Статистика

Конспект курса

Мат-Мех, ПМИ

5–7 семестры (2015–2016)

$Revision: 1.62 $

Содержание

1. Оценки характеристик и параметров распределения

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

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

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

---

[1] ANalysis Of VAriation

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

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

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

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

Замечание. Подходящее определение выбирается по контексту.

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

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

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: $\{[e_0, e_1), \ldots, [e_{m-1}, e_m)\}$ и считать число наблюдений $\nu_i = \nu_i[e_{i-1}, e_i)$, попавших в интервал.

Следствие. *По ЗБЧ (теореме Бернулли),*

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

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

Виды признаков Виды признаков случайной величины $\xi : (\Omega, \mathcal{F}, \mathsf{P}) \to (V, \mathfrak{A})$ 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.

1.2. 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 $T(\hat{\mathcal{P}}_n)$.

Определение. *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

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

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

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

Более того, если cdf_ξ непрерывна, эта сходимость имеет порядок $1/\sqrt{n}$ по теореме Колмогорова:

$$\sqrt{n} \sup_{x \in \mathbb{R}} \left| \widehat{\mathrm{cdf}}_\xi(x) - \mathrm{cdf}_\xi(x) \right| \xrightarrow{\mathrm{d}} \mathcal{P}_{\text{K.S.}},$$

где $\mathcal{P}_{\text{K.S.}}$ — распределение Колмогорова-Смирнова.

Замечание. Поскольку $\widehat{\mathrm{cdf}}_\xi(x) = \omega_x$, где ω_x — частота попадания наблюдений в интервал в $(-\infty, x)$, а $\mathrm{cdf}_\xi(x) = \mathsf{P}(\xi \in (-\infty, x))$ — 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 $\hat{\mathcal{P}}_n$.*

1.3. Характеристики распределений и их оценки

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

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

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

1.3.1. 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* (квантилью уровня p) называется

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

Квартиль есть квантиль уровня, кратного 1/4; дециль — 1/10; перцентиль — 1/100.

Замечание. $\sup$ берется для учета случая не непрерывных функций распределения.

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 $\mathbf{x} = \left(x_{(1)}, \ldots, x_{(n)}\right)$ 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.

1.3.2. 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 $\mathsf{E}\xi = \mu$ известно, то дополнительно вводится

$$s_\mu^2 := \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^n (x_i - \mu)^2 \\ \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^n x_i^2\right) - \mu^2. \end{cases}$$

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[2].

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.

SD позволяет оценивать стандартное отклонение распределения ξ. Пусть $\hat{\theta}_n$ — статистика. Она имеет какое-то своё распределение, стандартное отклонение которого можно также оценить.

Определение. *Стандартная ошибка* оценки есть

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

Это показатель разброса оценки случайной величины.

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

$$\mathrm{SE} := \mathrm{SE}(\bar{\mathbf{x}}) = \sqrt{\widehat{\mathsf{D}\bar{\mathbf{x}}}} = \sqrt{\frac{\widehat{\mathsf{D}\xi}}{n}} = \frac{s}{\sqrt{n}}.$$

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

Пусть $c_\gamma = \mathrm{qnt}_{\mathrm{N}(0,1)}\,\gamma$.

---

[2] `http://mathworld.wolfram.com/SampleVarianceDistribution.html`

Пример (С мостом и машинами)**.** При возведении моста требуется, чтобы под ним могли проехать, условно, 95% машин. Чтобы эту высоту вычислить, достаточно собрать выборку высоты кузова проезжающих машин. Тогда нахождение искомой величины можно наглядно представить как выбор такой квантили гистограммы выборки, что суммирование соответствующих вероятностей даст 0.95. В предположении, что выборка из нормального распределения, с более устойчивой оценкой квантили, интервал будет иметь вид

$$(\bar{\mathbf{x}}, \mathrm{SD} \cdot c_\gamma + \bar{\mathbf{x}})\,.$$

SE как показатель разброса среднего использовать по смыслу нельзя.

Пример (С паромом)**.** Число машин, которое способен перевезти паром, есть Грузоподъемность/$\mathsf{E}\xi$, где ξ — вес машины. Поскольку оценка $\bar{\mathbf{x}}$ всегда считается с погрешностью относительно истинного значения, интервал допустимого числа машин будет иметь вид

$$\frac{\text{Грузоподъемность}}{\bar{\mathbf{x}} \pm \mathrm{SE} \cdot c_\gamma}.$$

1.3.3. Анализ характера разброса

Определение. *Коэффициент асимметрии Пирсона* («скошенности»[3])

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

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

1.3.4. Характеристики зависимости

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

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

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

---

[3] «Skewness».

Замечание (Важное). $\xi_1 \nparallel \xi_2 \implies \operatorname{cov}(\xi_1, \xi_2)$, но $\operatorname{cov}(\xi_1, \xi_2) \not\Longrightarrow \xi_1 \nparallel \xi_2$. 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 $\mathsf{m}_{2k} \geq 0$ 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.

1.4. Точечная оценка параметров распределения

1.4.1. Метод моментов

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

$$\begin{cases} \mathsf{E}g_1(\xi) = \phi_1(\theta_1, \ldots, \theta_r) \\ \qquad \vdots \\ \mathsf{E}g_r(\xi) = \phi_r(\theta_1, \ldots, \theta_r) \end{cases} \implies \begin{cases} \theta_1 = f_1(\mathsf{E}g_1(\xi), \ldots, \mathsf{E}g_r(\xi)) \\ \ \vdots \\ \theta_r = f_r(\mathsf{E}g_1(\xi), \ldots, \mathsf{E}g_r(\xi)). \end{cases}$$

Примем

$$\theta_i^* = f_i(\hat{\mathsf{E}}g_1(\xi), \ldots, \hat{\mathsf{E}}g_r(\xi)).$$

Часто, $g_i(\xi) = \xi^i$.

Замечание. Как правило, эти оценки смещенные. Несмещенность означала бы выполнение для всех $\theta \in \Theta$ равенства

$$\mathsf{E}\theta^* = \mathsf{E}\phi^{-1}(\hat{\mathsf{E}}g(\xi)) = \theta = \phi^{-1}(\mathsf{E}\hat{\mathsf{E}}g(\xi)).$$

Но часто ϕ^{-1} — выпуклая, так что имеем, на самом деле, неравенство Йенсена.

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 ($r = 1$)**.** $\xi \sim \mathrm{U}(0, \theta)$.

- Оценка по 1-му моменту: $g(\xi) = \xi$ и

$$\mathsf{E}\xi = \int_0^\theta \frac{1}{\theta} x \,\mathrm{d}x = \frac{1}{\theta} \left. \frac{x^2}{2} \right|_0^\theta = \frac{\theta}{2} \implies \theta = 2\mathsf{E}\xi,\ \theta^* = 2\bar{\mathbf{x}}.$$

- Оценка по k-му моменту: $g(\xi) = \xi^k$ и

$$\mathsf{E}\xi^k = \frac{1}{\theta} \int_0^\theta x^k \,\mathrm{d}x = \frac{1}{\theta} \left. \frac{x^{k+1}}{k+1} \right|_0^\theta = \frac{\theta^k}{k+1} \implies \theta^* = \sqrt[k]{(k+1)\frac{1}{n}\sum_{i=1}^n x_i^k}.$$

Пример 2 ($r = 1$)**.** Пусть $\xi \sim \mathrm{Exp}(\lambda)$. Тогда $\mathsf{E}\xi = \lambda$ и $\bar{\mathbf{x}} = \lambda$.

Пример 3 ($r = 2$)**.** Пусть $\mathcal{P}_\xi(\theta_1, \theta_2) = \mathrm{Bin}(m, p)$. Тогда $g_1(\xi) = \xi$, $g_2(\xi) = (\xi - \mathsf{E}\xi)^2$ и

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

1.4.2. Метод оценки максимального правдоподобия

Утверждение. Пусть $\mathbf{x}$ — выборка. В качестве оценки максимального правдоподобия[4] $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ следует взять

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

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:

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.

1.5. Свойства оценок

1.5.1. Несмещенность

Определение. *Смещение*[5] есть

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

Определение. *Среднеквадратичная ошибка*[6] есть

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

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

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

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 $\mathsf{E}\xi$.

---

[4] Maximum likelihood estimate (MLE).

[5] Bias.

[6] Mean square error (MSE).

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:

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

Пример. $\widehat{\operatorname{cdf}}_\xi$ — несмещенная оценка cdf_ξ.

Доказательство. FIXME □

1.5.2. 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,

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

□

Предложение. $\hat{\mathsf{m}}_k$ *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.

Предложение. $\hat{\mathsf{m}}_k^{(0)}$ *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. ☐

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

Доказательство. Либо по (1.5.2) 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 (1.5.2) получаем утверждение. ☐

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

Доказательство. По (1.5.2) с $k = 2$. ☐

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

Доказательство. Введем случайную величину

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

y_i независимы ($\operatorname{cdf}_{y_i}(x)\operatorname{cdf}_{y_j}(y) = \operatorname{cdf}_{y_i,y_j}(x, y)$) и одинаково распределены; кроме того, их математическое ожидание конечно:

$$\mathsf{E}y_i = 1 \cdot \mathsf{P}(x_i < x) + 0 \cdot \mathsf{P}(x_i \geq x) = \mathsf{P}(x_i < x) = \operatorname{cdf}_\xi(x) < \infty.$$

Тогда по теореме Хинчина о ЗБЧ,

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

☐

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.

Предложение. *Оценки, полученные по методу моментов, являются состоятельными.*

Доказательство. Поскольку f_i — непрерывные функции и при непрерывных преобразованиях сходимость не портится:

$$\theta^* = \phi^{-1}(\hat{\mathsf{E}}g(\xi)) \xrightarrow{\mathsf{P}} \phi^{-1}(\mathsf{E}g(\xi)) = \phi^{-1}(\phi(\theta)) = \theta.$$

□

Предложение. *Оценки $\hat{\boldsymbol{\theta}}_{\text{MLE}}$, полученные по методу максимального правдоподобия, являются состоятельными.*

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:

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

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

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

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

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

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

1.5.3. Асимптотическая нормальность

Определение. Оценка $\hat{\theta}_n$ называется *асимптотически нормальной* оценкой параметра θ с коэффициентом $\sigma^2(\theta)$ если

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma^2(\theta)).$$

Пример. $\bar{\mathbf{x}}$ — асимптотически нормальная оценка, если $\mathsf{D}\xi < \infty,\ \mathsf{D}\xi \neq 0$:

$$\sqrt{n}\left(\bar{\mathbf{x}} - \mathsf{E}\xi\right) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \mathsf{D}\xi).$$

Доказательство. По ЦПТ,

$$\sqrt{n}\left(\bar{\mathbf{x}} - \mathsf{E}\xi\right) = \frac{\sum_{i=1}^{n} x_i - n\mathsf{E}\xi}{\sqrt{n}} \xrightarrow{\mathrm{d}} \mathrm{N}(0, \mathsf{D}\xi).$$

□

Пример. $\hat{\mathsf{m}}_k$ — асимптотически нормальная оценка.

Доказательство. FIXME □

Пример. s^2 и $\tilde{s}^2$ — асимптотически нормальные оценки, если $0 \neq \mathsf{D}(\xi - \mathsf{E}\xi)^2 < \infty$:

$$\sqrt{n}\left(s^2 - \mathsf{D}\xi\right) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \mathsf{D}(\xi - \mathsf{E}\xi)^2).$$

Доказательство. Пусть $\eta = \xi - \mathsf{E}\xi$ и $y_i = x_i - \mathsf{E}\xi$. Тогда

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{\mathbf{x}})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mathsf{E}\xi - (\bar{\mathbf{x}}-\mathsf{E}\xi))^2 = \hat{\mathsf{E}}\eta^2 - \bar{\mathbf{y}}^2\\ \sqrt{n}(s^2-\mathsf{D}\xi) &= \sqrt{n}(\hat{\mathsf{E}}\eta^2 - \bar{\mathbf{y}}^2 - \mathsf{D}\xi) = \sqrt{n}(\hat{\mathsf{E}}\eta^2 - \underbrace{\mathsf{D}\xi}_{=\mathsf{D}\eta=\mathsf{E}\eta^2}) - \sqrt{n}\bar{\mathbf{y}}^2\\ &= \underbrace{\frac{\sum_{i=1}^{n} y_i^2 - n\mathsf{E}\eta^2}{\sqrt{n}}}_{\xrightarrow{\mathrm{d}}\mathrm{N}(0,\mathsf{D}\eta^2)} - \underbrace{\bar{\mathbf{y}}}_{\xrightarrow{\mathrm{d}}0}\ \underbrace{\sqrt{n}\bar{\mathbf{y}}}_{\xrightarrow{\mathrm{d}}\mathrm{N}(0,\mathsf{D}\xi)} \xrightarrow{\mathrm{d}} \mathrm{N}(0, \mathsf{D}(\xi-\mathsf{E}\xi)^2). \end{aligned}$$

□

Пример. $\widehat{\mathrm{cdf}}_\xi$ — асимптотически нормальная оценка:

$$\sqrt{n}\left(\widehat{\mathrm{cdf}}_\xi(x) - \mathrm{cdf}_\xi(x)\right) \xrightarrow{\mathrm{d}} \mathrm{N}(0,\hat{\sigma}^2), \quad \text{где } \hat{\sigma}^2 = \mathrm{cdf}(x)(1-\mathrm{cdf}(x)).$$

Доказательство. FIXME □

Пример. При определенных условиях, $\widehat{\mathrm{med}}\xi$ является асимптотически нормальной оценкой $\mathrm{med}\,\xi$.

Доказательство. FIXME □

Предложение. *Асимптотически нормальная оценка состоятельна.*

Доказательство. Действительно,

$$\hat{\theta}_n - \theta = \frac{1}{\sqrt{n}}\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{d}} 0 \cdot \eta = 0, \quad \eta \sim \mathrm{N}(0, \sigma^2(\theta)).$$

Слабая сходимость же к константе влечет слабую сходимость по вероятности, откуда $\hat{\theta}_n \xrightarrow{\mathsf{P}} \theta$. □

Следствие. *Асимптотически нормальные оценки сходятся к оцениваемому параметру со скоростью* $1/\sqrt{n}$.

Замечание. Если оценка не асимптотически нормальная, она *может* сходится к параметру быстрее $1/\sqrt{n}$.

Определение. Пусть $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}$ — асимптотически нормальные оценки с соответствующими коэффициентами $\left(\sigma^{(1)}\right)^2(\theta)$ и $\left(\sigma^{(2)}\right)^2(\theta)$. Говорят, что $\hat{\theta}^{(1)}$ *лучше* $\hat{\theta}^{(2)}$ *в смысле асимптотического подхода*, если

$$\left(\sigma^{(1)}\right)^2(\theta) \leq \left(\sigma^{(2)}\right)^2(\theta), \quad \forall\theta\in\Theta$$

(и хотя бы при одном θ это неравенство строгое).

1.5.4. Эффективность

Определение. Говорят, что оценка $\hat{\theta}^{(1)}$ *лучше* $\hat{\theta}^{(2)}$ *в среднеквадратичном смысле*, если

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

Замечание. Оценка $\hat{\theta}^*$ является лучшей в среднеквадратичном смысле в классе всевозможных оценок только если она совпадает с самим оцениваемым параметром, $\hat{\theta}^* = \theta$. Значит, в невырожденном с точки зрения статистики случае наилучшей среднеквадратичной оценки не существует.

Замечание. Для несмещенных оценок определение эквивалентно, конечно,

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

Определение. Поскольку в классе всех оценок наилучшей не существует, класс этот разбивается на $\mathcal{K}_b = \left\{\hat{\theta} \mid \mathsf{E}\hat{\theta} = \theta + \operatorname{bias}\theta = \theta + b\right\}$ — классы всех оценок с одинаковым смещением. Оценка $\hat{\theta}^* \in \mathcal{K}_b$ называется эффективной, если её среднеквадратичная ошибка меньше всех других оценок в этом классе:

$$\mathsf{E}(\hat{\theta}^* - \theta)^2 \leq \mathsf{E}(\hat{\theta} - \theta)^2, \quad \forall \hat{\theta} \in \mathcal{K}_b,\ \theta \in \Theta.$$

Определение. Эффективная оценка в классе $\mathcal{K}_0$ называется просто *эффективной.*

Предложение. *Эффективная оценка в классе $\mathcal{K}_b$ единственна.*

Доказательство. FIXME □

Пример. Для $\xi \sim \mathrm{U}(0, \theta)$, оценка $\hat{\theta}_n^{(1)} = \max(\mathbf{x})$ более эффективна, чем $\hat{\theta}_n^{(2)} = 2\bar{\mathbf{x}}$.

Доказательство. FIXME □

Пример. Пусть $\xi \sim \operatorname{Pois}(\lambda)$. 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 $\hat{\theta}_{\mathrm{MLE}}$).

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.

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

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

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

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.

1.6. Проверка оценок на эффективность

Пусть $\mathcal{P}_\xi(\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^\mathsf{T}$ — 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 4. Пусть $\xi \sim \mathrm{N}(\mu, \sigma^2)$. По независимости x_i, $p_{\boldsymbol{\theta}}(\mathbf{x})$ распадается в произведение:

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

Пример 5. $\xi \sim \mathrm{Pois}(\lambda)$,

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

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

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

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,

$$I_n(\theta) = -\sum_{i=1}^{n} \mathsf{E}(\ln p_\theta(x_i))'' = n \cdot i(\theta), \quad \text{где } i(\theta) = -\mathsf{E}(\ln p_\theta(\xi))''.$$

Определение. $C \subset \mathbb{R}$ есть *носитель* параметрического семейства распределений $\mathcal{P}(\theta)$, если

$$\xi \sim \mathcal{P}(\theta) \implies \mathsf{P}(\xi \in C) = 1, \ \forall \theta \in \Theta.$$

Определение. Условие регулярности:

- Существует C — носитель распределения $\mathcal{P}(\theta)$ такой, что $\forall y$ $\sqrt{p_\theta(y)}$ непрерывно дифференцируема по θ всюду в области Θ.
- Существует $I_n(\theta) > 0$, непрерывное по θ всюду в области Θ.

Пример. $\mathrm{Exp}(\lambda)$ — регулярное семейство; $\mathrm{U}(0, \theta)$ — не является регулярным.

