (X_1, \dots, X_n)$ — выборка

$T(X_1, \dots, X_n)$ — статистика

Задача: оценить распределение $T(X)$ или функционал $V(T(X))$.

$$\left. \begin{array}{ll} X_{11}^*, \dots, X_{1n}^* & \longrightarrow T(X_1^*) \\ & \vdots \\ X_{b1}^*, \dots, X_{bn}^* & \longrightarrow T(X_b^*) \\ & \vdots \\ X_{B1}^*, \dots, X_{Bn}^* & \longrightarrow T(X_B^*) \end{array} \right\} v_{boot} — \text{бутстрепная оценка } v = V(T(X))$$

Пример

$x = (5, 1, 3, 6, 4)$ — реализация выборки

$T(X_1, X_2, X_3, X_4, X_5) = \bar{X}$ — статистика

$T(5, 1, 3, 6, 4) = 3.8$ — реализация статистики

Задача: оценить дисперсию статистики, т.е. $v = V(T(X)) = D T(X)$.

$$\left. \begin{array}{l} 5, 4, 3, 4, 6 \longrightarrow 4.4 \\ 3, 1, 4, 6, 5 \longrightarrow 3.8 \\ 6, 5, 6, 1, 6 \longrightarrow 4.8 \\ 4, 1, 5, 6, 4 \longrightarrow 4.0 \\ 1, 1, 4, 6, 5 \longrightarrow 3.4 \\ 6, 4, 1, 5, 5 \longrightarrow 4.2 \\ 6, 5, 6, 3, 6 \longrightarrow 5.2 \end{array} \right\} v_{boot} = 0.317$$

Зоопарк: оценить дисперсию выборочного среднего

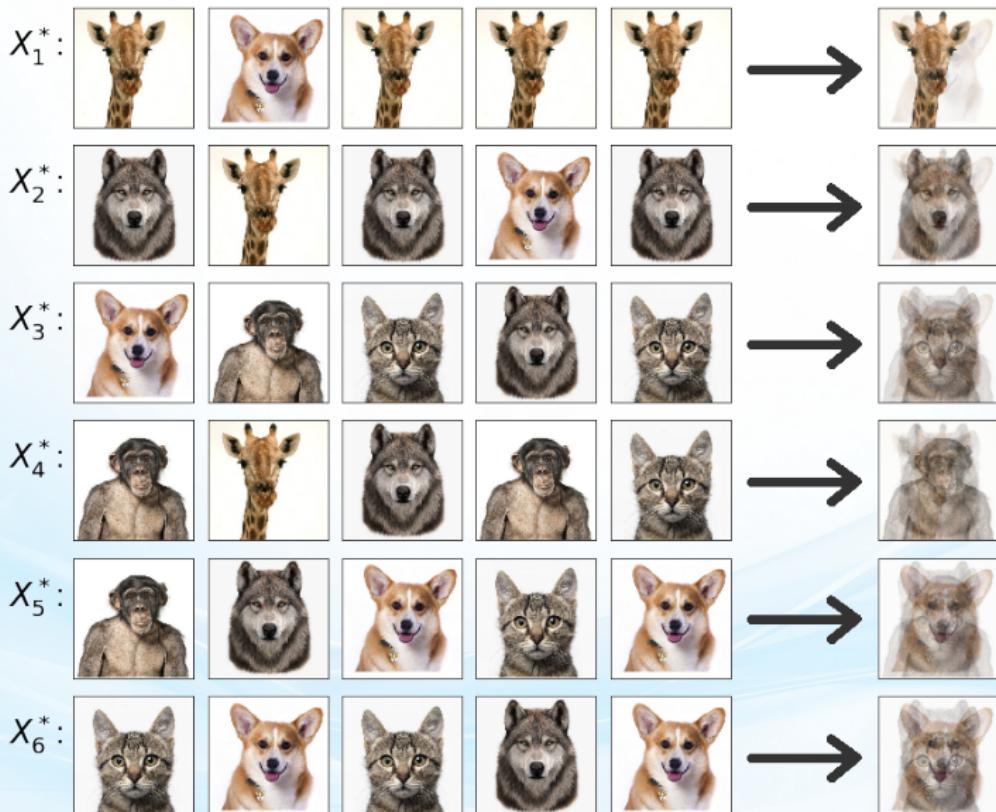
Выборка:



Задача:

