Статистика

Конспект курса

Мат-Мех, ПМИ

5–7 семестры (2015–2016)

$Revision: 1.51 $

Содержание

1. Описательная статистика

1.1. Выборка и эмпирическая случайная величина

Пусть $\xi \sim \mathcal{P}$ — случайная величина с распределением $\mathcal{P}$.

Определение. *Повторной независимой выборкой* объема n *(до эксперимента)* называется набор

$$\mathbf{x} = (x_1, \dots, x_n)\,, \quad x_i \sim \mathcal{P}\ \forall i \in 1:n,\ x_1 \perp\!\!\!\perp \dots \perp\!\!\!\perp x_n$$

независимых в совокупности одинаково распределенных случайных величин с распределением $\mathcal{P}$.

Определение. *Повторной независимой выборкой* объема n *(после эксперимента)* называется набор реализаций, т.е. конкретных значений ξ, случайных величин x_i:

$$\mathbf{x} = (x_1, \dots, x_n)\,, \quad x_i \in \operatorname{im}\xi\ \forall i \in 1:n.$$

Определение. *Эмпирической случайной величиной* $\hat{\xi}_n$ называется случайная величина с дискретным распределением

$$\hat{\xi}_n \sim \hat{\mathcal{P}}_n : \begin{pmatrix} x_1 & \dots & x_n \\ 1/n & \dots & 1/n \end{pmatrix}.$$

Замечание. Подходящее определение выбирается по контексту.

Если ξ имеет дискретное распределение, то выборку можно *сгруппировать*; тогда получим случайную величину $\hat{\xi}_m$ с распределением

$$\hat{\mathcal{P}}_m : \begin{pmatrix} x_1^* & \dots & x_m^* \\ \omega_1 & \dots & \omega_m \end{pmatrix} \quad \omega_i = \frac{\nu_i}{n},$$

где x_i^* — уникальные значения из выборки $\mathbf{x}$, а ν_i — число x_i^* в $\mathbf{x}$ (т.н. «абсолютная частота»; тогда ω_i — «относительная частота»). В противном случае, можно разбить интервал всевозможных значений выборки на m подынтервалов: $\{[e_0, e_1), \dots, [e_{m-1}, e_m)\}$ и считать число наблюдений $\nu_i = \nu_i[e_{i-1}, e_i)$, попавших в интервал.

Следствие. *По ЗБЧ (теореме Бернулли),*

$$\omega_i \xrightarrow{\mathsf{P}} p_i = \mathsf{P}(e_{i-1} \le \xi < e_i),$$

т.е. относительная частота является хорошей оценкой вероятности на больших объемах выборки.

1.2. Виды признаков

Виды признаков случайной величины $\xi : \big(\Omega, \mathcal{F}, \mathsf{P}\big) \to \big(V, \mathfrak{A}\big)$ характеризуются тем, что из себя представляет множество V и что можно делать с его элементами.

Количественные признаки: $V \subset \mathbb{R}$

По типу операций:

- Аддитивные: заданы, т.е. имеют смысл в контексте данного признака, операции $+, -$
- Мультипликативные: заданы операции $\cdot, /$; признак принимает не отрицательные значения.

По типу данных:

- Непрерывные
- Дискретные

Порядковые признаки V — упорядоченное множество, определены отношения $>, =$.

Качественные признаки на V заданы отношения $=, \neq$

Пример. Цвет глаз, имена, пол.

2. Точечное оценивание

2.1. Характеристики распределений и метод подстановки

Определение. *Статистика* — измеримая функция от выборки.

Обобщением статистики является понятие характеристики.

Определение. *Характеристика* — функционал от распределения:

$$T : \{\mathcal{P}\} \to V.$$

Где V — измеримое пространство, чтобы на нём можно было завести σ-алгебру.

Замечание. Чаще всего, $V = \mathbb{R}$.

Определение. Выделяют *генеральные* характеристики $T(\mathcal{P}) =: \theta$ и *выборочные* характеристики $T(\hat{\mathcal{P}}_n)$. *Оценка* — выборочная характеристика $T(\hat{\mathcal{P}}_n) =: \hat{\theta}_n$, не зависящая от генеральной характеристики θ.

Следствие. *Выражения для вычисления генеральных и выборочных характеристик отличаются только используемыми мерами ($\mathcal{P}$ и $\hat{\mathcal{P}}_n$ соответственно).*

Определение. Пусть $\hat{\mathcal{P}}_n$ — распределение эмпирической случайной величины. Тогда *эмпирическая функция распределения* есть

$$\widehat{\mathrm{cdf}}_\xi(x) = \mathrm{cdf}_{\hat{\xi}_n}(x) = \hat{\mathcal{P}}_n((-\infty, x)) = \int_{-\infty}^{x} \mathrm{d}\hat{\mathcal{P}}_n = \sum_{x_i : x_i \leq x} \frac{1}{n} = \frac{|\{x_i \in \mathbf{x} : x_i < x\}|}{n}.$$

Утверждение. Пусть $\widehat{\mathrm{cdf}}_\xi$ — эмпирическая функция распределения, cdf_ξ — функция распределения ξ. Тогда, по теореме Гливенко-Кантелли,

$$\sup_x \left| \widehat{\mathrm{cdf}}_\xi(x) - \mathrm{cdf}_\xi(x) \right| \xrightarrow{\text{a.s.}} 0.$$

Замечание. Поскольку $\widehat{\mathrm{cdf}}_\xi(x) = \omega_x$, где ω_x — частота попадания наблюдений в интервал в $(-\infty, x)$, а $\mathrm{cdf}_\xi(x) = \mathsf{P}\left(\xi \in (-\infty, x)\right)$ — вероятность того же события, то можно применить теорему Бернулли (ЗБЧ):

$$\widehat{\mathrm{cdf}}_\xi(x) \xrightarrow{\mathsf{P}} \mathrm{cdf}_\xi(x).$$

Следствие. *Значит, при достаточно больших n, в качестве интересующей характеристики θ распределения $\mathcal{P}$ можем брать ее оценку $\hat{\theta}_n$ — аналогичную характеристику $\hat{\mathcal{P}}_n$.*

2.2. Моменты

Определение. Генеральные и соответствующие им выборочные характеристики *k-го момента* и *k-го центрального момента*:

$$\mathsf{m}_k = \int_{\mathbb{R}} x^k \,\mathrm{d}\mathcal{P} \qquad\qquad \hat{\mathsf{m}}_k = \int_{\mathbb{R}} x^k \,\mathrm{d}\hat{\mathcal{P}}_n = \frac{1}{n}\sum_{i=1}^n x_i^k$$

$$\mathsf{m}_k^{(0)} = \int_{\mathbb{R}} (x - \mathsf{m}_1)^k \,\mathrm{d}\mathcal{P} \qquad\qquad \hat{\mathsf{m}}_k^{(0)} = \int_{\mathbb{R}} (x - \hat{\mathsf{m}}_1)^k \,\mathrm{d}\hat{\mathcal{P}}_n = \frac{1}{n}\sum_{i=1}^n (x_i - \hat{\mathsf{m}}_1)^k.$$

2.2.1. Характеристики положения

В качестве характеристики положения выделяется 1-й момент — математическое ожидание и *выборочное среднее*:

$$\mathsf{m}_1 = \mathsf{E}\xi, \qquad \hat{\mathsf{m}}_1 =: \bar{\mathbf{x}} = \widehat{\mathsf{E}\xi} = \mathsf{E}\hat{\xi}_n.$$

Замечание. В случае мультипликативных признаков можно посчитать среднее геометрическое; часто логарифмируют и считают среднее арифметическое.

Определение. Пусть $p \in [0,1]$ и $\mathrm{cdf} = \mathrm{cdf}_{\mathcal{P}}$. *$p$-квантилью* называется

$$\mathrm{qnt}_{\mathcal{P}}(p) =: z_p = \sup\left\{z : \mathrm{cdf}(z) \le p\right\}.$$

Квартиль есть 1/4- или 3/4-квантиль.

Определение. *Медиана* есть 1/2-квантиль:

$$\mathrm{med}\,\xi = z_{1/2}.$$

Определение. Мода ($\mathrm{mode}\,\xi$) есть точка локального максимума плотности.

По методу подстановки можем получить аналогичные выборочные характеристики.

Определение. *Выборочная p-квантиль* есть такая точка $\hat{z}_p$, что она больше по значению $|\mathbf{x}|\cdot p = np$ точек из выборки:

$$\hat{z}_p = \sup\left\{z : \widehat{\mathrm{cdf}}_\xi(z) \le p\right\} = x_{\lfloor np \rfloor + 1}.$$

Определение. *Выборочная медиана* упорядоченной выборки $\mathbf{x} = \big(x_{(1)}, \ldots, x_{(n)}\big)$ есть

$$\hat{z}_{1/2} = \widehat{\mathrm{med}} = \begin{cases} x_{(k+1)} & n = 2k+1 \\ \dfrac{x_{(k)} + x_{(k+1)}}{2} & n = 2k \end{cases}$$

Определение. *Выборочная мода* ($\widehat{\mathrm{mode}}$) есть значение из выборки, которое чаще всего встречается.

2.2.2. Характеристики разброса

В качестве характеристики разброса выделяется 2-й центральный момент — дисперсия и выборочная дисперсия:

$$\mathsf{m}_2^{(0)} = \mathsf{D}\xi \qquad \hat{\mathsf{m}}_2^{(0)} =: s^2 = \widehat{\mathsf{D}\xi} = \mathsf{D}\hat{\xi}_n = \begin{cases} \mathsf{E}\left(\hat{\xi}_n - \mathsf{E}\hat{\xi}_n\right)^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^n (x_i - \bar{\mathbf{x}})^2 \\ \mathsf{E}\hat{\xi}_n^2 - \left(\mathsf{E}\hat{\xi}_n\right)^2 = \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^n x_i^2\right) - \bar{\mathbf{x}}^2. \end{cases}$$

Замечание. Если среднее $\mathsf{E}\xi = a$ известно, то дополнительно вводится

$$s_a^2 := \begin{cases} \dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i - a)^2 \\ \left(\dfrac{1}{n}\sum\limits_{i=1}^{n} x_i^2\right) - a^2. \end{cases}$$

Пример (Оценка дисперсии оценки мат. ожидания)**.** Пусть строится оценка мат. ожидания $\bar{\mathbf{x}}$. Может интересовать точность построенной оценки. Вычислим дисперсию теоретически, после чего оценим точность по выборке:

$$\mathsf{D}\bar{\mathbf{x}} = \mathsf{D}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{D}x_i = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{D}\xi = \frac{\mathsf{D}\xi}{n},$$

откуда

$$\widehat{\mathsf{D}\bar{\mathbf{x}}} = \frac{s^2}{n}.$$

Пример (Дисперсия оценки дисперсии)**.** См. по ссылке[1].

Определение (Энтропия)**.** Количество информации, необходимое для выявления объекта из n-элементного множества вычисляется по *формуле Хартли*:

$$H = \log_2 n$$

(множество это следует итеративно разбивать пополам, откуда и оценка). Пусть теперь множество не равновероятно, т.е. задано дискретное распределение

$$\mathcal{P}_\xi : \begin{pmatrix} x_1 & \dots & x_n \\ p_1 & \dots & p_n \end{pmatrix}.$$

Тогда количество информации $H(\xi)$, которую нужно получить, чтобы узнать, какой исход эксперимента осуществлен, вычисляется по формуле Шеннона и называется *энтропией*:

$$H(\xi) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}.$$

Замечание. В случае равномерного дискретного распределения, конечно, $H = H(\xi)$.

Определение. *Выборочное стандартное отклонение* есть

$$\mathrm{SD} := \sqrt{\widehat{\mathsf{D}\xi}} = s.$$

Это показатель разброса случайной величины; показатель того, насколько элементы выборки отличаются от выборочного среднего по значению.

Определение. *Стандартная ошибка* оценки есть

$$\mathrm{SE} := \sqrt{\widehat{\mathsf{D}\hat{\theta}}}.$$

Это показатель разброса оценки случайной величины.

Замечание. В частном случае $\theta = \mathsf{E}\xi$, $\hat{\theta} = \bar{\mathbf{x}}$ получаем *выборочную стандартную ошибку среднего*

$$\mathrm{SE} = \sqrt{\frac{\mathsf{D}\xi}{n}} = \frac{s}{\sqrt{n}}.$$

Это, в свою очередь, показатель того, насколько выборочное среднее отличается от истинного.

Пример (Пример с мостом и машинами)**.** FIXME

[1] http://mathworld.wolfram.com/SampleVarianceDistribution.html

2.2.3. Анализ характера разброса

Определение. *Коэффициент асимметрии Пирсона* («скошенности»[2])

$$\gamma_3 = \mathsf{A}\xi = \frac{\mathsf{m}_3^{(0)}}{\sigma^3} = \frac{\mathsf{E}\xi - \operatorname{med}\xi}{\sigma}.$$

Замечание. Не зависит от линейных преобразований.

Определение. *Коэффициент эксцесса* («крутизны», «kurtosis»):

$$\gamma_4 = \mathsf{K}\xi = \frac{\mathsf{m}_4^{(0)}}{\sigma^4} - 3.$$

Замечание. Величина $\mathsf{m}_4^{(0)}/\sigma^4 = 3$ соответствует стандартному нормальному распределению. Так что можно сравнивать выборку и γ_4 $\mathrm{N}(0,1)$.

Замечание. При замене $z := (\xi - \mathsf{E}\xi)/\sigma$ величину $\mathsf{m}_4^{(0)}/\sigma^4 = \mathsf{E}(z^4)$ можно интерпретировать как ожидание четвертой степени центрированных и нормированных данных. Точки выборки, лежащие внутри $\mathsf{E}\xi \pm \sigma$ из-за малости по модулю не будут увеличивать значение коэффициента, в то время как аутлаеры будут или «тяжелые хвосты» плотности распределения будут. Поэтому γ_4 принимает большие значения на распределениях с «тяжелыми хвостами» или выборках с некоторым количеством аутлаеров.

Замечание. Справедлива оценка

$$\gamma_3^2 + 4 \le \gamma_4 + 3 \le \infty,$$

где минимум достигается $\mathrm{Ber}(1/2)$.

Определение. Пусть $(\xi_1, \xi_2) \sim \mathcal{P}$ и $(x_1, y_1), \ldots, (x_n, y_n) \sim \mathcal{P}(du \times dv)$. Тогда можно записать две другие важные характеристики: *ковариацию* и *коэффициент корреляции*:

$$\operatorname{cov}(\xi_1, \xi_2) = \iint_{\mathbb{R}^2} (u - \mathsf{m}_1(u))(v - \mathsf{m}_1(v))\mathcal{P}(du \times dv) \qquad \widehat{\operatorname{cov}}(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})(y_i - \bar{\mathbf{y}})$$

$$\operatorname{cor}(\xi_1, \xi_2) = \frac{\operatorname{cov}(\xi_1, \xi_2)}{\sigma_{\xi_1}\sigma_{\xi_2}} \qquad \widehat{\operatorname{cor}}(\mathbf{x}, \mathbf{y}) = \frac{\widehat{\operatorname{cov}}(\mathbf{x}, \mathbf{y})}{s(\mathbf{x})s(\mathbf{y})}.$$

Замечание (Важное). $\xi_1 \not\parallel \xi_2 \implies \operatorname{cov}(\xi_1, \xi_2)$, но $\operatorname{cov}(\xi_1, \xi_2) \not\Longrightarrow \xi_1 \not\parallel \xi_2$. Необходимость и достаточность выполняется только в случае нормального распределения.

Замечание (Проблема моментов). Для заданной последовательности моментов $\mathsf{m}_1, \mathsf{m}_2, \ldots$ не обязано существовать подходящее распределение. Помимо требований $\mathsf{m}_{2k} \ge 0$ и взаимосвязи между соседними моментами по неравенству Гёльдера, существенно, что ряд Тейлора по m_ℓ, в который, как известно, раскладывается характеристическая функция, должен сходиться равномерно.

2.3. Свойства оценок

2.3.1. Несмещенность

Определение. *Смещение*[3] есть

$$\operatorname{bias}\hat{\theta}_n := \mathsf{E}\hat{\theta}_n - \theta \quad \forall \theta \in \Theta.$$

Определение. *Среднеквадратичная ошибка*[4] есть

$$\operatorname{MSE}\hat{\theta}_n := \mathsf{E}(\hat{\theta}_n - \theta)^2.$$

[2] «Skewness».
[3] Bias.
[4] Mean square error (MSE).

Замечание. Поскольку

$$\mathsf{D}\hat{\theta}_n = \mathsf{D}(\hat{\theta}_n - \theta) = \mathsf{E}(\hat{\theta}_n - \theta)^2 - (\mathsf{E}(\hat{\theta}_n - \theta))^2,$$

то

$$\underbrace{\mathsf{E}(\hat{\theta}_n - \theta)^2}_{\text{MSE}} = \mathsf{D}\hat{\theta}_n + \underbrace{(\mathsf{E}(\hat{\theta}_n - \theta))^2}_{\text{bias}^2}.$$

Определение. Оценка называется *несмещенной*, если $\operatorname{bias}\hat{\theta}_n = 0$, т.е.

$$\mathsf{E}\hat{\theta}_n = \theta.$$

Пример. $\bar{\mathbf{x}}$ — несмещенная оценка $\mathsf{E}\xi$.

Доказательство. Пусть $\theta = \mathsf{E}\xi$, $\hat{\theta}_n = \mathsf{E}\hat{\xi}_n = \bar{\mathbf{x}}$. Тогда

$$\mathsf{E}\bar{\mathbf{x}} = \mathsf{E}\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}x_i = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}\xi = \mathsf{E}\xi \implies \mathsf{E}\hat{\theta}_n = \mathsf{E}\theta,\ \operatorname{bias}\hat{\theta}_n = 0.$$

□

Пример. s^2 является только *асимптотически* несмещенной оценкой $\mathsf{D}\xi$.

Доказательство. Поскольку дисперсия не зависит от сдвига, обозначим $\eta = \xi - \mathsf{E}\xi$ и $y_i = x_i - \mathsf{E}\xi$; тогда

$$\begin{aligned}
\mathsf{E}s^2 &= \mathsf{E}\widehat{\mathsf{D}\xi} = \mathsf{E}\widehat{\mathsf{D}\eta} = \mathsf{E}\left(\widehat{\mathsf{E}\eta^2} - \left(\widehat{\mathsf{E}\eta}\right)^2\right) = \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2\right) \\
&= \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{1}{n^2}\sum_{i,j=1}^{n} y_i y_j\right) = \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{1}{n^2}\sum_{i,j=1}^{n} y_i^2\right) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}y_i^2 - \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{E}y_i^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}\left(x_i - \mathsf{E}\xi\right)^2 - \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{E}\left(x_i - \mathsf{E}\xi\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\mathsf{D}x_i - \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{D}x_i = \mathsf{D}\xi - \frac{1}{n}\mathsf{D}\xi \\
&= \frac{n-1}{n}\mathsf{D}\xi \xrightarrow[n\to\infty]{} \mathsf{D}\xi.
\end{aligned}$$

□

Определение. *Исправленная* дисперсия:

$$\tilde{s}^2 := \frac{n}{n-1}s^2.$$

2.3.2. Состоятельность

Определение. Оценка называется *состоятельной в среднеквадратичном смысле*, если

$$\operatorname{MSE}\hat{\theta}_n \xrightarrow[n\to\infty]{} 0.$$

Определение. Оценка называется *состоятельной*, если

$$\hat{\theta}_n \xrightarrow{\mathsf{P}} \theta.$$

Предложение. *Если оценка несмещенная и состоятельная в среднеквадратичном смысле, то она состоятельная.*

Доказательство. В самом деле, по неравенству Чебышева,

$$\mathsf{P}(|\hat{\theta}_n - \theta| > \epsilon) = \mathsf{P}(|\hat{\theta}_n - \mathsf{E}\hat{\theta}_n| > \epsilon) \leq \frac{\mathsf{D}\hat{\theta}_n}{\epsilon^2} = \frac{\operatorname{MSE}\hat{\theta}_n}{\epsilon^2} \xrightarrow[n\to\infty]{} 0.$$

□

Пример. $\hat{\mathsf{m}}_k$ является состоятельной оценкой m_k.

Доказательство. Докажем для $\hat{\mathsf{m}}_1$. По определению выборки до эксперимента, $x_i \sim \mathcal{P}$. Тогда, по теореме Хинчина о ЗБЧ,

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^n x_i}{n} \xrightarrow{\mathsf{P}} \mathsf{m}_1(\mathcal{P}).$$

Для k-го момента доказывается аналогично заменой $y_i := x_i^k$. □

Замечание. Для $\mathsf{m}_k^{(0)}$ доказательство не пройдет, потому что x_i и $\bar{\mathbf{x}}$ не будут независимыми.

Пример. $\hat{\mathsf{m}}_k^{(0)}$ является состоятельной оценкой $\mathsf{m}_k^{(0)}$.

Утверждение. Пусть $\xi_n \xrightarrow{\mathsf{P}} c$ и $f \in C(U_\epsilon(c))$. Тогда $f(\xi_n) \xrightarrow{\mathsf{P}} f(c)$.

Доказательство предложения. Докажем для s^2. Пусть $f : (x, y) \mapsto x - y^2$. Устроим последовательность $(\hat{\mathsf{m}}_2, \hat{\mathsf{m}}_1) \xrightarrow{\mathsf{P}} (\mathsf{m}_2, \mathsf{m}_1)$. Тогда

$$f(\hat{\mathsf{m}}_2, \hat{\mathsf{m}}_1) = \hat{\mathsf{m}}_2 - \hat{\mathsf{m}}_1^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{\mathbf{x}}^2 = s^2 \xrightarrow{\mathsf{P}} f(\mathsf{m}_2, \mathsf{m}_1) = \mathsf{D}\xi.$$

Для $\mathsf{m}_k^{(0)}$ доказывается аналогично. □

Пример. $\bar{\mathbf{x}}$ — состоятельная оценка $\mathsf{E}\xi$.