Доказательство. FIXME □

Утверждение. Для несмещенных оценок в условиях регулярности справедливо неравенство *Рао–Крамера*:

$$\mathsf{D}\hat{\theta}_n \geq \frac{1}{I_n(\theta)}.$$

Для смещенных оценок,

$$\mathsf{D}\hat{\theta}_n \geq \frac{(1 + \mathrm{bias}'(\theta))^2}{I_n(\theta)}.$$

Следствие. *Несмещенная оценка является эффективной, если:*

$$\mathsf{D}\hat{\theta}_n = \frac{1}{I_n(\theta)}.$$

Упражнение (Хорошее)**.** Показать, что $\bar{\mathbf{x}}$ является эффективной оценкой μ в модели $\xi \sim \mathrm{N}(\mu, \sigma^2)$.

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

1.7. Построение эффективных оценок

Эффективную оценку можно построить как функцию от полной и достаточной статистики.

Определение. Пусть $x_i \sim \mathcal{P}_\theta,\ \theta \in \Theta$. Статистика $T(\mathbf{x})$ называется *достаточной* для параметра θ, если при любом t и $B \in \mathfrak{B}(\mathbb{R}^n)$ распределение $\mathsf{P}(\mathbf{x} \in B \mid T = t)$ не зависит от θ.

Таким образом, известное и фиксированное значение достаточной статистики дает всю информацию о параметре (и выборка тогда не нужна).

Утверждение (Факторизация Неймана–Фишера)*.* T достаточна, тогда и только тогда, когда

$$\mathcal{L}(\theta \mid \mathbf{x}) \overset{\text{æ}}{=} h(\mathbf{x})\Psi(T, \theta).$$

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

$$\mathcal{L}(\lambda \mid \mathbf{x}) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = \underbrace{\frac{1}{\prod_{i=1}^{n} x_i!}}_{h(\mathbf{x})}\underbrace{e^{-n\lambda}\lambda^{n\bar{\mathbf{x}}}}_{\Psi(n\bar{\mathbf{x}},\lambda)} \implies T = n\bar{\mathbf{x}}.$$

Пример. Пусть $\xi \sim \mathrm{U}(0, \theta)$, тогда

$$\mathcal{L}(\lambda \mid \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\theta}\mathbf{1}_{[0,\theta]}(x_i) = \underbrace{\frac{1}{\theta^n}\mathbf{1}(x_{(n)} < \theta)}_{\Psi(x_{(n)},\theta)}\underbrace{\mathbf{1}(x_{(1)} \geq 0)}_{h(\mathbf{x})} \implies T = x_{(n)}.$$

Пример. Пусть $\xi \sim \mathrm{N}(\mu, \sigma^2)$, тогда проверяется, что для $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)^\mathsf{T}$

$$\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}) = \Psi\left(\left(n\overline{\mathbf{x}^2}, n\bar{\mathbf{x}}\right), \boldsymbol{\theta}\right).$$

Определение. Статистика T называется *полной*, если для борелевской g

$$\mathsf{E}g(T) = 0 \implies g(T) = 0, \quad \forall\theta \in \Theta.$$

Следствие. *Оценка, являющаяся функцией от S, единственна в классе оценок с таким же смещением.*

Доказательство. Пусть их две: $\hat{\theta}_n^{(1)}(T)$ и $\hat{\theta}_n^{(2)}(T)$. Тогда $\mathsf{E}\left(\hat{\theta}_n^{(1)}(T) - \hat{\theta}_n^{(2)}(T)\right) = \mathsf{E}g(T) = 0$, значит $g(T) = 0 = \hat{\theta}_n^{(1)}(T) - \hat{\theta}_n^{(2)}(T)$. □

Теорема (Рао–Блэкуэлл–Колмогоров)**.** *Если T — полная и достаточная, то оценка $\hat{\theta}(T)$, являющаяся функцией от T, эффективна в классе оценок с таким же смещением.*

Доказательство. Для любой $\hat\theta_n \in K_b$,

$$\begin{aligned}\mathrm{MSE}\,\hat\theta_n &= \mathsf{E}(\hat\theta_n-\theta)^2 = \mathsf{E}(\hat\theta_n-\hat\theta_n(T)+\hat\theta_n(T)-\theta)^2 = \mathsf{E}(\hat\theta_n-\hat\theta_n(T))^2+\mathsf{E}(\hat\theta_n(T)-\theta)^2\\ &\geq \mathsf{E}(\hat\theta_n(T)-\theta)^2 = \mathrm{MSE}\,\hat\theta_n(T).\end{aligned}$$

Использовано равенство

$$\begin{aligned}\mathsf{E}(\hat\theta_n-\hat\theta_n(T))(\hat\theta_n(T)-\theta) &= \mathsf{E}\left(\mathsf{E}\left\{(\hat\theta_n-\hat\theta_n(T))(\hat\theta_n(T)-\theta)\mid T\right\}\right)\\ &= \mathsf{E}\left((\hat\theta_n(T)-\theta)\mathsf{E}\left\{(\hat\theta_n-\hat\theta_n(T))\mid T\right\}\right)=0,\end{aligned}$$

потому что $\mathsf{E}\left\{\hat\theta_n \mid T\right\} = \hat\theta_n(T)$ по полноте T, так что и $\mathsf{E}\left\{(\hat\theta_n-\hat\theta_n(T))\mid T\right\}=0$. □

2. Некоторые распределения, связанные с нормальным

2.1. 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] $\chi^2(m)$

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

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

$$\mathrm{cdf}_{\chi^2(50)}(80)=0.9955\approx\Phi(3)=0.9986.$$

2.2. 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 $m>1$, $\mathsf{E}\xi=0$ по симметричности.
- При $m>2$, $\mathsf{D}\xi=m/(m-2)$.

---

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

- При $m > 3$, $\mathsf{A}\xi = 0$ по симметричности.
- При $m > 4$, $\mathsf{K}\xi = 6/(m-4)$.

Предложение. *Распределение Стьюдента сходится к стандартному нормальному:*

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

Соображения по поводу. $\mathsf{D}\xi \to 1,\ \mathsf{K}\xi \to 0.$ ☐

2.3. 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 $\chi^2(k) \xrightarrow[k\to\infty]{} 1$.

2.4. Квадратичные формы от нормально распределенных случайных величин

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

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 $\boldsymbol{\xi}^\mathsf{T}\mathbf{B}\boldsymbol{\xi} \perp\!\!\!\perp \boldsymbol{\xi}^\mathsf{T}\mathbf{C}\boldsymbol{\xi} \iff \mathbf{BC} = \mathbf{0}$.

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

$$\bar{\mathbf{x}}^2 = \frac{1}{n^2}\left(\sum_{i=1}^n x_i\right)^2 = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n x_i x_j = \frac{1}{n}\mathbf{x}\underbrace{\begin{pmatrix} 1/n & \ldots & 1/n \\ \vdots & \ddots & \vdots \\ 1/n & \ldots & 1/n \end{pmatrix}}_{\mathbf{B}}\mathbf{x}^\mathsf{T}$$

$$s^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \frac{1}{n}\mathbf{xBx}^\mathsf{T} = \frac{1}{n}\left(\mathbf{xI}_n\mathbf{x}^\mathsf{T} - \mathbf{xBx}^\mathsf{T}\right) = \frac{1}{n}\mathbf{x}\underbrace{\begin{pmatrix} 1-1/n & \ldots & -1/n \\ \vdots & \ddots & \vdots \\ -1/n & \ldots & 1-1/n \end{pmatrix}}_{\mathbf{C}=\mathbf{I}_n-\mathbf{B}}\mathbf{x}^\mathsf{T}.$$

Таким образом, $n\bar{\mathbf{x}}^2 = \mathbf{xBx}^\mathsf{T}$ и $ns^2 = \mathbf{xCx}^\mathsf{T}$. Но

$$\mathbf{BC} = \mathbf{B}(\mathbf{I}_n - \mathbf{B}) = \mathbf{B} - \mathbf{B}^2 = \mathbf{0},$$

так как

$$\mathbf{B}^2 = \begin{pmatrix} 1/n & \ldots & 1/n \\ \vdots & \ddots & \vdots \\ 1/n & \ldots & 1/n \end{pmatrix}^2 = \begin{pmatrix} n\cdot 1/n & \ldots & n\cdot 1/n \\ \vdots & \ddots & \vdots \\ n\cdot 1/n & \ldots & n\cdot 1/n \end{pmatrix} = \mathbf{B}.$$

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

Видно, что $\sigma^{-2}\boldsymbol{\xi}^\mathsf{T}\mathbf{I}_p\boldsymbol{\xi} \sim \chi^2(p)$. 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

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

Пример. $n\sigma^{-2}s^2 \sim \chi^2(p-1)$. Воспользуемся представлением из предыдущего примера: $ps^2 = \mathbf{x}^\mathsf{T}\mathbf{Cx}$. 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 $\boldsymbol{\xi} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_p)$, $\boldsymbol{\xi}^\mathsf{T}\boldsymbol{\xi} = \sum_i Q_i$, 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. $\mathbf{B}_i\mathbf{B}_j = \mathbf{0}$.

3. Проверка гипотез

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

3.1. Построение критерия

3.1.1. Понятие гипотезы и критерия

Пусть $x_1, \ldots, x_n \sim \mathcal{P}$, $\mathscr{P}$ — множество всех распределений. О $\mathcal{P}$ возможно делать утверждения вида $\mathcal{P} \in \mathscr{P}' \subset \mathscr{P}$. Стоит задача выбрать такое утверждение, что оно некоторым наилучшим образом соответствует выборке.

Определение. *Модель* — это предположение о выделенном классе $\mathscr{P}_M \subset \mathscr{P}$, которому принадлежит $\mathcal{P}$ (допустим, $\mathscr{P}_M = \{\mathrm{N}(\mu, \sigma_0)\}$, где σ_0 — фиксированное значение). Иными словами, это 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:

$$H : \mathcal{P} = \mathcal{P}_0 \in \mathscr{P}_M$$

(например, $\mathcal{P}_0 = \mathrm{N}(\mu_0, \sigma_0)$) или *сложной*, если целому множеству:

$$H : \mathcal{P} \in \mathscr{P}' \subset \mathscr{P}_M$$

(например, $\mathscr{P}' = \{\mathrm{N}(\mu, \sigma_0) : \mu > 0\}$).

Очень часто возникает (и далее рассматривается) случай выдвижения только лишь двух гипотез: $H_0 : \mathcal{P} \in \mathscr{P}_0 \subset \mathscr{P}$ — *нулевой*, основной и $H_1 : \mathcal{P} \in \mathscr{P}_1 \subset \mathscr{P}$ — *альтернативной.* H_1 учитывает отклонения от H_0, обнаружение которых желательно. Возможны варианты:

- простая H_0 и простая H_1;
- простая H_0 и сложная H_1;
- сложная H_0 и сложная H_1.

Определение. *Критерий* есть отображение

$$\varphi : \mathbf{x} \mapsto \{H_0, H_1\}.$$

Критерий «решает», противоречат или не противоречат выдвинутой гипотезе выборочные данные.

Определение. Говорят, что гипотеза *отвергается*, если $\varphi(\mathbf{x}) = H_1$ и *не отвергается* иначе.

Так как в заданной постановке любой критерий принимает не более двух значений, то $\operatorname{dom}\varphi$ разбивается на два дизъюнктных множества $\mathscr{A}_{\text{крит}}$ и $\mathscr{A}_{\text{дов}}$, называемых *критической* и *доверительной* областями, таких, что

$$\varphi(\mathbf{x}) = \begin{cases} H_0 & \mathbf{x} \in \mathscr{A}_{\text{дов}} \\ H_1 & \mathbf{x} \in \mathscr{A}_{\text{крит}}. \end{cases}$$

Поскольку выборка конечного объема позволяет делать только вероятностные заключения, со статистическим критерием ассоциированы ошибки i-ых родов.

Определение. Говорят, что произошла *ошибка i-го рода* критерия φ, если критерий отверг верную гипотезу H_{i-1}. Соответствующие вероятности обозначаются

$$\alpha_i(\varphi) = \mathsf{P}_{H_{i-1}}(\varphi(\mathbf{x}) \neq H_{i-1}).$$

Поскольку в рассмотрение введены только H_0 и H_1, возможны ошибки *I-го рода* — принятие случайного различия за систематическое и *II-го рода* — принятие наблюдаемого различия за случайный эффект с соответствующими вероятностями

$$\begin{aligned} \alpha_{\text{I}} &= \mathsf{P}_{H_0}(\varphi(\mathbf{x}) \neq H_0) = \mathsf{P}_{H_0}(\mathbf{x} \in \mathscr{A}_{\text{крит}}) \\ \alpha_{\text{II}} &= \mathsf{P}_{H_1}(\varphi(\mathbf{x}) \neq H_1) = \mathsf{P}_{H_1}(\mathbf{x} \in \mathscr{A}_{\text{дов}}). \end{aligned}$$

Замечание. Если H_i — сложная гипотеза, то $\alpha_{i+1}(\varphi)$ будет зависеть от того, на каком именно распределении $\mathcal{P}$, отвечающем H_i, вычисляется эта вероятность.

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

$$\beta = 1 - \alpha_{\text{II}} = 1 - \mathsf{P}_{H_1}(\varphi(\mathbf{x}) = H_0) = \mathsf{P}_{H_1}(\varphi(\mathbf{x}) = H_1).$$

Иными словами, это способность критерия отличать H_0 от H_1.

3.1.2. Построение оптимальных критериев

Рассмотрим пример.

Пример. Пусть $\xi \sim \mathrm{N}(\mu, 1)$ и $\mathbf{x} = \{x\}$. Выдвинем $H_0 : \mu = 0$ против простой альтернативы $H_1 : \mu = 1$ Рассмотрим критерий

$$\varphi_b(\mathbf{x}) = \begin{cases} H_0 & x < b \\ H_1 & x \geq b. \end{cases}$$

(FIXME: Рисунок) Ясно, что из всего множества критериев $\{\varphi_b\}$, не все одинаково хорошо описывают нормальную выборку: так, тождественный $\varphi_\infty(\mathbf{x}) \equiv H_0$ будет вести себя хуже любого φ_b, $b < \infty$.

Выбирать оптимальный критерий можно тремя способами: минимаксным, Байесовским подходами и выбором наиболее мощного критерия.

Определение. $\varphi^{(1)}$ не хуже $\varphi^{(1)}$ в *минимаксном* смысле, если

$$\max(\alpha_{\text{I}}(\varphi^{(1)}), \alpha_{\text{II}}(\varphi^{(1)})) \leq \max(\alpha_{\text{I}}(\varphi^{(2)}), \alpha_{\text{II}}(\varphi^{(2)})).$$

Если ϕ^* не хуже всех остальных в этом смысле, то он называется *минимаксным.*

Пример. $\varphi_{1/2}$ минимаксный.

Определение. Пусть известны $r = \mathsf{P}(H_0)$, $s = 1 - r = \mathsf{P}(H_1)$ (или задана линейная функция потерь, равная r в случае ошибки 1-го рода и s — второго). Тогда $\varphi^{(1)}$ не хуже $\varphi^{(1)}$ в *байесовском* смысле, если

$$r\alpha_{\text{I}}(\varphi^{(1)}) + s\alpha_{\text{II}}(\varphi^{(1)}) \leq r\alpha_{\text{I}}(\varphi^{(2)}) + s\alpha_{\text{II}}(\varphi^{(2)}).$$

Если ϕ^* не хуже всех остальных в этом смысле, то он называется *байесовским.*

Замечание. Короче говоря, по правилу полной вероятности это вероятность ошибки критерия:

$$\mathsf{P}(H_0)\mathsf{P}(\underbrace{\varphi(\mathbf{x}) = H_1}_{\text{Err}} \mid H_0) + \mathsf{P}(H_1)\mathsf{P}(\underbrace{\varphi(\mathbf{x}) = H_0}_{\text{Err}} \mid H_1) = \mathsf{P}(\text{Err}).$$

Пример. $\varphi_{1/2}$ байесовский с $s = r$.

Определение. Пусть до эксперимента зафиксирован *уровень значимости*[8] критерия $\alpha \in [0, 1]$. Критерий

$$\varphi^* \in K_\alpha = \{\varphi(\mathbf{x}) \mid \alpha_{\mathrm{I}}(\varphi) \leq \alpha\}$$

называется наиболее мощным критерием, если

$$\alpha_{\mathrm{II}}(\varphi^*) \leq \alpha_{\mathrm{II}}(\varphi), \quad \forall \phi \in K_\alpha.$$

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

Все три подхода могут быть сведены к универсальному критерию — критерию отношения правдоподобия.

В примере 3.1.2 интуитивно ясно, что следует выбрать ту гипотезу, значение плотности которой в точке x больше другой. В случае, если $n > 1$, справедливо взять произведение плотностей — т.е. функций правдоподобия и рассматривать их отношение

$$L(\mathbf{x}) = \frac{\mathcal{L}_2(\boldsymbol{\theta} \mid \mathbf{x})}{\mathcal{L}_1(\boldsymbol{\theta} \mid \mathbf{x})}.$$

Выбор гипотезы затем делать по тому, больше или меньше L единицы. Однако, чтобы учесть произвольный уровень ошибки, следует сравнивать не с 1, а с константой c.

Определение. Пусть $\mathcal{P}_0, \mathcal{P}_1$ либо одновременно дискретны, либо непрерывны. Пусть также $\neg\exists c_0 : \mathsf{P}_{H_0}(L(\mathbf{x}) = c_0) = 0$ (в противном случае критерий не сможет различить гипотезы на множестве не-нулевой меры) — т.е., $\mathsf{P}_{H_0}(L(\mathbf{x}) \geq c)$ непрерывна по $c > 0$. Тогда *критерием отношения правдоподобия* называется

$$\varphi_c(\mathbf{x}) = \begin{cases} H_0 & L(\mathbf{x}) < c \\ H_1 & L(\mathbf{x}) \geq c. \end{cases}$$

Явный вид оптимальных критериев

Утверждение. В предположениях из определения, критерий отношения правдоподобия является

1. минимаксным при $c : \alpha_{\mathrm{I}}(\varphi_c) = \alpha_{\mathrm{II}}(\varphi_c)$;
2. байесовским при заданных $r, s : c = r/s$;
3. (лемма Неймана-Пирсона) наибольшей мощности при заданном $\alpha : \alpha_{\mathrm{I}}(\varphi_c) = \alpha$.