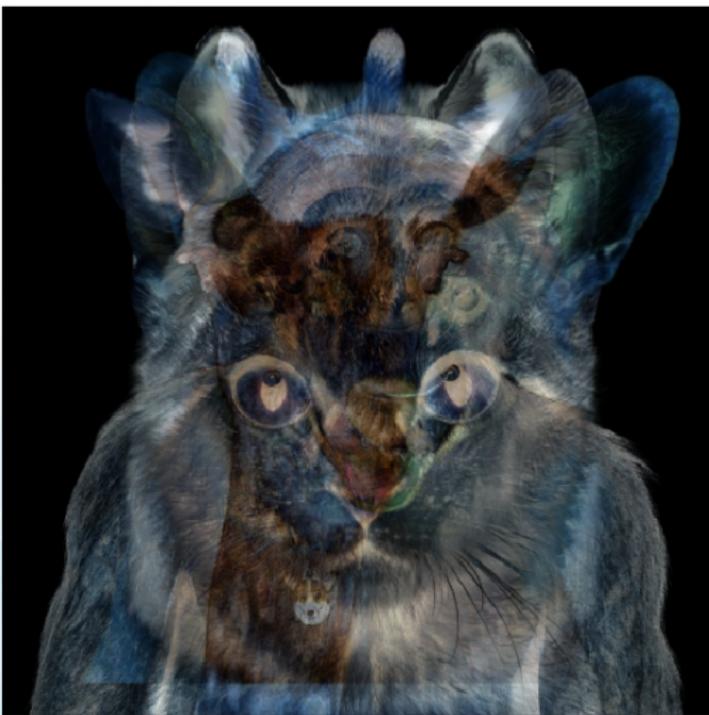
для каждого пикселя и каждого цветового канала
оценить дисперсию выборочного среднего.

Зоопарк: оценить дисперсию выборочного среднего



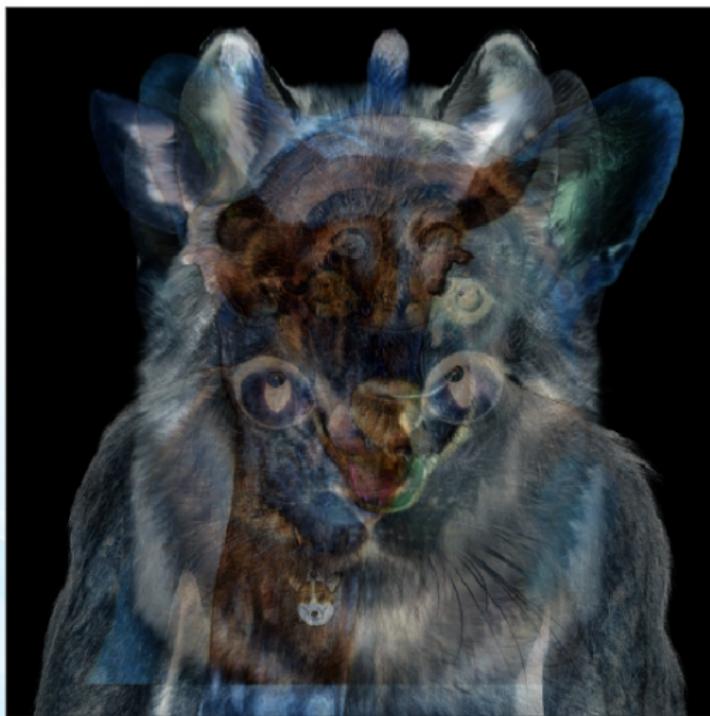
Зоопарк: оценить дисперсию выборочного среднего

Дисперсия по бутстрепной выборке средних:



Зоопарк: оценить дисперсию выборочного среднего

При большем количестве бутстрепных выборок:



Особенности

- ▶ Два этапа аппроксимации

$$V \approx \hat{V} \approx \hat{V}_{boot}.$$

метод подстановки Монте-Карло

Точность аппроксимации м. подстановки: $1/\sqrt{n}$

Точность аппроксимации м. Монте-Карло: $1/\sqrt{B}$

- ▶ Число B стоит брать как можно больше.
- ▶ Размер бутстрепной выборки **всегда тот же**, что и у исходной.

При генерации выборок иного размера распределение статистики T , вообще говоря, может быть другим.
Например, дисперсия выборочного среднего зависит от размера выборки.
- ▶ Генерация бутстр. выборки проводится независимо с повторами.

Иначе полученный набор даже не является выборкой.

Бутстрепные доверительные интервалы

1. Нормальный интервал

Пусть $\hat{\theta}$ — а.н.о. θ с ас. дисп. $\sigma^2(\theta)$.

\hat{v}_{boot} — бутстрепная оценка дисперсии.