Доказательство. Либо по (2.3.2) для $k = 1$, либо из того факта, что $\operatorname{bias}\bar{\mathbf{x}} = 0$, значит

$$\operatorname{MSE}\bar{\mathbf{x}} = \mathsf{D}\bar{\mathbf{x}} = \frac{\mathsf{D}\xi}{n} \xrightarrow[n\to\infty]{} 0,$$

и по (2.3.2) получаем утверждение. □

Пример. s^2 — состоятельная оценка $\mathsf{D}\xi$.

Доказательство. По (2.3.2) с $k = 2$. □

Пример. $\widehat{\operatorname{cdf}}_\xi$ — состоятельная оценка cdf в каждой точке.

Доказательство. Введем

$$y_i := \mathbf{1}_{\{x_i < x\}} = \begin{cases} 1 & x_i < x \\ 0 & x_i \geq x. \end{cases}$$

Тогда по теореме Хинчина о ЗБЧ,

$$\widehat{\operatorname{cdf}}_\xi(x) = \frac{|\{x_i \in \bar{\mathbf{x}} : x_i < x\}|}{n} = \frac{\sum_{i=1}^n y_i}{n} \xrightarrow{\mathsf{P}} \mathsf{E}y_i = \mathsf{E}\mathbf{1}_{\{x_i<x\}} = \mathsf{P}(x_i < x) = \operatorname{cdf}(x).$$

Независимость y_i очевидна, если расписать $\operatorname{cdf}_{y_i}(x)\operatorname{cdf}_{y_j}(y) = \operatorname{cdf}_{y_i,y_j}(x, y)$. □

Замечание. Помимо ранее упомянутых, больше нет состоятельных выборочных характеристик.

Утверждение. Пусть $\exists! p_0 : \operatorname{cdf}(x) = p_0$ и $\operatorname{cdf}(x)$ монотонно возрастает в окрестности p_0. Тогда $\bar{z}_{p_0} \xrightarrow{\mathsf{P}} z_{p_0}$, т.е. является состоятельной оценкой.

2.3.3. Асимптотическая нормальность

Определение. Оценка называется *асимптотически нормальной*, если

$$\frac{\hat{\theta}_n - \mathsf{E}\hat{\theta}_n}{\sqrt{\mathsf{D}\hat{\theta}_n}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1).$$

Замечание. Если $\xi \sim \mathrm{N}(a, \sigma^2)$, то $\bar{\mathbf{x}}$ — просто нормальная оценка (как линейная комбинация нормальных случайных величин).

Все рассмотренные прежде оценки — асимптотически нормальные.

2.3.4. Эффективность

Определение. Соответствующая определению 2.5.2 оценка называется *эффективной*.

Определение. Оценка $\hat{\theta}^{(1)}$ называется *эффективной* в сравнении с $\hat{\theta}^{(2)}$, если

$$\mathrm{MSE}\,\hat{\theta}^{(1)} < \mathrm{MSE}\,\hat{\theta}^{(2)}.$$

Замечание. Для несмещенных оценок это эквивалентно, конечно,

$$\mathsf{D}\hat{\theta}^{(1)} < \mathsf{D}\hat{\theta}^{(2)}.$$

Пример (Сравнение оценок мат. ожидания симметричного распределения)**.** Пусть $\mathcal{P}$ *симметрично* — в этом случае $\widehat{\operatorname{med}\xi} = \bar{\mathbf{x}}$ и имеет смысл сравнить две этих характеристики.

$$\begin{aligned}
\mathsf{D}\bar{\mathbf{x}} &= \frac{\mathsf{D}\xi}{n} \\
\mathsf{D}\widehat{\operatorname{med}\xi} &\sim \frac{1}{4n\,\mathrm{pdf}^2_{\mathrm{N}(a,\sigma^2)}(\operatorname{med}\xi)} \quad \text{при } n \to \infty.
\end{aligned}$$

Так, если $\xi \sim \mathrm{N}(a, \sigma^2)$, то

$$\mathrm{pdf}^2_{\mathrm{N}(a,\sigma^2)}(\operatorname{med}\xi) = \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{(\operatorname{med}\xi - a)^2}{\sigma^2}\right\} = \frac{1}{2\pi\sigma^2},$$

откуда

$$\mathsf{D}\widehat{\operatorname{med}\xi} = \frac{\pi}{2}\frac{\sigma^2}{n} > \frac{\sigma^2}{n} = \mathsf{D}\bar{\mathbf{x}},$$

значит $\widehat{\operatorname{med}\xi}$ эффективнее $\bar{\mathbf{x}}$.

Замечание. В то же время, $\widehat{\operatorname{med}\xi}$ более устойчива к аутлаерам, чем $\bar{\mathbf{x}}$, и этим лучше.

2.4. Метод моментов

Пусть $\mathcal{P}(\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^{\mathrm{T}}$ — параметрическая модель. Найдем оценки для параметров $\hat{\theta}_i$, $i \in \overline{1:r}$, для чего составим и решим систему уравнений:

$$\begin{cases}
\mathsf{m}_1 = \phi_1(\theta_1, \ldots, \theta_r) \\
\quad\vdots \\
\mathsf{m}_r = \phi_r(\theta_1, \ldots, \theta_r)
\end{cases}
\Longrightarrow
\begin{cases}
\theta_1 = f_1(\mathsf{m}_1, \ldots, \mathsf{m}_r) \\
\quad\vdots \\
\theta_r = f_r(\mathsf{m}_1, \ldots, \mathsf{m}_r).
\end{cases}$$

Примем

$$\hat{\theta}_i = f_i(\hat{\mathsf{m}}_1, \ldots, \hat{\mathsf{m}}_r).$$

Замечание. Поскольку f_i — непрерывные функции и при непрерывных преобразованиях сходимость не портится, оценки $\hat{\theta}_i$ являются состоятельными.

Как правило, эти оценки смещенные.

Замечание. Случается, что решение находится вне пространства параметров. На практике, если пространство параметров компактное, можно взять точку, ближайшую к полученной оценке. Однако это свидетельствует о том, что модель плохо соответствует данным.

Пример 1 $(r=1)$**.** Пусть $\mathcal{P}_\xi(\lambda) = \mathrm{Exp}(\lambda)$. Тогда $\mathsf{E}\xi = 1/\lambda$ и $\bar{\mathbf{x}} = 1/\lambda$.

Пример 2 $(r=2)$**.** Пусть $\mathcal{P}_\xi(\theta_1, \theta_2) = \mathrm{Bin}(m, p)$. Тогда

$$\begin{cases} \mathsf{E}\xi = mp \\ \mathsf{D}\xi = mp(1-p) \end{cases} \quad \begin{cases} m = \dfrac{\mathsf{E}\xi}{p} \\ \mathsf{D}\xi = \mathsf{E}\xi - \mathsf{E}\xi p \end{cases} \quad \begin{cases} p = \dfrac{\mathsf{E}\xi - \mathsf{D}\xi}{\mathsf{E}\xi} \\ m = \dfrac{(\mathsf{E}\xi)^2}{\mathsf{E}\xi - \mathsf{D}\xi} \end{cases} \implies \begin{cases} \hat{p} = \dfrac{\bar{\mathbf{x}} - s^2}{\bar{\mathbf{x}}} \\ \hat{m} = \dfrac{\bar{\mathbf{x}}^2}{\bar{\mathbf{x}} - s^2}. \end{cases}$$

2.5. Метод оценки максимального правдоподобия

2.5.1. Функция правдоподобия и метод

Пусть $\mathcal{P}_\xi(\boldsymbol{\theta}),\ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^{\mathrm{T}}$ — параметрическая модель.

Определение. *Функция правдоподобия:*

$$\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) = \mathsf{P}(\mathbf{y} \mid \boldsymbol{\theta}) = \begin{cases} \mathsf{P}_{\boldsymbol{\theta}}(x_1 = y_1, \ldots, x_n = y_n) & \mathcal{P}_\xi(\boldsymbol{\theta}) \text{ дискретно;} \\ p_{\boldsymbol{\theta}}(\mathbf{y}) & \mathcal{P}_\xi(\boldsymbol{\theta}) \text{ абсолютно непрерывно.} \end{cases}$$

Пример 3. Пусть $\xi \sim \mathrm{N}(a, \sigma^2)$. По независимости x_i, $p_{\boldsymbol{\theta}}(\mathbf{x})$ распадается в произведение:

$$\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}) = p_{\boldsymbol{\theta}}(\mathbf{x}) = \prod_{i=1}^n p_{\boldsymbol{\theta}}(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i - a)^2}{2\sigma^2}\right\} = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - a)^2\right\}.$$

Пример 4. $\xi \sim \mathrm{Pois}(\lambda)$,

$$\mathsf{P}(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda} \implies \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}) = \prod_{i=1}^n \frac{1}{x_i!} \lambda^{x_i} e^{-\lambda} = \frac{1}{\prod_{i=1}^n x_i!} \lambda^{n\bar{\mathbf{x}}} e^{-n\lambda}.$$

Утверждение. Пусть $\mathbf{x}$ — выборка. В качестве оценки максимального правдоподобия[5] $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ следует взять

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \ln \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}).$$

Предложение. $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ *является состоятельной оценкой.*

Доказательство. Пусть $\boldsymbol{\theta}_0$ — истинный параметр $\mathcal{P}(\boldsymbol{\theta})$. По УЗБЧ,

$$\frac{1}{n} \ln \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \ln p_{\boldsymbol{\theta}}(x_i) \xrightarrow{\mathsf{P}} \mathsf{E} \ln p_{\boldsymbol{\theta}}(x_i) = \int_{\mathbb{R}} \ln\left(p_{\boldsymbol{\theta}}(x)\right) p_{\boldsymbol{\theta}_0}(x)\, \mathrm{d}x.$$

Навесим на обе стороны argmax в условии, что это непрерывное преобразование:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} \leftarrow \operatorname*{argmax}_{\boldsymbol{\theta}} \frac{1}{n} \ln \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}) \xrightarrow{\mathsf{P}} \operatorname*{argmax}_{\boldsymbol{\theta}} \int_{\mathbb{R}} \ln\left(p_{\boldsymbol{\theta}}(x)\right) p_{\boldsymbol{\theta}_0}(x)\, \mathrm{d}x \to \boldsymbol{\theta}^*.$$

Тогда в предположении непрерывности $p_{\boldsymbol{\theta}}$ по $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} \xrightarrow{\mathsf{P}} \boldsymbol{\theta}^*$. Покажем, что $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$. Поделим на $p_{\boldsymbol{\theta}_0}$ — константу по $\boldsymbol{\theta}$:

$$\frac{1}{n} \sum_{i=1}^n \ln \frac{p_{\boldsymbol{\theta}}}{p_{\boldsymbol{\theta}_0}}(x_i) \xrightarrow{\mathsf{P}} \int_{\mathbb{R}} \ln\left(\frac{p_{\boldsymbol{\theta}}(x)}{p_{\boldsymbol{\theta}_0}(x)}\right) p_{\boldsymbol{\theta}_0}(x)\, \mathrm{d}x = \mathsf{E} \ln \frac{p_{\boldsymbol{\theta}}}{p_{\boldsymbol{\theta}_0}} \le \ln \mathsf{E} \frac{p_{\boldsymbol{\theta}}}{p_{\boldsymbol{\theta}_0}} = \ln \int_{\mathbb{R}} \frac{p_{\boldsymbol{\theta}}}{p_{\boldsymbol{\theta}_0}}(x) p_{\boldsymbol{\theta}_0}(x)\, \mathrm{d}x = \ln 1 = 0$$

[5] Maximum likelihood estimate (MLE).

по неравенству Єнсена $\mathsf{E}g(\xi) \le g(\mathsf{E}\xi)$, для выпуклой вверх $g(x) = \log(x)$. Таким образом,

$$\int_{\mathbb{R}} \ln\left(\frac{p_{\boldsymbol{\theta}}(x)}{p_{\boldsymbol{\theta}_0}(x)}\right) p_{\boldsymbol{\theta}_0}(x)\,\mathrm{d}x = 0 \iff \ln\left(\frac{p_{\boldsymbol{\theta}}(x)}{p_{\boldsymbol{\theta}_0}(x)}\right) = 0 \iff \frac{p_{\boldsymbol{\theta}}(x)}{p_{\boldsymbol{\theta}_0}(x)} = 1 \iff p_{\boldsymbol{\theta}}(x) = p_{\boldsymbol{\theta}_0}(x).$$

В предположении свойства *идентифицируемости* задачи ($\theta_1 \ne \theta_2 \implies \mathcal{P}_{\theta_1} \ne \mathcal{P}_{\theta_2}$), получаем $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. □

Пример. $\xi \sim \mathrm{Pois}(\lambda)$.

$$\ln \mathcal{L}(\lambda \mid \mathbf{x}) = -\sum_{i=1}^{n} x_i! - n\lambda + \ln(\lambda^{n\bar{\mathbf{x}}}) \implies \frac{\mathrm{d}\ln\mathcal{L}(\lambda \mid \mathbf{x})}{\mathrm{d}\lambda} = -n + \lambda^{-n\bar{\mathbf{x}}} n\bar{\mathbf{x}}\lambda^{n\bar{\mathbf{x}}-1} = -n + \frac{n\bar{\mathbf{x}}}{\lambda}$$

откуда

$$\frac{\mathrm{d}\ln\mathcal{L}(\lambda \mid \mathbf{x})}{\mathrm{d}\lambda} = 0 \iff -n + \frac{n\bar{\mathbf{x}}}{\lambda} = 0,\ n\bar{\mathbf{x}} - n\lambda = 0,\ \lambda = \bar{\mathbf{x}}.$$

Утверждение. В условиях регулярности[6]:

1. Существует один глобальный максимум, так что
$$\left.\frac{\mathrm{d}\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x})}{\mathrm{d}\boldsymbol{\theta}}\right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\mathrm{MLE}}} = 0.$$
2. $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ обладает всеми свойствами:
    a) Состоятельность;
    b) Асимптотическая несмещенность;
    c) Асимптотическая нормальность;
    d) Эффективность.

2.5.2. Информационное количество Фишера и эффективность оценки MLE

Пусть $r = 1$.

Определение. *Информанта n-го порядка*:

$$S_n(\mathbf{x}, \theta) = \frac{\mathrm{d}^n \ln \mathcal{L}(\theta \mid \mathbf{x})}{\mathrm{d}\theta^n}.$$

Определение. *Информационное количество Фишера*:

$$I_n(\theta) := -\mathsf{E}S_2(\mathbf{x}, \theta).$$

Утверждение.

$$I_n(\theta) = \mathsf{E}S_1^2(\mathbf{x}, \theta).$$

Пример. $\xi \sim \mathrm{Pois}(\lambda)$.

$$S_1(\mathbf{x}, \theta) = -n + \frac{n\bar{\mathbf{x}}}{\lambda}, \quad S_2(\mathbf{x}, \theta) = -\frac{n\bar{\mathbf{x}}}{\lambda^2} \implies I_n(\lambda) = \mathsf{E}\frac{n\bar{\mathbf{x}}}{\lambda^2} = \frac{n}{\lambda^2}\mathsf{E}\bar{\mathbf{x}} = \frac{n}{\lambda}.$$

[6] Условия регулярности таковы:
- $\{x : p_\theta(x) > 0\}$ не зависит от θ,
- $\int_{V^n} \hat{\theta}\mathcal{L}(\theta \mid \mathbf{x})\,\mathrm{d}\mathbf{x}$ и $\int_{V^n} \mathcal{L}(\theta \mid \mathbf{x})\,\mathrm{d}\mathbf{x}$ можно дифференцировать по θ под знаком интеграла,
- $I_n(\theta) > 0$.

Замечание.
$$\ln \mathcal{L}(\theta \mid \mathbf{x}) = \sum_{i=1}^{n} \ln p_\theta(x_i) \implies S_2 = \frac{\mathrm{d}^2 \ln \mathcal{L}(\theta \mid \mathbf{x})}{\mathrm{d}\theta^2} = \sum_{i=1}^{n} (\ln p_\theta(x_i))'',$$
откуда, для повторной независимой выборки,
$$I_n(\theta) = -\sum_{i=1}^{n} \mathsf{E}(\ln p_\theta(x_i))'' = n \cdot i(\theta), \quad \text{где } i(\theta) = -\mathsf{E}(\ln p_\theta(\xi))''.$$

Утверждение. Для несмещенных оценок в условиях регулярности справедливо неравенство *Рао–Крамера*:
$$\mathsf{D}\hat{\theta}_n \geq \frac{1}{I_n(\theta)}.$$

Определение. *Эффективная* оценка:
$$\mathsf{D}\hat{\theta}_n = \frac{1}{I_n(\theta)}.$$

Замечание. Для несмещенных оценок неравенство Рао–Крамера указывает точную нижнюю границу дисперсий оценок.

Пример. Пусть $\xi \sim \mathrm{Pois}(\lambda)$. Поскольку
$$\begin{aligned} \mathsf{D}\hat{\lambda}_n &= \mathsf{D}\bar{\mathbf{x}} = \mathsf{E}\xi/n = \lambda/n \\ I_n(\lambda) &= n/\lambda, \end{aligned}$$
то $\hat{\lambda}_n$ — эффективная оценка (как и ожидалась по свойствам $\hat{\theta}_{\mathrm{MLE}}$).

Определение. Пусть $\hat{\theta}_n$ — асимптотически несмещенная оценка. Тогда $\hat{\theta}_n$ — *асимптотически эффективная*, если
$$\mathsf{D}\hat{\theta}_n \cdot I_n \xrightarrow[n\to\infty]{} 1.$$

Пример. Пусть $\xi \sim \mathrm{N}(a, \sigma^2)$. Можно посчитать, что s^2 является только асимптотически эффективной оценкой σ^2; $\tilde{s}^2$ — просто эффектвной.

3. Некоторые распределения, связанные с нормальным

3.1. Распределение $\chi^2(m)$

Определение (Распределение $\chi^2(m)$). η имеет распределение χ^2 с m степенями свободны:
$$\eta \sim \chi^2(m) \iff \eta = \sum_{i=1}^{m} \zeta_i^2, \quad \zeta_i \sim \mathrm{N}(0,1),\ \zeta_i \text{ независимы}.$$

Свойства[7] $\chi^2(m)$

$$\begin{aligned} \mathsf{E}\eta &= \sum_{i=1}^{m} \mathsf{E}\beta_i^2 = m \\ \mathsf{D}\eta &= 2m. \end{aligned}$$

Утверждение. Пусть $\eta_m \sim \chi^2(m)$. Тогда, по ЦПТ,
$$\frac{\eta_m - \mathsf{E}\eta_m}{\sqrt{\mathsf{D}\eta_m}} = \frac{\eta_m - m}{\sqrt{2m}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1).$$

Пример. $m = 50$, $\eta_m = 80$. Тогда
$$\frac{80-50}{10} = 3$$
и можно посчитать, к примеру, $\Phi(3) \approx 0.9986$.

[7] Вычисление $\mathsf{D}\eta$: https://www.statlect.com/probability-distributions/chi-square-distribution

3.2. Распределение Стьюдента $\mathrm{t}(m)$

Определение (Распределение $\mathrm{t}(m)$)**.** ξ имеет распределение Стьюдента с m степенями свободы, если

$$\xi \sim \mathrm{t}(m) \iff \xi = \frac{\zeta}{\sqrt{\eta/m}}, \quad \zeta \sim \mathrm{N}(0,1),\ \eta \sim \chi^2(m).$$

Свойства $\mathrm{t}(m)$

- При $m = 1$ это распределение Коши.
- При $m > 1$, $\mathsf{E}\zeta = 0$ по симметричности.
- При $m > 2$, $\mathsf{D}\zeta = m/(m-2)$.
- При $m > 3$, $\mathsf{A}\zeta = 0$ по симметричности.
- При $m > 4$, $\mathsf{K}\zeta = 6/(m-4)$.

Предложение. *Распределение Стьюдента сходится к стандартному нормальному:*

$$\mathrm{t} \Rightarrow \mathrm{N}(0,1).$$

Соображения по поводу. $\mathsf{D}\zeta \to 1,\ \mathsf{K}\zeta \to 0.$ □

3.3. Распределение Фишера

Определение. Распределение Фишера имеет вид

$$\mathrm{F}(m,k) = \frac{\chi^2(m)/m}{\chi^2(k)/k}.$$

Замечание. $\mathrm{F}(1,k) \sim \mathrm{t}^2(k)$; $m \cdot \mathrm{F}(m,\infty) = \chi^2(m)$ потому что $\chi^2(k) \xrightarrow[k\to\infty]{} 1$.

3.4. Квадратичные формы от нормально распределенных случайных величин

Пусть $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)^{\mathrm{T}} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_p)$, $\mathbf{B}$ — симметричная, неотрицательно определенная матрица. Найдем распределение $\boldsymbol{\xi}^T\mathbf{B}\boldsymbol{\xi}$.

Утверждение. Пусть $\boldsymbol{\xi} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_p)$, $\mathbf{B}, \mathbf{C}$ — симметричные матрицы размерности $p \times p$. Тогда $\boldsymbol{\xi}^{\mathrm{T}}\mathbf{B}\boldsymbol{\xi} \perp\!\!\!\perp \boldsymbol{\xi}^{\mathrm{T}}\mathbf{C}\boldsymbol{\xi} \iff \mathbf{BC} = \mathbf{0}$.