Пример. Пусть $\xi \sim \mathrm{N}(\mu, 1)$. $H_0 : \mu = \mu_0$, $H_1 : \mu = \mu_1 > \mu_0$. Критическая область задается неравенством

$$L(\mathbf{x}) = \frac{\mathcal{L}_2(\boldsymbol{\theta} \mid \mathbf{x})}{\mathcal{L}_1(\boldsymbol{\theta} \mid \mathbf{x})} = \exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu_0)^2 - (x_i - \mu_1)^2\right) \geq c,$$

упрощая которое получают

$$\bar{\mathbf{x}} \geq \frac{\ln c}{(\mu_1 - \mu_0)n} + \frac{\mu_1 - \mu_0}{2}.$$

- Пусть критерий байесовский с $r = 1/4$, $s = 3/4$; тогда $c = 1/3$.

---

[8] Неформально, α обратно пропорциональна «строгости» критерия, выбираемой экспериментатором.

- Пусть критерий наибольшей мощности с заданным α, критическая область задается $\bar{\mathbf{x}} \geq c_0$. Тогда

$$\alpha_{\mathrm{I}} = \mathsf{P}_{H_0}(\bar{\mathbf{x}} \geq c_0) = \mathsf{P}_{H_0}(\sqrt{n}(\bar{\mathbf{x}} - \mu_0) \geq \sqrt{n}\,(c_0 - \mu_0)) = 1 - \mathrm{cdf}_{\mathrm{N}(0,1)}\left(\sqrt{n}\,(c_0 - \mu_0)\right) = \alpha$$

  если

$$c_0 = \frac{z_{1-\alpha}}{\sqrt{n}} + \mu_0.$$

- Пусть критерий минимаксный. Тогда

$$\alpha_{\mathrm{II}} = \mathrm{cdf}_{\mathrm{N}(0,1)}\left(\sqrt{n}(c_0 - \mu_1)\right) = 1 - \mathrm{cdf}_{\mathrm{N}(0,1)}(\sqrt{n}(c_0 - \mu_0)) = \mathrm{cdf}_{\mathrm{N}(0,1)}(\sqrt{n}(\mu_0 - c_0))$$

  если

$$c_0 - \mu_1 = \mu_0 - c_0, \quad c_0 = \frac{\mu_0 + \mu_1}{2}.$$

Определение. Критерий называется *состоятельным* против альтернативы H_1, если $\forall \mathcal{P}_1 \in \mathscr{P}_1$

$$\beta(\varphi, \mathcal{P}_1) = 1 - \mathsf{P}_{\mathcal{P}_1}(\varphi(\mathbf{x}) = H_0) \xrightarrow[n\to\infty]{} 1.$$

Пример. Критерий наибольшей мощности

$$\varphi(\mathbf{x}) = \begin{cases} H_0 & \bar{\mathbf{x}} < \dfrac{z_{1-\alpha}}{\sqrt{n}} + \mu_0 \\ H_1 & \text{иначе} \end{cases}$$

является состоятельным, потому что

$$\alpha_{\mathrm{II}}(\varphi) = \mathsf{P}_{H_1}\left(\bar{\mathbf{x}} - \frac{z_{1-\alpha}}{\sqrt{n}} < \mu_0\right),$$

но, при верной H_1,

$$\xi_n = \bar{\mathbf{x}} - \frac{z_{1-\alpha}}{\sqrt{n}} \xrightarrow{\mathsf{P}} \mu_1,$$

из чего следует слабая сходимость — сходимость $\mathrm{cdf}_{\xi_n}(x) \to \mathrm{cdf}_{\mu_1}(x) = \mathsf{P}(\mu_1 < x)$ по всех точках; учитывая $\mathrm{cdf}_{\mu_1}(\mu_0) = 0$,

$$\alpha_{\mathrm{II}}(\varphi) = F_{\xi_n}(\mu_0) \to F_{\mu_1}(\mu_0) = 0.$$

Определение. Если

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

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

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

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

О постановке H_0 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 (С гранатами)**.** FIXME

Об отвержении гипотезы 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.

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. 3.1.3). 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 $\mathscr{A}_{\text{крит}}$.

3.1.3. Построение критерия при помощи статистики критерия

Определение. *Статистика критерия* есть отображение

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

такое, что при верной H_0, $T \xrightarrow{\text{d}} \mathscr{Q}$, где $\mathscr{Q}$ — полностью известное непрерывное распределение, а при верной H_1 известно поведение T.

Поскольку распределение T при верной H_0 известно, она должна вести себя как любая другая случайная величина из $\mathscr{Q}$ — попадание в некоторые области менее вероятно, чем в другие. Поэтому разумно разбить $\operatorname{im} T$ по уровню значимости α на $\mathscr{A}_{\text{крит}} \sqcup \mathscr{A}_{\text{дов}}$ так, что попадание в $\mathscr{A}_{\text{крит}}$ происходит с заранее зафиксированной (малой) вероятностью α. Значит, если $T(\mathbf{x}) \in \mathscr{A}_{\text{крит}}$, то с некоторой же вероятностью можно заявлять об отвержении H_0. Таким образом, T измеряет то, насколько выборка соответствует гипотезе.

Разберем на примере построение разбиения. Пусть $\xi \sim \mathrm{N}(\mu, \sigma^2)$, $H_0 : \mu = \mu_0$ и фиксирован α. По **??**, используется статистика

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

В зависимости от H_1, возможны варианты.

Простая альтернатива Пусть $H_1 : \mu = \mu_1$, причем $\mu_1 > \mu_0$. Тогда, поскольку при верной H_1, $\mathsf{E}\bar{\mathbf{x}} = 1/n \cdot \sum_{i=1}^{n} \xi_i = n/n \cdot \mu_1$, то

$$\mathsf{E}T = \frac{\sqrt{n}\,(\mu_1 - \mu_0)}{\sigma} \implies T \sim \mathrm{N}\left(\frac{\sqrt{n}\,(\mu_1 - \mu_0)}{\sigma}, 1\right) \text{ при верной } H_1.$$

(дисперсия, конечно, не меняется при сдвиге).

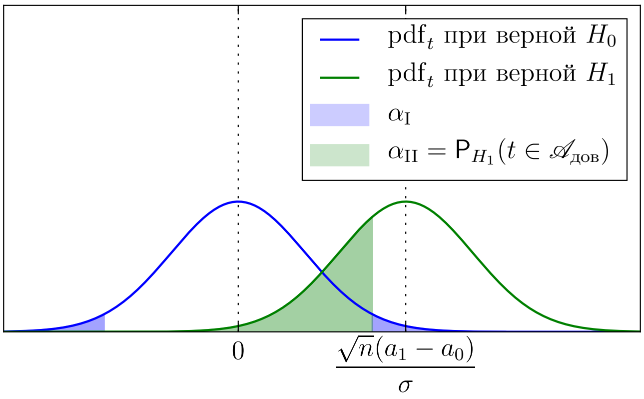


Рис. 1: Плотности распределения z (неоптимальное разбиение)

Чтобы минимизировать α_{II}, логично определить $\mathscr{A}_{\text{крит}}$ только на одном хвосте — с той стороны, где находится альтернатива. 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 μ_1 дальше от μ_0, тем α_{II} меньше.

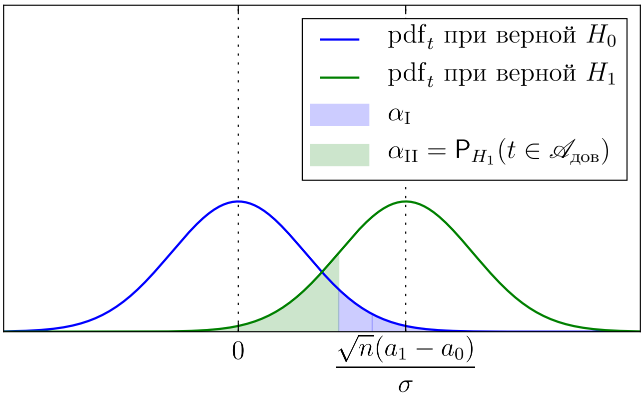


Рис. 2: Плотности распределения z (оптимальное разбиение)

Таким образом, $\mathscr{A}_{\text{крит}} = (C, +\infty)$, $C = z_{1-\alpha}$.

Разумеется, если $\mu_1 < \mu_0$, то $\mathscr{A}_{\text{крит}} = (-\infty, C)$, $C = z_\alpha$.

Односторонний критерий (сложная альтернатива) В общем случае, пусть $H_1 : \mu = \mu_1\ \forall \mu_1 > \mu_0$; тогда по ЗБЧ $\bar{\mathbf{x}} - \mu_0 \to \mu_1 - \mu_0 > 0$ и $T \to +\infty$. Следовательно, чтобы максимизировать величину $\beta = \mathsf{P}_{H_1}(\mathbf{x} \in \mathscr{A}^{(\alpha)}_{\text{крит}})$, следует разместить $\mathscr{A}$ на правом хвосте плотности: $\mathscr{A}_{\text{крит}} = (z_{1-\alpha}, \infty)$. Для $H_1 : \mu = \mu_1\ \forall \mu_1 < \mu_0$ аналогично $\mathscr{A}_{\text{крит}} = (-\infty, z_\alpha)$.

Двусторонний критерий Пусть $H_1 : \mathsf{E}\xi = \mu_1 \neq \mu_0$; тогда по ЗБЧ $\bar{\mathbf{x}} - \mu_0 \to \mu_1 - \mu_0$ и $|T| \xrightarrow[n\to\infty]{} \infty$, откуда $\mathscr{A}_{\text{крит}} = \mathbb{R} \setminus \left(z_{\alpha/2}, z_{1-\alpha/2}\right) = \mathbb{R} \setminus (-C, C)$, где $C = -z_{\alpha/2}$.

В общем виде, с использованием статистики, критерий может быть определен как

$$\varphi(\mathbf{x}) = \begin{cases} H_0 & |T(\mathbf{x})| < C \\ H_1 & |T(\mathbf{x})| \geq C, \end{cases}$$

где *критическое значение* C определяется из уравнения $\alpha = \mathsf{P}(|T| \geq C)$.

Иногда вместо сравнения значения T с критическим вычисляют *реально достигнутый уровень значимости критерия* («*p-value*»).

Определение. *p-value* значения статистики T на выборке $\mathbf{x}$ есть вероятность, взяв выборку из распределения H_0, получить по ней большее отклонение $|T(\mathbf{x})|$ эмпирического от истинного распределения, чем получено по проверяемой выборке:

$$p\text{-value} = \alpha^* = \mathsf{P}_{H_0}(|T| \geq |T(\mathbf{x})|).$$

Значит, критерий может быть задан и как

$$\varphi(\mathbf{x}) = \begin{cases} H_0 & \alpha^* > \alpha \\ H_1 & \alpha^* \leq \alpha. \end{cases}$$

Замечание. p-value обратно пропорционален «существенности» результата.

Замечание. Пусть $\alpha^* = 0.05$. Это значит, что в среднем всего лишь 5% «контрольных» выборок, удовлетворяющих основной гипотезе, будут обладать большим отклонением $|T(\mathbf{x})|$ по сравнению с тестируемой выборкой — последняя ведет себя не хуже, чем 5% «правильных» выборок.

3.1.4. Схема построения критерия с помощью статистики

1. Фиксируют предположение относительно данных.
2. Выдвигают H_0 и H_1.
   - H_0 формулируется согласно замечанию 3.1.2.
   - H_1 ставится по смыслу задачи (см. далее).
3. Выбирают подходящий критерий и статистику T.
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.
6. Считают значение статистики и принимают решение об отвержении H_0 одним из способов:

$$\varphi(\mathbf{x}) = \begin{cases} H_0 & |T(\mathbf{x})| < C \\ H_1 & |T(\mathbf{x})| \geq C, \end{cases} \qquad \varphi(\mathbf{x}) = \begin{cases} H_0 & \alpha^* > \alpha \\ H_1 & \alpha^* \leq \alpha. \end{cases}$$

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 $\xi \sim \mathrm{N}(\mu, 4)$.
2. Выдвинута $H_0 : \mathsf{E}\xi = \mu_0 = 0$ — если гипотеза опровергнется, то холодильник не купят; $H_1 : \mathsf{E}\xi = \mu_1 \neq \mu_0$.
3. Поскольку модель нормальная и известная σ^2, выберем статистику **??** («z-test»):

   $$z = \frac{\sqrt{n}\,(\bar{\mathbf{x}} - \mu_0)}{\sigma} \sim \mathrm{N}(0, 1) \text{ при верной } H_0.$$

   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 $\mu_1 \neq \mu_0$, 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
$$z(\mathbf{x}) = \frac{\sqrt{n}\,(\bar{\mathbf{x}} - \mu_0)}{\sigma} = \frac{5(0.7-0)}{2} = 1.75.$$
Дальнейшее принятие решения возможно на основании критического значения или p-value.

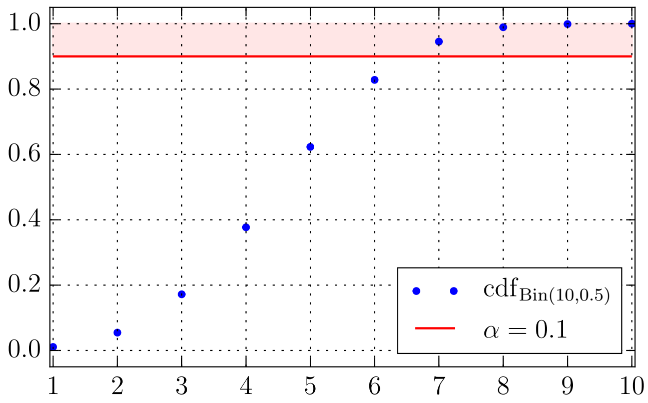
   - По вычислению критического значения:
     - ◇ $z \in \mathscr{A}_{\text{крит}}^{(\alpha^{(1)})}$, H_0 отвергается, холодильник не покупают.
     - ◇ $z \in \mathscr{A}_{\text{дов}}^{(\alpha^{(2)})}$, H_0 не отвергается, холодильник, быть может, покупают.
   - Можно посчитать p-value:
$$2 \cdot (1 - \mathrm{cdf}_{\mathrm{N}(0,1)}(1.75)) \approx 0.08.$$
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 3.1.3, 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 $\beta = \mathsf{P}_{H_1}(\mathbf{x} \in \mathscr{A}^{(\alpha)}_{\text{крит}}) = \mathsf{P}_{H_1}(p\text{-value} < \alpha)$ считать выборку с параметрами H_1, а T относительно H_0.

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

$H_0 : \mathsf{E}\xi = \mu = \mu_0$. Соответствие оценки математического ожидания гипотезе удобно выражать разницей $\bar{\mathbf{x}} - \mu_0$ с «идеальным» значением 0. Отнормировав эту разницу, получим статистику, распределение которой известно.

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

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

$$t = \sqrt{n}\frac{(\bar{\mathbf{x}} - \mu_0)}{\sigma} \xrightarrow[n\to\infty]{} \mathrm{N}(0,1)$$

Доказательство. По ЦПТ. □

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

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

Доказательство.

$$z = \frac{\bar{\mathbf{x}} - \mu_0}{\sqrt{\mathsf{D}\bar{\mathbf{x}}}} = \sqrt{n}\,\frac{\bar{\mathbf{x}} - \mu_0}{\sigma} \sim \mathrm{N}(0,1).$$

□

Разбиение

$H_1 : \mathsf{E}\xi \neq \mu_0 \;\; \mathscr{A}_{\text{крит}} = \mathbb{R} \setminus \left(z_{\alpha/2}, z_{1-\alpha/2}\right)$

$H_1 : \mathsf{E}\xi > \mu_0 \;\; \mathscr{A}_{\text{крит}} = (z_{1-\alpha}, \infty)$

$H_1 : \mathsf{E}\xi < \mu_0 \;\; \mathscr{A}_{\text{крит}} = (-\infty, z_{\alpha})$

3.2.2. $\mathsf{D}\xi$ неизвестна

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

$$t = \sqrt{n-1}\,\frac{\bar{\mathbf{x}} - \mu_0}{s} = \frac{\sqrt{n-1}(\bar{\mathbf{x}} - \mu_0)}{\sqrt{n-1}/\sqrt{n} \cdot \tilde{s}} = \sqrt{n}\,\frac{\bar{\mathbf{x}} - \mu_0}{\tilde{s}} \xrightarrow[n\to\infty]{} \mathrm{N}(0,1).$$

Предложение. *При условии нормальности данных,*

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

Доказательство.

$$t = \frac{\sqrt{n-1}\,(\bar{\mathbf{x}} - \mu_0)}{s} = \frac{\sqrt{n-1}\left(\dfrac{\bar{\mathbf{x}} - \mu_0}{\sigma}\right)}{s/\sigma} = \frac{\left(\dfrac{\bar{\mathbf{x}} - \mu_0}{\sigma}\right)}{\sqrt{\dfrac{s^2/\sigma^2}{n-1}}} = \frac{\dfrac{\sqrt{n}(\bar{\mathbf{x}} - \mu_0)}{\sigma}}{\sqrt{\dfrac{ns^2/\sigma^2}{n-1}}} = \frac{\beta}{\sqrt{\eta/(n-1)}} \sim \mathrm{t}(n-1),$$

поскольку

$$\beta = \frac{\sqrt{n}(\bar{\mathbf{x}} - \mu_0)}{\sigma} \sim \mathrm{N}(0,1), \quad \eta = \frac{ns^2}{\sigma^2} \sim \chi^2(n-1).$$

□

Разбиение

$H_1 : \mathsf{E}\xi \neq \mu_0 \;\; \mathscr{A}_{\text{крит}} = \mathbb{R} \setminus \Big(\mathrm{qnt}_{\mathrm{t}(n-1)}(\alpha/2), \mathrm{qnt}_{\mathrm{t}(n-1)}(1-\alpha/2)\Big)$

$H_1 : \mathsf{E}\xi > \mu_0 \;\; \mathscr{A}_{\text{крит}} = (\mathrm{qnt}_{\mathrm{t}(n-1)}(1-\alpha), \infty)$

$H_1 : \mathsf{E}\xi < \mu_0 \;\; \mathscr{A}_{\text{крит}} = (-\infty, \mathrm{qnt}_{\mathrm{t}(n-1)}\,\alpha)$

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.

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

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

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

3.3.1. $\mathsf{E}\xi = \mu < \infty$

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

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

Доказательство.