Бутстрепный дов. интервал для параметра θ имеет вид

$$\left(\hat{\theta} - z_{(1+\alpha)/2} \sqrt{\hat{v}_{boot}}, \quad \hat{\theta} + z_{(1+\alpha)/2} \sqrt{\hat{v}_{boot}} \right)$$

2. Центральный интервал

$\theta = G(P)$ и $\hat{\theta} = G(\hat{P}_n)$ — оценка методом подстановки.

$\theta_1^*, \dots, \theta_B^*$ — оценки по бутстрепным выборкам.

Бутстрепный доверительный интервал имеет вид

$$C^* = \left(2\hat{\theta} - \theta_{(\lceil B(1+\alpha)/2 \rceil)}^*, \quad 2\hat{\theta} - \theta_{(\lfloor B(1-\alpha)/2 \rfloor)}^* \right).$$

Бутстрепные доверительные интервалы

3. Квантильный интервал

$\hat{\theta}$ — некоторая оценка θ .

$\theta_1^*, \dots, \theta_B^*$ — оценки по бутстрепным выборкам.

Бутстрепный доверительный интервал имеет вид

$$C^* = \left(\theta_{(\lfloor B(1-\alpha)/2 \rfloor)}^*, \quad \theta_{(\lceil B(1+\alpha)/2 \rceil)}^* \right).$$

Утв. Если существует монотонное преобразование φ ,

для которого $\varphi(\hat{\theta}) \sim \mathcal{N}(\varphi(\theta), \sigma^2)$, то $P(\theta \in C^*) = \alpha$.

На практике такое преобразование существует редко, но при

этом часто может существовать приближенное преобразование.

Пример: построение дов. интервалов для θ

$x = (5, 1, 3, 6, 4)$ — реализация выборки

$\theta = EX_1$ — параметр, $\hat{\theta} = \bar{X}$ — оценка, $\hat{\theta} = 3.8$ — реализация оценки

Реализации оценки параметра по бутстрепным выборкам ($B = 100$):

4.2, 4.2, 2.6, 3.2, 4.2, 3.8, 3.2, 3.6, 3.6, 3.4, 3.8, 4.4, 3.6, 3.2, 4.6, 4.2, 3.0, 3.2, 4.0, 3.0, 3.2, 3.0, 2.6, 3.0, 3.6, 3.4, 5.0, 4.8, 3.4, 2.6, 2.6, 3.6, 3.2, 4.2, 3.2, 3.4, 4.4, 4.2, 4.4, 3.4, 4.0, 2.4, 3.4, 3.8, 2.0, 3.0, 4.6, 3.2, 3.6, 3.6, 4.0, 3.8, 4.0, 3.4, 3.8, 3.8, 4.2, 3.2, 2.8, 4.0, 3.2, 3.4, 3.0, 4.0, 3.6, 3.4, 3.8, 3.2, 3.8, 2.6, 3.4, 5.0, 3.6, 3.0, 4.8, 4.2, 3.4, 5.2, 5.0, 3.4, 3.2, 3.6, 4.2, 3.4, 3.2, 3.8, 3.6, 3.8, 3.0, 2.8, 3.0, 4.0, 3.2, 3.6, 2.6, 3.2, 2.4, 3.6, 4.0, 4.2

1. Нормальный интервал

$$\hat{\theta} = 3.8, v_{boot} = 0.394, z_{0.975} = 1.96$$

$$(3.8 \pm 1.96 \cdot \sqrt{0.394}) = (2.57, 5.03)$$

2. Центральный интервал

$$B(1 + \alpha)/2 = 100 \cdot 0.975 = 97.5, B(1 - \alpha)/2 = 100 \cdot 0.025 = 2.5$$

$$\theta^*_{(\lceil 97.5 \rceil)} = 5, \quad \theta^*_{(\lfloor 2.5 \rfloor)} = 2.4$$

$$(2 \cdot 3.8 - 5, 2 \cdot 3.8 - 2.4) = (2.6, 5.2)$$

3. Квантильный интервал

$$(2.4, 5)$$

5.4. Критерии согласия

Общие критерии согласия

Задача. Пусть X_1, \dots, X_n — выборка из **нормального распределения** с параметрами a, σ . Найти оценку максимального правдоподобия.

Задача. Пусть X_1, \dots, X_n — выборка из **гамма-распределения** с параметрами α, β . Найти доверительный интервал для параметра α .

Задача. Пусть X_1, \dots, X_n — выборка из **пуассоновского распределения** с параметром θ . Построить асимптотический критерий проверки гипотез $H_0: \theta = \theta_0$ vs. $H_1: \theta \neq \theta_0$.