Пример (Независимость $\bar{\mathbf{x}}^2$ и s^2)**.** Пусть

$$\mathbf{B} = \frac{1}{p}\begin{pmatrix} 1 & \ldots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \ldots & 1 \end{pmatrix};$$

тогда

$$\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x} = \frac{1}{p}(x_1, \ldots, x_p)\begin{pmatrix} \sum_{i=1}^p x_i \\ \vdots \\ \sum_{i=1}^p x_i \end{pmatrix} = \frac{1}{p}\sum_{j=1}^p\sum_{i=1}^p x_i x_j = \frac{1}{p}\left(\sum_{i=1}^p x_i\right)^2 = p\bar{\mathbf{x}}^2.$$

В то же время, пусть

$$\mathbf{C} = \begin{pmatrix} 1-1/p & \ldots & -1/p \\ \vdots & \ddots & \vdots \\ -1/p & \ldots & 1-1/p \end{pmatrix};$$

тогда

$$\begin{aligned}\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x} &= (x_1,\ldots,x_p)\begin{pmatrix}(1-1/p)\,x_1 - 1/p\cdot\sum_{i=2}^{p} x_i\\ \vdots\\ -1/p\cdot\sum_{i=1}^{p-1} x_i + (1-1/p)\,x_p\end{pmatrix} = (x_1,\ldots,x_p)\begin{pmatrix}x_1 - 1/p\cdot\sum_{i=1}^{p} x_i\\ \vdots\\ x_p - 1/p\cdot\sum_{i=1}^{p} x_i\end{pmatrix}\\ &= \sum_{j=1}^{p}\left(x_j^2 - \frac{1}{p}\sum_{i=1}^{p} x_i x_j\right) = \sum_{j=1}^{p} x_j^2 - p\bar{\mathbf{x}}^2 = ps^2.\end{aligned}$$

Но, учитывая $\mathbf{B}^2 = \mathbf{B}$ (легко проверяется),

$$\mathbf{C} = \mathbf{I}_p - \mathbf{B} \implies \mathbf{B}\mathbf{C} = \mathbf{B} - \mathbf{B}^2 = \mathbf{0},$$

откуда $\bar{\mathbf{x}}^2 \perp\!\!\!\perp s^2$.

Видно, что $\sigma^{-2}\boldsymbol{\xi}^{\mathrm{T}}\mathbf{I}_p\boldsymbol{\xi} \sim \chi^2(p)$. На самом деле справедливо

Утверждение. Пусть $\boldsymbol{\xi} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_p)$, $\mathbf{B}$ — симметричная, неотрицательно неопределенная матрица размерности $p \times p$ и $\operatorname{rk}\mathbf{B} = r$. Тогда

$$\sigma^{-2}\boldsymbol{\xi}^{\mathrm{T}}\mathbf{B}\boldsymbol{\xi} \sim \chi^2(r) \iff \mathbf{B}^2 = \mathbf{B}.$$

Пример. $n\sigma^{-2}s^2 \sim \chi^2(p-1)$. Воспользуемся представлением из предыдущего примера: $ps^2 = \mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x}$. Но $\operatorname{rk}\mathbf{C} = \operatorname{rk}(\mathbf{I}_p - \mathbf{B}) = p-1$; $\mathbf{B}^2 = \mathbf{B}$, значит $p\sigma^{-2}s^2 \sim \chi^2(p-1)$.

Утверждение (Cochran). Пусть $\boldsymbol{\xi} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_p)$, $\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi} = \sum_i Q_i$, где Q_i — квадратичная форма, заданная $\mathbf{B}_i$, $\operatorname{rk}\mathbf{B}_i = r_i$. Тогда следующие утверждения эквивалентны:

1. $\sum r_i = p$
2. $Q_i \sim \chi^2(r_i)$
3. $Q_i \perp\!\!\!\perp Q_j, \quad \forall i \neq j$, т.е. $\mathbf{B}_i\mathbf{B}_j = \mathbf{0}$.

4. Проверка гипотез

Замечание. Этот раздел иногда называется «Confirmatory Data Analysis» в противовес «Exploratory Data Analysis», не включающему в себя понятие *гипотезы*.

4.1. Построение критерия

Пусть $x_1, \ldots, x_n \sim \mathcal{P}$.

Определение. *Модель* — утверждение о $\mathcal{P}$, которое считается верным и не проверяется.

Определение. *Гипотеза* — утверждение о $\mathcal{P}$, требующее проверки. Гипотеза называется *простой*, если она соответствует только одному распределению в рамках рассматриваемой модели, или — в противном случае — *сложной*.

Определение. *Нулевая гипотеза* H_0 — утверждение об интересующем экспериментатора свойстве распределений.

Замечание. Обыкновенно, H_0 — простая гипотеза.

Помимо H_0, может вводиться *альтернативная гипотеза* H_1, учитывающая отклонения от H_0, обнаружение которых желательно. Обыкновенно, это сложная гипотеза.

Определение. Пусть $\mathscr{P}$ — множество всех распределений. Поставленная H_0 задает разбиение $\mathscr{P} = \mathscr{P}_0 \sqcup \mathscr{P}_1$, где $\mathscr{P}_0$ — распределения, удовлетворяющие утверждению. Стоит задача проверить по выборке, что $\mathcal{P} \in \mathscr{P}_0$, т.е. пусть задан «оракул»

$$\varphi(\mathbf{x}) = \begin{cases} 0 & H_0 \text{ верна}, \mathcal{P} \in \mathscr{P}_0 \\ 1 & H_0 \text{ не верна}, \mathcal{P} \in \mathscr{P}_1. \end{cases}$$

Будем говорить, что гипотеза *отвергается*, если $\varphi(\mathbf{x}) = 1$ и *не отвергается* иначе.

Поскольку выборка конечного объема позволяет делать только вероятностные заключения, со статистическим критерием ассоциированы ошибки двух родов:

Определение. *Ошибка I-го рода* есть отвержение верной H_0*:*

$$H_0 \text{ верна, но } \varphi(\mathbf{x}) = 1.$$

Иными словами, это принятие случайного различия за систематическое.

Соответствующая вероятность обозначается

$$\alpha_{\mathrm{I}} = \mathsf{P}(\varphi(\mathbf{x}) = 1 \mid \mathcal{P} \in \mathscr{P}_0) = \mathsf{P}_{H_0}(\varphi(\mathbf{x}) = 1).$$

Определение. *Ошибка II-го рода* есть вероятность ошибочно не отвергнуть H_0*:*

$$H_0 \text{ не верна, но } \varphi(\mathbf{x}) = 0.$$

Иными словами, это принятие наблюдаемого различия за случайный эффект.

Соответствующая вероятность обозначается

$$\alpha_{\mathrm{II}} = \mathsf{P}(\varphi(\mathbf{x}) = 0 \mid \mathcal{P} \in \mathscr{P}_1) = \mathsf{P}_{H_1}(\varphi(\mathbf{x}) = 0).$$

Определение. *Мощность* критерия против альтернативы это вероятность справедливо отвергнуть H_0:

$$\beta = 1 - \alpha_{\mathrm{II}} = 1 - \mathsf{P}_{H_1}(\varphi(\mathbf{x}) = 0) = \mathsf{P}_{H_1}(\varphi(\mathbf{x}) = 1).$$

Иными словами, это способность критерия отличать H_0 от H_1.

Определение. Критерий называется *состоятельным*, если $\beta \to 1$.

Определение. До эксперимента фиксируется величина $\alpha \in [0, 1]$ называемая *уровнем значимости* критерия. Неформально, α обратно пропорциональна «строгости» критерия, выбираемой экспериментатором.

Замечание. Стандартные уровни значимости: $\alpha = 0.05$ или $\alpha = 0.01$.

Замечание. В идеале, хотят, чтобы $\alpha_{\mathrm{I}}, \alpha_{\mathrm{II}} \xrightarrow[n \to \infty]{} 0$. Однако отрицание H_0 порождает распределение, которое сложно вычислить.

Определение. Для выборки

$$\mathbf{x} = (x_1, \dots, x_n) : (\Omega, \mathcal{F}, \mathsf{P}) \to (V^n, \mathscr{A}^n)$$

критерий есть разбиение всевозможных значений выборки на доверительную и критическую области

$$V^n = \mathscr{A}^{(\alpha)}_{\text{крит}} \sqcup \mathscr{A}^{(\alpha)}_{\text{дов}}$$

с учетом желаемого уровня значимости α, т.е. такое, что

$$\mathsf{P}_{H_0}(\mathbf{x} \in \mathscr{A}^{(\alpha)}_{\text{крит}}) = \mathsf{P}_{H_0}(\varphi(\mathbf{x}) = 1) = \alpha_{\mathrm{I}} = \alpha.$$

Разумеется, $\mathscr{A}^{(\alpha)}_{\text{крит}} = \{\mathbf{x} \in V^n : \varphi(\mathbf{x}) = 1\}$ и $\mathscr{A}^{(\alpha)}_{\text{дов}} = \{\mathbf{x} \in V^n : \varphi(\mathbf{x}) = 0\}$.

Определение. Если

$\alpha = \alpha_{\mathrm{I}}$ то критерий называется *точным*,

$\alpha \xrightarrow[n\to\infty]{} \alpha_{\mathrm{I}}$ *асимптотическим*,

$\alpha < \alpha_{\mathrm{I}}$ *радикальным* (т.е. отвергает гипотезу чаще, чем точный),

$\alpha > \alpha_{\mathrm{I}}$ *консервативным* (если гипотеза отвергнута, то уж наверняка).

Замечание. $\mathscr{A}_{\text{крит}}$ может быть корректно построена бесчисленным множеством способов. Следует выбрать тот из них, что гарантирует наибольшую мощность.

Замечание. Задача может допускать две постановки; в этом случае, поскольку α_{I} ($=\alpha$ для правильно построенного критерия) контролируется экспериментатором, проверяется отрицание эффекта, который хотят подтвердить: к примеру, что новое лекарство *не* лучше старого; если H_0 отвергнется, это будет означать, что новое лекарство-таки лучше старого с вероятностью не ниже, чем $1-\alpha$.

Пример (Экстрасенс угадывает карты)**.** Некто заявляет о способности не глядя угадывать масти предлагаемых карт. Выдвигаем гипотезу H_0: человек *не* экстрасенс, так что угадывает с вероятностью $p = 1/4$. Предлагаем отгадать масть 25 карт; число угаданных карт есть x. Наибольшей «строгости» будет соответствовать $x = 25$, таким образом

$$\mathsf{P}_{H_0}(x \in \mathscr{A}_{\text{крит}}) = \mathsf{P}_{H_0}(x = 25) = \left(\frac{1}{4}\right)^{25} \approx 10^{-15} = \alpha.$$

Если же для нас достаточно 24 угаданных карт из 25, то α будет побольше: $\mathsf{P}_{H_0}(x \geq 24) = \mathsf{P}_{H_0}(x = 24) + \mathsf{P}_{H_0}(x = 25)$ и т.д. Зафиксированная до эксперимента α определяет количество карт c такое, что при угадывании большего количества H_0 будет считаться опровергнутой:

$$\alpha_{\mathrm{I}} = \mathsf{P}_{H_0}(x \geq c) \leq \alpha.$$

Среди всех таких c следует выбрать наименьшее, чтобы минимизировать ошибку второго рода.

Пример (С гранатами)**.** FIXME

Замечание. Утверждать об *отвержении* гипотезы можно с вероятностью ошибки α (достаточно малой и произвольно задаваемой экспериментатором); утверждать о *принятии* гипотезы можно с вероятностью ошибки α_{II} — не контролируемой и потенциально довольно большой. Иными словами, попадание в доверительную область может означать как то, что H_0 верна, так и то, что верна H_1, но для распознания этого не хватило мощности. Поэтому безопасно гипотезу можно только отвергать или не отвергать. Можно и принять, если известна мощность критерия против всех возможных альтернатив, экспериментатора устраивающая.

Замечание. При высокой вероятности ошибки II-го рода возможна ситуация не отвержении заведомо ложной гипотезы. Это, в свою очередь, может произойти из-за маленького объема выборки (критерий не находит разницу, см. 5). Чем больше объем выборки, тем мощность больше, но возможна ситуация, когда критерий чувствителен настолько, что находит разницу там, где не должен — например, при генерации «идеальным» датчиком случайных чисел, начиная с какого-то объема заведомо истинная гипотеза может быть отвергнута из-за ошибок в точности представления чисел с плавающей точкой.

Определение. *Критерий* называется *одно- (двух-) сторонним* по тому, где находится альтернатива.

Определение. *Критическая область* называется *одно- (дву-) сторонней* по тому, где формально располагается $\mathscr{A}_{\text{крит}}$.

4.2. Статистика критерия

Строить разбиение многомерного пространства V^n неудобно, поэтому вводится статистика критерия.

Определение. *Статистика критерия* есть отображение

$$T : \mathbf{x} \mapsto y \in \mathbb{R}.$$

Утверждение. С введенной статистикой критерия можно разбивать на критическую и доверительную области не V^n, а образ T:

$$\operatorname{im} T = \mathscr{A}^{(\alpha)}_{\text{крит}} \sqcup \mathscr{A}^{(\alpha)}_{\text{дов}}.$$

Остальные связанные с критерием понятия остаются в силе с соответствующей заменой $\mathbf{x}$ на T.

Определение. *p-value* значения статистики T есть вероятность получить такое же или большее по модулю значение T при верной H_0.

p-value обратно пропорционален «существенности» результата.

Если p-value меньше α, то H_0 отвергается, иначе нет.

Схема построения критерия с помощью статистики критерия на примере t-критерия

1. Фиксируют предположение относительно данных.

   $\xi \sim \mathrm{N}(a, \sigma^2)$.

2. Выдвигают H_0 и H_1.

   - H_0 формулируется согласно замечанию 4.1.
   - H_1 ставится по смыслу задачи (см. далее).

   $H_0 : a = a_0$.

3. Выбирают подходящий критерий и статистику T. T должна измерять то, насколько выборка соответствует гипотезе. В этом случае должно быть известно «идеальное соответствие» T ($\mathsf{E}\xi$ логично измерять разницей, $\bar{\mathbf{x}} - a_0$ — тогда идеальным соответствием будет 0.) Распределение T при верной H_0 должно быть известно хотя бы асимптотически. Кроме того, должно быть известно поведение статистики в случае верной H_1 (указать распределение в этом случае не всегда возможно, потому что часто H_1 соответствует целый класс распределений).

   $$T = z = \sqrt{n}\frac{\bar{\mathbf{x}} - a_0}{\sigma} \sim \mathrm{N}(0, 1).$$

   (по **??**).

4. Фиксируют уровень значимости α.

5. Строят разбиение $\operatorname{im} T$ с помощью квантилей распределения T (при верной H_0) так, чтобы $\alpha_{\mathrm{I}} = \alpha$; положение квантилей выбирают из известного поведения статистики при верной H_1:

   Двусторонний критерий Пусть $H_1 : \mathsf{E}\xi = a_1 \neq a_0$; тогда по ЗБЧ $\bar{\mathbf{x}} - a_0 \to a_1 - a_0$ и $|T| \xrightarrow[n\to\infty]{} \infty$. Следовательно, чтобы максимизировать величину $\beta = \mathsf{P}_{H_1}(\mathbf{x} \in \mathscr{A}^{(\alpha)}_{\text{крит}})$, следует разместить $\mathscr{A}_{\text{крит}}$ на правом и левом хвостах (симметрично) распределения статистики: $\mathscr{A}_{\text{крит}} = \mathbb{R} \setminus \big(z_{\alpha/2}, z_{1-\alpha/2}\big)$.

Односторонний критерий Пусть $H_1 : \mathsf{E}\xi = a_1 > a_0$; тогда по ЗБЧ $\bar{\mathbf{x}} - a_0 \to a_1 - a_0 > 0$ и $T \to +\infty$, откуда $\mathscr{A}_{\text{крит}} = (z_{1-\alpha}, \infty)$. Для $H_1 : \mathsf{E}\xi = \alpha_1 < a_0$ аналогично $\mathscr{A}_{\text{крит}} = (-\infty, z_\alpha)$.

Простая альтернатива Пусть $H_1 : \mathsf{E}\xi = a_1$; тогда, поскольку при верной H_1, $\mathsf{E}\bar{\mathbf{x}} = 1/n \cdot \sum_{i=1}^n \xi_i = n/n \cdot a_1$, то

$$\mathsf{E}z = \frac{\sqrt{n}\,(a_1 - a_0)}{\sigma} \implies z \sim \mathrm{N}\left(\frac{\sqrt{n}\,(a_1 - a_0)}{\sigma}, 1\right) \text{ при верной } H_1.$$

(дисперсия, конечно, не меняется при сдвиге).

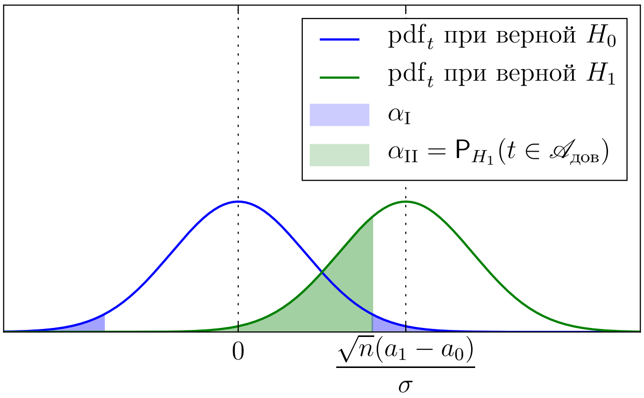


Рис. 1: Плотности распределения z (неоптимальное разбиение)

Чтобы минимизировать α_{II}, логично определить $\mathscr{A}_{\text{крит}}$ только на одном хвосте — с той стороны, где находится альтернатива.

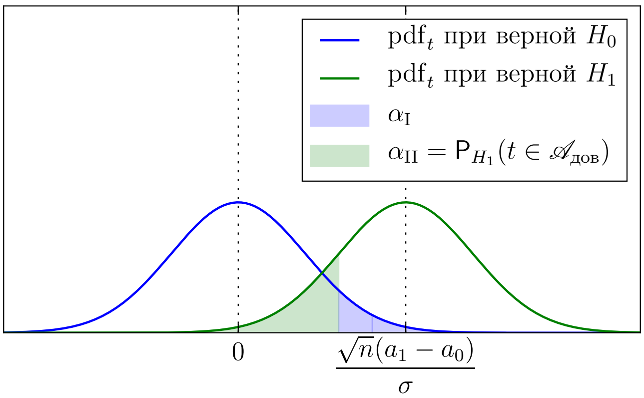


Рис. 2: Плотности распределения z (оптимальное разбиение)

Замечание. Помимо этого, по рисунку видно, что минимизировать α_{II} (согласившись на бóльшую ошибку первого рода) можно сдвинув вправо центр второй плотности, увеличив n.

Аналогично, чем a_1 дальше от a_0, тем α_{II} меньше.

6. Считают значение статистики и принимают решение об отвержении H_0 в зависимости от того, попадает ли оно в $\mathscr{A}_{\text{крит}}$ или $\mathscr{A}_{\text{дов}}$:

$$\varphi(\mathbf{x}) = \begin{cases} 0 & T(\mathbf{x}) \in \mathscr{A}_{\text{дов}} \\ 1 & T(\mathbf{x}) \in \mathscr{A}_{\text{крит}} \end{cases}$$

Альтернативно, можно посчитать p-value, на основании чего сделать вывод.

Пример (Средняя температура в холодильнике)**.** Хотят купить холодильник, такой, чтобы температура держалась в окрестности 0. Известно количество измерений $n = 25$ и $\bar{\mathbf{x}} = 0.7$.

1. Пусть $\xi \sim \mathrm{N}(a, 4)$.

2. Выдвинута $H_0 : \mathsf{E}\xi = a_0 = 0$ — если гипотеза опровергнется, то холодильник не купят; $H_1 : \mathsf{E}\xi = a_1 \neq a_0$.

3. Поскольку модель нормальная и известная σ^2, выберем статистику **??** («z-test»):

$$z = \frac{\sqrt{n}\,(\bar{\mathbf{x}} - a_0)}{\sigma} \sim \mathrm{N}(0,1) \text{ при верной } H_0.$$

   Идеальное значение статистики — 0.

4. Зафиксируем два уровня значимости: $\alpha^{(1)} = 0.2$ (храним петрушку) и $\alpha^{(2)} = 0.01$ (храним дорогую красную икру).

5. Построим разбиение. Поскольку $a_1 \neq a_0$, то $\mathscr{A}_{\text{крит}} = \mathbb{R} \setminus (z_{\alpha/2}, z_{1-\alpha/2})$. Для введенных уровней значимости это означает

   a) $\mathscr{A}_{\text{крит}}^{(\alpha^{(1)})} \approx \mathbb{R} \setminus (-1.28, 1.28)$.

   b) $\mathscr{A}_{\text{крит}}^{(\alpha^{(2)})} \approx \mathbb{R} \setminus (-2.576, 2.576)$.

6. Посчитаем

$$z(\mathbf{x}) = \frac{\sqrt{n}\,(\bar{\mathbf{x}} - a_0)}{\sigma} = \frac{5(0.7-0)}{2} = 1.75.$$

   Дальнейшее принятие решения возможно на основании критического значения или p-value.