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

☐

Разбиение

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

$H_1 : \mathsf{D}\xi > \sigma_0^2$ $\mathscr{A}_{\text{крит}} = (\mathrm{qnt}_{\chi^2(n)}(1-\alpha), \infty)$

$H_1 : \mathsf{D}\xi < \sigma_0^2$ $\mathscr{A}_{\text{крит}} = (0, \mathrm{qnt}_{\chi^2(n)}\,\alpha)$

3.3.2. $\mathsf{E}\xi$ 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. (2.4). ☐

Альтернативное доказательство. По определению запишем

$$\underbrace{\mathsf{D}\hat{\xi}_n}_{s^2} = \mathsf{D}(\hat{\xi}_n - \mu) = \underbrace{\mathsf{E}(\hat{\xi}_n - \mu)^2}_{s_\mu^2} - \underbrace{(\mathsf{E}(\hat{\xi}_n - \mu))^2}_{(\bar{\mathbf{x}}-\mu)^2}.$$

Домножив обе части на n/σ_0^2, получим

$$\frac{ns^2}{\sigma_0^2} = \frac{ns_\mu^2}{\sigma_0^2} - \frac{n(\bar{\mathbf{x}} - \mu)^2}{\sigma_0^2} = \underbrace{\frac{ns_\mu^2}{\sigma_0^2}}_{\sim\chi^2(n)} - \underbrace{\left(\frac{\sqrt{n}\,(\bar{\mathbf{x}} - \mu)}{\sigma_0}\right)^2}_{\sim\chi^2(1)} \Longrightarrow \frac{ns^2}{\sigma_0^2} \sim \chi^2(n-1).$$

☐

Замечание. Для строгого доказательства, нужно использовать независимость $\bar{\mathbf{x}}^2$ и s^2 (см. 2.4).

Разбиение

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

$H_1 : \mathsf{D}\xi > \sigma_0^2$ $\mathscr{A}_{\text{крит}} = (\mathrm{qnt}_{\chi^2(n-1)}(1-\alpha), \infty)$

$H_1 : \mathsf{D}\xi < \sigma_0^2$ $\mathscr{A}_{\text{крит}} = (0, \mathrm{qnt}_{\chi^2(n-1)}\,\alpha)$

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:

$$p\text{-value}/2 = \mathrm{cdf}_{\chi^2(100)}(96.96) = \texttt{pchisq}(96.96, 100) \approx 0.43 \implies p\text{-value} \approx 0.86.$$

Замечание. Можно посчитать и по таблицам для нормального распределения. Раз

$$\frac{\eta_m - \mathsf{E}\eta_m}{\sqrt{\mathsf{D}\eta_m}} \xrightarrow[m\to\infty]{\mathrm{d}} \mathrm{N}(0,1),$$

то

$$\frac{96.96 - 100}{\sqrt{200}} \approx -0.215 \implies p\text{-value}/2 = \Phi(-0.215) \approx 0.415.$$

⌟

3.4. 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»).*

3.4.1. 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 $\mathscr{A}_{\text{крит}} = \left(\mathrm{qnt}_{\chi^2(k-1)}(1-\alpha), \infty\right)$ — 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

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

⌟

3.4.2. Распределение с неизвестными параметрами

В случае сложной гипотезы $\mathcal{P} \in \{\mathcal{P}(\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\Theta}$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_r)^\mathsf{T}$, 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 $T = 0$.

3.4.3. 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

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

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.

3.5. Критерий Колмогорова-Смирнова согласия с видом распределения

3.5.1. 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.

3.5.2. Нормальное распределение

Пусть $H_0 : \mathcal{P}_\xi \in \left\{\mathrm{N}(\mu, \sigma^2)\right\}$. Как известно, критерий Колмогорова-Смирнова используется для непрерывных непараметрических распределений. Им можно воспользоваться и для данной H_0, если вместо μ, 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 $\mathrm{N}(\mu, \sigma^2)$ и $\mathrm{Exp}(\lambda)$ получаем распределение D_n, не зависящее от параметров (так что поправку можно делать вне зависимости от параметров; к примеру, $\mathrm{N}(\mu, \sigma^2)$ 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[9] для проверки $H_0 : \mathcal{P}_\xi = \mathrm{N}(\mu, \sigma^2)$ 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')
> xx <- rt(1000, df=20)
> lillie.test(xx)
        Lilliefors (Kolmogorov-Smirnov) normality test
data:  xx
D = 0.023412, p-value = 0.2032
> shapiro.test(xx)
        Shapiro-Wilk normality test
data:  xx
W = 0.99669, p-value = 0.03409
```

3.6. 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.

Замечание. Из всех тестов для тестирования согласия с нормальным распределением наибольшей мощностью при любых объемах выборки почти всегда обладает Shapiro-Wilk, см. `http://www.de.ufpb.br/~ulisses/disciplinas/normality_tests_comparison.pdf`

3.7. Визуальное определение согласия с распределением

3.7.1. 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)
}
```

3.7.2. 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/`

3.8. 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

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

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

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

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.

3.8.1. Двухвыборочный тест Колмогорова–Смирнова

Рассматривается $H_0 : \mathcal{P}_{\xi_1} = \mathcal{P}_{\xi_2}$ против $H_1 : \mathcal{P}_{\xi_1} \neq \mathcal{P}_{\xi_2}$ и оба распределения абсолютно непрерывны. В качестве статистики используется

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

3.9. Равенство математических ожиданий для независимых выборок

3.9.1. 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*[10], $(x_1, \ldots, x_{n_1}), (y_1, \ldots, y_{n_2})$, $n = n_1 + n_2$. 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,* $\widehat{\mathsf{D}(\bar{\mathbf{x}} - \bar{\mathbf{y}})} = s_1^2/n_1 + s_2^2/n_2$, *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*[11].

---

[10] Случай зависимой выборки рассматривается в другом параграфе.

[11] Behrens-Fisher problem.

Разбиение

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

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

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

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

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

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

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

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 $n_1 \approx n_2$.

3.9.2. 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, $H_0 : \mathcal{P}_{\xi_1} = \mathcal{P}_{\xi_2}$ против $H_1 : \mathcal{P}_{\xi_1} \neq \mathcal{P}_{\xi_2}$ (что выборки получены из одной «генеральной совокупности») в случае абсолютно непрерывных распределений.

3.9.3. 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

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

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

Иногда в качестве статистики берут количество инверсий в объединенной выборке.

3.9.4. Критерий Mann-Whitney (U test)

Используется статистика

$$U := \max\left(n_1n_2+\frac{n_1(n_1+1)}{2}-W_1, n_1n_2+\frac{n_2(n_2+1)}{2}-W_2\right).$$

При верной H_0, $\mathsf{P}(\xi_1 < \xi_2) = 1/2$. В этом случае,

$$\mathsf{E}U = \frac{n_1n_2}{2}, \qquad \mathsf{D}U = \frac{n_1n_2(n_1+n_2+1)}{12}.$$

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 $\operatorname{med}\xi$).

3.9.5. 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.

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

3.10. Равенство математических ожиданий для парных (зависимых) выборок

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

3.10.1. 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.

3.10.2. 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).

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

Увеличить мощность предыдущего критерия можно, учтя больше информации:

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

Для симметрии можно рассмотреть статистику

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

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.

3.11. Равенство дисперсии для двух распределений

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

3.11.1. 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.

3.11.2. 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).

3.11.3. Критерий Brown–Forsythe

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

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

4. Доверительное оценивание

4.1. 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 «$\xrightarrow[n\to\infty]{}$».

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

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

4.2.1. Доверительный интервал для μ

- Пусть σ^2 известно. Свяжем μ_0 с выборкой:

  $$\gamma = \mathsf{P}(c_1 < T < c_2) = \mathsf{P}\left(c_1 < \sqrt{n}\frac{(\bar{\mathbf{x}} - \mu_0)}{\sigma} < c_2\right) = \mathsf{P}\left(\mu_0 \in \left(\bar{\mathbf{x}} - \frac{\sigma c_2}{\sqrt{n}}, \bar{\mathbf{x}} - \frac{\sigma c_1}{\sqrt{n}}\right)\right).$$

  Решений уравнения $\mathsf{P}\left(c_1 < \sqrt{n}(\bar{\mathbf{x}} - \mu_0)/\sigma < c_2\right) = \Phi(c_2) - \Phi(c_1) = \gamma$ 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,

  $$\mathsf{P}\left(\mu_0 \in \left(\bar{\mathbf{x}} \pm \frac{\sigma}{\sqrt{n}} x_{\frac{1+\gamma}{2}}\right)\right) = \gamma.$$

- Пусть σ^2 неизвестно. По аналогии,

  $$\gamma = \mathsf{P}\left(c_1 < \frac{\sqrt{n-1}(\bar{\mathbf{x}} - \mu_0)}{s} < c_2\right) = \mathsf{P}\left(\mu_0 \in \left(\bar{\mathbf{x}} \pm \frac{c_\gamma s}{\sqrt{n-1}}\right)\right), \quad c_\gamma = \operatorname{cdf}^{-1}_{\mathrm{t}(n-1)}\left(\frac{1+\gamma}{2}\right)$$

  и

  $$\mathsf{P}\left(\mu_0 \in \left(\bar{\mathbf{x}} \pm \frac{\tilde{s}}{\sqrt{n}} x_{\frac{1+\gamma}{2}}\right)\right) = \gamma.$$

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

Решение.

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

┘

4.2.2. Доверительный интервал для σ^2

- Пусть μ 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

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

- Пусть μ 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

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

4.3. 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 $\mathsf{E}\xi = \mu$. Пусть $\{x_i\}$ i.i.d., тогда

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

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

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

Пример. Доверительный интервал для параметра $\mathrm{Exp}(\lambda)$. FIXME

4.4. 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 1.4.2, $\hat{\lambda}_{\mathrm{MLE}} = \bar{\mathbf{x}}$, по 1.6 $I_n(\lambda) = n/\lambda = n/\bar{\mathbf{x}}$ 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 $[0,1]$.

4.5. 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

$$\mathsf{P}(\theta \in [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})]) = \gamma = 1 - \alpha \implies \alpha = 1 - \mathsf{P}(\theta \in [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})]) = \mathsf{P}(\theta \notin [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})])$$

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

$$\varphi(\mathbf{x}) = \begin{cases} H_0 & \theta \notin [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})] \\ H_1 & \theta \in [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})]\,. \end{cases}$$

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

4.6. Использование SE для построения доверительных интервалов

Пусть оценка $\hat{\theta}_n$ имеет какое-то симметричное распределение хотя бы асимптотически. Как и для любой другой случайной величины (с симметричным распределением), доверительный интервал уровня γ (т.е. такой интервал, в котором лежит γ всех значений величины) задается как

$$\mathsf{E}\hat{\theta}_n \pm c_\gamma \sqrt{\mathsf{D}\hat{\theta}_n},$$

где $c_\gamma = \mathrm{qnt}\,\gamma$. К примеру, для $\mathrm{N}(0,1)$ и 95%-квантили это был бы интервал$(-1.96; 1.96)$, а так нужно передвинуть его на среднее и растянуть на корень из дисперсии.

Но стандартное отклонение $\sqrt{\mathsf{D}\hat{\theta}_n}$ распределения $\hat{\theta}_n$ можно оценить как SE. Значит доверительный интервал будет иметь вид

$$\mathsf{E}\hat{\theta}_n \pm c_\gamma \,\mathrm{SE}\,.$$

4.7. Доверительный интервал для двумерного параметра

Пусть $\hat{\theta}_1, \hat{\theta}_2 \sim \mathrm{N}(0,1)$. Доверительная область с уровнем доверия γ $C \subset \mathbb{R}^2 : \mathsf{P}((\hat{\theta}_1, \hat{\theta}_2) \in C) = \gamma$ может быть построена в виде квадрата или диска с центром в начале координат (по симметричности нормального распределения):

- По независимости,

  $$\mathsf{P}((\hat{\theta}_1, \hat{\theta}_2) \in C) = \mathsf{P}(\hat{\theta}_1 \in (-x_0, x_0))\mathsf{P}(\hat{\theta}_2 \in (-y_0, y_0)) = \gamma,$$

  откуда

  $$x_0 = y_0 = \mathrm{cdf}^{-1}_{\mathrm{N}(0,1)}\left(\frac{1+\sqrt{\gamma}}{2}\right).$$

- Найдем

  $$x : \mathsf{P}(\underbrace{\hat{\theta}_1^2 + \hat{\theta}_2^2}_{r^2} \le x) = \gamma.$$

  $r^2 \sim \chi^2(2) = \mathrm{Exp}(\lambda)$, где $\lambda = 1/2$, поскольку $\mathsf{E}\chi^2(2) = 2$. Значит,

  $$\mathsf{P}(r^2 < x) = \mathrm{cdf}_{\mathrm{Exp}(1/2)}(x) = 1 - e^{-1/2x} = \gamma \implies x = -2\ln(1-\gamma),$$

  и радиус получившегося круга есть

  $$r = \sqrt{x} = \sqrt{-2\ln(1-\gamma)}.$$

Находя радиус и сторону квадрата для разных γ, можно посчитать площади получившихся областей:

γ	Площадь квадрата	Площадь круга
0.99	31.5	28.94
0.9	15.19	14.47
0.8	10.48	10.11
⋮	⋮	⋮
0.3	2.26	2.24
0.2	1.41	1.40

Так как площадь круга всегда меньше площади квадрата, он является предпочтительным как доверительная область.

5. Корреляционный анализ

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

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

5.1. Вероятностная независимость

5.1.1. Визуальное определение независимости

- Поскольку при $p_\eta(y_0) \neq 0$

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

  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

5.1.2. 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

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

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[12] 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 $\alpha_{\text{I}} \to \alpha$.

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.

5.2. Линейная / полиномиальная зависимость

Пусть теперь ξ, η — количественные признаки.

Определение. Определим

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

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

---

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

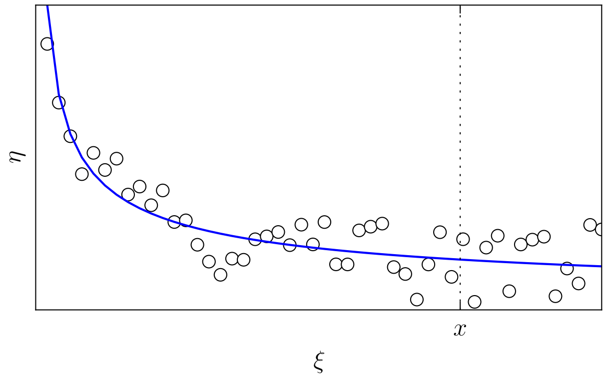


Рис. 3: Нелинейная зависимость

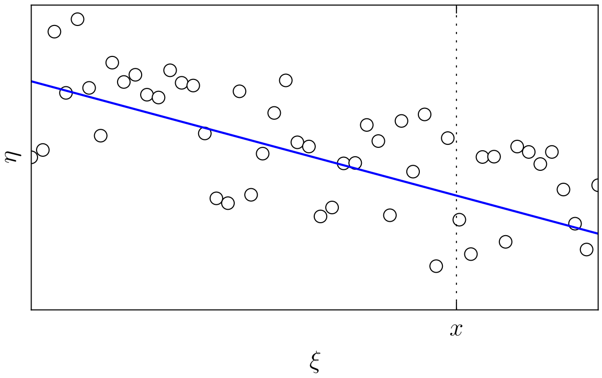


Рис. 4: Линейная зависимость

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).

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

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

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

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

☐

О соотношении ρ и коэффициента линейной регрессии По (6.2), 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 $H_0 : \rho = 0$.

Чаще, чем $H_0 : \rho = \rho_0$, проверяют $H_0 : \rho > \rho_0$. Если $\rho_0 = 0$, то $(\xi, \eta)^\mathsf{T} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ и, по ЦПТ,

$$T = \frac{\sqrt{n-2}\hat{\rho}_n}{\sqrt{1-\hat{\rho}_n^2}} \sim \mathrm{t}(n-2).$$

Идеальное значение — 0, два хвоста.

Если $\rho_0 \neq 0$, то ЦПТ не работает, тогда распределение $\hat{\rho}$ неизвестно. Тогда применяется z-преобразование Фишера

$$z = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}, \quad z_0 = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}.$$

Тогда ЦПТ работает и, если $(\xi, \eta)^\mathsf{T} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

$$T = \sqrt{n-3}(z - z_0) \xrightarrow{\mathrm{d}} \mathrm{N}(0, 1).$$

5.3. 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(1.5.2), 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:

$$\hat{\eta}^* = \underset{\hat{\eta}\in K}{\operatorname{argmin}} \|\eta - \hat{\eta}\|^2 = \underset{\hat{\eta}\in K}{\operatorname{argmin}} \mathsf{E}\left(\eta - \hat{\eta}\right)^2 = \mathsf{E}\left\{\eta \mid \phi(\xi)\right\}.$$

$\hat{\eta}^*$ называется *наилучшим среднеквадратичным приближением в классе* K.

5.4. 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*:

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

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. $(\xi, \eta)^{\mathsf{T}} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies r^2_{\eta|\xi} = \rho^2(\eta, \xi)$.

Выборочное корреляционное отношение По разложению дисперсии,

$$\mathsf{D}\eta = \mathsf{E}(\eta - \mathsf{E}\eta)^2 = \underbrace{\mathsf{E}(\mathsf{E}(\eta \mid \xi) - \mathsf{E}\eta)^2}_{\mathsf{DE}(\eta|\xi)} + \mathsf{E}(\eta - \mathsf{E}(\eta \mid \xi))^2.$$

Перейдем на выборочный язык. Пусть дана выборка

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}.$$

Сгруппируем её:

$$\begin{array}{c|ccc} x_1^* & y_{11} & \cdots & y_{1n_1} \\ \vdots & \vdots & \ddots & \vdots \\ x_k^* & y_{k1} & \cdots & y_{kn_k} \end{array}$$

Пусть ξ — дискретная случайная величина со значениями $(x_1^*, \ldots, x_k^*)$. Тогда, учитывая

$$\bar{y}_i = \bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_1} y_{ij} = \hat{\mathsf{E}}(\eta \mid \xi = x_i^*),$$

на выборочном языке получаем (домножив на n):

$$\begin{aligned} \underbrace{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2}_{\text{total sum of squares}} \quad &= \quad \underbrace{\sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}\right)^2}_{\text{межгрупповой разброс}} \quad + \quad \underbrace{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}_{\text{внутригрупповой разброс}} \\ n s_y^2 \quad &= \quad n s^2_{y|x} + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2. \end{aligned}$$

Отсюда, так как, $r^2_{\eta|\xi} = \mathsf{DE}(\eta \mid \xi) / \mathsf{D}\eta$,

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

5.5. 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.