Откуда **это** известно?



Критерии согласия

$H_0: P = P_0$

т.е. $H_0: P \in \mathcal{P}_0$, где $\mathcal{P}_0 = \{P_0\}$.

$H_0: P$ — нормальное распределение

т.е. $H_0: P \in \mathcal{P}_0$, где $\mathcal{P}_0 = \{P_{a,\sigma} \text{ — распределение } \mathcal{N}(a, \sigma^2)\}$.

$H_0: P$ — гамма-распределение

т.е. $H_0: P \in \mathcal{P}_0$, где $\mathcal{P}_0 = \{P_{\alpha,\beta} \text{ — распределение } \Gamma(\alpha, \beta)\}$.

Соответствующие критерии называются **критериями согласия**.

Если данные не противоречат проверяемым свойствам распределения (семейства распределений),
то можно считать, что выборка **согласуется** с основной гипотезой.

Критерий Колмогорова

$X = (X_1, \dots, X_n)$ — выборка из неизвестного распределения с ф.р. F .

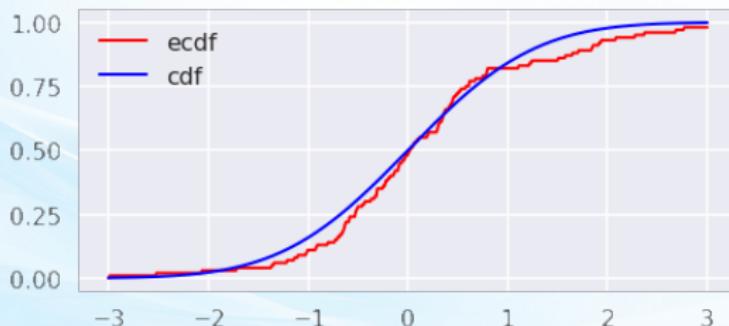
$$H_0: F = F_0$$

$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$ — статистика критерия, где \hat{F}_n — ЭФР

Теорема Колмогорова:

Если F_0 непрерывна $\sqrt{n}D_n \xrightarrow{d_0} \xi \sim K$ — распр. Колмогорова.

Критерий: $\{\sqrt{n}D_n \geq c_\alpha\}$, где c_α — $(1-\alpha)$ -квантиль распр. K .



Критерий Колмогорова

$X = (X_1, \dots, X_n)$ — выборка из неизвестного распределения с ф.р. F .

$$H_0: F = F_0$$

$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$ — статистика критерия, где \hat{F}_n — ЭФР

Критерий: $\{\sqrt{n}D_n \geq c_\alpha\}$, где c_α — $(1-\alpha)$ -квантиль распр. K .

Свойства:

1. распред. статистики D_n не зависит от конкретного F_0 ;
2. имеет низкую мощность;
3. не чувствителен к различиям на хвостах;
4. применим при $n \geq 20$;
5. выполнена состоятельность: $\beta(P) \rightarrow 1$ при $n \rightarrow \infty$ и $P \neq P_0$.

Другие критерии согласия, основанные на различиях между F_0 и \hat{F}_n .

1. Джини: $\int \left| \hat{F}_n(x) - F_0(x) \right| dx;$
2. Крамера-фон Мизеса: $\int \left(\hat{F}_n(x) - F_0(x) \right)^2 dx;$
3. Смирнова-Крамера-фон Мизеса: $\int \left(\hat{F}_n(x) - F_0(x) \right)^2 dF_0(x);$
4. Андерсона-Дарлинга: $\int \frac{\left(\hat{F}_n(x) - F_0(x) \right)^2}{F_0(x)(1-F_0(x))} dF_0(x);$
5. Купера: $\sup_{x \in \mathbb{R}} (\hat{F}_n(x) - F_0(x)) + \sup_{x \in \mathbb{R}} (F_0(x) - \hat{F}_n(x));$
6. Ватсона: $\int \left(\hat{F}_n(x) - F_0(x) - \int \left(\hat{F}_n(x) - F_0(x) \right) dF_0(x) \right) dF_0(x);$
7. Фроцини: $\int \left| \hat{F}_n(x) - F_0(x) \right| dF_0(x).$

Пусть $\mathcal{P} = \{P_{a,\sigma}\}$ — семейство распределений,

где a — параметр сдвига, σ — параметр масштаба.

Для проверки $H_0: P \in \mathcal{P}$ существуют модификации этих критериев, которые используют \hat{a} и $\hat{\sigma}$ — состоятельные оценки.

Зачем столько критериев?