   - По вычислению критического значения:
     - ◇ $z \in \mathscr{A}_{\text{крит}}^{(\alpha^{(1)})}$, H_0 отвергается, холодильник не покупают.
     - ◇ $z \in \mathscr{A}_{\text{дов}}^{(\alpha^{(2)})}$, H_0 не отвергается, холодильник, быть может, покупают.
   - Можно посчитать p-value:

$$2 \cdot (1 - \mathrm{cdf}_{\mathrm{N}(0,1)}(1.75)) \approx 0.08.$$

     Поэтому при уровне значимости $\alpha^{(1)} = 0.2 > 0.08$ H_0 отвергается, а при $\alpha^{(2)} = 0.01 < 0.08$ не отвергается.

Пример (С мышой)**.** В одном из рукавов Т-образного лабиринта лежит морковка. К развилке по лабиринту бежит мышь и 7 раз из 10 поворачивает в направлении морковки. На основании этих данных хотим сделать вывод, что мышь чует морковь на расстоянии, после чего написать научную статью.

- $\xi \sim \mathrm{Ber}(p)$. Выдвинем гипотезу, что мышь *не* чует морковку, $H_0 : p = p_0 = 0.5$. Поскольку $\mathsf{E}\xi = p$, воспользуемся критерием для проверки гипотезы о значении среднего с идеальным значением 0; учитывая $\mathsf{D}\xi = p(1-p)$,

$$\begin{aligned} T &= \sqrt{n}\frac{\bar{\mathbf{x}} - p_0}{\sqrt{p_0(1-p_0)}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1). \\ &= \frac{\sqrt{10}\cdot 0.2}{0.5} \approx 1.2649 \implies p\text{-value} = 2\cdot(1 - \mathrm{cdf}_{\mathrm{N}(0,1)}(1.2649)) \approx 0.2. \end{aligned}$$

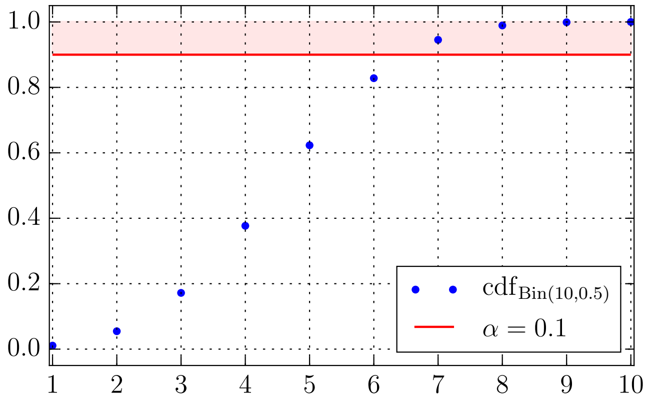
  Значит, с уровнем значимости 0.2 гипотеза не отвергается. Хочется иметь, конечно, один из стандартных уровней значимости, например 0.1.

- Увеличим мощность критерия, введя альтернативную гипотезу, что мышь чует морковку (в предположении, что все мыши любят морковь и к ней бегут), $H_1 : p_1 > p_0$. По 5, можем устроить односторонний критерий, так что p-value теперь 0.1. Однако пользуемся асимптотическим критерием при $n = 10$.

- Воспользуемся точным односторонним критерием со статистикой

$$T := n\bar{\mathbf{x}} = \sum_{i=1}^{n} x_i \sim \mathrm{Bin}(n, p_0)$$

и идеальным значением np_0. Тогда $T = 10 \cdot 0.7 = 7$. При уровне значимости $\alpha = 0.1$ успешно попадаем в критическую область, вследствие чего H_0 отвергается, и можем публиковаться.



Замечание. Исторически существовало два подхода к проверке гипотез: Фишера («significance test») и Неймана-Пирсона («hypothesis testing»).

Фишер Выдвигается H_0. Подсчитывается и сообщается точное p-value. Если результат «незначительный», не делается никаких выводов об отвержении H_0, но делается возможным дополнительный сбор данных.

Нейман-Пирсон Выдвигаются H_1, H_2, фиксируются $\alpha_{\mathrm{I}}, \alpha_{\mathrm{II}}$ и n. На этом основании определяются $\mathscr{A}_{\text{крит}}$ для каждой гипотезы. Если данные попали в $\mathscr{A}_{\text{крит}}$ H_1 — предпочитается H_2, иначе H_2.

Современная теория проверки гипотез есть смесь двух этих подходов, не всегда консистентная. Вводные курсы по статистике формулируют теорию, похожую на significance testing Фишера; при повышенных требованиях к математической строгости, пользуются теорией Неймана-Пирсона.

Замечание (О графике p-values). Поскольку

$$\alpha_{\mathrm{I}} \leftarrow \mathsf{P}_{H_0}(T \in \mathscr{A}_{\text{крит}}) = \mathsf{P}_{H_0}(p\text{-value} < \alpha),$$

то p-value по распределению стремятся к $\mathrm{U}(0, 1)$ при верной H_0. Это соображение позволяет визуально проверить истинность гипотезы: достаточно несколько (много) раз произвести эксперимент, для каждой выборки $\bar{\mathbf{x}}^{(i)}$ посчитать свой p-value, построить график и убедиться, что получилась прямая. Для подсчета мощности $\beta = \mathsf{P}_{H_1}(\mathbf{x} \in \mathscr{A}_{\text{крит}}^{(\alpha)}) = \mathsf{P}_{H_1}(p\text{-value} < \alpha)$ считать выборку с параметрами H_1, а T относительно H_0.

4.3. Проверка гипотезы о значении мат. ожидания (t-критерий)

$H_0 : \mathsf{E}\xi = a = a_0$. Соответствие оценки математического ожидания гипотезе удобно выражать разницей $\bar{\mathbf{x}} - a_0$ с «идеальным» значением 0. Отнормировав эту разницу, получим статистику, распределение которой известно.

Замечание (Важное). Критерий работает только в нормальной модели и не становится асимптотически нормальным в ином случае!

4.3.1. $\mathsf{D}\xi = \sigma^2 < \infty$

Предложение. *Пусть* $\mathsf{D}\xi = \sigma^2 < \infty$*; тогда используется следующая статистика*

$$t = \sqrt{n}\frac{(\bar{\mathbf{x}} - a_0)}{\sigma} \xrightarrow[n\to\infty]{} \mathrm{N}(0,1)$$

Доказательство. По ЦПТ. □

Предложение. *При условии нормальности данных, t-критерий называется «z-критерием», причем*

$$t = z \sim \mathrm{N}(0,1).$$

Доказательство.

$$z = \frac{\bar{\mathbf{x}} - a_0}{\sqrt{\mathsf{D}\bar{\mathbf{x}}}} = \sqrt{n}\,\frac{\bar{\mathbf{x}} - a_0}{\sigma} \sim \mathrm{N}(0,1).$$

□

Разбиение

$H_1 : \mathsf{E}\xi \neq a_0$ $\mathscr{A}_{\text{крит}} = \mathbb{R} \setminus \big(z_{\alpha/2}, z_{1-\alpha/2}\big)$

$H_1 : \mathsf{E}\xi > a_0$ $\mathscr{A}_{\text{крит}} = (z_{1-\alpha}, \infty)$

$H_1 : \mathsf{E}\xi < a_0$ $\mathscr{A}_{\text{крит}} = (-\infty, z_{\alpha})$

4.3.2. $\mathsf{D}\xi$ неизвестна

Предложение. *Пусть* $\mathsf{D}\xi$ *неизвестна; тогда используется следующая статистика*

$$t = \sqrt{n-1}\,\frac{\bar{\mathbf{x}} - a_0}{s} = \frac{\sqrt{n-1}(\bar{\mathbf{x}} - a_0)}{\sqrt{n-1}/\sqrt{n}\cdot\tilde{s}} = \sqrt{n}\,\frac{\bar{\mathbf{x}} - a_0}{\tilde{s}} \xrightarrow[n\to\infty]{} \mathrm{N}(0,1).$$

Предложение. *При условии нормальности данных,*

$$t \sim \mathrm{t}(n-1).$$

Доказательство.

$$t = \frac{\sqrt{n-1}\,(\bar{\mathbf{x}} - a_0)}{s} = \frac{\sqrt{n-1}\left(\dfrac{\bar{\mathbf{x}} - a_0}{\sigma}\right)}{s/\sigma} = \frac{\left(\dfrac{\bar{\mathbf{x}} - a_0}{\sigma}\right)}{\sqrt{\dfrac{s^2/\sigma^2}{n-1}}} = \frac{\dfrac{\sqrt{n}(\bar{\mathbf{x}} - a_0)}{\sigma}}{\sqrt{\dfrac{ns^2/\sigma^2}{n-1}}} = \frac{\beta}{\sqrt{\eta/(n-1)}} \sim \mathrm{t}(n-1),$$

поскольку

$$\beta = \frac{\sqrt{n}(\bar{\mathbf{x}} - a_0)}{\sigma} \sim \mathrm{N}(0,1), \quad \eta = \frac{ns^2}{\sigma^2} \sim \chi^2(n-1).$$

□

Разбиение

$H_1 : \mathsf{E}\xi \neq a_0$ $\mathscr{A}_{\text{крит}} = \mathbb{R} \setminus \Big(\mathrm{qnt}_{\mathrm{t}(n-1)}(\alpha/2), \mathrm{qnt}_{\mathrm{t}(n-1)}(1-\alpha/2)\Big)$

$H_1 : \mathsf{E}\xi > a_0$ $\mathscr{A}_{\text{крит}} = (\mathrm{qnt}_{\mathrm{t}(n-1)}(1-\alpha), \infty)$

$H_1 : \mathsf{E}\xi < a_0$ $\mathscr{A}_{\text{крит}} = (-\infty, \mathrm{qnt}_{\mathrm{t}(n-1)}\,\alpha)$

Замечание. При нормальной аппроксимации $\mathrm{qnt}_{\mathrm{t}(n-1)}$ заменить на $\mathrm{N}(0,1)$.

z-критерий для пропорции в модели Бернулли Пусть $\xi \sim \mathrm{Ber}(p)$. Поскольку $\mathsf{E}\xi = p$, можно воспользоваться только что введенной статистикой; учитывая $\mathsf{D}\xi = p(1-p)$,

$$T = \sqrt{n}\frac{\bar{\mathbf{x}} - p_0}{\sqrt{p_0(1-p_0)}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1).$$

Разбиение будет таким же, как и в случае известной дисперсии.

4.4. Проверка гипотезы о значении дисперсии в нормальной модели (критерий χ^2)

Пусть $\xi \sim \mathrm{N}(a, \sigma^2)$. $H_0 : \mathsf{D}\xi = \sigma^2 = \sigma_0^2$. Соответствие оценки дисперсии гипотезе удобно выражать отношением s^2/σ_0^2 (или s_a/σ_0^2 если a известно) с «идеальным» значением 1. Домножив на n, получим статистику, распределение которой известно.

4.4.1. $\mathsf{E}\xi = a < \infty$

Предложение. *Пусть* $\mathsf{E}\xi = a < \infty$*; При условии нормальности данных используется следующая статистика:*

$$\chi^2 = n\frac{s_a^2}{\sigma_0^2} \sim \chi^2(n).$$

Доказательство.

$$\chi^2 = \frac{ns_a^2}{\sigma_0^2} = \frac{n \cdot 1/n \cdot \sum_{i=1}^n (x_i - a)^2}{\sigma_0^2} = \sum_{i=1}^n \left(\frac{x_i - a}{\sigma_0}\right)^2 \sim \chi^2(n).$$

□

Разбиение

$H_1 : \mathsf{D}\xi \neq \sigma_0^2$ $\mathscr{A}_{\text{крит}} = \mathbb{R}_+ \setminus \left(\mathrm{qnt}_{\chi^2(n)}(\alpha/2), \mathrm{qnt}_{\chi^2(n)}(1-\alpha/2)\right)$

$H_1 : \mathsf{E}\xi > a$ $\mathscr{A}_{\text{крит}} = (\mathrm{qnt}_{\chi^2(n)}(1-\alpha), \infty)$

$H_1 : \mathsf{E}\xi < a$ $\mathscr{A}_{\text{крит}} = (0, \mathrm{qnt}_{\chi^2(n)}\,\alpha)$

4.4.2. $\mathsf{E}\xi$ неизвестно

Предложение. *Пусть* $\mathsf{E}\xi$ *неизвестно. При условии нормальности данных используется следующая статистика:*

$$\chi^2 = n\frac{s^2}{\sigma_0^2} = (n-1)\frac{\tilde{s}^2}{\sigma_0^2} \sim \chi^2(n-1).$$

Доказательство. См. (3.4). □

Альтернативное доказательство. По определению запишем

$$\mathsf{D}\hat{\xi}_n = \mathsf{D}(\hat{\xi}_n - a) = \mathsf{E}\left(\hat{\xi}_n - a\right)^2 - (\mathsf{E}\left(\hat{\xi}_n - a\right))^2.$$

Но

$$\begin{aligned} \mathsf{D}\hat{\xi}_n &= \mathsf{E}(\hat{\xi}_n - \mathsf{E}\hat{\xi}_n)^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{\mathbf{x}})^2 = s^2 \\ \mathsf{E}(\hat{\xi}_n - a)^2 &= \frac{1}{n}\sum_{i=1}^n (x_i - a)^2 = s_a^2 \\ (\mathsf{E}\left(\hat{\xi}_n - a\right))^2 &= (\bar{\mathbf{x}} - a)^2, \end{aligned}$$

откуда
$$s^2 = s_a^2 - (\bar{\mathbf{x}} - a)^2 \,.$$
Домножив обе части на n/σ_0^2, получим
$$\frac{ns^2}{\sigma_0^2} = \frac{ns_a^2}{\sigma_0^2} - \frac{n(\bar{\mathbf{x}} - a)^2}{\sigma_0^2} = \underbrace{\frac{ns_a^2}{\sigma_0^2}}_{\sim\chi^2(n)} - \underbrace{\left(\frac{\sqrt{n}\,(\bar{\mathbf{x}} - a)}{\sigma_0}\right)^2}_{\sim\chi^2(1)} \implies \frac{ns^2}{\sigma_0^2} \sim \chi^2(n-1).$$
□

Замечание. Для строгого доказательства, нужно использовать независимость $\bar{\mathbf{x}}^2$ и s^2 (см. 3.4).

Разбиение

$H_1 : \mathsf{D}\xi \neq \sigma_0^2$ $\mathscr{A}_{\text{крит}} = \mathbb{R}_+ \setminus \left(\text{qnt}_{\chi^2(n-1)}(\alpha/2), \text{qnt}_{\chi^2(n-1)}(1-\alpha/2)\right)$

$H_1 : \mathsf{E}\xi > a$ $\mathscr{A}_{\text{крит}} = (\text{qnt}_{\chi^2(n-1)}(1-\alpha), \infty)$

$H_1 : \mathsf{E}\xi < a$ $\mathscr{A}_{\text{крит}} = (0, \text{qnt}_{\chi^2(n-1)}\,\alpha)$

Упражнение. $s^2 = 1.44, \bar{\mathbf{x}} = 55, n = 101$. Проверить гипотезу $\sigma_0^2 = 1.5$ в нормальной модели.

Решение. Воспользуемся статистикой
$$\chi^2 = \frac{ns^2}{\sigma_0^2} = 101 \cdot 0.96 = 96.96.$$
«Идеальные» значения близки к $\mathsf{E}\xi_{\chi^2(100)} = 100$, так что определим критическую область на концах плотности:
$$p\text{-value}/2 = \text{cdf}_{\chi^2(100)}(96.96) = \texttt{pchisq(96.96, 100)} \approx 0.43 \implies p\text{-value} \approx 0.86.$$
Замечание. Можно посчитать и по таблицам для нормального распределения. Раз
$$\frac{\eta_m - \mathsf{E}\eta_m}{\sqrt{\mathsf{D}\eta_m}} \xrightarrow[m\to\infty]{\text{d}} \text{N}(0,1),$$
то
$$\frac{96.96 - 100}{\sqrt{200}} \approx -0.215 \implies p\text{-value}/2 = \Phi(-0.215) \approx 0.415.$$
⌟

4.5. Критерий χ^2 согласия с видом распределения

По выборке возможно проверить гипотезу о виде распределения случайной величины, реализацией которой является выборка.

Утверждение. Для проверки гипотезы согласия с видом произвольного *дискретного* распределения используется асимптотический *критерий* χ^2 *(«chi-squared test for goodness of fit»).*

4.5.1. Распределение с известными параметрами

Пусть

$$H_0 : \mathcal{P} = \mathcal{P}_0, \text{ где } \mathcal{P}_0 : \begin{pmatrix} x_1^* & \dots & x_k^* \\ p_1 & \dots & p_k \end{pmatrix}.$$

Сгруппируем $\mathbf{x}$; каждому x_i^* сопоставим *эмпирическую* абсолютную частоту ν_i; тогда np_i — *ожидаемая* абсолютная частота.

В качестве меры расхождения между эмпирическим и генеральным распределением рассматривается величина

$$\sum_{i=1}^{k} c_i \left(\frac{\nu_i}{n} - p_i\right)^2, \quad c_i = \frac{n}{p_i},$$

откуда записывается статистика критерия

$$T = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i}$$

с идеальным значением 0.

Утверждение. $T \xrightarrow{\mathrm{d}} \chi^2(k-1)$.

Разбиение $\mathscr{A}_{\text{крит}} = \Big(\mathrm{qnt}_{\chi^2(k-1)}(1-\alpha), \infty\Big)$ — гипотеза отвергается, если расстояние между предполагаемым и наблюдаемым распределениями большое.

Упражнение. $n = 100$,

$$\begin{pmatrix} \diamondsuit & \heartsuit & \clubsuit & \spadesuit \\ 20 & 30 & 10 & 40 \end{pmatrix}.$$

Проверить гипотезу, что колода полная.

Решение. $H_0 : \mathcal{P}_\xi = \mathrm{U}(1/4)$. Поскольку речь идет о согласии с дискретным не параметризованным распределением, напрямую воспользуемся критерием χ^2. Раз все $np_i = 100 \cdot 1/4 = 25$,

$$\chi^2 = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i} = 1 + 1 + \frac{15^2}{25} + \frac{15^2}{25} = 2 + 2 \cdot 9 = 20.$$

Так как $\chi^2 \sim \chi^2(k-1) = \chi^2(3)$ со средним 3, и «идеальное» значение 0, определим критическую область в правом конце плотности. Из этих соображений

$$p\text{-value} = 1 - \mathrm{cdf}_{\chi^2(3)}(20) = 1 - \texttt{pchisq}(20, 3) \approx 0.00017.$$

┘

4.5.2. Распределение с неизвестными параметрами

В случае сложной гипотезы $\mathcal{P} \in \{\mathcal{P}(\boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta}$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_r)^{\mathrm{T}}$, следует найти оценку $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ (или $\hat{\boldsymbol{\theta}} : \hat{\boldsymbol{\theta}} \to \hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$) по методу максимального правдоподобия. При подстановке оценок вместо истинных параметров критерий становится консервативным. Чтобы этого избежать, необходимо сделать поправку на количество параметров — отнять r. Что приятно, одна и та же поправка работает для всех распределений; в этом случае,

$$T \xrightarrow{\mathrm{d}} \chi^2(k - r - 1).$$

Упражнение 1. 60 человек купило подарок сразу, 10 со второго раза, 20 с третьего, 10 с четвертого:

$$\begin{pmatrix} 0 & 1 & 2 & 3 \\ 60 & 10 & 20 & 10 \end{pmatrix}.$$

Проверить гипотезу о том, что это выборка из геометрического распределения.

Решение. $H_0 : \mathcal{P}_\xi = \mathrm{Geom}(p)$. Воспользуемся критерием χ^2 для параметризированного распределения $\mathrm{Geom}(\hat{p}_{\mathrm{MLE}})$.

Найдем
$$\hat{p}_{\mathrm{MLE}} = \operatorname*{argmax}_p \log \mathcal{L}(\mathbf{x}; p) \Longleftarrow \frac{\mathrm{d}}{\mathrm{d}p} \log \mathcal{L}(\mathbf{x}; \hat{p}_{\mathrm{MLE}}) = 0.$$

Так как $\mathrm{pdf}_{\mathrm{Geom}(p)}(k) = (1-p)^k p$,
$$\begin{aligned}\log \mathcal{L}(\mathbf{x}; p) &= \log \prod_{k=1}^{n} (1-p)^k p = \log(1-p)^{n\bar{x}} p^n = n\bar{\mathbf{x}} \log(1-p) + n \log p \\ &= n(\bar{\mathbf{x}} \log(1-p) + \log p)\end{aligned}$$

откуда
$$\frac{\mathrm{d}}{\mathrm{d}p} \log \mathcal{L}(\mathbf{x}; p) = n \left(-\frac{\bar{\mathbf{x}}}{1-p} + \frac{1}{p} \right) = 0 \iff 1 - p - p\bar{\mathbf{x}} = 0 \iff p = \frac{1}{1+\bar{\mathbf{x}}}.$$

Учитывая
$$\bar{\mathbf{x}} = 0.1 + 2 \cdot 0.2 + 3 \cdot 0.1 = 0.8,$$

найдем
$$\hat{p}_{\mathrm{MLE}} = \frac{1}{1+0.8} \approx 0.55.$$

Посчитаем статистику χ^2, найдя соответствующие p_i:
$$p_0 = \mathsf{P}_{\mathrm{Geom}(0.55)}(0) = 0.55, \quad p_1 \approx 0.26, \quad p_2 \approx 0.11, \quad p_3 \approx 0.09.$$

Тогда
$$\chi^2 = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i} = \frac{25}{55} + \frac{16^2}{26} + \frac{81}{11} + \frac{1}{9} \approx 17.77.$$

Наконец, поскольку $\chi^2 \xrightarrow[n\to\infty]{} \chi^2(k-r-1)$,
$$p\text{-value} = 1 - \mathrm{cdf}_{\chi^2(2)}(17.77) \approx 0.00014.$$

$\lrcorner$

Определение. Критерий *применим*, если $\alpha \to \alpha_{\mathrm{I}}$.