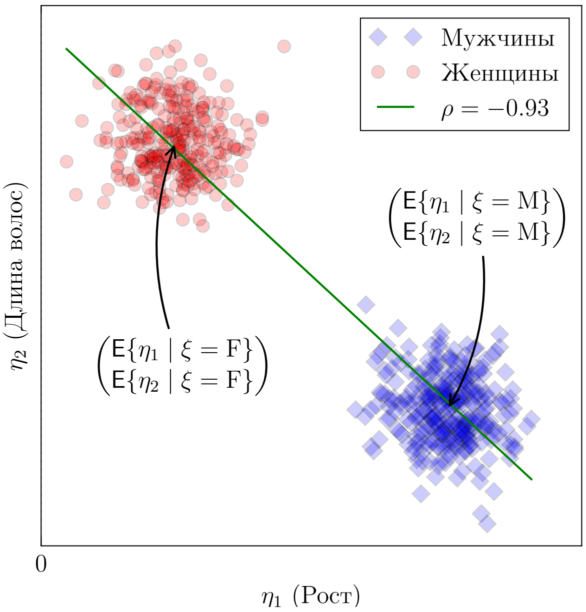


Рис. 5: Исходные данные (бимодальность)

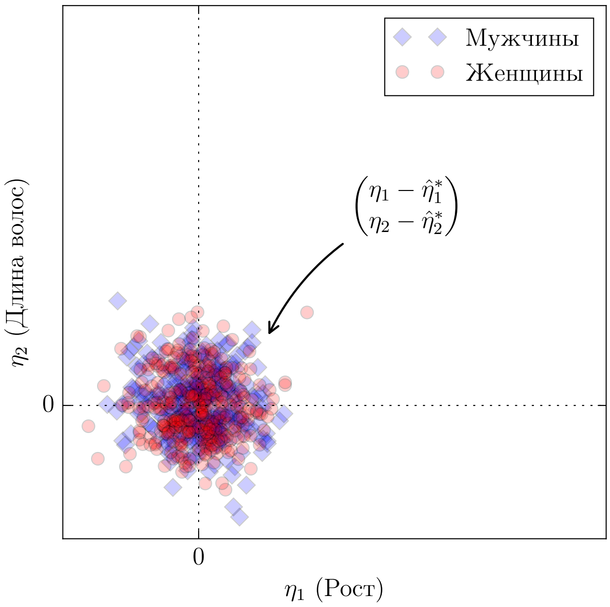


Рис. 6: Центрированные данные

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

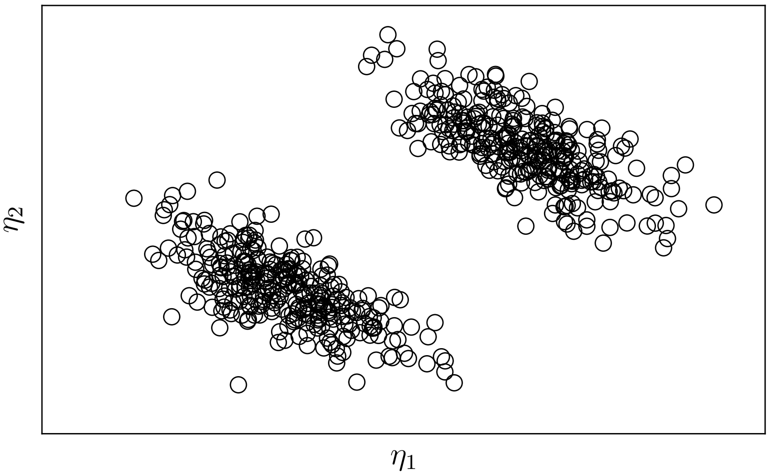


Рис. 7: $\rho(\eta_1, \eta_2) > 0$, но $\rho(\eta_1, \eta_2 \mid \xi) < 0$

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).

5.6. 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).

5.6.1. 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)

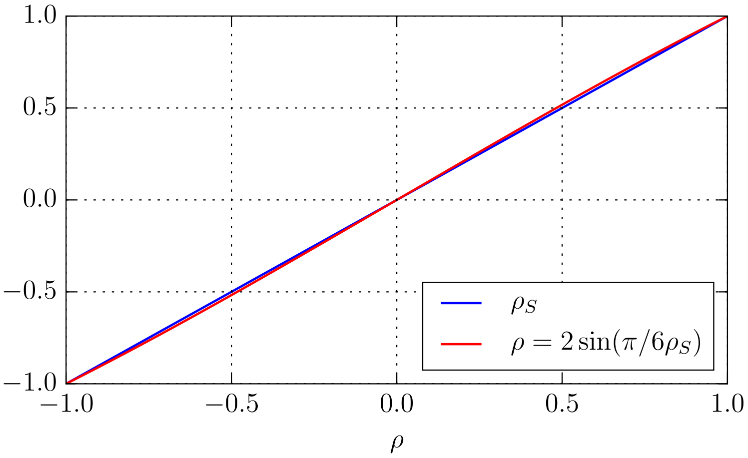


Рис. 8: $\hat{\rho} \approx \hat{\rho}_S$

- 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.

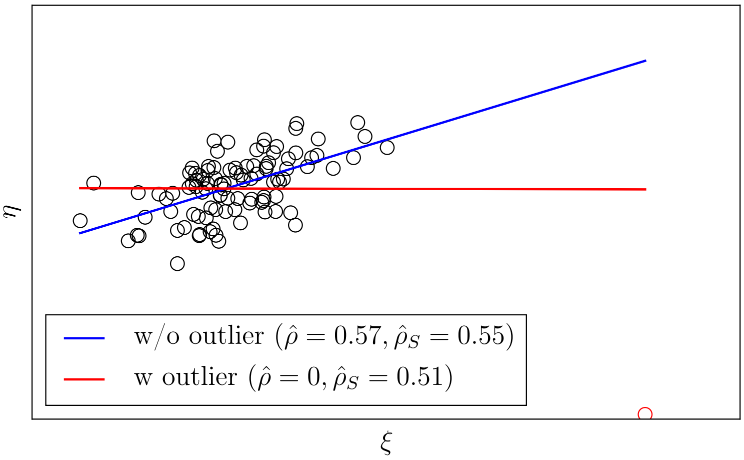


Рис. 9: $\hat{\rho}$ до и после добавления аутлаера

- 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.

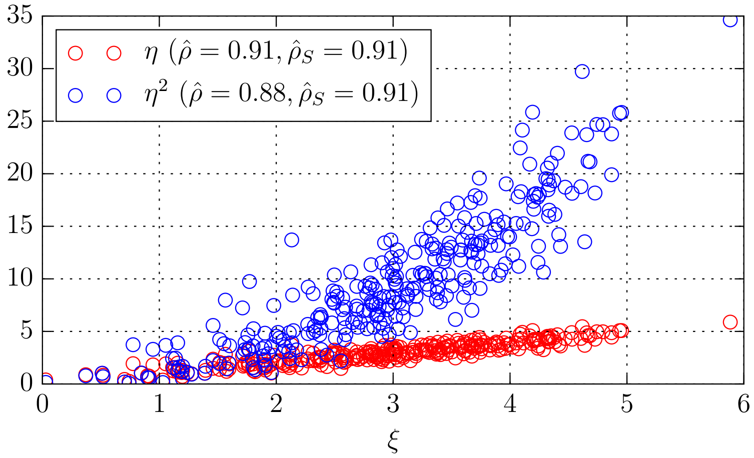


Рис. 10: Монотонное преобразование нормальных данных

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

Определение. Пусть $(\xi_1,\eta_1)^\mathsf{T} \perp\!\!\!\perp (\xi_2,\eta_2)^\mathsf{T} \sim \mathcal{P}_{\xi,\eta} \sim (\xi,\eta)^\mathsf{T}$; 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 $(\xi,\eta)^\mathsf{T} \sim \mathrm{N}(\boldsymbol{\mu},\Sigma)$, 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).

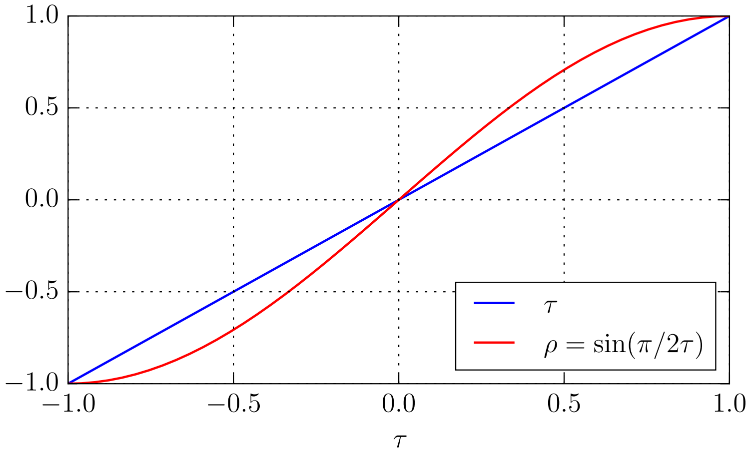


Рис. 11: ρ и τ

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.

5.7. 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.

6. Регрессионный анализ

6.1. 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).

6.2. 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,

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

Замечание. Найти минимум ϕ можно, как обычно, решив систему $\partial\phi/\partial\beta_i = 0$[13].

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:

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

---

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

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

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

6.2.1. 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:

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

6.2.2. 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[14], 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,

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

---

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

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 $c_\gamma = \mathrm{cdf}^{-1}_{\mathrm{t}(n-2)}\left((1+\gamma)/2\right)$:

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

Аналогично, для β_2:

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

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.

6.3. Множественная линейная регрессия

6.3.1. Псевдо-обратные матрицы

Определение. Матрица $\mathbf{A}^{-}$ называется *псевдо-обратной*, если

1. По аналогии с $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \implies \mathbf{A}\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}$ и $\mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}$, выполняется
$$\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}, \quad \mathbf{A}^{-}\mathbf{A}\mathbf{A}^{-} = \mathbf{A}^{-}.$$

2. *(Псевдо-обратная по Муру-Пенроузу)* По аналогии со случаем, когда $\mathbf{A}^{-1} = \mathbf{A}^{\mathsf{T}}$, $(\mathbf{A}^{-1}\mathbf{A})^{\mathsf{T}} = \mathbf{A}^{\mathsf{T}}\left(\mathbf{A}^{-1}\right)^{\mathsf{T}} = \mathbf{A}^{-1}\mathbf{A}$, выполняется
$$\mathbf{A}^{-}\mathbf{A} = (\mathbf{A}^{-}\mathbf{A})^{\mathsf{T}}, \quad \mathbf{A}\mathbf{A}^{-} = (\mathbf{A}\mathbf{A}^{-})^{\mathsf{T}}.$$

Свойства

1. Если столбцы $\mathbf{A}$ линейно-независимы, то существует $(\mathbf{A}^{\mathsf{T}}\mathbf{A})^{-1}$ и
$$\mathbf{A}^{-} = (\mathbf{A}^{\mathsf{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{T}}.$$

2. Пусть ищут решение $\mathbf{X}\mathbf{b} = \mathbf{y}$ относительно $\mathbf{b}$
   a) Если уравнение не имеет решений, то на $\mathbf{b} = \mathbf{X}^{-}\mathbf{y}$ достигается минимум невязки между левой и правой частями:
$$\mathbf{b}^{*} = \mathbf{X}^{-}\mathbf{y} = \operatorname*{argmin}_{\mathbf{b}} \|\mathbf{X}\mathbf{b} - \mathbf{y}\|^{2}.$$
   b) Если решение не единственно, то $\mathbf{b} = \mathbf{X}^{-}\mathbf{y}$ есть решение с минимальной нормой.

6.3.2. Проекторы на подпространства

Пусть $\mathscr{L}_d \subset \mathbb{R}^m$ — линейное подпространство размерности d, натянутое на $\{\mathbf{p}_1, \ldots, \mathbf{p}_d\}$, $\mathbf{P} = [\mathbf{p}_1 : \ldots : \mathbf{p}_d]$. Тогда проектор на $\mathscr{L}_d$ будет задан как
$$\operatorname*{proj}_{\mathscr{L}_d} = \mathbf{\Pi} = \mathbf{P}(\mathbf{P}^{\mathsf{T}}\mathbf{P})^{-1}\mathbf{P}^{\mathsf{T}} = \mathbf{P}\mathbf{P}^{-}.$$

Если $\{\mathbf{p}_i\}_{i=1}^{d}$ — ортонормированная система, то
$$\mathbf{\Pi} = \mathbf{P}\mathbf{I}\mathbf{P}^{\mathsf{T}} = \mathbf{P}\mathbf{P}^{\mathsf{T}}.$$

Кроме того,
$$\operatorname*{proj}_{\mathscr{L}_d^{\perp}} = \underset{m\times m}{\mathbf{I}} - \mathbf{P}\mathbf{P}^{\mathsf{T}}.$$

(т.е., чтобы получить ортогональное пространство к проекции, нужно из исходного вектора вычесть проекцию).

Свойства

1. $\mathbf{\Pi}\mathbf{\Pi} = \mathbf{\Pi}$
2. $(\mathbf{I} - \mathbf{\Pi})(\mathbf{I} - \mathbf{\Pi}) = \mathbf{I} - \mathbf{\Pi}$
3. $\mathbf{\Pi}^{\mathsf{T}} = (\mathbf{P}\mathbf{P}^{\mathsf{T}})^{\mathsf{T}} = \mathbf{\Pi}$.

6.3.3. Ordinary and Total Least Squares

Пусть

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{1n} & \dots & x_{nk} \end{pmatrix}$$

матрица данных с n индивидами [15] по столбцам, каждый из которых описывается k признаками;

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

вектор наблюдений[16];

$$\mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix}$$

вектор неизвестных коэффициентов.

OLS Пусть $\mathbf{y} \in \mathbb{R}^n,\ \mathbf{X} \in \mathbb{R}^{n \times m},\ \mathbf{b} \in \mathbb{R}^m,\ \operatorname{rk}\mathbf{X} = m$. Пусть допускаются ошибки в наблюдениях такие, что $\mathsf{E}\epsilon_i = 0,\ \epsilon_i \perp\!\!\!\perp \epsilon_j,\ \mathsf{D}\epsilon_i = \sigma^2 \implies \operatorname{cov}\boldsymbol{\epsilon} = \sigma^2\mathbf{I}$. Тогда в модели

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon},$$

найти

$$\hat{\mathbf{b}} = \operatorname*{argmin}_{\mathbf{b}} \|\mathbf{X}\mathbf{b} - \mathbf{y}\|^2 = \operatorname*{argmin}_{\tilde{\mathbf{y}}} \|\tilde{\mathbf{y}} - \mathbf{y}\|^2 = \mathbf{X}^{-}\mathbf{y} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}, \quad \tilde{\mathbf{y}} := \mathbf{X}\mathbf{b}.$$

Откуда регрессией будет[17]

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} = \underbrace{\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}}_{\substack{\text{proj} \\ \text{colspace}\,\mathbf{X}}}\mathbf{y} = \mathbf{H}\mathbf{y}.$$

Можно посчитать остатки — разницу между наблюдениями и предсказанием по регрессии:

$$\text{residuals} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{M}\mathbf{y}.$$

TLS Модель допускает ошибки $\boldsymbol{\Delta}$ также и в $\mathbf{X}$,

$$\mathbf{y} = (\mathbf{X} + \boldsymbol{\Delta})\,\mathbf{b} + \boldsymbol{\epsilon}$$

(где известны $\mathbf{y},\ \tilde{\mathbf{X}} := \mathbf{X} + \boldsymbol{\Delta}$, а $\mathbf{X}$ — нет). Найти

$$\operatorname*{argmin}_{\mathbf{b}:\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\mathbf{b}} \left(\left\|\tilde{\mathbf{X}} - \mathbf{X}\right\|_F^2 + \|\tilde{\mathbf{y}} - \mathbf{y}\|^2 \right), \quad \|\mathbf{A}\|_F^2 = \sum_{i.j} a_{ij}^2.$$

Дальше рассматривается OLS.

---

[15] Также «predictors», «regressors», «controlled variables», «explanatory variables», «features», «inputs».

[16] Также «regressands», «response», «explaining variables», «outcome», «experimented variables».

[17] $\mathbf{H}$ — «hat matrix».

6.3.4. Свободный член

Видно, что $\mathbf{X}\mathbf{b} = \mathbf{y}$ задает СЛАУ, где каждое уравнение — прямая, проходящая через 0. Чтобы иметь возможность описывать случаи не-центрированных данных, пригодны два варинаты:

1. Ввести фиктивный столбец из единиц:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \in \mathbb{R}^{n\times m}, \quad m = k+1.$$

2. Центрировать признаки.

Предложение. *Оба способы эквиваленты.*

Теорема (О делении регрессоров)**.** *Пусть* $\mathbf{X}$ *матрица данных с признаками («регрессорами») по столбцам,* $\mathbf{X} = [\mathbf{X}_1 : \mathbf{X}_2]$, $\hat{\mathbf{b}} = (\hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2)^\mathsf{T}$, $\mathbf{M}_1 = \mathbf{I} - \mathbf{H}_1$, $\mathbf{H}_1 = \operatorname{proj}_{\operatorname{colspace} \mathbf{X}_1}$. *Тогда* $\hat{\mathbf{b}}_2$ *можно получить как регрессию* $\mathbf{M}_1\mathbf{y}$ *на* $\mathbf{M}_1\mathbf{X}_2$. *Остатки регрессии* $\mathbf{M}_1\mathbf{y}$ *будут такими же как остатки исходной.*

Доказательство. Без доказательства. □

Пусть $\mathbf{b} \in \mathbb{R}^m$, $\hat{\mathbf{b}} = \mathbf{X}^{-}\mathbf{y} = \left(\hat{b}_0, \hat{b}_1, \dots, \hat{b}_k\right)$. Центрируем $\mathbf{X}$, вычитая среднее по каждому столбцу: $\mathbf{X}^{(c)} \in \mathbb{R}^{n\times k}$. Центрируем $\mathbf{y}$: $\mathbf{y}^{(c)} \in \mathbb{R}^n$; тогда $\hat{\mathbf{b}}^{(c)} = (\mathbf{X}^{(c)})^{-}\mathbf{y}^{(c)}$ и по теореме

$$\hat{\mathbf{b}}^{(c)} = \begin{pmatrix} \hat{b}_1^{(c)} \\ \vdots \\ \hat{b}_k^{(c)} \end{pmatrix} = \begin{pmatrix} \hat{b}_1 \\ \vdots \\ \hat{b}_k \end{pmatrix}, \quad \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y}^{(c)} - \hat{\mathbf{y}}^{(c)}.$$

Следствие.

$$\hat{b}_0 = \bar{\mathbf{y}} - \sum_{i=1}^{k} \hat{b}_i \bar{\mathbf{x}}_i.$$

6.3.5. Стандартизованные признаки

Если признаки изначально измерены в разных шкалах, то коэффициенты перед признаками можно интерпретировать как «важность».