Пусть \mathcal{P} — все парнокопытные звери.

H_0 : P — лошадь

S — критерий, проверяющий некоторое свойство парнокопытного.

Если результат не свойственен лошади,
то гипотезу следует отклонить.



Животное 1



Животное 1



Животное 1



Животное 1



Животное 1



Животное 2



Животное 2



Животное 2



Животное 3



Животное 3



Животное 3



Животное 3



Животное 3



Животное 3



Животное 4



Животное 4



Выводы

- ▶ Разные критерии проверяют выполнение разных свойств распределения (или класса распределений).
Если свойство не выполняется, критерий отвергает гипотезу.
Лошади обычно не пятнистые.
- ▶ Иначе, возможно, критерий недостаточно мощный.
Копыта есть у всех.
- ▶ Чем больше разных критериев, тем лучше.
- ▶ Комбинирование происходит с помощью МПГ.

QQ plot — графический способ

Пусть $\mathcal{P} = \{P_{a,\sigma}\}$ — семейство распределений,

a — параметр сдвига,

σ — параметр масштаба.

$X = (X_1, \dots, X_n)$ — выборка из неизв. распр. P .

Верно ли, что $P \in \mathcal{P}$?

Пусть $F_{a,\sigma}$ — функция распр. $P_{a,\sigma}$.

По свойствам параметров: $F_{a,\sigma}(x) = F_{0,1}\left(\frac{x-a}{\sigma}\right)$.

Утверждение: $F_{a,\sigma}(X_i) \sim U[0, 1]$

$$\Rightarrow F_{a,\sigma}(X_{(1)}) = F_{0,1}\left(\frac{X_{(1)} - a}{\sigma}\right), \dots, F_{a,\sigma}(X_{(n)}) = F_{0,1}\left(\frac{X_{(n)} - a}{\sigma}\right)$$

— вариационный ряд, построенный по выборке из $U[0, 1]$.

QQ plot — графический способ

$$F_{a,\sigma}(X_{(1)}) = F_{0,1}\left(\frac{X_{(1)} - a}{\sigma}\right), \dots, F_{a,\sigma}(X_{(n)}) = F_{0,1}\left(\frac{X_{(n)} - a}{\sigma}\right)$$

— вариационный ряд, построенный по выборке из $U[0, 1]$.

$$\Rightarrow X_{(i)} = a + \sigma \cdot F_{0,1}^{-1}(F_{a,\sigma}(X_{(i)}))$$

Вместо выборочных квантилей $F_{a,\sigma}(X_{(i)})$ подставим теоретические квантили $E F_{a,\sigma}(X_{(i)}) = \frac{i}{n+1}$ (из ДЗ-1):

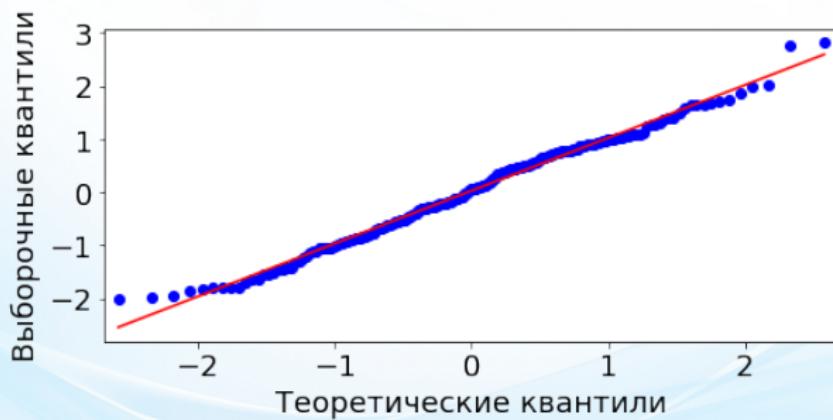
$$X_{(i)} \approx a + \sigma \cdot F_{0,1}^{-1}(E F_{a,\sigma}(X_{(i)})) = a + \sigma \cdot F_{0,1}^{-1}\left(\frac{i}{n+1}\right)$$

Вывод: точки $(X_{(i)}, F_{0,1}^{-1}\left(\frac{i}{n+1}\right))$ примерно лежат на одной прямой.

Q-Q plot — график точек $(X_{(i)}, F_{0,1}^{-1}\left(\frac{i}{n+1}\right))$.

QQ plot — графический способ

Q-Q plot — график точек $(X_{(i)}, F_{0,1}^{-1}\left(\frac{i}{n+1}\right))$.