Замечание. Поскольку критерий асимптотический, с достаточной степенью точностью он применим в случае, если

1. $n \geq 50$;
2. $np_i \geq 5$.

Замечание. Если условие $np_i \geq 5$ не выполняется, следует объединить состояния, например, с краев или слева направо; если в хвосте оказалось < 5, то следует присоединить к последнему.

Пример (С монеткой)**.** Пусть $n = 4040$, $\#\mathrm{H} = 2048$, $\#\mathrm{T} = 1092$. Проверим $H_0 : \mathcal{P} = \mathrm{Ber}(0.5)$ с $\alpha = 0.1$. Условия критерия выполняются, поэтому посчитаем
$$T = \frac{(2048-2020)^2}{2020} + \frac{(1092-2020)^2}{2020} = \frac{28^2+28^2}{2020} \approx 0.78 \sim \chi^2(1),$$

откуда
$$p\text{-value} = 1 - \mathrm{cdf}_{\chi^2(1)}(0.78) \approx 0.38.$$

$0.38 > 0.1$, значит H_0 не отвергается.

Замечание. Прохождение критерия не достаточно. Так, альтернирующая (и явно не случайная) последовательность $\mathbf{x} = (0, 1, 0, 1, \ldots)$ имеет $T = 0$.

4.5.3. Согласие с нормальным распределением

Для проверки гипотезы $H_0 : \mathcal{P}_\xi = \mathrm{N}(a, \sigma^2)$ также можно воспользоваться статистикой критерия χ^2 для сложной гипотезы. В этом случае, нужно дискретизировать нормальное распределение, так, что

$$\mathcal{P}_0 = \begin{pmatrix} x_1^* & \dots & x_k^* \\ p_1(\hat{\boldsymbol{\theta}}) & \dots & p_k(\hat{\boldsymbol{\theta}}) \end{pmatrix}, \quad \hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathrm{MLE}}.$$

Тем не менее, нужно иметь в виду две теоретических неточности:

1. Построение $\mathcal{P}_0$ происходит случайно, в результате объединения элементов выборки после того, как она получена.

2. Оценка параметров $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ должна быть посчитана для $\mathcal{P}_0$, а не для исходного (нормального) распределения — не $\bar{\mathbf{x}}, s^2$. Однако на практике на этот момент не обращают внимания.

Существует два возможных способа дискретизации:

1. Гистограмма: одинаковые интервалы, но разные вероятности.

2. Неравные интервалы с равными вероятностями.

```
N <- length(xs)
xs <- sort(xs)
probs <- pnorm(xs, mean=mean(xs), sd=sd(xs))
i <- 1; j <- i+1
while (N * (probs[j] - probs[i]) < 5) {
  j <- j+1
}
mean(xs[i:j])  # our x_1^*; n_1 = j - i + 1
```

Этот способ разбиения предпочтителен, потому что:

- Можно разбить максимально часто — так, чтобы $np_i = 5\ \forall i$, следовательно и мощность будет максимальна.
- Он оказывается точнее первого на практике.
- Получается единственное p-value.

Замечание. Следует иметь в виду, что этот способ не годится для непрерывных, но плохо дискретизированных данных.

4.6. Критерий Колмогорова-Смирнова согласия с видом распределения

4.6.1. Произвольное абсолютно непрерывное распределение

$H_0 : \xi \sim \mathcal{P} = \mathcal{P}_0$.

Утверждение. Для проверки гипотезы согласия с видом произвольного *абсолютно непрерывного* распределения с известными параметрами используется асимптотический критерий Колмогорова-Смирнова со следующей статистикой:

$$D_n = \sup_{x \in \mathbf{x}} \left| \widehat{\mathrm{cdf}}_n(x) - \mathrm{cdf}_0(x) \right|,$$

где cdf_0 — функция распределения $\mathcal{P}_0$ нулевой гипотезы.

Альтернатива только одна: $H_1 : \xi \not\sim \mathcal{P}_0$; $\mathscr{A}_{\text{крит}} = (\mathrm{qnt}_{\text{K-S}}(1-\alpha), \infty)$.

Замечание. Критерий является не асимптотическим, но *точным.* Значит им пользоваться и при маленьких объемах выборки (мощность, при этом, останется низкой все-равно).

Замечание. $\sup_x \sqrt{n} \left|\widehat{\mathrm{cdf}}_n(x) - \mathrm{cdf}_0(x)\right| \xrightarrow{\mathrm{d}} \mathcal{P}_{\text{K.S.}}$, где $\mathcal{P}_{\text{K.S.}}$ — распределение Колмогорова. Значит, при больших объемах выборки для такой статистики критерия можно пользоваться таблицами распределения Колмогорова.

Упражнение. Проверить гипотезу, что $\mathbf{x} = (0.1, 0.2, 0.4, 0.3, 0.1)$ есть выборка из $\mathrm{U}[0, 1]$.

Решение. $D_n = 0.6$, p-value ≈ 0.05 (по таблицам или компьютером). Таким образом, при $\alpha > 0.05$ гипотеза отвергается, при $\alpha < 0.05$ — нет.

```
> ks.test(c(0.1,0.2,0.4,0.3,0.1), 'punif')
        One-sample Kolmogorov-Smirnov test
data:  c(0.1, 0.2, 0.4, 0.3, 0.1)
D = 0.6, p-value = 0.05465
```

⌟

Замечание. Критерий Колмогорова-Смирнова консервативный — значение p-value завышено. Поэтому если гипотеза отвергается, то наверняка.

4.6.2. Нормальное распределение

Пусть $H_0 : \mathcal{P}_\xi \in \left\{\mathrm{N}(a, \sigma^2)\right\}$. Как известно, критерий Колмогорова-Смирнова используется для непрерывных непараметрических распределений. Им можно воспользоваться и для данной H_0, если вместо a, σ^2 подставить соответствующие оценки — в таком случае критерий будет консервативным. По аналогии с χ^2 хотелось бы сделать поправку на количество параметров — такая поправка осуществляется путем моделирования распределения тестовой статистики. Для $\mathrm{N}(a, \sigma^2)$ и $\mathrm{Exp}(\lambda)$ получаем распределение D_n, не зависящее от параметров (так что поправку можно делать вне зависимости от параметров; к примеру, $\mathrm{N}(a, \sigma^2)$ можно привести к $\mathrm{N}(0, 1)$ непрерывным преобразованием):

Критерий Бартлетта есть критерий Колмогорова-Смирнова для $H_0 : \mathcal{P}_\xi = \mathrm{Exp}(\lambda)$.

Критерий Лиллиефорса[8] для проверки $H_0 : \mathcal{P}_\xi = \mathrm{N}(a, \sigma^2)$ считается статистикой D_n с $\mathrm{cdf}_0(x) = \mathrm{cdf}_{\mathrm{N}(\bar{\mathbf{x}}, s^2)}(x)$, сходящейся к распределению Лиллиефорса (Колмогорова-Смирнова с учетом подстановки оценок).

Критерий Шапиро-Уилка $T \approx \rho^2$, т.е. ведет себя примерно как квадрат коэффициента корреляции на normal probability plot.

Замечание. Распределения Лиллиефорса и Колмогорова-Смирнова были получены путем моделирования.

Пример. В R:

```
> require('nortest')
> lillie.test(rt(1000, df=10))
        Lilliefors (Kolmogorov-Smirnov) normality test
data:  rt(1000, df = 10)
D = 0.03042, p-value = 0.02953
```

4.7. Критерий типа ω^2

Определение. Статистика

$$Q = n \int_{\mathbb{R}} \left(\mathrm{cdf}_n(x) - \mathrm{cdf}_0(x)\right)^2 w(x) \,\mathrm{d}\,\mathrm{cdf}_0(x),$$

где $w(x)$ — весовая функция.

Замечание. Статистика может быть проинтерпретирована как площадь разницы между соответствующими функциями распределения.

Cramer von Mises Q с $w \equiv 1$.

Anderson-Darling Q с

$$w(x) = \frac{1}{\mathrm{cdf}_0(x)(1-\mathrm{cdf}_0(x))}.$$

Замечание. Весовая функция критерия Anderson-Darling присваивает большой вес значениям на хвостах распределения, поэтому сам критерий является мощным против разницы на хвостах, но и менее мощным при сдвиге.

Замечание. Все эти критерии точны.

Замечание. Распределение статистики в каждом случае не зависит от cdf_0 и все эти критерии состоятельные против любой альтернативы, поэтому не очень мощные.

4.8. Визуальное определение согласия с распределением

4.8.1. P-P plot

Определение. *P-P plot* есть график

$$\left\{\left(\mathrm{cdf}_0(x_i)+\frac{1}{2n}, \widehat{\mathrm{cdf}}_n(x_i)\right)\right\}_{i=1}^n.$$

Пример. В R:

```
pp.plot <- function(xs, cdf.0=pnorm, n.knots=1000) {
  knots <- seq(min(xs), max(xs), length.out=n.knots)
  plot(cdf.0(knots), ecdf(xs)(knots))
  abline(0, 1)
}
```

4.8.2. Q-Q plot

Определение. *Q-Q plot* есть график

$$\left\{\left(x_i, \mathrm{cdf}_0^{-1}\left(\widehat{\mathrm{cdf}}_n(x_i)+\frac{1}{2n}\right)\right)\right\}_{i=1}^n.$$

Определение. Частный случай Q-Q plot для $\mathrm{cdf}_0^{-1} = \mathrm{cdf}_{\mathrm{N}(0,1)}^{-1}$ называется *normal probability plot.*

Пример. В R:

```
qq.plot <- function(xs, qf.0=qnorm, n.ppoints=1000) {
  qs <- ppoints(n.ppoints)
  plot(qf.0(qs), unname(quantile(xs, probs=qs)))
  abline(mean(xs), sd(xs))
}
```

Замечание. Если $\hat{\mathcal{P}}_n \to \mathcal{P}_\xi$, то оба графика будут стремиться к $y = x$. Референсной прямой normal probability plot будет $y = \widehat{\mathsf{D}\xi} \cdot x + \widehat{\mathsf{E}\xi}$.

Замечание. Больше о различии Q-Q и P-P plots, см. http://v8doc.sas.com/sashtml/qc/chap8/sect9.htm

Замечание. Различные интерпретации параметров распределения по Q-Q plot можно посмотреть в интерактивном приложении: https://xiongge.shinyapps.io/QQplots/

4.9. Гипотеза о равенстве распределений

$H_0 : \mathcal{P}_{\xi_1} = \mathcal{P}_{\xi_2}$.

Возможно рассматривать два случая:

Независимые выборки Две группы индивидов, на которых измеряется один и тот же признак. Формально: пусть $\zeta \in \{1, 2\}$ — номер группы, ξ — признак. Тогда $\xi_1 \sim \mathcal{P}_{\xi|\zeta=1}$, $\xi_2 \sim \mathcal{P}_{\xi|\zeta=2}$ и $\xi_1 \perp\!\!\!\perp \xi_2$. В этом случае выборка имеет вид

$$((x_1, x_2, \ldots, x_{n_1}), (y_1, y_2, \ldots, y_{n_2})).$$

Зависимые выборки Одна группа индивидов, на каждом из которых измеряются две характеристики (либо же «до» и «после»). В этом случае выборка имеет вид

$$((x_1, y_1), \ldots, (x_n, y_n)).$$

Замечание. Для одной и той же гипотезы могут существовать разные критерии; их возможно сравнить по мощности, но только если они состоятельны против одной и той же альтернативы.

Замечание. Непараметрические критерии хороши тем, что основаны на рангах, значит устойчивы к аутлаерам; плохи тем, что не используют всю информацию о значении — только порядок, из-за чего обладают меньшей мощностью.

4.10. Равенство математических ожиданий для независимых выборок

4.10.1. Двухвыборочный t-критерий

$H_0 : \mathsf{E}\xi_1 = \mathsf{E}\xi_2$.

Определение. И для зависимых, и для независимых выборок используется *двухвыборочный t-критерий*

$$t = \frac{\bar{\mathbf{x}} - \bar{\mathbf{y}}}{\sqrt{\mathsf{D}(\bar{\mathbf{x}} - \bar{\mathbf{y}})}} \xrightarrow{\sim} \mathrm{N}(0, 1).$$

Пусть выборка *независима*[9], $(x_1, \ldots, x_{n_1}), (y_1, \ldots, y_{n_2})$, $n = n_1 + n_2$. Значит $\mathsf{D}(\bar{\mathbf{x}} - \bar{\mathbf{y}}) = \mathsf{D}\bar{\mathbf{x}} + \mathsf{D}\bar{\mathbf{y}}$.

Двухвыборочный t-критерий для независимых выборок с $\sigma_1^2 \neq \sigma_2^2$ (Welch t-test)

Предложение. *Если дисперсия известна,* $\mathsf{D}(\bar{\mathbf{x}} - \bar{\mathbf{y}}) = \mathsf{D}\bar{\mathbf{x}} + \mathsf{D}\bar{\mathbf{y}} = \sigma_1^2/n_1 + \sigma_2^2/n_2$ *и*

$$t = \frac{\bar{\mathbf{x}} - \bar{\mathbf{y}}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \xrightarrow[n_1, n_2 \to \infty]{} \mathrm{N}(0, 1).$$

Если данные нормальные, то

$$t \sim \mathrm{N}(0, 1).$$

Предложение. *Если дисперсия неизвестна,* $\mathsf{D}\widehat{(\bar{\mathbf{x}} - \bar{\mathbf{y}})} = s_1^2/n_1 + s_2^2/n_2$, *откуда*

$$t = \frac{\bar{\mathbf{x}} - \bar{\mathbf{y}}}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \xrightarrow[n_1, n_2 \to \infty]{} \mathrm{N}(0, 1).$$

Замечание. Точное распределение неизвестно, примерно равно t с дробным числом степеней свободы (что вычисляется интерполяцией по соседним степеням). Всегда ожидается, что если данные нормальны, то распределение известно. Это противоречие носит название *проблемы Беренса-Фишера*[10].

[9] Случай зависимой выборки рассматривается в другом параграфе.

[10] Behrens-Fisher problem.

Разбиение

$H_1 : \mathsf{E}\xi_1 \neq \mathsf{E}\xi_2$ $\mathscr{A}_{\text{крит}} = \mathbb{R} \setminus \left(z_{\alpha/2}, z_{1-\alpha/2}\right)$

$H_1 : \mathsf{E}\xi_1 > \mathsf{E}\xi_2$ $\mathscr{A}_{\text{крит}} = (z_{1-\alpha}, \infty)$

$H_1 : \mathsf{E}\xi_1 < \mathsf{E}\xi_2$ $\mathscr{A}_{\text{крит}} = (-\infty, z_\alpha)$

Двухвыборочный t-критерий для независимых выборок с $\sigma_1^2 = \sigma_2^2$ (pooled t-test)

Предложение. *Если дисперсия известна,*

$$\mathsf{D}(\bar{\mathbf{x}} - \bar{\mathbf{y}}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right),$$

откуда

$$t = \frac{\bar{\mathbf{x}} - \bar{\mathbf{y}}}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \xrightarrow[n_1, n_2 \to \infty]{} \mathrm{N}(0,1).$$

Если данные нормальные, то

$$t \sim \mathrm{N}(0,1).$$

Предложение. *Если дисперсия неизвестна,*

$$t = \frac{\bar{\mathbf{x}} - \bar{\mathbf{y}}}{\tilde{s}_{1,2}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \xrightarrow[n_1, n_2 \to \infty]{} \mathrm{N}(0,1).$$

Если данные нормальные, то

$$t \sim \mathrm{t}(n_1 + n_2 - 2).$$

Доказательство. Оценку дисперсии можно найти по объединенной и центрированной выборке (т.е. если H_0 верна, то $\mathsf{E}\xi_1 = \mathsf{E}\xi_2$ и можно думать как про одну выборку):

$$\begin{aligned} s_{1,2}^2 &= \frac{\overbrace{\sum_{i=1}^n (x_i - \bar{\mathbf{x}})^2}^{\sim\chi^2(n_1-1)} + \overbrace{\sum_{i=1}^n (y_i - \bar{\mathbf{y}})^2}^{\sim\chi^2(n_2-1)}}{n_1 + n_2} = \frac{n_1 \cdot s_1^2}{n_1 + n_2} + \frac{n_2 \cdot s_2^2}{n_1 + n_2} \\ \tilde{s}_{1,2}^2 &= \frac{\sum_{i=1}^n (x_i - \bar{\mathbf{x}})^2 + \sum_{i=1}^n (y_i - \bar{\mathbf{y}})^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)\tilde{s}_1^2}{n_1 + n_2 - 2} + \frac{(n_2 - 1)\tilde{s}_2^2}{n_1 + n_2 - 2}, \end{aligned}$$

где в последнем случае оценка несмещенная и $\mathsf{E}\tilde{s}_{1,2}^2 = \sigma^2$. □

Замечание. Этот вариант более точен, чем в случае $\sigma_1 \neq \sigma_2$.

Разбиение

$H_1 : \mathsf{E}\xi_1 \neq \mathsf{E}\xi_2$ $\mathscr{A}_{\text{крит}} = \mathbb{R} \setminus \left(\mathrm{qnt}_{\mathrm{t}(n_1+n_2-2)}(\alpha/2), \mathrm{qnt}_{\mathrm{t}(n_1+n_2-2)}(1-\alpha/2)\right)$

$H_1 : \mathsf{E}\xi_1 > \mathsf{E}\xi_2$ $\mathscr{A}_{\text{крит}} = (\mathrm{qnt}_{\mathrm{t}(n_1+n_2-2)}(1-\alpha), \infty)$

$H_1 : \mathsf{E}\xi_1 < \mathsf{E}\xi_2$ $\mathscr{A}_{\text{крит}} = (-\infty, \mathrm{qnt}_{\mathrm{t}(n_1+n_2-2)}\,\alpha)$

Испытания Бернулли Пусть $\xi_i \sim \mathrm{Ber}(p_i)$, $i \in \{1, 2\}$. Рассмотрим $H_0 : p_1 = p_2$ против $H_1 : p_1 \neq p_2$. Поскольку $\mathsf{E}\xi_i = p_i$, применим двух-выборочный t-критерий. Объедим выборки и запишем:

$$\mathsf{D}\left(\bar{\mathbf{x}} - \bar{\mathbf{y}}\right) = \frac{\hat{p}(1-\hat{p})}{n_1 + n_2} \implies t = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}} \xrightarrow[n_1, n_2 \to \infty]{} \mathrm{N}(0,1), \quad \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

Разбиение Аналогично с $\mathrm{qnt}_{\mathrm{N}(0,1)}$.

Определение. Выборка обладает *сбалансированным дизайном*, если $n_1 = n_2$.

Если дизайн сбалансирован, то

$$s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)=\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2},$$

т.е. даже если дисперсии разные, результат одинаковый. Это остается справедливым даже при $n_1 \approx n_2$.

4.10.2. Непараметрический t-критерий

Можно использовать обычный t-критерий, но примененный к рангам.

Пусть, как и прежде, дана выборка $(\mathbf{x},\mathbf{y})$. Следующие два критерия — Wilcoxon и Mann-Whitney — проверяют гипотезу $H_0 : \mathsf{P}(\xi_1 > \xi_2) = \mathsf{P}(\xi_1 < \xi_2)$ или, альтернативно, что выборки получены из одной генеральной совпокупности.

4.10.3. Критерии суммы рангов Wilcoxon

Следует сопоставить каждой выборке соответствующие её элементам ранги в *объединенной выборке*:

$$\begin{aligned}(x_1,\dots,x_{n_1}) &\mapsto (R_1,\dots,R_{n_1})\\ (y_1,\dots,y_{n_2}) &\mapsto (T_1,\dots,T_{n_2}).\end{aligned}$$

Ясно, что если в целом элементы одной выборки окажутся больше другой, то нельзя будет говорить об их однородности. Определим

$$W_1 := \sum_{i=1}^{n_1} R_i, \qquad W_2 := \sum_{i=1}^{n_2} T_i.$$

В качестве статистики можно было бы использовать либо W_1, либо W_2, однако, ни той, ни другой статистике невозможно априорно отдать предпочтение. Поэтому используется статистика

$$W := \max\left(W_1, W_2\right),$$

не имеющая аналитического выражения (но для которого посчитаны соответствующие таблицы).

Иногда в качестве статистики берут количество инверсий в объединенной выборке.

4.10.4. Критерий Mann-Whitney (U test)

Используется статистика

$$U := \max\left(n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_1, n_1 n_2 + \frac{n_2(n_2+1)}{2} - W_2\right).$$

Замечание. $\mathsf{E}U - n_1 n_2/2 \xrightarrow[n_1,n_2\to\infty]{} 0$. Асимптотически,

$$\frac{U-\mathsf{E}U}{\sqrt{\mathsf{D}U}} \xrightarrow[n_1,n_2\to\infty]{} \mathrm{N}(0,1),$$

но для малых объемов выборки можно посчитать и точные распределения.

Замечание. Критерий состоятельный против альтернативы

$$H_1 : \mathsf{P}(\xi_1 > \xi_2) \neq \mathsf{P}(\xi_1 < \xi_2).$$

Если формы распределений одинаковы, то эта альтернатива обозначает сдвиг. Для симметричных распределений это условие обозначает равенство медиан (а для нормального — математических ожиданий). Поэтому критерий устойчив к аутлаерам, хоть и за счет небольшой ($\approx 5\%$) потери мощности.