Определение. Чтобы стандартизировать наблюдения, следует поделить центрированные столбцы на нормы каждого столбца, получится $\mathbf{X}^{(s)} \in \mathbb{R}^{n\times k}$; $\mathbf{y}^{(s)} = \mathbf{y}^{(c)} / \left\|\mathbf{y}^{(c)}\right\|$. Тогда

$$\hat{\mathbf{b}}^{(s)} = (\mathbf{X}^{(s)})^{-}\mathbf{y}^{(s)} = \left(\left(\mathbf{X}^{(s)}\right)^\mathsf{T} \mathbf{X}^{(s)}\right)^{-1} \left(\mathbf{X}^{(s)}\right)^\mathsf{T} \mathbf{y}^{(s)} = \hat{\boldsymbol{\beta}}, \quad \hat{\beta}_i = \frac{\left\|\mathbf{x}_i^{(c)}\right\|}{\left\|\mathbf{y}^{(c)}\right\|} \hat{b}_i.$$

Вектор $\hat{\boldsymbol{\beta}}$ имеет такой вид, потому что по ходу вычислений два раза поделили и один раз умножили на $\left\|\mathbf{x}_i^{(c)}\right\|$, и умножили на $\left\|\mathbf{y}^{(c)}\right\|$.

6.3.6. Свойства оценки $\hat{\mathbf{b}}$

1. *Несмещенность* (по $\mathsf{E}\boldsymbol{\epsilon} = \mathbf{0}$):

$$\mathsf{E}\hat{\mathbf{b}} = \mathsf{E}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathsf{E}\mathbf{y} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathsf{E}\left(\mathbf{X}\mathbf{b} + \boldsymbol{\epsilon}\right) = \mathbf{b}.$$

2. *Состоятельность*: если оценка несмещенная и состоятельная в среднеквадратичном смысле, то она несмещенная; однако ситуация

$$\mathrm{MSE}\,\hat{\mathbf{b}} = \mathsf{D}\hat{\mathbf{b}} \xrightarrow[n\to\infty]{} 0$$

   невозможна в текущей постановке, потому что $\mathbf{X}$ — фиксированная матрица наблюдений.

   Предложение (О состоятельности оценки)**.** *Пусть* $\mathbf{X}_n \in \mathbb{R}^{n\times m}$ *— последовательность (случайных) матриц,* $\boldsymbol{\epsilon}_n \in \mathbb{R}^n$*; кроме того,*

   a) Существует вероятностный предел в виде невырожденной матрицы

$$\boldsymbol{\Sigma}_{\mathbf{xx}} := \frac{1}{n}\left(\mathbf{X}_n\right)^{\mathsf{T}}\mathbf{X}_n$$

   b) Ошибки независимы с регрессорами

$$\frac{1}{n}\mathbf{X}_n^{\mathsf{T}}\boldsymbol{\epsilon}_n = \mathbf{0}_m.$$

   Тогда оценка

$$\operatorname*{plim}_{n\to\infty}\hat{\mathbf{b}}_{\mathrm{OLS},n}$$

   является состоятельной.

3. Ковариационная матрица:

$$\operatorname{cov}\hat{\mathbf{b}} = \operatorname{cov}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\operatorname{cov}\boldsymbol{\epsilon}\,\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}.$$

6.3.7. Свойства $\hat{\mathbf{b}}^{(c)}$ и $\hat{\mathbf{b}}^{(s)}$

1. На самом деле, $\left(\mathbf{X}^{(c)}\right)^{\mathsf{T}}\mathbf{X}^{(c)} = \mathbf{S}_{\mathbf{xx}}/n$, $\left(\mathbf{X}^{(c)}\right)^{\mathsf{T}}\mathbf{y}^{(c)} = \mathbf{S}_{\mathbf{xy}}/n$ — суть выборочные ковариационные матрицы; тогда в их терминах

$$\hat{\mathbf{b}}^{(c)} = (\hat{b}_1,\ldots,\hat{b}_k)^{\mathsf{T}} = \left(n\left(\mathbf{X}^{(c)}\right)^{\mathsf{T}}\mathbf{X}^{(c)}\right)^{-1} n\left(\mathbf{X}^{(c)}\right)^{\mathsf{T}}\mathbf{y}^{(c)} = \mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{S}_{\mathbf{xy}}.$$

   Видно, что чем более скоррелированы признаки, тем более пропорциональны столбцы $\mathbf{S}_{\mathbf{xx}}$ и тем более вырождена $\mathbf{S}_{\mathbf{xx}}$, значит «больше» $\mathbf{S}_{\mathbf{xx}}^{-1}$, следовательно $\hat{\mathbf{b}}^{(c)}$ и разница $\mathbf{y}^{(c)} - \hat{\mathbf{y}}^{(c)} = \mathbf{y}^{(c)} - \mathbf{X}^{(c)}\hat{\mathbf{b}}^{(c)}$.

2. Ковариационная матрица:

$$\begin{aligned}
\operatorname{cov}\hat{\mathbf{b}}^{(c)} &= \operatorname{cov}\left(\mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{S}_{\mathbf{xy}}\right) = \mathbf{S}_{\mathbf{xx}}^{-1}\operatorname{cov}\left(\mathbf{S}_{\mathbf{xy}}\right)\left(\mathbf{S}_{\mathbf{xx}}^{-1}\right)^{\mathsf{T}} = \mathbf{S}_{\mathbf{xx}}^{-1}\operatorname{cov}\left(n\left(\mathbf{X}^{(c)}\right)^{\mathsf{T}}\mathbf{y}^{(c)}\right)\mathbf{S}_{\mathbf{xx}}^{-1} \\
&= \mathbf{S}_{\mathbf{xx}}^{-1}\left(\mathbf{X}^{(c)}\right)^{\mathsf{T}} n\operatorname{cov}\left(\mathbf{y}^{(c)}\right)\mathbf{X}^{(c)}\mathbf{S}_{\mathbf{xx}}^{-1} = \sigma^2\mathbf{S}_{\mathbf{xx}}^{-1}n\left(\mathbf{X}^{(c)}\right)^{\mathsf{T}}\mathbf{X}^{(c)}\mathbf{S}_{\mathbf{xx}}^{-1} \\
&= \sigma^2\left(n\left(\mathbf{X}^{(c)}\right)^{\mathsf{T}}\mathbf{X}^{(c)}\right)^{-1}\left(n\left(\mathbf{X}^{(c)}\right)^{\mathsf{T}}\mathbf{X}^{(c)}\right)\mathbf{S}_{\mathbf{xx}}^{-1} = \sigma^2\mathbf{S}_{\mathbf{xx}}^{-1} \neq \frac{\sigma^2}{n}\cdot\mathbf{S}_{\mathbf{xx}}^{-1} \xrightarrow[n\to\infty]{} 0.
\end{aligned}$$

3. Аналогично,

$$\hat{\mathbf{b}}^{(s)} = \mathbf{R}_{\mathbf{xx}}^{-1}\mathbf{R}_{\mathbf{xy}}$$

   и

$$\operatorname{cov}\hat{\mathbf{b}}^{(s)} = \frac{\sigma^{(s)2}}{n}\mathbf{R}_{\mathbf{xx}}^{-1}, \quad \sigma^{(s)} = \frac{\sigma}{\left\|\mathbf{y}^{(c)}\right\|}.$$

6.3.8. Сравнение оценок

По аналогии с одномерным случаем, *наилучшая оценка* — с минимально возможной дисперсией; аналог дисперсии — ковариационная матрица. Порядок вводится следующим образом:

Определение. $\mathbf{A} < \mathbf{B} \iff \mathbf{A} - \mathbf{B}$ положительно определена, т.е.

$$\forall \gamma \ \gamma^{\mathsf{T}}(\mathbf{A} - \mathbf{B})\gamma < 0.$$

Замечание. Пусть $\boldsymbol{\gamma}^{(i)} = (0, \ldots, \underbrace{1}_{i}, \ldots, 0)^{\mathsf{T}}$; тогда $a_{ii} < b_{ii}$.

Теорема (Гаусс-Марков). *В условиях* $\mathsf{E}\epsilon_i = 0,\ \mathsf{D}\epsilon_i = \sigma^2,\ \epsilon_i \perp\!\!\!\perp \epsilon_j$, $\hat{\mathbf{b}}_{\mathrm{OLS}}$ *является «BLUE»: «best linear unbiased estimate».*

Следствие. *Если* $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$, *то*

$$\hat{\mathbf{b}}_{\mathrm{OLS}} = \hat{\mathbf{b}}_{\mathrm{MLE}}.$$

Доказательство. MLE оценка есть

$$\begin{aligned} \mathsf{P}(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} = \sigma^2\mathbf{I}) &= \frac{1}{(2\pi)^{n/2}\sqrt{\det \sigma^2\mathbf{I}}} \exp\left\{-\frac{1}{2}\left(\mathbf{y} - \boldsymbol{\mu}\right)\sigma^{-2}\mathbf{I}(\mathbf{y} - \boldsymbol{\mu})^{\mathsf{T}}\right\} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\left\|\mathbf{X}\mathbf{b} - \boldsymbol{\mu}\right\|^2\right\} \underset{\mathbf{b}}{\to} \max \end{aligned}$$

что аналогично,

$$\|\mathbf{X}\mathbf{b} - \boldsymbol{\mu}\|^2 \underset{\mathbf{b}}{\to} \min.$$

Это же есть постановка задачи OLS. ☐

6.3.9. Оценка σ^2

Разложение дисперсии

$$\mathsf{D}\eta = \mathsf{E}(\eta - \mathsf{E}\eta)^2 = \underbrace{\mathsf{E}(\mathsf{E}(\eta \mid \xi) - \mathsf{E}\eta)^2}_{\mathsf{DE}(\eta|\xi)} + \mathsf{E}(\eta - \mathsf{E}(\eta \mid \xi))^2$$

на выборочном языке будет иметь вид

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{\mathbf{y}})^2}_{\text{SSTotal}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{\mathbf{y}})^2}_{\text{SSRegr}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SSError}}.$$

Замечание. Иногда также пишут

$$\text{SSTotal} = \text{SSEffect} + \text{SSResidual},$$

что ведет к неиллюзорной путанице!

Пусть $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Тогда, по теореме Cochran,

$$\frac{\text{SST}}{\sigma^2} \sim \chi^2(n-1), \quad \frac{\text{SSR}}{\sigma^2} \sim \chi^2(\underbrace{m-1}_{k}), \quad \frac{\text{SSE}}{\sigma^2} \sim \chi^2(\underbrace{n-m}_{n-k-1})$$

и SSE $\perp\!\!\!\perp$ SSR. ОМП оценка (асимптотическая!) для σ^2 — SSE$/n$; несмещенной оценкой (с поправкой на размерность) будет

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-m}.$$

6.3.10. Проверка значимости коэффициентов линейной регрессии и доверительных интервалов

Определение. Коэффициент b_i *значим*, если отвергается $H_0 : b_i = 0$. Если коэффициент значим, значит признак существеннен для регрессии.

Для построения точного критерия, предполагают $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Значит, поскольку $\hat{\mathbf{b}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})$, $\hat{\mathbf{b}}$ имеет тоже нормальное распределение со средним $\mathbf{0}$ (по несмещенности), но какой-то ковариационной матрицей: $\hat{\mathbf{b}} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Тогда $\mathsf{E}\hat{b}_i = b_i = 0,\ \mathsf{D}\hat{b}_i = \sigma_i^2$ и

$$t = \frac{\hat{b}_i - b_i}{\sqrt{\mathsf{D}\hat{b}_i}} = \frac{\hat{b}_i}{\sqrt{\sigma^2((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1})_{ii}}} = \frac{\hat{b}_i}{\sqrt{\sigma^2/n \cdot \left(\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{ii}}} = \sqrt{n}\frac{\hat{b}_i}{\sigma\left(\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{ii}^{1/2}} \sim \mathrm{N}(0,1).$$

Подставляя оценку σ, получают

$$t = \sqrt{n}\frac{\hat{b}_i}{\hat{\sigma}\left(\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{ii}^{1/2}} = \sqrt{n}\frac{\hat{b}_i}{\sqrt{\dfrac{\mathrm{SSE}}{(n-m)}}\left(\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{ii}^{1/2}} = \frac{\dfrac{\sqrt{n}\hat{b}_i}{\sigma\left(\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{ii}^{1/2}}}{\sqrt{\dfrac{\mathrm{SSE}}{(n-m)\,\sigma^2}}} = \frac{\mathrm{N}(0,1)}{\sqrt{\dfrac{\chi^2(n-m)}{n-m}}} \sim \mathrm{t}(n-m).$$

Расстояние Махалонобиса Если на прямой разброс удобно измерять стандартных отклонениях σ, то в многомерном пространстве аналогом такой характеристики является расстояние Махалонобиса.

Определение. Пусть $\mathbf{V}$ — неотрицательно определенная симметричная матрица; тогда *расстояние Махалонобиса* есть

$$r_M^2(\mathbf{x}, \mathbf{y}; \mathbf{V}) = (\mathbf{x} - \mathbf{y})^\mathsf{T}\mathbf{V}^{-1}(\mathbf{x} - \mathbf{y}).$$

Замечание. Если $\boldsymbol{\xi} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$, то

$$\mathrm{pdf}_{\boldsymbol{\xi}}(\mathbf{x}) = C \cdot \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} = C \cdot \exp\left\{-\frac{1}{2}r_M^2(\mathbf{x}, \boldsymbol{\mu}; \mathbf{V})\right\}.$$

Для любых двух $\mathbf{x}_1, \mathbf{x}_2$ на линии уровня, $\mathrm{pdf}_{\boldsymbol{\xi}}(\mathbf{x}_1) = \mathrm{pdf}_{\boldsymbol{\xi}}(\mathbf{x}_2)$. Значит, $r_M^2(\mathbf{x}_1, \boldsymbol{\mu}; \mathbf{V}) = r_M^2(\mathbf{x}_2, \boldsymbol{\mu}; \mathbf{V})$, в то время, как Евклидово расстояние не обязано быть одинаковым из-за разной выраженности главных компонент. Однако $r_M^2(\mathbf{x}, \mathbf{y}; \mathbf{I}) = \|\mathbf{x} - \mathbf{y}\|_2^2$. Таким образом, r_M^2 — это Екклидово расстояние с поправкой на ковариацию, задаваемую $\mathbf{V}$.

Замечание. Если $\boldsymbol{\xi} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$, то

$$r_M^2(\mathbf{x}, \boldsymbol{\mu}; \mathbf{V}) = (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2(m)$$

как сумма квадратов центрированных и нормированных нормальных случайных величин. Кроме того,

$$\boldsymbol{\eta} = \mathbf{V}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \sim \mathrm{N}(0, \mathbf{I}) \implies r_M^2(\boldsymbol{\eta}, \mathbf{0}; \mathbf{I}) = \boldsymbol{\eta}^\mathsf{T}\boldsymbol{\eta} \sim \chi^2(m).$$

Доверительный эллипсоид В одномерном случае симметричного распределения, область носителя, где лежит γ всех значений распределения определяется равенством

$$\mathsf{P}(|\mathsf{E}\xi - x| < \sqrt{\mathsf{D}\xi}c_\gamma) = \gamma.$$

Т.е. как такое множество значений, что расстояние их от среднего с учетом стандартного отклонения меньше квантиля уровня γ. В случае оценки среднего μ_0, например, получают стандартное

$$\mathsf{P}\left(\frac{|\bar{\mathbf{x}} - \mu_0|}{\mathrm{SE}} < c_\gamma\right) = \mathsf{P}\left(-c_\gamma < \sqrt{n}\frac{\bar{\mathbf{x}} - \mu_0}{\sigma} < c_\gamma\right), \quad \sqrt{n}\frac{\bar{\mathbf{x}} - \mu_0}{\sigma} \sim \mathrm{N}(0,1)$$

так что $c_\gamma = \operatorname{qnt}_{\mathrm{N}(0,1)} \gamma$.

Аналогично можно нарисовать эллипсоид, в который помещается выборка с точностью γ. Расстояние с учетом ковариации будет задаваться соответственно параметризованным расстоянием Махалонобиса:

$$\mathsf{P}(r_M^2(\mathbf{x}, \boldsymbol{\mu}; \mathrm{SD}) < c_\gamma) = \gamma.$$

В случае, если $\hat{\boldsymbol{\theta}}_n \xrightarrow{\mathrm{d}} \mathrm{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, по предыдущему,

$$r_M^2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n) \sim \chi^2(m).$$

Значит,

$$\mathsf{P}(r_M^2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n) < c_\gamma) = \gamma, \quad c_\gamma = \operatorname{qnt}_{\chi^2(m)} \gamma.$$

6.3.11. Значимость регрессии

Можно проверить тремя способами:

1. Аналогично парной регрессии: $H_0 : \mathbf{b}^{(c)} = \mathbf{0}$. Критерий

$$t = r_M^2(\hat{\mathbf{b}}^{(c)}, \mathbf{0}; \mathrm{SE}(\hat{\mathbf{b}}^{(c)})) \sim \chi^2(k)$$

а именно,

$$t = \left(\hat{\mathbf{b}}^{(c)}\right)^\mathsf{T} \mathrm{SE}^{-1}(\hat{\mathbf{b}}^{(c)})\hat{\mathbf{b}}^{(c)} = \left(\hat{\mathbf{b}}^{(c)}\right)^\mathsf{T} \left(\frac{\sigma^2}{n} \cdot \mathbf{S}_{\mathbf{xx}}^{-1}\right)^{-1} \hat{\mathbf{b}}^{(c)} = \frac{n\left(\hat{\mathbf{b}}^{(c)}\right)^\mathsf{T} \mathbf{S}_{\mathbf{xx}}\hat{\mathbf{b}}^{(c)}}{\sigma^2}.$$

Неизвестный σ^2 следует оценить как

$$s^2 = \frac{\mathrm{SSE}}{n-(k+1)};$$

тогда

$$\frac{n\left(\hat{\mathbf{b}}^{(c)}\right)^\mathsf{T} \mathbf{S}_{\mathbf{xx}}\hat{\mathbf{b}}^{(c)}/k}{s^2} \sim \mathrm{F}(k, n-(k+1)).$$

2. Через ANOVA:

$$t = \frac{\mathrm{SSR}/k}{\mathrm{SSE}/(n-k+1)} \sim \mathrm{F}(k, n-(k+1))$$

Замечание. У этой статистики с предыдущей совпадает также и числитель, хотя это не оче-видно.