5.4. Критерии согласия

Критерии проверки нормальности

Критерий Жарка-Бера

$X = (X_1, \dots, X_n)$ — выборка из неизвестного распределения P .

H_0 : P — нормальное распределение.

$\kappa = \frac{E(\xi - a)^3}{\sigma^3}$ — коэффициент асимметрии (skewness).

$\gamma = \frac{E(\xi - a)^4}{\sigma^4} - 3$ — коэффициент эксцесса (kurtosis).

Для нормального распределения $\kappa = 0, \gamma = 0$.

Их оценки методом подстановки:

$$\hat{\kappa} = \frac{1}{nS^3} \sum_{i=1}^n (X_i - \bar{X})^3, \quad \hat{\gamma} = \frac{1}{nS^4} \sum_{i=1}^n (X_i - \bar{X})^4 - 3.$$

Статистика критерия $T(X) = \frac{n}{6} (\hat{\kappa}^2 + \hat{\gamma}^2 / 4) \sim \chi^2_2$.

Критерий: $\{T(x) > \chi^2_{2,1-\alpha}\}$.

p-value: $p(t) = 1 - F(t)$, где F — функция распр. χ^2_2 .

Критерий Шапиро-Уилка

$X = (X_1, \dots, X_n)$ — выборка из неизвестного распределения P .

H_0 : P — нормальное распределение.

Статистика критерия

$$W(X) = \left(\sum_{i=1}^n a_i X_{(i)} \right)^2 / n S^2,$$

$$a = \frac{m^T V^{-1}}{(m^T V^{-2} m)^{1/2}},$$

m и V — вектор средних и матрица ковариаций вар. ряда $\mathcal{N}(0, 1)$.

При H_0 имеет табличное распределение.

На практике обычно он самый мощный для проверки нормальности.

О проверке нормальности

- ▶ Часто в предположении нормальности выборки применимы более мощные методы.
- ▶ Перед использованием таких методов нужно проверить нормальность.
- ▶ Методы могут быть устойчивыми к небольшим отклонениям от нормальности.
- ▶ Если метод устойчив к отклонениям, обычно достаточно посмотреть на Q-Q plot.
- ▶ Если метод неустойчив к отклонениям, желательно проверить нормальность критерием Шапиро-Уилка.



Критерии проверки нормальности [Кобзарь]

- 3.2. Критерии нормальности распределения
- 3.2.1. Общие критерии согласия, модифицированные для проверки нормальности распределения (231). 3.2.1.1. Модифицированный критерий χ^2 (231).
 - 3.2.1.2. Критерии типа Колмогорова–Смирнова (233). 3.2.1.3. Критерий Френини (235). 3.2.2. Специальные критерии нормальности (235). 3.2.2.1. Критерий Шапиро–Уилка (238). 3.2.2.2. Энтропийный критерий нормальности (критерий Васичека) (241). 3.2.2.3. Критерий Хегази–Грина (243).
 - 3.2.2.4. Критерий Али–Чёrgo–Ревеса (244). 3.2.2.5. Корреляционный критерий Филибена (245). 3.2.2.6. Регрессионный критерий нормальности Ла Бре-ка (248). 3.2.2.7. Критерий нормальности Локка–Спурье (252). 3.2.2.8. Критерий нормальности Оя (254). 3.2.2.9. Критерий среднего абсолютного отклонения (критерий Гири) (257). 3.2.2.10. Критерий Дэвида–Хартли–Пирсона (258).
 - 3.2.2.11. Комбинированный критерий Шпигельхальтера (260). 3.2.2.12. Критерий нормальности Саркади (261). 3.2.2.13. Критерий нормальности Лина–Мудхолкара (263). 3.2.2.14. Критерий нормальности Мартинеса–Иглевича (265). 3.2.2.15. Критерий нормальности Д’Агостино (266). 3.2.2.16. Критерии асимметрии и эксцесса (268). 3.2.2.17. Критерий характеристической функции (критерий Муроты–Такеучи) (272). 3.2.2.18. Критерии проверки нормальности распределения по совокупности независимых выборок малого объема (273). 3.2.2.18.1. Применение критерия Шапиро–Уилка (274). 3.2.2.18.2. Применение критерия Саркади (274). 3.2.2.18.3. Критерий Смирнова (275). 3.2.2.19. Сравнительная мощность различных критериев нормальности (277).