Замечание. Критерии Манна-Уитни и Вилкоксона *эквивалентны* — в том смысле, что выделяют один и тот же p-value. Тем не менее, проверяют они разные гипотезы ($\mathsf{E}\xi$ не то же, что $\operatorname{med}\xi$).

4.10.5. Критерий серий (runs)

Следует объединить выборку и в качестве статистики выбрать количество серий, т.е. подряд идущих элементов из одной выборки. Эта статистика имеет специально подобранное распределение.

4.10.6. Критерий равенства распределений

Рассматривается $H_0 : \mathcal{P}_{\xi_1} = \mathcal{P}_{\xi_2}$ против $H_1 : \mathcal{P}_{\xi_1} \neq \mathcal{P}_{\xi_2}$. Этот критерий мощный против неравенства распределений (а не отличия математического ожидания). В качестве статистики используется

$$D = \sup_x \left| \widehat{\operatorname{cdf}}_{\xi_1}(x) - \widehat{\operatorname{cdf}}_{\xi_2}(x) \right|.$$

Замечание. Все эти критерии подразумевают отсутствие повторяющихся наблюдений для избежания появления дробных рангов.

4.11. Равенство математических ожиданий для парных (зависимых) выборок

Выборка представлена набором пар $\{(x_i, y_i)\}_{i=1}^n$.

4.11.1. t-критерий

Пусть ξ_1, ξ_2 заданы на одном $(\Omega, \mathcal{F}, \mathsf{P})$. Тогда гипотезу $H_0 : \mathsf{E}\xi_1 = \mathsf{E}\xi_2$ можно свести к $H_0 : \mathsf{E}(\xi_1 - \xi_2) = \mathsf{E}\eta = 0$ использовать не-парный t-тест.

Замечание (Мощность и зависимость). Сравним статистику для сбалансированного дизайна:

- Независимая выборка
$$t_{\text{indep}} = \frac{\bar{\mathbf{x}} - \bar{\mathbf{y}}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \frac{\sqrt{n}\,(\bar{\mathbf{x}} - \bar{\mathbf{y}})}{\sqrt{\sigma_1^2 + \sigma_2^2}}.$$

- Зависимая выборка:
$$\begin{aligned} \mathsf{D}(\bar{\mathbf{x}} - \bar{\mathbf{y}}) &= \mathsf{D}\bar{\mathbf{x}} + \mathsf{D}\bar{\mathbf{y}} - 2\operatorname{cov}(\bar{\mathbf{x}}\bar{\mathbf{y}}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - 2\rho\sqrt{\mathsf{D}\bar{\mathbf{x}}}\sqrt{\mathsf{D}\bar{\mathbf{y}}} \\ &= \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - 2\rho\frac{\sigma_1}{\sqrt{n}}\frac{\sigma_2}{\sqrt{n}} = \frac{1}{n}\left(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\right), \end{aligned}$$
откуда
$$t_{\text{dep}} = \frac{\sqrt{n}\,(\bar{\mathbf{x}} - \bar{\mathbf{y}})}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho}}.$$

При $\rho > 0$, $t_{\text{dep}} > t_{\text{indep}}$. Значит, статистика чаще попадает в критическую область и критерий лучше находит различия (и мощность, следовательно, выше). Значит, тот же эксперимент на зависимых выборках мощнее.

Пример. Проверяют гипотезу, что белый свет лам влияет на решение задач.

- При тестировании на разных индивидах, должна быть уверенность, что они одинаковы по критичным параметрам (IQ, например).
- При тестировании на одинаковых индивидах следует составлять разные, но одинаковы по сложности задачи (второй раз одну и ту же задачу решать не займет много времени!). Мощность этого эксперимента будет выше.

4.11.2. Непараметрический тест знаков (Sign test)

$H_0 : \mathsf{P}(\xi_1 < \xi_2) = \mathsf{P}(\xi_1 > \xi_2)$. Используется статистика

$$W = \sum_{i=1}^{n} \psi_i, \quad \psi_i = \begin{cases} 1 & x_i > y_i \\ 0 & x_i < y_i. \end{cases}$$

Если при подсчете статистики $x_i = y_i$, эта пара игнорируется вместе с соответствующим уменьшением объема выборки.

Пусть после удаления всех пар, таких, что $x_i = y_i$, объем выборки стал равен m. Тогда $W \sim \mathrm{Bin}(m, 0.5)$ и для построения разбиения можно пользоваться $\mathrm{qnt}_{\mathrm{Bin}(m,0.5)}$.

Замечание. Критерий применим к порядковым признакам.

Замечание. Критерий очень устойчив к аутлаерам (но и очень низкомощен поэтому).

4.11.3. Непараметрический критерий (Paired Wilcoxon; Wilcoxon signed-rank test)

Увеличить мощность предыдущего критерия можно, учтя больше информации:

$$W := \sum_{i=1}^{n} R_i \psi_i, \quad R_i := \mathrm{rk}\,|x_i - y_i|\,.$$

Для симметрии можно рассмотреть статистику

$$W = \sum_{i=1}^{n} R_i \operatorname{sign}(x_i - y_i)$$

с идеальным значением 0. При верной H_0, распределение W не имеет простого аналитического выражения (но может быть посчитана по таблицам), при этом $\mathsf{E}W = 0$, $\mathsf{D}W = n(n+1)(2n+1)/6$. Кроме того, $W \xrightarrow{\mathrm{d}} \mathrm{N}(0, \mathsf{D}W)$, так что уже при $n \geq 10$ можно полагать, что $z = W/\sqrt{\mathsf{D}W} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1)$ и строить разбиение соответственно.

Замечание. Критерий уже не применим к порядковым признакам.

4.12. Равенство дисперсии для двух распределений

$H_0 : \mathsf{D}\xi_1 = \mathsf{D}\xi_2$, $\xi_1 \perp\!\!\!\perp \xi_2$, $\xi_i \sim \mathrm{N}(a, \sigma_i)^2$.

4.12.1. Критерий Фишера

$H_0 : \sigma_1^2 = \sigma_2^2$. Естественно использовать отношение s_1^2/s_2^2 с идеальным значением 1. Поделив на число степеней свободы, получим статистику

$$F := \frac{\tilde{s}_1^2}{\tilde{s}_2^2} \sim \mathrm{F}(|\mathbf{x}| - 1, |\mathbf{y}| - 1).$$

Замечание. При отклонении от нормальности не становится асимптотическим.

4.12.2. Критерий Левена (Levene's test)

Так как $\mathsf{D}\xi_i = \mathsf{E}\left(\xi_i - \mathsf{E}\xi_i\right)^2$, то критерий о равенстве дисперсий можно было бы свести к критерию о равенстве математических ожиданий; в этом случае применили бы t-критерий (подразумевающий разные дисперсии) к выборкам $\left\{(x_i - \bar{\mathbf{x}})^2\right\}$ и $\left\{(y_i - \bar{\mathbf{y}})^2\right\}$. Однако при возведении в квадрат распределение стало бы несимметричным и потребовался бы больший объем выборки. Кроме того, значительно бы усилились аутлаеры.

Вместо этого используют гипотезу $H_0 : \mathsf{E}\left|\xi_1 - \mathsf{E}\xi_1\right| = \mathsf{E}\left|\xi_2 - \mathsf{E}\xi_2\right|$ вместе с t-критерием, подразумевающим равенство дисперсий (для нормальных данных; иначе с разными).

4.12.3. Критерий Brown-Forsythe

Критерий Brown-Forsythe — это t-критерий для гипотезы $H_0 : \mathsf{E}\left|\xi_1 - \operatorname{med}\xi_1\right| = \mathsf{E}\left|\xi_2 - \operatorname{med}\xi_2\right|$.

Замечание. Устойчив к аутлаерам из-за использования $\operatorname{med}\xi_i$.

5. Доверительное оценивание

5.1. Мотивация и определение

Для построенных оценок может понадобиться оценка точности. Так, даже состоятельная оценка может не быть в полном смысле «точной»: пусть $\theta_n^* \xrightarrow{\mathsf{P}} \theta_0$; тогда

$$\hat{\theta}_n' = \begin{cases} c & n < N \gg 1 \\ \hat{\theta}_n & \text{иначе} \end{cases}$$

все-равно будет, конечно, состоятельной.

$\mathsf{D}\hat{\theta}_n$ может быть не всегда просто вычислить и использовать.

Определение. $[c_1, c_2]$ — *доверительный интервал* для параметра θ_0 с уровнем доверия $\gamma \in [0,1]$, если $\forall \theta_0$

$$\mathsf{P}(\theta_0 \in [c_1, c_2]) = \gamma, \quad \text{где } c_1 = c_1(\mathbf{x}), c_2 = c_2(\mathbf{x}) \text{ — статистики.}$$

Замечание. Если выборка из дискретного распределения, то c_1, c_2 — тоже. Поэтому наперед заданную точность получить может не получиться; в таких случаях знак «$=$» заменяют «$\geq$». Аналогично с заменой на «$\xrightarrow[n\to\infty]{}$».

5.2. Доверительные интервалы для математического ожидания и дисперсии в нормальной модели

Предположение. *Пусть* $\xi \sim \mathrm{N}(a, \sigma^2)$.

5.2.1. Доверительный интервал для a

- Пусть σ^2 известно. Свяжем a_0 с выборкой:

$$\gamma = \mathsf{P}(c_1 < T < c_2) = \mathsf{P}\left(c_1 < \sqrt{n}\frac{(\bar{\mathbf{x}} - a_0)}{\sigma} < c_2\right) = \mathsf{P}\left(a_0 \in \left(\bar{\mathbf{x}} - \frac{\sigma c_2}{\sqrt{n}}, \bar{\mathbf{x}} - \frac{\sigma c_1}{\sqrt{n}}\right)\right).$$

  Решений уравнения $\mathsf{P}\left(c_1 < \sqrt{n}(\bar{\mathbf{x}} - a_0)/\sigma < c_2\right) = \Phi(c_2) - \Phi(c_1) = \gamma$ бесконечно много. Чем $[c_1, c_2]$ короче, тем лучше. Поскольку Φ симметрична и унимодальна,

$$\begin{aligned} c_1 &= -c_\gamma \\ c_2 &= c_\gamma, \end{aligned} \quad \text{где } c_\gamma = \operatorname{cdf}^{-1}_{\mathrm{N}(0,1)}\left(\gamma + \frac{1-\gamma}{2}\right) = x_{\frac{1+\gamma}{2}}.$$

  Наконец,

$$\mathsf{P}\left(a_0 \in \left(\bar{\mathbf{x}} \pm \frac{\sigma}{\sqrt{n}} x_{\frac{1+\gamma}{2}}\right)\right) = \gamma.$$

- Пусть σ^2 неизвестно. По аналогии,

$$\gamma = \mathsf{P}\left(c_1 < \frac{\sqrt{n-1}(\bar{\mathbf{x}} - a_0)}{s} < c_2\right) = \mathsf{P}\left(a_0 \in \left(\bar{\mathbf{x}} \pm \frac{c_\gamma s}{\sqrt{n-1}}\right)\right), \quad c_\gamma = \mathrm{cdf}^{-1}_{\mathrm{t}(n-1)}\left(\frac{1+\gamma}{2}\right)$$

и

$$\mathsf{P}\left(a_0 \in \left(\bar{\mathbf{x}} \pm \frac{\tilde{s}}{\sqrt{n}} x_{\frac{1+\gamma}{2}}\right)\right) = \gamma.$$

Упражнение. Пусть $s^2 = 1.21, \bar{\mathbf{x}} = 1.9, n = 36$. Построить 95% доверительный интервал для $\mathsf{E}\xi$.

Решение.

$$c_\gamma = \mathtt{qt}(0.975, 35) \approx 2.03 \implies \left(1.9 \pm \frac{2.03 \cdot \sqrt{1.21}}{\sqrt{35}}\right) = (1.52; 2.28).$$

⌟

5.2.2. Доверительный интервал для σ^2

- Пусть a известно. Поскольку плотность χ^2 становится все более симметричной с ростом n, примем

$$c_1 = \mathrm{cdf}^{-1}_{\chi^2(n)}\left(\frac{1-\gamma}{2}\right), \; c_2 = \mathrm{cdf}^{-1}_{\chi^2(n)}\left(\frac{1+\gamma}{2}\right).$$

Тогда

$$\mathsf{P}\left(c_1 < \frac{ns_a^2}{\sigma_0^2} < c_2\right) = \gamma \iff \mathsf{P}\left(\sigma_0^2 \in \left(\frac{ns_a^2}{x_{(1+\gamma)/2}}, \frac{ns_a^2}{x_{(1-\gamma)/2}}\right)\right) = \gamma.$$

- Пусть a неизвестно. Тогда аналогично

$$\mathsf{P}\left(\sigma_0^2 \in \left(\frac{ns^2}{x_{(1+\gamma)/2}}, \frac{ns^2}{x_{(1+\gamma)/2}}\right)\right) = \gamma,$$

где $x_{(1\pm\gamma)/2} = \mathrm{cdf}^{-1}_{\chi^2(n-1)}((1 \pm \gamma)/2)$.

Определение. Случайная величина $g(x_1, \ldots, x_n, \theta)$ называется *центральной статистикой параметра* θ, если

1. Её распределение («центральное распределение») не зависит от распределения θ.
2. G_n (функция распределения центрального распределения) непрерывна.
3. $\forall z_1, z_2$ и $\mathcal{P}_\theta$-почти всюду

$$z_1 < g(x_1, \ldots, x_n, \theta) < z_2$$

монотонно разрешимо относительно θ, т.е.

$$\exists f_1, f_n : f_1(x_1, \ldots, x_n, \theta, z_1, z_2) < \theta < f_2(x_1, \ldots, x_n, \theta, z_1, z_2).$$

Рассмотрим всегда разрешимое

$$\begin{aligned} \gamma &= G_n(z_2) - G_n(z_1) = \mathsf{P}(z_1 < g(x_1, \ldots, x_n, \theta) < z_2) \\ &= \mathsf{P}(\underbrace{f_1(z_1, z_2, x_1, \ldots, x_n)}_{c_1} < \theta < \underbrace{f_2(z_1, z_2, x_1, \ldots, x_n)}_{c_2}). \end{aligned}$$

5.3. Асимптотический доверительный интервал для математического ожидания в модели с конечной дисперсией

Если модель неизвестна, но известно, что $\mathsf{D}\xi < \infty$, можно построить доверительный интервал для $\mathsf{E}\xi = a$. Пусть $\{x_i\}$ i.i.d., тогда

$$t = \frac{\sqrt{n}\,(\bar{\mathbf{x}} - a)}{\sigma} \xrightarrow[n\to\infty]{} \mathrm{N}(0,1).$$

Если положить $\sigma := s$, то сходимость не испортится, потому что s^2 — состоятельная оценка σ^2. Тогда

$$\mathsf{P}\left(\mathsf{E}\xi \in \left(\bar{\mathbf{x}} \pm \frac{sc_\gamma}{\sqrt{n}}\right)\right) \xrightarrow[n\to\infty]{} \gamma, \quad c_\gamma = \mathrm{cdf}^{-1}_{\mathrm{t}(n-1)}\left(\frac{1+\gamma}{2}\right).$$

Альтернативно замену σ на s можно обосновать по теореме Слуцкого.

Утверждение (Слуцкий). Если $\xi_n \xrightarrow{\mathrm{d}} \xi$, $\eta_n \xrightarrow{\mathsf{P}} c$, то $\xi_n + \eta_n \xrightarrow{\mathrm{d}} \xi + c$ и $\xi_n\eta_n \xrightarrow{\mathrm{d}} c\xi$.

Используя тот факт, что $s \xrightarrow{\mathsf{P}} \sigma$, запишем

$$\mathsf{P}\left(c_1 < \frac{\sqrt{n}\,(\bar{\mathbf{x}} - a)}{\sigma}\frac{\sigma}{s} < c_2\right) \xrightarrow[n\to\infty]{} \Phi(c_2) - \Phi(c_1).$$

5.4. Асимптотический доверительный интервал для параметра на основе MLE

Если умеем находить $\hat\theta_{\mathrm{MLE}}$, т,о по асимптотической нормальности,

$$\frac{\hat\theta_{\mathrm{MLE}} - \mathsf{E}\hat\theta_{\mathrm{MLE}}}{\sqrt{\mathsf{D}\hat\theta_{\mathrm{MLE}}}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1),$$

по асимптотической несмещенности,

$$\frac{\hat\theta_{\mathrm{MLE}} - \theta}{\sqrt{\mathsf{D}\hat\theta_{\mathrm{MLE}}}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1),$$

и, учитывая асимптотическую эффективность ($\mathsf{D}\hat\theta_{\mathrm{MLE}} I_n(\theta) \xrightarrow[n\to\infty]{} 1$), запишем статистику

$$T = \left(\hat\theta_{\mathrm{MLE}} - \theta\right)\sqrt{I_n(\theta)} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1).$$

Чтобы по аналогии с предыдущим выразить θ в $\mathsf{P}(c_1 < T < c_2) = \mathsf{P}(|T| < c_\gamma) = \gamma$, необходимо выразить θ из $I_n(\theta)$. Для Pois и Ber это эквивалентно решению квадратного уравнения.

В общем случае, можно вместо θ в $I_n(\theta)$ подставить $\hat\theta_{\mathrm{MLE}}$ (при $n \to \infty$ это не должно сильно испортить дело), откуда

$$\mathsf{P}(|T| < c_\gamma) = \gamma \iff \mathsf{P}\left(-c_\gamma < \left(\hat\theta_{\mathrm{MLE}} - \theta\right)\sqrt{I_n(\theta)} < c_\gamma\right) = \gamma \iff \mathsf{P}\left(\theta \in \left(\hat\theta_{\mathrm{MLE}} \pm \frac{c_\gamma}{\sqrt{I_n(\theta)}}\right)\right) = \gamma,$$

где

$$T \xrightarrow{\mathrm{d}} \mathrm{N}(0,1) \implies c_\gamma = \mathrm{cdf}^{-1}_{\mathrm{N}(0,1)}\left(\frac{1+\gamma}{2}\right).$$

Пример. $\xi \sim \mathrm{Pois}(\lambda)$. По 2.5.1, $\hat\lambda_{\mathrm{MLE}} = \bar{\mathbf{x}}$, по 2.5.2 $I_n(\lambda) = n/\lambda = n/\bar{\mathbf{x}}$ откуда

$$\mathsf{P}\left(\lambda \in \left(\bar{\mathbf{x}} \pm \mathrm{cdf}^{-1}_{\mathrm{N}(0,1)}\left(\frac{1+\gamma}{2}\right)\frac{\sqrt{\bar{\mathbf{x}}}}{\sqrt{n}}\right)\right) = \gamma.$$

Пример. $\xi \sim \mathrm{Ber}(p)$. $p = \mathsf{E}\xi$. $\hat p = \bar{\mathbf{x}}$, откуда

$$\mathsf{P}\left(p \in \left(\hat p \pm c_\gamma\frac{\sqrt{\hat p(1-\hat p)}}{\sqrt{n}}\right)\right) \xrightarrow[n\to\infty]{} \gamma.$$

Замечание. Этот доверительный интервал не очень хорош, потому что не принадлежит $[0,1]$.

5.5. Доверительный интервал для проверки гипотезы о значении параметра

Зафиксируем $H_0 : \theta = \theta_0$ и $\gamma = 1 - \alpha$, где α играет роль уровня значимости. По определению доверительного интервала, $\mathsf{P}(\theta \in [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})]) = \gamma$. Тогда разбиением будет

$$\mathscr{A}_{\text{дов}} = [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})]\,,\ \mathscr{A}_{\text{крит}} = \mathbb{R} \setminus \mathscr{A}_{\text{дов}}\,,$$

причем

$$\mathsf{P}(\theta_0 \notin [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})]) = \alpha.$$

Иными словам, попадание в критическую область происходит с уровнем значимости α, что соответствует определению критерия.

5.5.1. Использование SE для построения доверительных интервалов

Пусть $\hat\theta$ — асимптотически нормальная оценка параметра θ. Чтобы сходимость оставалась верной, даже если подставляем оценку, нужна состоятельность. Тогда доверительный интервал для среднего с уровнем доверия γ имеет вид

$$\mathsf{E}\hat\theta \pm c_\gamma\sqrt{\mathsf{D}\hat\theta} \approx \mathsf{E}\hat\theta \pm \mathrm{cdf}^{-1}_{\mathrm{N}(0,1)}\left(\frac{1+\gamma}{2}\right)\cdot \mathrm{SE}\,.$$

6. Корреляционный и регрессионный анализы

Определение. *Мера зависимости* — это функционал $r : (\xi, \eta) \mapsto x \in [-1, 1]$ со свойствами:

1. $|r| \le 1$.
2. $\xi \perp\!\!\!\perp \eta \implies r(\xi, \eta) = 0$.
3. Если ξ и η «максимально зависимы», то $r(\xi, \eta) = 1$.

6.1. Вероятностная независимость

6.1.1. Визуальное определение независимости

- Поскольку при $p_\eta(y_0) \ne 0$

  $$\xi \perp\!\!\!\perp \eta \iff p_{\xi|\eta}(x \mid y_0) = \frac{p_{\xi,\eta}(x, y_0)}{p_\eta(y_0)} = p_\xi(x),$$

  то срезы графика совместной плотности при фиксированном y_0 после нормировки $p_\eta(y_0)$ должны выглядеть одинаково для всех y_0.