3. Через коэффициент детерминации регрессии: известно выражение для множественного коэффициента корреляции:

$$R^2(\eta, \xi_1, \ldots, \xi_k) = \frac{\mathsf{E}(\hat{\eta}^* - \mathsf{E}\eta)^2}{\mathsf{D}\eta}, \quad \mathsf{D}\eta = \mathsf{E}(\eta - \mathsf{E}\eta)^2 = \mathsf{E}(\hat{\eta}^* - \mathsf{E}\eta)^2 + \mathsf{E}(\eta - \hat{\eta}^*)^2;$$

на выборочном языке для множественной линейной регрессии получают

$$\begin{aligned} R^2 &= \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\mathrm{SST}-\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} \\ \operatorname{adjusted} R^2 &= 1 - \frac{\mathrm{SSE}/(n-(k+1))}{\mathrm{SST}/(n-1)} \xrightarrow[n\to\infty]{} R^2. \end{aligned}$$

Замечание. При удалении даже незначимого признака R^2 уменьшится; adjusted R^2 не обязательно в силу поправки $n-(k+1)$, действующей как штраф за количество переменных.

Приводя к виду ANOVA-критерия, имеют

$$t = \frac{\dfrac{\mathrm{SSR}}{k}}{\dfrac{\mathrm{SSE}}{n-(k+1)}} = \frac{\dfrac{\mathrm{SSR}}{k\,\mathrm{SST}}\mathrm{SST}}{\dfrac{(\mathrm{SST}-\mathrm{SSE}+\mathrm{SST})\mathrm{SST}}{\dfrac{\mathrm{SST}}{n-(k+1)}}} = \frac{R^2/k}{(1-R^2)/(n-(k+1))}.$$

6.3.12. Анализ оценок коэффициентов

Для анализа оценок коэффициентов можно посмотреть на попарные срезы доверительного эллипсоида. Для пары $\hat{\beta}_i, \hat{\beta}_j$ его можно нарисовать (самостоятельно), в качестве центра взяв в качестве центра точку $(\hat{\beta}_i, \hat{\beta}_j)^\mathsf{T}$, в качестве наклона главной оси и вытянутости — величину $\operatorname{corr}(\hat{\beta}_i, \hat{\beta}_j)$. Если центр достаточно далек от 0, то линии уровня не должны пересекать оси, потому что гипотеза $\beta_i = \beta_j = 0$ отвергается.

- Чем дальше от начала координат центр эллипсоида, тем больше значимость признаков.
- Чем больше корреляция тем менее адекватно центр отражает ситуацию.
- Возможны два случая: когда эллипсоид сонаправлен или перпендикулярен прямым $y = \pm x$; в первом случае («хорошем») коэффициенты значимы совокупности (даже если один близок к 0, то второй вполне далек и наоборот), во втором оба коэффициента могут быть одновременно как значимы, так и нет (и, значит, и сильно, и слабо влиять на результат).

Корреляция между оценками коэффициентов При возрастании корреляции признаков:

- дисперсия оценок коэффициентов стремится к бесконечности;
- становится сложно оценить вклад каждого признака в регрессию.

Пример. Пусть $k = 2$, $\eta = b_0 + b_1\xi_1 + b_2\xi_2$. Пусть также матрица корреляций есть

$$\mathbf{R}_{\mathbf{xx}} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Тогда

$$\operatorname{cov} \hat{\boldsymbol{\beta}} = \frac{\sigma^{(s)2}}{n} \mathbf{R}_{\mathbf{xx}}^{-1} = \frac{\sigma^{(s)2}}{n} \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

Значит, $\mathsf{D}\hat{\beta}_i \xrightarrow[\rho \to 1]{} \infty$.

С этой проблемой можно бороться, удаляя подходящие признаки из анализа[18] по следующим критериям:

1. Множественный коэффициент корреляции

$$R^2(\xi_i; \{\xi_j,\ j \neq i\}).$$

   Чем он больше, тем скорее i-й признак нужно удалить.

2. Допустимость i-го признака:

$$\text{tolerance}_i = 1 - R^2(\xi_i; \{\xi_j,\ j \neq i\}).$$

   Чем он меньше, тем скорее i-й признак нужно удалить. Помимо предыдущего соотношения справедливо

$$\mathsf{D}\hat{b}_i = \frac{\sigma^2}{\sum_{\ell=1}^n (x_\ell - \bar{\mathbf{x}}_i)^2} \frac{1}{\text{tolerance}_i}, \qquad \frac{1}{\text{tolerance}_i} \text{ — Variance Inflation Factor,}$$

   так что при большой допустимости дисперсия мала.

3. Частные корреляции

$$\rho(\xi_i, \eta \mid \{\xi_j,\ j \neq i\}) = \rho(\xi_i - \hat{\xi}_i, \eta - \hat{\eta})$$

   Чем i-я частная корреляция больше, тем больше вклад признака в регрессию (тем менее он предпочтителен для удаления).

---

[18] Нет признака — нет проблемы.

4. Полу-частные корреляции

$$\rho(\xi_i - \hat{\xi}_i, \eta).$$

Пусть $\mathbf{b} = (b_0, \ldots, b_{k-r}, \underbrace{b_{k-r+1}, \ldots, b_k}_{r\text{ штук}})^\mathsf{T}$. Если $H_0 : \mathbf{b}_{k-r+1,k} = \mathbf{0}$ не отвергается, значит последние r признаков не влияют на модель и следует выбрать более простую модель — без этих коэффициентов. Можно использовать расстояние Махалонобиса до 0 в метрике $\operatorname{cov}(\mathbf{b}_{k-r+1,k})$:

$$\begin{aligned} t &= r_M^2(\hat{\mathbf{b}}_{k-r+1,k}, \mathbf{0}; \operatorname{cov}(\mathbf{b}_{k-r+1,k})) \sim \chi^2(r) \\ &= \hat{\mathbf{b}}_{k-r+1,k}^\mathsf{T}((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1})_{(\mathrm{IV})}\hat{\mathbf{b}}_{k-r+1,k}/\sigma^2, \end{aligned}$$

где $((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1})_{(\mathrm{IV})}$ — IV квадрант $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$. Если σ^2 неизвестна, то

$$\begin{aligned} t &= \frac{\hat{\mathbf{b}}_{k-r+1,k}^\mathsf{T}((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1})_{(\mathrm{IV})}\hat{\mathbf{b}}_{k-r+1,k}/r}{\hat{\sigma}^2} \sim \mathrm{F}(r, n-(k+1)) \\ &= \frac{(R_{1,k}^2 - R_{1,k-r}^2)/r}{(1-R_{1,k}^2)/(n-m)}. \end{aligned}$$

Выбор оптимального набора признаков можно производить автоматически, по одному добавляя признаки («Forward stepwise regression») или убирая их («Backward»). Пусть вариант Forward. На шаге i добавляется тот признак, что максимизирует

$$R_{1,i+1}^2 - R_{1,i}^2;$$

остановиться следует, когда $\left|R_{1,i+1}^2 - R_{1,i}^2\right|$ достаточно мало. $H_0 : R_{1,i+1}^2 - R_{1,i}^2 = 0$, т.е. $b_{i+1} = 0$ перед добавленным признаком.

$$t = \frac{\hat{b}_i}{\mathrm{SE}(\hat{b}_i)} \sim \mathrm{t}(n-m).$$

Тогда $k = i+1$, $r = 1$ и статистика будет иметь вид

$$t = \frac{(R_{1,i+1}^2 - R_{1,i}^2)}{(1-R_{1,i+1}^2)/(n-(i+2))} \sim \mathrm{F}(1, n-(i+2)).$$

По сути, это есть перемасштабированное значение разницы $R_{1,i+1}^2 - R_{1,i}^2$.

Замечание. Однако признак выбран «лучший» (а не случайный), значит распределение не F.

- Полное решение задачи — выбрать ℓ признаков из k перебором.
- Жадный алгоритм — последовательно выбирать наиболее подходящие признаки.

Замечание. Если большое количество заполнить средними, то искусственно уменьшится ширина доверительных интервалов.

6.3.13. Анализ аутлаеров

Matrix plot Аутлаеров можно найти «на глаз» при помощи стандартного matrix plot данных.

Deleted residuals можно применить технику кросс-валидации: удалить признак, построить модель, сравнить. Если индивид является аутлаером, то наблюдение y_i на нём «перетягивает» на себя регрессионную прямую. Тогда явно «большой» будет разница

$$r_i^{(i)} = \hat{y}_i^{(i)} - y_i$$

между $\hat{y}_i^{(i)}$ — значением регрессии на i-м индивиде без этого индивида и y_i — наблюдении на i-м индивиде. $r_i^{(i)}$ будет «большой» также по сравнению с $r_i = \hat{y}_i - y_i$. Напротив, если i-й индивид аутлаером не является, то будет справедливо приближенное равенство $r_i^{(i)} \approx r_i$, так что графиком $(r_i, r_i^{(i)})$ будет прямая.

Studentized residuals Справедливо

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \implies (\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

откуда

$$\operatorname{cov}(\mathbf{y} - \hat{\mathbf{y}}) = \operatorname{cov}(\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})^{\mathsf{T}} \operatorname{cov}\mathbf{y}(\mathbf{I} - \mathbf{H}) = \sigma^2(\mathbf{I} - \mathbf{H})^{\mathsf{T}}(\mathbf{I} - \mathbf{H}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

потому что $\mathbf{I} - \mathbf{H}$ — матрица проектора. Тогда,

$$\mathsf{D}(y_i - \hat{y}_i) = \mathsf{D}r_i = \sigma^2(1 - h_{ii}).$$

Как следствие, $\mathsf{D}\epsilon_i = \sigma^2 \geq \mathsf{D}r_i$.

Определение. h_{ii} — рычаг[19].

Чем больше i-й рычаг, тем больше ошибка на i-м индивиде.

Определение. Стандартизированные остатки:

$$\frac{r_i}{\sqrt{\mathsf{D}r_i}} = \frac{r_i}{\sigma\sqrt{1 - h_{ii}}}.$$

Можно рассмотреть $\hat{\sigma}^{(i)}$ — оценку дисперсии без i-го индивида; тогда, при нормально распределенных ошибках наблюдения,

$$\frac{r_i}{\hat{\sigma}^{(i)}\sqrt{1 - h_{ii}}} \sim \mathrm{t}(n - m - 1)$$

(«-1» потому что меньше на одного индивида).

Замечание. Полученную величину можно сравнивать со «средним»

$$\sum_{i=1}^{m} h_{ii} = \operatorname{tr}\mathbf{H} = k + 1$$

(как след идемпотентной матрицы, равный её рангу[20]: след есть сумма собственных чисел, однако у идемпотента два возможных собственных числа: 0 и 1, а кратность 1 в точности равна рангу).

Расстояние по Куку Пусть $\hat{\mathbf{b}}^{(i)}$ — оценка без i-го индивида. Если расстояние между $\hat{\mathbf{b}}^{(i)}$ и $\hat{\mathbf{b}}$ «большое», то i-й индивид есть аутлаер:

$$r_M^2(\hat{\mathbf{b}}, \hat{\mathbf{b}}^{(i)}; \operatorname{cov}\hat{\mathbf{b}}) = (\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)})^{\mathsf{T}} \operatorname{cov}^{-1}(\hat{\mathbf{b}})(\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)}) = \frac{1}{\sigma^2}(\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)})\mathbf{X}^{\mathsf{T}}\mathbf{X}(\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)})$$

так что расстояние по Куку определяется как

$$\frac{(\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)})\mathbf{X}^{\mathsf{T}}\mathbf{X}(\hat{\mathbf{b}} - \hat{\mathbf{b}}^{(i)})/m}{\hat{\sigma}^2}.$$

Можно сравнить с расстоянием Махалонобиса в пространстве независимых признаков: если x_i — i-й индивид, $\bar{\mathbf{x}}$ — вектор средних, то аутлаером можно назвать индивида, для которого велико

$$r_M^2(x_i, \bar{\mathbf{x}}; \mathbf{S}_{\mathbf{xx}}).$$

Замечание. Если индивид не аутлаер по Куку, но аутлаер по Махалонобису, то велика дисперсия и $\mathbf{S}_{\mathbf{xx}}$ оценивается неправильно.

	Аутлаер по Куку	Не аутлаер по Куку
Аутлаер по Махалонобису	Далеко от линии регрессии	Далеко от $\bar{\mathbf{x}}$ на линии регрессии
Не аутлаер по Махалонобису	Близко к $\bar{\mathbf{x}}$ по одной координате и далеко по другой	Близко к $\bar{\mathbf{x}}$

---

[19] «Leverage».

[20] http://math.stackexchange.com/a/101515

6.3.14. Проверка правильности и выбор модели

- Если известно, что ошибки нормально распределены (например, в случае измерений прибора), то если остатки не имеют нормального распределения, то модель не является правильной.
- Если исходные данные имеют нелинейную зависимость, то и расположение остатков по линейной регрессии на графике будет отражать характер этой зависимости.
- Модель с наименьшим количеством параметров при прочих равных является предпочтительной, поэтому если заранее известно, что среднее 0, то свободный член из модели лучше удалить.

Замечание.

$$\eta = \mathsf{E}(\eta \mid \xi_1, \dots, \xi_i) + \mathsf{E}(\eta \mid \xi_{i+1}, \dots, \xi_k).$$

Значит все модели «верны» и можно среди них выбрать наилучшую.

6.3.15. Сведение нелинейной модели к линейной

Пусть

$$\eta = \phi(\xi_1, \dots, \xi_k) + \epsilon$$

и ϕ — нелинейная функция.

- ϕ — многочлен. Можно свести к линейной, добавляя признаки $\xi, \xi^2, \dots$ и для этих признаков строить модель.
- ξ — качественный признак. Можно ввести $k-1$ штук[21] фиктивных признаков со значениями $\{0, 1\}$ и для них строить модель. $A_1, \dots, A_k$ — градации ϕ.

6.3.16. Другие странные замечания

- $\eta = \phi(\xi) + \epsilon,\ \mathsf{E}\epsilon = 0, \epsilon \perp\!\!\!\perp \xi$. Если ϕ не линейная, то в ϵ войдет кусочек ξ и независимости не будет.
- Остатки всегда ортогональны т.к. проектор $\implies$ график — горизонтальная прямая всегда
- Графике $\hat{y}_i$ против $\hat{y}_i - y_i$ может быть наклонной прямой в случае pairwise MD deletion (и ковариационная матрица не соответствует данным).

7. Дисперсионный анализ

7.1. Однофакторный дисперсионный анализ (One-way ANOVA[22])

Задача может быть поставлена двумя эквивалентными образами:

1. Пусть $\eta_i \sim \mathcal{P}_i,\ i \in 1 : k$. Проверить гипотезу, что все распределения равны:

$$H_0 : \mathcal{P}_1 = \dots = \mathcal{P}_k.$$

2. Пусть дан двумерный вектор $\begin{pmatrix}\xi & \eta\end{pmatrix}^{\mathsf{T}}$, причем ξ («фактор») принимает k значений $A_1, \dots, A_k$. Рассмотрим $\eta_i \sim \mathcal{P}_i = \mathcal{P}_{\eta|\xi=A_i}$. Проверить гипотезу

$$H_0 : \mathcal{P}_{\eta|\xi=A_1} = \dots = \mathcal{P}_{\eta|\xi=A_k}.$$

---

[21] При добавлении вектора из единиц к k признакам получается вырожденная матрица.

[22] ANalysis Of VAriation

Пусть теперь $\eta_i \sim \mathrm{N}(\mu_i, \sigma^2)$. Разумеется,

$$H_0 : \mu_1 = \dots = \mu_k \iff H_0 : \mathsf{E}\eta_1 = \dots = \mathsf{E}\eta_k$$
$$\iff H_0 : \mathsf{E}(\eta \mid \xi = A_1) = \dots = \mathsf{E}(\eta \mid \xi = A_k) \iff H_0 : \mathsf{DE}(\eta \mid \xi) = 0.$$

Для построения критерия, вспомним разложение дисперсии на выборочном языке:

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2}_{Q=\hat{\mathsf{D}}\eta} = \underbrace{\sum_{i=1}^{k} n_i\,(\bar{y}_i - \bar{y})^2}_{Q_1=\hat{\mathsf{D}}\mathsf{E}(\eta|\xi)} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}_{Q_2},$$

откуда в качестве критерия (следуя гипотезе) выберем Q_1 с идеальным значением 0. Однако Q_1 полезно отнормировать по Q_2 для учета различных внутригрупповых разбросов. Чтобы получить статистику с известным распределением, вспомним, что по теореме Cochran, $Q_1 \perp\!\!\!\perp Q_2$,

$$\frac{Q}{\sigma^2} \sim \chi^2(n-1), \quad \frac{Q_1}{\sigma^2} \sim \chi^2(k-1), \quad \frac{Q_2}{\sigma^2} \sim \chi^2(n-k)$$

и

$$t = \frac{Q_1/(k-1)}{Q_2/(n-k)} \sim \mathrm{F}((k-1), (n-k)).$$

Замечание. Это обобщение статистики для проверки гипотезы о равенстве математических ожиданий независимых двумерных выборок с равными дисперсиями (с $k = 2$, то есть):

$$t = \frac{\bar{\mathbf{x}} - \bar{\mathbf{y}}}{\tilde{s}_{1,2}\sqrt{1/n_1 - 1/n_2}}$$

с $\tilde{s}^2_{1,2} = Q_2/(n-2)$. Дело в том, что статистики распределены одинаковы — по определению,

$$t^2(n-2) = \mathrm{F}(1, n-2).$$

Чтобы воспользоваться полученным критерием, должно убедиться, что дисперсии одинаковые. Как и в случае $k = 2$, это можно проверить по тесту Левена, только многомерному, т.е. проверить равенство математических ожиданий $\mathsf{E}\,|\xi - \mathsf{E}\xi_i|$ $\forall i \in 1 : k$, $|y_{ij} - \bar{y}_i|$ — опять же, через саму ANOVA.