Сравнение критериев проверки нормальности [Кобзарь]

Наименование критерия (раздел)	Характер альтернативного распределения					Ранг
	асимметричное		симметричное		≈ нормальное	
	$\alpha_4 < 3$	$\alpha_4 > 3$	$\alpha_4 < 3$	$\alpha_4 > 3$	$\alpha_4 \approx 3$	
Критерий Шапиро–Уилка (3.2.2.1)	1	1	3	2	2	1
Критерий K^2 (3.2.2.16)	7	8	10	6	4	2
Критерий Дарбина (3.1.2.7)	11	7	7	15	1	3
Критерий Д'Агостино (3.2.2.14)	12	9	4	5	12	4
Критерий α_4 (3.2.2.16)	14	5	2	4	18	5
Критерий Васичека (3.2.2.2)	2	14	8	10	10	6
Критерий Дэвида–Хартли–Пирсона (3.2.2.10)	21	2	1	9	1	7
Критерий χ^2 (3.1.1.1)	9	20	9	8	3	8
Критерий Андерсона–Дарлинга (3.1.2.4)	18	3	5	18	7	9
Критерий Филлибена (3.2.2.5)	3	12	18	1	9	10
Критерий Колмогорова–Смирнова (3.1.2.1)	16	10	6	16	5	11
Критерий Мартинеса–Иглевича (3.2.2.14)	10	16	13	3	15	12
Критерий Лина–Мудхолкара (3.2.2.13)	4	15	12	12	16	13
Критерий α_3 (3.2.2.16)	8	6	21	7	19	14
Критерий Шпигельхальтера (3.2.2.11)	19	13	11	11	8	15
Критерий Саркади (3.2.2.12)	5	18	15	14	13	16
Критерий Смирнова–Крамера–фон Мизеса (3.1.2.2)	17	11	20	17	6	17
Критерий Локка–Спурье (3.2.2.7)	13	4	19	21	17	18
Критерий Оя (3.2.2.8)	20	17	14	13	14	19
Критерий Хегази–Грина (3.2.2.3)	6	19	16	19	21	20
Критерий Муроты–Такеучи (3.2.2.17)	15	21	17	20	20	21

Другие распределения [Кобзарь]

- 3.3. Критерии проверки экспоненциальности распределения
 3.3.1. Критерий Шапиро–Уилка (279). 3.3.2. Критерии типа Колмогорова–Смирнова (282). 3.3.3. Критерии типа Смирнова–Крамера–фон Мизеса для цензурированных данных (286). 3.3.4. Критерий Фроцини (288). 3.3.5. Корреляционный критерий экспоненциальности (288). 3.3.6. Регрессионный критерий Брейна–Шапиро (290). 3.3.7. Критерий Кимбера–Мичела (292). 3.3.8. Критерий Фишера (293). 3.3.9. Критерий Бартлетта–Морана (294). 3.3.10. Критерий Климко–Антла–Радемакера–Рокетта (294). 3.3.11. Критерий Холлендера–Прошана (295). 3.3.12. Критерий Кочара (298). 3.3.13. Критерий Эпса–Палли–Чёрго–Уэлча (299). 3.3.14. Критерий Бергмана (301). 3.3.15. Критерий Шермана (303). 3.3.16. Критерий наибольшего интервала (304). 3.3.17. Критерий Хартли (305). 3.3.18. Критерий показательных меток (305). 3.3.19. Ранговый критерий независимости интервалов (306). 3.3.20. Критерии, основанные на трансформации экспоненциального распределения в равномерное (308). 3.3.20.1. Критерий \bar{U} (308). 3.3.20.2. Критерий \hat{U} (309). 3.3.20.3. Критерий Гринвуда (309). 3.3.21. Критерий Манн–Фертига–Шуера для распределения Вейбулла (311). 3.3.22. Критерий Дешпанда (316). 3.3.23. Критерий Лоулесса (317).
- 3.4. Критерии согласия для равномерного распределения
 3.4.1. Критерий Шермана (319). 3.4.2. Критерий Морана (320). 3.4.3. Критерий Ченга–Спиринга (322). 3.4.4. Критерий Саркади–Косика (323). 3.4.5. Энтропийный критерий Дудевича–ван дер Мюлена (324). 3.4.6. Критерий Хегази–Грина (326). 3.4.7. Критерий Янга (328). 3.4.8. Критерии типа Колмогорова–Смирнова (330). 3.4.9. Критерий Фроцини (331). 3.4.10. Критерий Гринвуда–Кэсенберри–Миллера (332). 3.4.11. „Сглаженный“ критерий Неймана–Бартона (333).