- Для выборки независимость можно попытаться определить по *таблицам сопряженности*: сгруппируем $\{(x_i, y_i)\}_{i=1}^n$ и сопоставим каждой уникальной паре абсолютную частоту ν_{ij}:

  $$\begin{array}{cccc} & y_1^* & \cdots & y_s^* \\ x_1^* & \nu_{11} & \cdots & \nu_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ x_k^* & \nu_{k1} & \cdots & \nu_{ks} \end{array}$$

  Тогда признаки с большей чем случайной вероятностью будут независимы при пропорциональных строчках / столбцах. Более формально, признаки независимы, если

  $$\frac{\nu_{ij}}{\sum_k \nu_{kj}} = \frac{\nu_{ij}}{\nu_{\cdot j}} = \hat p_{i|j} \propto \hat p_{i|\ell},$$

  т.е. вероятности условного распределения не зависят от выбора строки.

Пример. Таблица сопряженности похожей на независимую выборки:

1	3	2
2	5	3
9	20	11

6.1.2. Критерий независимости χ^2

По определению, для двумерных дискретных распределений, независимость есть

$$\xi \perp\!\!\!\perp \eta \iff \underbrace{\mathsf{P}(\xi = i, \eta = j)}_{p_{ij}} = \underbrace{\mathsf{P}(\xi = i)}_{p_i}\underbrace{\mathsf{P}(\eta = j)}_{p_j} = \underbrace{\sum_{k=1}^{K} \mathsf{P}(\xi = i, \eta = k)}_{p_{i\cdot}} \cdot \underbrace{\sum_{s=1}^{S} \mathsf{P}(\xi = s, \eta = j)}_{p_{\cdot j}}.$$

Проверим $H_0 : \xi \perp\!\!\!\perp \eta$.

Утверждение. ОМП оценкой будет $\hat{p}_{i\cdot} = \nu_{i\cdot}/n$ и $\hat{p}_{\cdot j} = \nu_{\cdot j}/n$.

Следовательно,

$$\xi \perp\!\!\!\perp \eta \iff \hat{p}_{ij} = \frac{\nu_{ij}}{n} = \hat{p}_{i\cdot}\hat{p}_{\cdot j} = \frac{\nu_{i\cdot}}{n} \cdot \frac{\nu_{\cdot j}}{n}.$$

Это равенство удается получить редко; важно определить, не является ли это нарушение случайным.

Запишем статистику

$$\chi^2 = \sum_{i=1}^{K}\sum_{j=1}^{S} \frac{(\nu_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}} = \sum_{i=1}^{K}\sum_{j=1}^{S} \frac{(\nu_{ij} - \nu_{i\cdot}\nu_{\cdot j}/n)^2}{\nu_{i\cdot}\nu_{\cdot j}/n} \xrightarrow{\mathrm{d}} \chi^2((k-1)(s-1))$$

Количество параметров таково, потому что если $\xi \parallel \eta$, то всего $ks - 1$ параметров (-1 потому что $\sum_{ij} p_{ij} = 1$); если $\xi \perp\!\!\!\perp \eta$, то $k + s - 2$ (-2 потому что $\sum_i p_{ij} = 1$ и $\sum_j p_{ij} = 1$). Значит $ks - 1 - k - s + 2 = (k-1)(s-1)$.

Пример. Дано S кубиков. Проверить гипотезу, что кубики одинаковы.

Решение. FIXME ⌟

Замечание. На маленьких выборках ($n < 40$, $np_{ij} < 5$) возникают проблемы со сходимостью, потому что можно объединять только столбцы / строки и каждый раз терять сразу $S-1$ ($K-1$) степень свободы. В этих случаях используют критерием с перестановкой[11] или, в случае таблиц сопряженности 2×2, точным критерием Фишера.

Замечание. Критерий верен для количественных, порядковых и качественных признаков, потому что нигде не участвуют значения из выборки.

Замечание. Критерий асимптотический, поэтому $\alpha_{\mathrm{I}} \to \alpha$.

Замечание. Критерий не удовлетворяет 1-му пункту определения меры зависимости ($\chi^2 \notin [-1, 1]$). Это обычно исправляют так: рассматривают *среднеквадратичную сопряженность*

$$r^2 := \frac{\chi^2}{n}$$

или коэффициент сопряженности Пирсона

$$p^2 := \frac{\chi^2}{\chi^2 + n}$$

(тогда 1 никогда не достигается). Могли бы работать с $1 - p$-value, но так почему-то никогда не делают.

[11] https://en.wikipedia.org/wiki/Resampling_(statistics)#Permutation_tests

6.2. Линейная / полиномиальная зависимость

Пусть теперь ξ, η — количественные признаки.

Определение. Определим

$$\phi(x) := \mathsf{E}\{\eta \mid \xi = x\}.$$

Тогда назовем зависимость *линейной*, если $\phi(x)$ — линейная функция, *квадратичной* — если квадратичная и т.д.

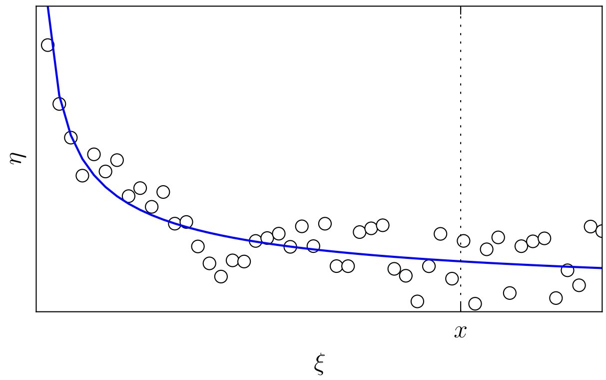


Рис. 3: Нелинейная зависимость

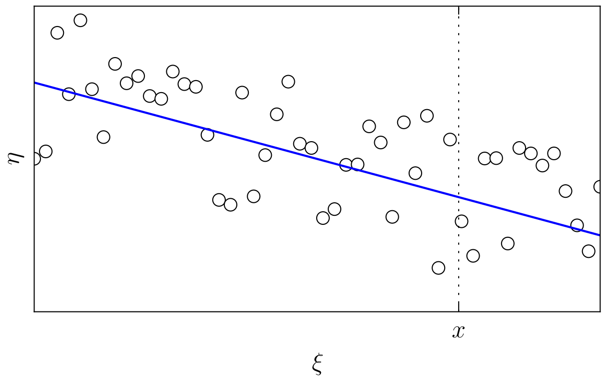


Рис. 4: Линейная зависимость

Определение. Мера *линейной* зависимости между случайным величинами ξ и η есть *коэффициент корреляции Пирсона*

$$\rho = \frac{\operatorname{cov}(\xi, \eta)}{\sqrt{\mathsf{D}\xi}\sqrt{\mathsf{D}\eta}}.$$

Замечание. Про ρ можно думать как про cos между векторами в соответствующем пространстве.

Замечание (Важное).

$$\begin{aligned} \xi \perp\!\!\!\perp \eta \quad &\Longrightarrow \quad \rho = 0 \\ \xi, \eta \sim \mathrm{N}(a, \sigma^2),\ \xi \perp\!\!\!\perp \eta \quad &\Longleftrightarrow \quad \rho = 0. \end{aligned}$$

Предложение. *Для линейно зависимых данных, конечно,* $\rho = 1$.

Доказательство. Пусть $\eta = a + b\xi$; тогда

$$\begin{aligned} \rho(\xi, \eta) &= \frac{\operatorname{cov}(\xi, a + b\xi)}{\sqrt{\mathsf{D}\xi}\sqrt{\mathsf{D}(a + b\xi)}} = \frac{\mathsf{E}\xi(a + b\xi) - \mathsf{E}\xi\mathsf{E}(a + b\xi)}{\sqrt{\mathsf{D}\xi}\sqrt{\mathsf{D}b\xi}} = \frac{\mathsf{E}\xi a + b\mathsf{E}\xi^2 - \mathsf{E}\xi\mathsf{E}a - \mathsf{E}\xi b\mathsf{E}\xi}{b\sqrt{\mathsf{D}\xi}\sqrt{\mathsf{D}\xi}} = \\ &= \frac{a\mathsf{E}\xi + b\mathsf{E}\xi^2 - a\mathsf{E}\xi - b(\mathsf{E}\xi)^2}{b\mathsf{D}\xi} = \frac{b(\mathsf{E}\xi^2 - (\mathsf{E}\xi)^2)}{b\mathsf{D}\xi} = 1. \end{aligned}$$

□

О соотношении ρ и коэффициента линейной регрессии По (6.5.1), если линейная регрессия уравнением $y = kx + b$, то

$$k = \rho\frac{\sigma_\eta}{\sigma_\xi}.$$

В общем случае, по виду прямой линейной регрессии ничего нельзя сказать о зависимости между случайными величинами. Так, если $\eta = a + b\xi$ есть линейная функция от ξ, то, по предыдущему, $\rho = 1$ и

$$k = 1 \cdot \frac{\sqrt{\mathsf{D}(a + b\xi)}}{\sqrt{\mathsf{D}\xi}} = b$$

и прямая может иметь произвольный, в зависимости от b, наклон.

Замечание. В то же время, поскольку для
$$\begin{pmatrix}\xi\\ \eta\end{pmatrix} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma), \quad \Sigma = \begin{pmatrix}\sigma_\xi^2 & \operatorname{cov}(\xi,\eta)\\ \operatorname{cov}(\xi,\eta) & \sigma_\eta^2\end{pmatrix}$$
справедливо, что
$$k = \rho\frac{\sigma_\eta}{\sigma_\xi} = \frac{\operatorname{cov}(\xi,\eta)}{\sigma_\xi\sigma_\eta}\cdot\frac{\sigma_\eta}{\sigma_\xi} = \frac{1}{\sigma_\xi^2}\operatorname{cov}(\xi,\eta),$$
то $k = 0 \iff \operatorname{cov}(\xi,\eta) = 0$, а для стандартно нормальных данных $k = \rho = \operatorname{cov}(\xi,\eta)$.

Значимость коэффициента корреляции

Определение. Коэффициент корреляции *значим*, если отвергается $H_0 : \rho = 0$.

Статистика для проверки значимости при $(\xi,\eta)^{\mathrm{T}} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (иначе статистика асимптотическая)
$$T = \frac{\sqrt{n-2}\hat{\rho}_n}{\sqrt{1-\hat{\rho}_n^2}} \sim \mathrm{t}(n-2).$$
Идеальное значение — 0, два хвоста.

6.3. Метод наименьших квадратов (Ordinary Least Squares)

Пусть $\eta, \xi \in L^2(\mathcal{F}, \mathsf{P})$ пространству $\mathcal{F}$-измеримых по мере P функций с конечным вторым моментом и скалярным произведением $(\eta,\xi) = \mathsf{E}\eta\xi$, причем $\hat{\eta} \in K = \{\phi(\xi)\} = \{\hat{\eta} : \sigma(\phi(\xi))\text{-измерима}\}$. По свойству УМО(2.3.2), вектор
$$\hat{\eta}^* = \mathsf{E}\left\{\eta \mid \phi(\xi)\right\}$$
будет ортогональной проекцией η на K, т.е. $(\eta - \hat{\eta}^*, \hat{\eta}) = 0 \ \forall \hat{\eta} \in K$. Значит, он минимизирует квадрат нормы расстояния от η до K:
$$\hat{\eta}^* = \operatorname*{argmin}_{\hat{\eta}\in K} \|\eta - \hat{\eta}\|^2 = \operatorname*{argmin}_{\hat{\eta}\in K} \mathsf{E}\left(\eta - \hat{\eta}\right)^2 = \mathsf{E}\left\{\eta \mid \phi(\hat{\eta})\right\}.$$
$\hat{\eta}^*$ называется *наилучшим среднеквадратичным приближением в классе* K.

6.4. Корреляционное отношение

Если $K = \mathcal{L} = \{a\xi + b\}$ — линейное пространство, то теорема Пифагора принимает вид
$$\mathsf{D}\eta = \mathsf{E}\left(\eta - \mathsf{E}\eta\right)^2 = \underbrace{\mathsf{E}\left(\hat{\eta}^* - \mathsf{E}\eta\right)^2}_{\text{объяснённая доля аппроксимации}} + \underbrace{\mathsf{E}\left(\eta - \hat{\eta}^*\right)^2}_{\text{ошибка аппроксимации}}.$$
Откуда можно записать меру аппроксимации
$$\frac{\mathsf{E}(\hat{\eta}^* - \mathsf{E}\eta)^2}{\mathsf{D}\eta} = 1 - \frac{\mathsf{E}\left(\eta - \hat{\eta}^*\right)^2}{\mathsf{D}\eta} = 1 - \frac{\min_{\hat{\eta}\in\mathcal{L}} \mathsf{E}\left(\eta - \hat{\eta}\right)^2}{\mathsf{D}\eta}.$$
Определение. Полученная величина называется коэффициентом корреляции ρ^2:
$$\rho^2 := 1 - \frac{\min_{\hat{\eta}\in\mathcal{L}} \mathsf{E}\left(\eta - \hat{\eta}\right)^2}{\mathsf{D}\eta}.$$
ρ — коэффициент корреляции Пирсона.

Определение. *Множественный коэффициент корреляции* есть полученная величина для МНК с $K = \mathcal{M} = \left\{\sum_{i=1}^k b_i\xi_i + b_0\right\}$.
$$R^2(\eta, \xi_1, \ldots, \xi_k) := 1 - \frac{\min_{\hat{\eta}\in\mathcal{M}} \mathsf{E}\left(\eta - \hat{\eta}\right)^2}{\mathsf{D}\eta}.$$

Замечание. $R^2 \geq \rho^2$; если же $R^2 = \rho^2$, то $\xi_1, \ldots, \xi_k$ все зависимы.

Определение. В общем случае, если $K = \{\phi(\xi) \text{ измеримые}\}$, то полученная величина называется *корреляционным отношением*:

$$r^2_{\eta|\xi} := 1 - \frac{\min_{\hat{\eta}\in K} \mathsf{E}\,(\eta - \hat{\eta})^2}{\mathsf{D}\eta} = \frac{\mathsf{DE}(\eta \mid \xi)}{\mathsf{D}\eta}.$$

Свойства корреляционного отношения

1. $r^2_{\eta|\xi} \in [0, 1]$.
2. $\eta \perp\!\!\!\perp \xi \implies r^2_{\eta|\xi} = 0$.
3. $\eta = \phi(\xi) \iff r^2_{\eta|\xi} = 1$.
4. Вообще говоря, $r^2_{\eta|\xi} \neq r^2_{\xi|\eta}$. К примеру, для любой не монотонной функции (так, чтобы не существовала обратная).
5. $r^2_{\eta|\xi} \geq \rho^2(\eta, \xi)$ (потому что минимум по всем функциям меньше, чем лишь по линейным, значит $1 - \min$ больше).
6. $(\xi, \eta)^{\mathrm{T}} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies r^2_{\eta|\xi} = \rho^2(\eta, \xi)$.

Выборочное корреляционное отношение

$$\hat{r}^2_{\eta|\xi} = \hat{r}^2_{y|x} = \frac{s^2_{y|x}}{s^2_y}.$$

6.5. Регрессия

Определение. *Регрессией* η по ξ называется $\mathsf{E}\,\{\eta \mid \xi\}$.

Замечание. Таким образом осуществляется предсказание η по ξ с минимальной среднеквадратичной ошибкой.

Определение. *Функция регрессии* есть $f(x) = \mathsf{E}\,\{\eta \mid \xi = x\}$.

Замечание. f находится по МНК для $K = \{\psi(\xi) : \psi \text{—измеримая}\}$.

Виды регрессий

- Нелинейными и линейными ($K = \{a\xi + b\}$);
- Парными (предсказывая величину по одной случайной величине) и множественными (по многим).

6.5.1. Парная линейная регрессия

Определение. Пусть $\xi, \eta \in L^2$. *Парной линейной регрессией η по ξ* называется наилучшее среднеквадратичное приближение $h_{\beta_1^*,\beta_2^*}(\xi) = \beta_1^*\xi + \beta_2^*$ в классе линейных по ξ функций $K = \mathcal{L} = \{\beta_1\xi + \beta_2\}$. Иными словами,

$$h_{\beta_1^*,\beta_2^*}(\xi) = \underset{\beta_1,\beta_2}{\operatorname{argmin}} \|\eta - h_{\beta_1,\beta_2}(\xi)\|^2 = \mathsf{E}\,\{\eta \mid h_{\beta_1,\beta_2}(\xi)\}\} = \underset{\beta_1,\beta_2}{\operatorname{argmin}} \underbrace{\mathsf{E}(\eta - (\beta_1\xi + \beta_2))^2}_{\phi(\beta_1,\beta_2)} = \beta_1^*\xi + \beta_2^*.$$

Замечание. Найти минимум ϕ можно, как обычно, решив систему $\partial\phi/\partial\beta_i = 0$[12].

Утверждение. β_1^*, β_2^* таковы, что

$$\frac{h(\xi) - \mathsf{E}\eta}{\sqrt{\mathsf{D}\eta}} = \rho\frac{\xi - \mathsf{E}\xi}{\sqrt{\mathsf{D}\xi}}.$$

Это уравнение задает линию регрессии. Иными словами,

$$h(\xi) = \underbrace{\rho\frac{\sqrt{\mathsf{D}\eta}}{\sqrt{\mathsf{D}\xi}}}_{\beta_1^*}\xi + \underbrace{\mathsf{E}\eta - \rho\frac{\sqrt{\mathsf{D}\eta}}{\sqrt{\mathsf{D}\xi}}\mathsf{E}\xi}_{\beta_2^*}.$$

Отсюда можно получить соотношение между коэффициентом линейной регрессии $\beta_1^* = k$ (наклоном регрессионной прямой) и коэффициентом корреляции:

$$k = \rho\frac{\sigma_\eta}{\sigma_\xi}.$$

Замечание. Подстановкой проверятся, что

$$\phi(\beta_1^*, \beta_2^*) = \min_{\hat\eta\in K}\mathsf{E}\,(\eta - \hat\eta)^2 = \mathsf{D}\eta(1-\rho^2),$$

откуда можно найти уже известное выражение для коэффициента корреляции Пирсона

$$\rho^2(\eta,\xi) = 1 - \frac{\phi(\beta_1^*,\beta_2^*)}{\mathsf{D}\eta} = 1 - \frac{\min_{\hat\eta\in H}\mathsf{E}\,(\eta-\hat\eta)^2}{\mathsf{D}\eta}, \quad \hat\eta := h(\xi).$$

Определение. Линейная регрессия *значима*, если $\beta_1^* \neq 0 \implies \rho \neq 0$. Значимость регрессии эквивалентна значимости предсказания по ней.

Определение. Величина *sum of squares residual* есть

$$\mathrm{SSR} = n\cdot\phi(\beta_1^*,\beta_2^*) = \sum_{i=1}^n (y_i - \hat y_i)^2, \quad \hat y_i = h_{\beta_1^*,\beta_2^*}(x_i).$$

6.5.2. Модель линейной регрессии

Можно описать выборку как

$$y_i = \beta_1 x_i + \beta_2 + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0,\sigma^2),\ \epsilon_i \perp\!\!\!\perp \epsilon_j.$$

σ^2 — мешающий параметр, который можно оценить через SSR/n. Но если $\epsilon_i \sim \mathrm{N}(0,\sigma^2)$, то

$$\hat\sigma^2 = \frac{\mathrm{SSR}}{n-2}$$

есть несмещенная оценка σ^2. Значит,

$$\frac{\sum_{i=1}^n (y_i - \hat y_i)^2}{\sigma^2} \sim \chi^2(n-2).$$

Замечание. МНК минимизирует разницу $y_i - \hat y_i$, что на графике соответствует вертикальным отрезкам, соединяющим y_i и $\hat y_i = h(x_i)$. Это не то же, что минимизация перпендикуляров от y_i на $h(x)$ — техники метода анализа главных компонент («PCA»).

Замечание. Существует три базовых модели, в которых функция регрессии линейная:

- $\eta = \beta_1\xi + \beta_2 + \epsilon$, $\epsilon \perp\!\!\!\perp \xi$, $\mathsf{E}\epsilon = 0$.
- $(\xi,\eta)^{\mathrm{T}} \sim \mathrm{N}(\mu,\sigma^2)$.
- ξ принимает всего два значения (возможно, как качественный признак).