Замечание. Если условия нормальности нарушаются, то критерий становится асимптотическим. Тогда вместо F следует использовать χ^2, так как $\mathrm{F}(k, m) \xrightarrow[m\to\infty]{} \chi^2(k)$.

Пример. Пусть дана выборка вида $\{(\xi = \text{пол}, \eta = \text{вес})\}$. Выдвинем H_0 : вес не зависит от пола. Очевидно, что ξ — категориальная случайная величина, а η — количественная. Значения ξ разобьют всю выборку на две (?) группы. Тогда проверка гипотезы сведется к проверке равенства распределений в двух группах, $\mathcal{P}_{\eta|\xi=s_1} = \mathcal{P}_{\eta|\xi=s_2}$. В предположении, что $(\eta \mid \xi = s_i) \sim \mathrm{N}(\mu_i, \sigma^2)$, равенство распределений будет следовать из равенства математических ожиданий.

7.2. Множественные сравнения

Пример. Проблема множественных сравнений возникает, например, в следующих ситуациях.

- Пусть одна группа испытуемых принимает лекарство, а вторая нет. По завершению эксперимента две группы сравниваются по m показателям. Однако чем больше показателей сравнивается, тем больше вероятность того, что *хотя бы по одному* показателю будет совпадение (в силу случайности).

- Испытывают $m = 100$ монет на честность сериями по $n = 10$ бросков: $\{(\xi_1^{(1)}, \ldots, \xi_{10}^{(1)}), \ldots, (\xi_1^{(100)}, \ldots, \xi_{10}^{(100)})\}$, иными словами $\left\{\eta^{(1)}, \ldots, \eta^{(100)}\right\}$, где $\eta^{(i)} \sim \mathrm{Bin}(10, p)$. Проверить m гипотез $H_0^{(i)} : \eta^{(i)} \sim \mathrm{Bin}(10, 1/2)$, $i \in 1 : m$. Зафиксируем $\alpha = 0.05$. Тогда, учитывая $\mathrm{pmf}_{\mathrm{Bin}(10,1/2)}(k) = C_{10}^k 2^{-10}$, $\mathsf{P}_{H_0^{(i)}}(\eta^{(i)} \geq 9) = 10 \cdot 2^{-10} + 1 \cdot 2^{-10} \approx 0.0107$, однако уже $\mathsf{P}_{H_0^{(i)}}(\eta^{(i)} \geq 8) \approx 0.0546$. Так что критерием наибольшей мощности будет $\eta^{(i)} \geq 9$:

$$\alpha_{\mathrm{I}} = \mathsf{P}_{H_0^{(i)}}(H_0^{(i)} \text{ отв}) = \mathsf{P}_{H_0^{(i)}}(\eta^{(i)} \geq 9) \approx 0.0107 \leq 0.05.$$

Но использование того же критерия для множественных сравнений сильно завышает α_{I}:

$$\begin{aligned}\mathsf{P}\left(\bigvee_{i=1}^{100} H_0^{(i)} \text{ отв}\right) &= 1 - \mathsf{P}\left(\bigwedge_{i=1}^{100} H_0^{(i)} \text{ не отв}\right) = 1 - \left(1 - \mathsf{P}(H_0^{(i)} \text{ отв})\right)^{100} \\ &= 1 - (1 - 0.0107)^{100} \approx 0.6589.\end{aligned}$$

Пусть проверяются гипотезы $H_0^{(1)}, \ldots, H_0^{(m)}$. Возможны такие ситуации:

	Retain H_0 (критерий не значим)	Reject H_0 (критерий значим)
True H_0	# True Negative	# False Discovery
False H_0	# False Negative	# True Discovery

Используя обозначения таблички,

$$\alpha_{\mathrm{I}} \approx \frac{\mathrm{FD}}{\mathrm{TN} + \mathrm{FD}}, \qquad \alpha_{\mathrm{II}} \approx \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TD}}.$$

Определение. Family-wise error rate (FWER):

$$\mathrm{FWER} = \mathsf{P}(\text{хотя бы один раз отвергнута верная гипотеза}) = \mathsf{P}\left(\bigvee_{i=1}^{m} H_0^{(i)} \text{ отв}\right).$$

Иными словами, FWER — это ошибка первого рода для всей совокупности экспериментов.

Требуется контролировать FWER на предзаданном уровне α, т.е. чтобы FWER $\sim \alpha$, где $\sim \in \{=, \leq, \to\}$. В *слабом* смысле это осуществляется, если FWER $\sim \alpha$ только если *все* $H_0^{(i)}$, $i \in 1 : m$ верны. В *сильном* смысле контроль FWER на уровне α гарантируется для *любой* конфигурации верных и не верных $H_0^{(j)}$.

Определение. Пусть $T_0 := \left\{i : H^{(i)} \text{ верна}\right\}$. Тогда

$$\mathrm{weak}\,\mathrm{FWER}_T = \mathsf{P}(\text{хотя бы один раз отвергнута верная гипотеза, если верны } H^{(i)},\ i \in T)$$

т.е. если $T_0 = T$.

Определение.

$$\mathrm{strong}\,\mathrm{FWER} = \max_{T : T \subset \{1, \ldots, m\}} \mathrm{weak}\,\mathrm{FWER}_T.$$

Обозначение. $\mathrm{FWER}_T := \mathrm{weak}\,\mathrm{FWER}_T$.

Это осуществляется двумя процедурами:

- Single

- Stepdown

7.2.1. Single

Каждая $H^{(i)}$ проверяется отдельно с уровнем значимости α_{I}. Задача сводится к тому, чтобы найти такое α_{I}, что FWER $\leq \alpha$ для какого-то нужного предзаданного α. Пусть $T = 1 : m$, т.е. будто все тесты верны; тогда

$$\alpha = \mathrm{FWER}_{\{1,\dots,m\}} = \mathsf{P}\left(\bigvee_{i=1}^{m} H_0^{(i)} \text{ отв}\right) \leq \sum_{i=1}^{m} \mathsf{P}(H^{(i)} \text{ отв}) = m\alpha_{\mathrm{I}} \implies \alpha_{\mathrm{I}} := \frac{\alpha}{m}.$$

Замечание. Из-за неравенства тест консервативный, т.е. FWER $\ll \alpha$. Значит не максимально мощный.

$$\begin{aligned} \text{strong FWER} \quad &= \quad \max_{T \subset \{1,\dots,m\}} \mathsf{P}(H^{(i)} \text{ отвергается, } i \notin T) \\ &\leq \quad \sum_{i \notin T_0} \mathsf{P}(H^{(i)} \text{ отвергается}) = |\{i : i \notin T\}|\, \alpha_1 = \alpha. \end{aligned}$$

Следствие. FWER *всегда хуже* strong FWER.

Определение. Поправка Бонферрони

$$\alpha_{\mathrm{I}} = \frac{\alpha}{m}.$$

Тест нужно проверять не с α_{I}, а с α/m. Так критерий будет консервативным (иначе — радикальным, что хуже).

Определение. Поправка Бонферрони для p-value:

$$p\text{-value} < \frac{\alpha}{m} \implies \text{отвергаем} \iff mp < \alpha \implies \text{отвергаем.}$$

7.2.2. Stepdown (Holm's algorithm)

Для увеличения мощности, применяется «Holm's algorithm»:

1. считаются все p-value $p_1, \dots, p_m$,
2. упорядочиваются: $p_{(1)} \leq \ldots \leq p_{(m)}$.
3. если $mp_{(1)} < \alpha$ то гипотеза отвергается, иначе и все последующие не отвергаются
4. в общем, если
   $$p_{(j)} < \frac{\alpha}{m - j + 1}$$
   то гипотеза отвергается, иначе и все последующие не отвергаются.

Замечание. Сей тест более мощный, потому что не всегда происходит умножение на m.

Замечание. Процедуру сложно повторить, потому что при упорядочивании гипотезы могут перемешиваться.

Предложение. FWER $\leq \alpha$.

Доказательство. Упорядочим p-value: $p_{(1)} \leq \dots \leq p_{(j)} \leq \ldots \leq p_{(m)}$. Пусть $I = \left\{i : H_0^{(i)} \text{ верна}\right\}$, $m_0 = |I|$ — количество верных гипотез, $j = \min_{k \in I}$. m_0 должно «поместиться до конца»:

$$j \leq m - m_0 + 1 \implies \frac{\alpha}{m - j + 1} \leq \frac{\alpha}{m_0}.$$

Значит

$$\begin{aligned} \mathrm{FWER}_I \quad &\leq \quad \mathsf{P}\left(p_{(j)} < \frac{\alpha}{m - j + 1}\right) \leq \mathsf{P}\left(p_{(j)} < \frac{\alpha}{m_0}\right) \leq \mathsf{P}\left(\min_{i \in I} p_i < \frac{\alpha}{m_0}\right) \\ &= \quad \mathsf{P}\left(\bigvee_{i \in I} p_i < \frac{\alpha}{m_0}\right) \leq \sum_{i \in I} \mathsf{P}\left(p_i < \frac{\alpha}{m_0}\right) = m_0 \frac{\alpha}{m_0} = \alpha \end{aligned}$$

$\square$

Частный случай Если все гипотезы и критерии независимы, то возможно точно посчитать FWER:

$$\begin{aligned}\mathrm{FWER}_{\{1,\ldots,m\}} &= \mathsf{P}\left(\bigvee_{i=1}^{m} H_0^{(i)} \text{ отв}\right) = 1 - \mathsf{P}\left(\bigwedge_{i=1}^{m} H_0^{(i)} \text{ не отв}\right) \\ &= 1 - (1-\alpha)^m = \alpha \implies \alpha_1 = 1 - \sqrt[m]{1-\alpha}\end{aligned}$$

Определение. *Поправка Šidák*:

$$\alpha_1 = 1 - \sqrt[m]{1-\alpha}.$$

7.3. ANOVA Post-Hoc Comparison

В случае отвержения гипотезы ANOVA, можно провести дополнительное выборочное тестирование выделенных групп.

7.3.1. Распределение размаха

Сопоставим $\xi_1, \ldots, \xi_n$ i.i.d. с $\mathrm{cdf}_{\xi_i}(x) = F(x)$ вариационный ряд $\xi_{(1)}, \ldots, \xi_{(n)}$.

Определение. *Размах* есть случайная величина

$$w_n = \xi_{(n)} - \xi_{(1)}$$

с функцией распределения

$$\mathsf{P}(w_n < w) = n \int_{\mathbb{R}} (F(x+w) - F(x))^{n-1}\, \mathrm{d}F(x)$$

($w_n < w \implies w_i < w$, $\mathsf{P}(w_i < w) = F(x+w) - F(x)$ — $n-1$ штук таких, плюс перебор разных минимумов по $1 : n$).

Замечание. В частном случае $F(x) = \mathrm{cdf}_{\mathrm{N}(0,\sigma^2)}(x)$, $\Phi(x) = \mathrm{cdf}_{\mathrm{N}(0,1)}(x)$ рассматривается *стандартизированный размах*

$$\mathsf{P}\left(\frac{w_n}{\sigma} < w\right) = n \int_{\mathbb{R}} (\Phi(x+w) - \Phi(x))^{n-1}\, \mathrm{d}\Phi(x).$$

Если σ неизвестна, то с подставленной оценкой $w/\tilde{s}$ называется стьюдентизированным размахом.

Утверждение. Пусть ℓ — некий параметр и — η^2 такая, что $\ell\eta^2/\sigma^2 \sim \chi^2(\ell)$; тогда

$$\frac{w_n}{\eta} \sim q(n, \ell),$$

где q — распределение стьюдентизированного размаха. Это распределение затабулировано.

Пример (Проверка выборки на outliers)**.** В нормальной модели, H_0 : нет outliers. Статистика

$$\frac{x_{(n)} - x_{(1)}}{\tilde{s}} \sim q(n, n-1)$$

потому что, естественно,

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

Замечание. Полученный критерий не очень мощный — если H_0 отвергается, то есть аутлаеры присутствуют, то $x_{(n)} - x_{(1)}$ есть большая величина, но аналогично большой является и $\tilde{s}$, поэтому всё значение статистики вырастет незначительно по сравнению со случаем не-отвержения H_0, когда аутлаеров нет. Мощность же тем больше, чем больше (по модулю) значение статистики в случае, когда требуется отвержение H_0. Это видно из того, что $\beta = \mathsf{P}_{H_1}(T(\mathbf{x}) \in \mathscr{A}_{\text{крит}})$; но мощность, как площадь под графиком плотности H_1 на критическом луче (которые располагаются на хвостах плотности H_0), тем больше, чем дальше плотность H_1 от H_0, т.е. чем больше значения статистики T в ситуации отвержения H_0.

Выход заключется в построении более устойчивых оценок для σ^2 — например, на основе медианы и абсолютного отклонения.

7.3.2. Least Significant Difference (LSD)

LSD test — это просто попарный t-test:

$$t = \frac{\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j}{\tilde{s}_{1,\dots,k}\sqrt{1/n_i + 1/n_j}} \sim \mathrm{t}(n-k),$$

где $\tilde{s}_{1,\dots,k}$ — это pooled по k группам standard deviation.

Замечание. Его стоит применять после множественного сравнения лишь к тем группам, важность которых была зафиксирована экспериментатором до проведения множественного сравнения.

Замечание. Критерий радикален. Значит, если он не нашел разницу, то и другие критерии тоже не найдут.

Замечание. Если групп немного, то можно применить поправку Бонферрони.

7.3.3. Tukey's Honest Signifficat Difference (HSD) Test

Предположение. *Модель нормальная с дисперсией* σ_0^2*, и дизайн сбалансирован:* $\mathrm{N}(\mu_i, \sigma_0^2),\ n_0 = n_i\ \forall i \in 1:k$.

По предложению 7.3.1,

$$t = \frac{\bar{\mathbf{y}}_{(k)} - \bar{\mathbf{y}}_{(1)}}{\sqrt{2\tilde{s}^2_{1,\dots,k}/n_0}} \sim q(k, n-k).$$

Тогда для проверки $H_0 : \mu_i = \mu_j$ используется HSD статистика

$$t_{ij} = \frac{|\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j|}{\tilde{s}_{1,\dots,k}\sqrt{2/n_0}},$$

а p-value считаются по $q(k, n-k)$ (таким образом, смотрят на каждую пару $(\bar{\mathbf{y}}_i, \bar{\mathbf{y}}_j)$ как на пару из размаха).

Предложение. *Это точный критерий.*

Доказательство. Действительно

$$\begin{aligned}\mathrm{FWER}_{\{1:m\}} &= \mathsf{P}\left(\bigvee_{i=1}^{m} H_0^{(i)} \text{ отв}\right) = 1 - \mathsf{P}\left(\bigwedge_{i=1}^{m} H_0^{(i)} \text{ не отв}\right) = 1 - \mathsf{P}\left(t_{ij} < t_\alpha\ \forall i,j\right) \\ &= 1 - \mathsf{P}\left(\max_{i,j} t_{ij} < t_\alpha\right) = 1 - \mathsf{P}\left(t_{k1} < t_\alpha\right) = 1 - \mathsf{P}\left(t_{k1} < F^{-1}(1-\alpha)\right) = 1 - (1-\alpha) = \alpha.\end{aligned}$$

□

7.3.4. Другие критерии

Newman-Keals stepdown вариант HSD.

Tukey-Cramer HSD вариант Tukey для несбалансированного дизайна

Dunnett сравнивает все группы с контрольной

7.3.5. Scheffé's Method

ANOVA гипотезу $H_0 : \mu_1 = \dots = \mu_k$ можно записать как

$$H_0 : \sum_{i=1}^{k} c_i \mu_i = 0, \quad \sum_{i=1}^{k} c_i = 0,$$

где $\{c_i\}_{i=1}^{k}$ — *«контраст»*.

Пример. Пусть две группы принимают k лекарств, в том числе — первым номером — плацебо. Сравнить все лекарства с плацебо *одним сравнением* можно сравнив с ним среднее арифметическое всех лекарств, для чего положить $c_1 = 1, \ c_2 = \dots c_k = -1/(k-1)$.

Полученную сумму следует отнормировать и получить статистику

$$t = \frac{\sum_{i=1}^{k} c_i \bar{\mathbf{y}}_i}{\sqrt{\mathsf{D}\left(\sum_{i=1}^{k} c_i \bar{\mathbf{y}}_i\right)}} = \frac{\sum_{i=1}^{k} c_i \bar{\mathbf{y}}_i}{\sigma \sqrt{\sum_{i=1}^{k} c_i^2 / n_i}} \sim \mathrm{N}(0,1).$$

При замене σ на $\tilde{s}$, получают, *как обычно*, $t \sim \mathrm{t}(n-k)$.

Пусть $c_1, \dots, c_d, \ d \leq k-1$ — наборы ортогональных контрастов. Тогда для любого вектора

$$t_j = \frac{\sum_{i=1}^{k} c_i^{(j)} \bar{\mathbf{y}}_i}{\sigma \sqrt{\sum_{i=1}^{k} (c_i^{(j)})^2 / n_i}}, \quad j \in 1 : d.$$

Линейная комбинация нормальных векторов с ортогональными коэффициентами независима. Следовательно, можно использовать поправки Šidák.

Сколько бы ни захотелось проверить контрастов, хочется уверенности, что FWER $\leq \alpha$. Статистика

$$\frac{t^2}{k-1} \sim \mathrm{F}(k-1, n-k).$$

Замечание. В HSD можно каждую пару рассматривать как конкретный набор контрастов. Следовательно, метод Шеффе менее мощный по сравнению с HSD (поскольку проверяет все).

7.3.6. Сравнение мощностей

Статистики всех критериев можно свести к одной с разными критическими значениями. Для примера, пусть $k = 4, n = 20, \alpha = 0.05$; тогда

Критерий	*Критическое значение*
LSD	2.09
Dunnett	2.54
Bonferroni с 3-мя плановыми сравнениями	2.63
HSD	2.8
Bonferroni с $6 = C_4^2$ сравнениями	2.93
Scheffé	3.05

Чем больше критическое значение, тем ниже мощность, конечно.

A. Свойства условного математического ожидания

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

   Замечание (Важное)*.* Таким образом, если выборка нормальная, то зависимость линейная *всегда*.

7. $\operatorname{argmin}_{\hat{\eta} \in K = \{\phi(\xi)\}} \mathsf{E}(\eta - \hat{\eta})^2 = \mathsf{E}\{\eta \mid \xi\} = \hat{\eta}^*$.