5.3 Критерии согласия

Критерий хи-квадрат

Критерий хи-квадрат

$X = (X_1, \dots, X_n)$ — выборка из неизвестного распределения P
со значениями в \mathcal{X} .

$$H_0: P = P_0$$

Рассмотрим разбиение $\mathcal{X} = \bigsqcup_{j=1}^k B_j$

$$\text{Статистика критерия } \hat{\chi} = \sum_{j=1}^k \frac{(\mu_j - np_j^0)^2}{np_j^0},$$

$$\text{где } \mu_j = \#\{i \mid X_i \in B_j\},$$

$$p_j^0 = P_0(X_1 \in B_j)$$

Теорема Пирсона: $\hat{\chi} \xrightarrow{d_0} \chi_{k-1}^2$.

Критерий: $\{\hat{\chi} \geq \chi_{k-1, 1-\alpha}^2\}$

Свойства:

1. $n \geq 50, \forall i : np_j^0 \geq 5$ — необходимо для хорошего приближения;
2. обычно берут $k \sim \log_2 n$;
3. применим во многих задачах;
4. имеет маленькую мощность;
5. неоднозначное разбиение на интервалы.

Горошок Менделя

Мендель рассматривал

следующую классификацию гороха:



Рассмотрим непосредственно горошины ($n = 556$, $k = 4$)

B_j — вид горошин	круглые желтые	морщин. желтые	круглые зеленые	морщин. зеленые
Вер-ти $p_j^0 = P_0(X_1 \in B_j)$	9/16	3/16	3/16	1/16
Частота $\mu_j = \#\{X_i \in B_j\}$	315	101	108	32

$$\widehat{\chi}^2 = \frac{(315 - 556 \cdot 9/16)^2}{556 \cdot 9/16} + \frac{(101 - 556 \cdot 3/16)^2}{556 \cdot 3/16} + \frac{(108 - 556 \cdot 3/16)^2}{556 \cdot 3/16} + \frac{(32 - 556 \cdot 1/16)^2}{556 \cdot 1/16} \approx 0.47$$

p-value $\approx 0.925 \implies$ гипотеза не отвергается.

Обобщенный критерий хи-квадрат

$X = (X_1, \dots, X_n)$ — выборка из неизв. распр. P со значениями в \mathcal{X} .

$H_0: P \in \mathcal{P}_0$, где $\mathcal{P}_0 = \{P_\theta \mid \theta \in \Theta \subset \mathbb{R}^d\}$

Рассмотрим разбиение $\mathcal{X} = \bigsqcup_{j=1}^k B_j$

Статистика критерия

$$\hat{\chi} = \sum_{j=1}^k \frac{(\mu_j - np_j^0(\hat{\theta}))^2}{np_j^0(\hat{\theta})},$$

где $\mu_j = \#\{i \mid X_i \in B_j\}$,

$$p_j^0(\theta) = P_\theta(X_1 \in B_j).$$

ОМП $\hat{\theta}$ получается из условия $\sum_{j=1}^k \mu_j \ln p_j^0(\theta) \rightarrow \max_{\theta}$.

Теорема: в условиях регулярности $\hat{\chi} \xrightarrow{d_0} \chi^2_{k-1-d}$.

Примеры на критерий хи-квадрат

Каждый человек имеет кровь, принадлежащую одной из четырех групп: 0, A, B, AB. Наследование управляет тремя генами: A, B, 0, при этом ген 0 подавляется генами A и B.

p, q, r — вероятности появления генов A, B, 0 у родителей.

Группа B_j	Гены родителей	Вероятность $p_j^0(\theta)$	Данные μ_j
0	00	r^2	121
A	AA, A0	$p^2 + 2pr$	120
B	BB, B0	$q^2 + 2qr$	79
AB	AB	$2pq$	33

H_0 : механизм наследования крови имеет место.

Примеры на критерий хи-квадрат

B_j — группы 0, A, B, AB, т.е. $k = 4$.

$d = 2$ — количество параметров (используем $r = 1 - p - q$).

ОМП для параметра $\theta = (p, q)$ ищем из условия:

$$\sum_{j=1}^k \mu_j \ln p_j^0(\theta) = 121 \ln r^2 + 120 \ln(p^2 + 2pr) + 79 \ln(q^2 + 2qr) + 33 \ln 2pq \rightarrow \max_{p,q}$$

Получаем $\hat{p} \approx 0.241$, $\hat{q} \approx 0.167$, $\hat{r} \approx 0.592$.

Оценки $p_j^0(\hat{\theta})$: $\hat{p}_1 = 0.343$, $\hat{p}_2 = 0.340$, $\hat{p}_3 = 0.224$, $\hat{p}_4 = 0.093$.

Отсюда $\hat{\chi} = 0.001$, $pvalue = 0.97$.

Theta



BCE !