[12]См. https://en.wikipedia.org/wiki/Simple_linear_regression

6.5.3. Доверительные интервалы для β_1 и β_2

Как обычно, помимо точечной оценки $\hat{\beta}_1$ и $\hat{\beta}_2$, интересуемся диапазоном значений, которые может принимать оценка с заданной вероятностью. Примем предположение о несмещенности оценки, т.е. $\mathsf{E}\hat{\beta}_i = \beta_i$. Поскольку в модели $y_i = \beta_1 x_i + \beta_2 + \epsilon_i$ ошибка $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ есть случайная величина, оценки $\hat{\beta}_i$ — тоже становятся случайными величинами: $\hat{\beta}_i \sim \mathrm{N}(\beta_i, \mathsf{D}\hat{\beta}_i)$. В курсе регрессионного анализа доказывается[13], что

$$\mathsf{D}\hat{\beta}_1 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{\mathbf{x}})^2}, \qquad \mathsf{D}\hat{\beta}_2 = \frac{\sigma^2}{n}.$$

Кроме того,

$$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\mathsf{D}\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{n} s_x} = \frac{\sqrt{\frac{\mathrm{SSR}}{n-2}}}{\sqrt{\sum_{i=1}^n (x_i - \bar{\mathbf{x}})^2}}, \qquad \mathrm{SE}(\hat{\beta}_2) = \mathrm{SE}(\hat{\beta}_1) \cdot s_x$$

Предложение. *Статистика*

$$t = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\frac{1}{n-2}\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (x_i - \bar{\mathbf{x}})^2}}} \sim \mathrm{t}(n-2).$$

Доказательство. Известно,

$$t \sim \mathrm{t}(m) \iff t = \frac{\xi}{\sqrt{\eta/m}}, \quad \xi \sim \mathrm{N}(0,1),\ \eta \sim \chi^2(m).$$

Ясно, что

$$\frac{\hat{\beta}_1 - \beta_1}{\dfrac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{\mathbf{x}})^2}}} \sim \mathrm{N}(0,1), \qquad \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sigma^2} \sim \chi^2(n-2).$$

Тогда

$$\frac{\left(\dfrac{\hat{\beta}_1 - \beta_1}{\left(\dfrac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{\mathbf{x}})^2}}\right)}\right)}{\left(\dfrac{\left(\dfrac{\sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2}}{\sigma}\right)}{\sqrt{n-2}}\right)} = \frac{\dfrac{(\hat{\beta}_1 - \beta_1)\sqrt{\sum_{i=1}^n (x_i - \bar{\mathbf{x}})^2}}{\sigma}}{\dfrac{\sqrt{\sum_{i=1}^n \epsilon_i^2}}{\sigma\sqrt{n-2}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\frac{1}{n-2}\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (x_i - \bar{\mathbf{x}})^2}}} \sim \mathrm{t}(n-2).$$

□

Используя статистику t, введем доверительные интервалы с $c_\gamma = \mathrm{cdf}^{-1}_{\mathrm{t}(n-2)}((1+\gamma)/2)$:

$$t \in (-c_\gamma, c_\gamma) \iff \frac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}(\hat{\beta}_1)} \in (-c_\gamma, c_\gamma) \iff \beta_1 \in \left(\hat{\beta}_1 - c_\gamma\,\mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + c_\gamma\,\mathrm{SE}(\hat{\beta}_1)\right).$$

Аналогично, для β_2:

$$\beta_2 \in \left(\hat{\beta}_2 - c_\gamma\,\mathrm{SE}(\hat{\beta}_2), \hat{\beta}_2 + c_\gamma\,\mathrm{SE}(\hat{\beta}_2)\right).$$

[13]См. https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares

Замечание. На картинке доверительные интервалы изображаются в виде «рукавов» вокруг графика линейной регрессии — т.е. область всевозможных положений прямой при варьировании β_1, β_2 в заданных интервалах.

Пример. Линейная регрессия как предсказательная модель может быть использована неправильно в следующих случаях:

- неправильная модель;
- применение к неоднородным данным (аутлаер или неоднородность);
- хотим построить предсказание в точке, далекой от данных (проблема — большая ошибка);
- не знаем какая модель там, где данных нет.

6.6. Частная корреляция

Определение. Частная корреляция случайных величин η_1, η_2 относительно $\{\xi_1, \ldots, \xi_k\}$ есть

$$\rho\left(\eta_1, \eta_2 \mid \{\xi_1, \ldots, \xi_k\}\right) := \rho\left(\eta_1 - \hat{\eta}_1^*, \eta_2 - \hat{\eta}_2^*\right), \quad \text{где } \hat{\eta}_i^* = \underset{\hat{\eta}_i \in \left\{\sum_{i=1}^k b_i \xi_i + b_0\right\}}{\operatorname{argmin}} \mathsf{E}\left(\eta_i - \hat{\eta}_i\right)^2.$$

Если регрессия линейна, то

$$\rho(\eta_1, \eta_2 \mid \xi_1, \ldots, \xi_k) = \rho(\eta_1 - \mathsf{E}\left\{\eta_1 \mid \xi_1, \ldots, \xi_k\right\}, \eta_2 - \mathsf{E}\left\{\eta_2 \mid \xi_1, \ldots, \xi_k\right\}).$$

Замечание (Важное). Пусть в эксперименте подсчитан ненулевой ρ. Это может означать, что один из факторов является причиной, а другой следствием; чтобы установить, что есть что, проводят эксперимент и смотрят, какой фактор в реальности влияет на какой. Это может также означать, что влияет сторонний фактор. Чтобы его исключить, считают частную корреляцию.

Пример. Возможна ситуация, когда $\rho(\eta_1, \eta_2) \neq 0$, но $\rho(\eta_1, \eta_2 \mid \xi) = 0$. Частная корреляция есть, по сути, корреляция на центрированных данных.

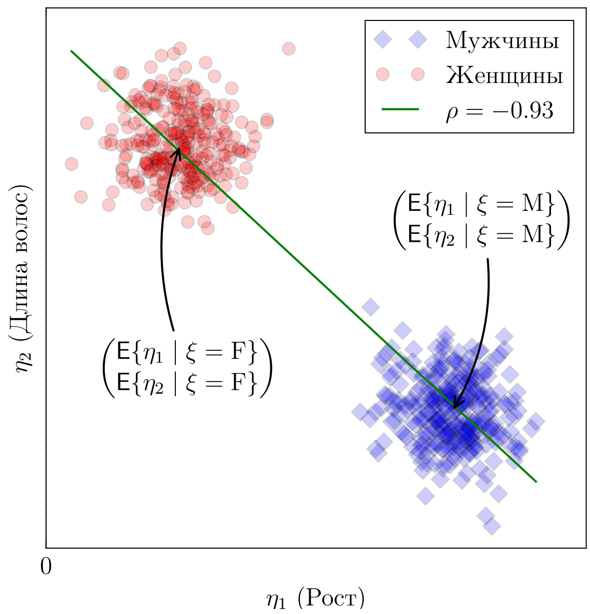


Рис. 5: Исходные данные (бимодальность)

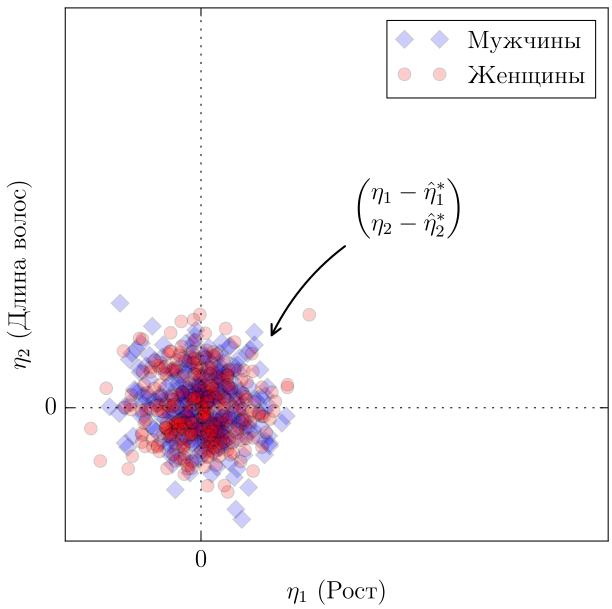


Рис. 6: Центрированные данные

Пример. Возможна и ситуация как на (7), где определенно $\rho(\eta_1, \eta_2) > 0$, но $\rho(\eta_1, \eta_2 \mid \xi) < 0$.

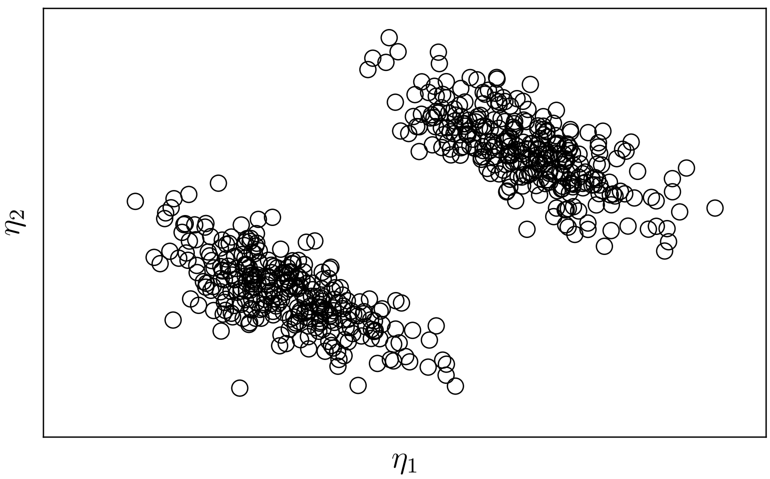


Рис. 7: $\rho(\eta_1, \eta_2) > 0$, но $\rho(\eta_1, \eta_2 \mid \xi) < 0$

Замечание. По аналогии с предыдущим примером, если $|\operatorname{im}\xi| \to \infty$, то графики (η_1, η_2) при фиксированном ξ образуют эллипсоид (в этом случае с положительной корреляцией).

6.7. Зависимость между порядковыми признаками

Пусть на выборке

$$\begin{pmatrix}\xi\\ \eta\end{pmatrix} \sim \begin{pmatrix}x_1\\ y_1\end{pmatrix}, \ldots, \begin{pmatrix}x_n\\ y_n\end{pmatrix}$$

задан только порядок. Тогда можем считать только эмпирическую функцию распределения.

Следующие коэффициенты основаны на рангах. Ранговые характеристики хорошо работают на выборках *без повторений* (чтобы, к примеру, не возникало дробных рангов).

6.7.1. Ранговый коэффициент Спирмана

Определение. *Ранговый коэффициент Спирмана* есть

$$\rho_S = \rho(\operatorname{cdf}_\xi(\xi), \operatorname{cdf}_\eta(\eta)).$$

Замечание. $\operatorname{cdf}_\xi(\xi) \sim \mathrm{U}(0,1)$, потому что $\mathsf{P}(\operatorname{cdf}_\xi(\xi) < x) = \mathsf{P}(\xi < \operatorname{cdf}_\xi^{-1}(x)) = \operatorname{cdf}_\xi(\operatorname{cdf}_\xi^{-1}(x)) = x$.

Определение. *Ранг* элемента из выборки есть его порядковый номер в упорядоченной выборке:

$$\operatorname{rk} x_{(i)} = i.$$

Обозначение. $\operatorname{rk} x_{(i)} =: R_i$, $\operatorname{rk} y_{(i)} =: T_i$.

Можем ввести эмпирическое распределение

$$\operatorname{cdf}_{\xi_n}(x_i) = \frac{\operatorname{rk} x_i}{n}, \quad \operatorname{cdf}_{\eta_n}(y_i) = \frac{\operatorname{rk} y_i}{n} = \frac{T_i}{n}.$$

Тогда будет справедливо следующее

Определение. *Выборочный коэффициент Спирмана* определяется как выборочный коэффициент корреляции Пирсона $\hat\rho$, но с заменой значений на ранги:

$$\hat\rho_S \begin{pmatrix}\xi_n\\ \eta_n\end{pmatrix} = \rho \begin{pmatrix}R_n\\ T_n\end{pmatrix} = \frac{1/n \cdot \sum_{i=1}^n R_i T_i - \bar R \bar T}{\sqrt{1/n \cdot \sum_{i=1}^n \left(R_i - \bar R\right)^2} \sqrt{1/n \cdot \sum_{i=1}^n \left(T_i - \bar T\right)^2}}.$$

Если нет повторяющихся наблюдений, то знаменатель будет одним и тем же у всех выборок объема n, значит его можно посчитать заранее. В этом (и только этом) случае, справедлива более простая формула:

$$\hat{\rho}_S = 1 - \frac{6\sum_{i=1}^n (R_i - T_i)^2}{n^3 - n}.$$

Замечание. Из последней формулы хорошо видно, что если x_i, y_i все идут в одном порядке, то $R_i - T_i = 0\ \forall i$ и $\hat{\rho}_S = 1$.

Замечание. ρ_S для *количественных* признаков есть мера монотонной зависимости:

$$\rho_S = 1 \iff (x_i > x_{i+1} \implies y_i > y_{i+1}\ \forall i)$$

(даже если зависимость нелинейная и $\rho \neq 1$). Иными словами, $\rho_S > 0$, если y имеет тенденцию к возрастанию с возрастанием x (и $\rho_S < 0$ иначе). Чем большое ρ_S, тем более явно выражена зависимость y от x в виде некоторой монотонной функции.

Согласованость ρ и ρ_S ρ_S не согласована с ρ в том же смысле, что ρ и $r_{\xi|\eta}$.

Утверждение. Если данные нормальные то справедлива формула

$$\rho = 2\sin\left(\frac{\pi}{6}\rho_S\right).$$

Значит, можем сравнить критерии между собой.

- С точностью до погрешности, по значению, $\hat{\rho}$ и $\hat{\rho}_S$ — это одно и то же (см. 8)

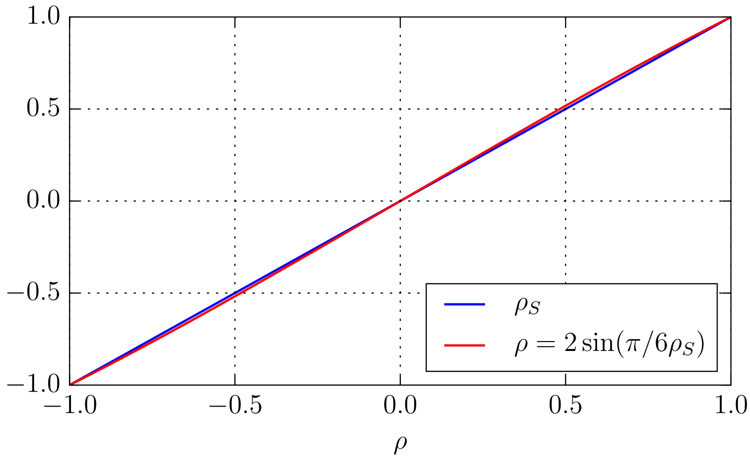


Рис. 8: $\hat{\rho} \approx \hat{\rho}_S$

- Обычный критерий оценки — выборочную дисперсию — посчитать сложно. Тем не менее, можем заметить, что $\hat{\rho}_S$ более устойчив к аутлаерам (см. 9). Всегда можно добавить аутлаер такой, что $\hat{\rho} = 0$; $\hat{\rho}_S$ же поменяется не сильно. Поэтому для нормальных данных, ρ_S — это оценка, что нет аутлаеров.

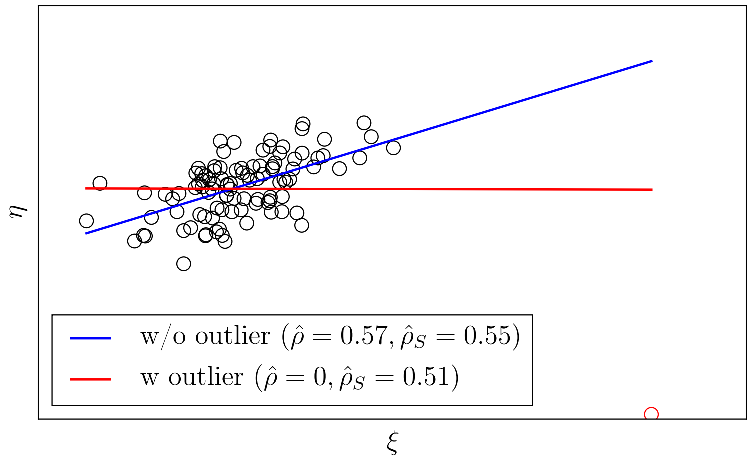


Рис. 9: $\hat{\rho}$ до и после добавления аутлаера

- Монотонным преобразованием можем всегда сделать так, чтобы ρ изменился (например, возведя в квадрат); при монотонном преобразовании, однако, не меняется ρ_S (см. 10). Значит, чтобы узнать ρ исходных (нормальных) данных, можно не выполнять обратного преобразования, а сразу посчитать ρ_S.

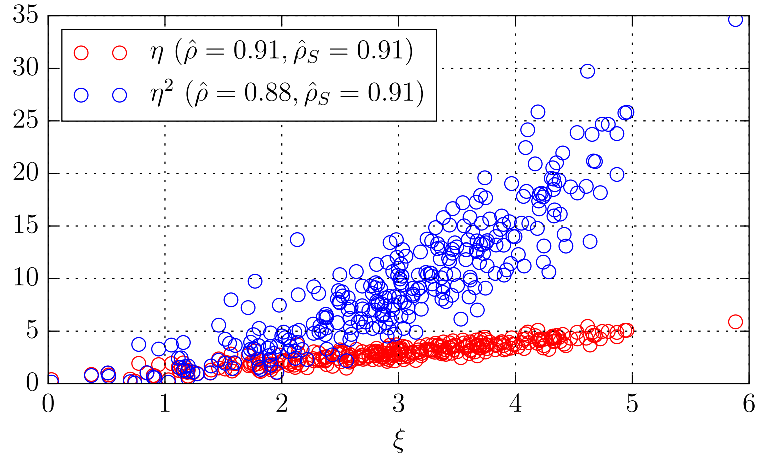


Рис. 10: Монотонное преобразование нормальных данных

6.7.2. Ранговый коэффициент Кэндалла $\tau(\xi, \eta)$

Определение. Пусть $(\xi_1, \eta_1)^{\mathrm{T}} \perp\!\!\!\perp (\xi_2, \eta_2)^{\mathrm{T}} \sim \mathcal{P}_{\xi,\eta} \sim (\xi, \eta)^{\mathrm{T}}$; тогда *ранговым коэффициентом Кэндалла* называется

$$\tau(\xi, \eta) = \rho(\operatorname{sign}(\xi_2 - \xi_1), \operatorname{sign}(\eta_2 - \eta_1)) = \mathsf{P}((\xi_2 - \xi_1)(\eta_2 - \eta_1) > 0) - \mathsf{P}((\xi_2 - \xi_1)(\eta_2 - \eta_1) < 0).$$

На выборочном языке, пусть дана выборка $(x_1, y_1), \ldots, (x_n, y_n)$; тогда

$$\tau = \frac{\#(\text{одинаково упорядоченных пар}) - \#(\text{по-разному упорядоченных пар})}{\#(\text{комбинаций пар})},$$

где пара $(x_i, y_i), (x_j, y_j)$ считается одинаково упорядоченной, если $\operatorname{sign}(x_i - x_j) = \operatorname{sign}(y_i - y_j)$, а $\#(\text{комбинаций пар}) = C_n^2 = n(n-1)/2$.

Утверждение. Если $(\xi, \eta)^{\mathrm{T}} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$, то справедлива формула

$$\rho = \sin\left(\frac{\pi}{2}\tau\right).$$

Из утверждения следует, что τ все время меньше ρ и ρ_S (по модулю).

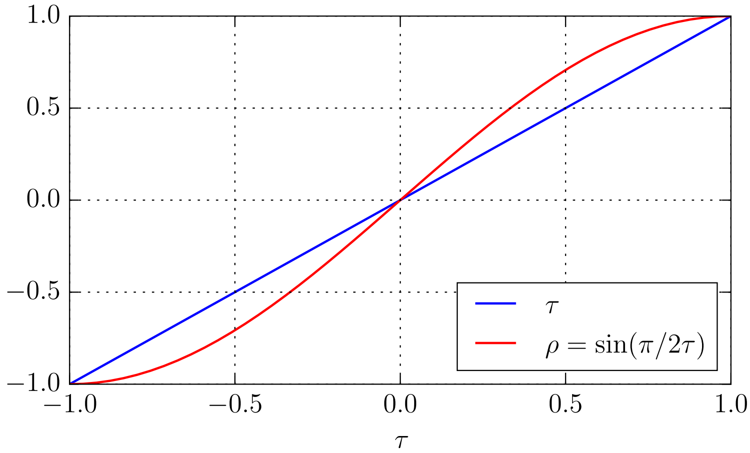


Рис. 11: ρ и τ

Пример (Проверка ряда на тренд)**.** Пусть ξ — номера точек, а η — значения ряда. Тогда H_0 : $\tau_0 = 0$ и если H_0 отвергается, то тренд присутствует.

6.8. Корреляционные матрицы

Если признаков много, то их наглядно характеризуют корреляционные матрицы. Улучшить наглядность можно переупорядочив признаки так, чтобы на диагонали матрицы стояли блоки корреляций признаков из «корреляционных плеяд».

Определение. Пусть ρ_0; корреляционная плеяда есть множество признаков, таких, что их попарная корреляция больше ρ_0.

Можно выделить и несколько уровней $\rho_i : \rho_0 < \rho_1 < \dots$. Тогда сначала следует составить плеяду по ρ_0, затем внутри полученного по ρ_1 и т.д.

A. Другие полезные распределения случайных величин

A.1. Пуассона

A.2. Логнормальное

B. Свойства условного математического ожидания

1. $\mathsf{E}\left\{a\xi + b\theta \mid \eta\right\} = a\mathsf{E}\left\{\xi \mid \eta\right\} + b\mathsf{E}\left\{\theta \mid \eta\right\}$.
2. $\mathsf{E}\mathsf{E}\left\{\eta \mid \xi\right\} = \mathsf{E}\eta$.
3. $\xi \perp\!\!\!\perp \eta \implies \mathsf{E}\left\{\xi \mid \eta\right\} = \mathsf{E}\xi$.
4. $\eta = f(\xi) \implies \mathsf{E}\left\{\eta \mid \xi\right\} = \mathsf{E}\left\{f(\xi) \mid \xi\right\} = f(\xi)$.
5. $\mathsf{E}(\eta f(\xi) \mid f(\xi)) = f(\xi)\mathsf{E}\left\{\eta \mid \xi\right\}$.
6. $(\xi, \eta)^{\mathrm{T}} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies \mathsf{E}(\xi \mid \eta) = a\eta + b$.

   Замечание (Важное). Таким образом, если выборка нормальная, то зависимость линейная *всегда*.
7. $\operatorname{argmin}_{\hat\eta \in K = \{\phi(\xi)\}} \mathsf{E}(\eta - \hat\eta)^2 = \mathsf{E}\left\{\eta \mid \xi\right\} = \hat\eta^*$.