Статистика

Конспект курса

Мат-Мех, ПМИ

5–6 семестры (2015–2016)

$Revision: 1.39 $

Содержание

Вопросы по курсу

1. Распределения и их свойства: нормальное, Стьюдента, χ^2, экспоненциальное, Пуассона, Бернулли, бионмиальное, отрицательно-биномиальное, геометрическое, логнормальное, 12, 40
3. Что такое повторная независимая выборка (два определения), 2

1. Выборка

1.1. Выборка и эмпирическая случайная величина

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

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

$$\mathbf{x} = (x_1, \ldots, x_n)$$

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

Определение. *Повторной независимой выборкой (после эксперимента)* называется набор реализаций, т.е. конкретных значений ξ, случайных величин x_i. Это позволяет ввести *эмпирическую случайную величину* $\bar{\xi}_n$ с дискретным распределением

$$\bar{P}_n : \begin{pmatrix} x_1 & \dots & x_n \\ 1/n & \dots & 1/n \end{pmatrix}.$$

Замечание. Подходящее определение выбирается по контексту.

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

$$\bar{P}_m : \begin{pmatrix} x_1^* & \dots & x_m^* \\ \nu_1 & \dots & \nu_m \end{pmatrix} \quad \nu_i = \frac{n_i}{n},$$

где x_i^* — уникальные значения из выборки $\mathbf{x}$, а n_i — число x_i^* в $\mathbf{x}$ (т.н. «абсолютная частота»; тогда ν_i — «относительная частота»).

1.2. 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.

2. Точечное оценивание

2.1. Характеристики распределений

2.1.1. Определения 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, $V = \mathbb{R}$.

Выделяют *генеральные* характеристики $T(\mathcal{P})$ и *выборочные* характеристики $T(\bar{P}_n)$. Выражения для вычисления генеральных и выборочных характеристик отличаются только используемыми мерами ($\mathcal{P}$ и $\bar{P}_n$ соответственно).

Определение. *Оценка* — статистика, не зависящая от оцениваемой характеристики θ. Оценка характеристики θ, полученная по выборке объема n, обозначается $\hat{\theta}_n$.

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

$$\bar{F}_n(x) = \bar{P}_n((-\infty, x)) = \int_{-\infty}^{x} \mathrm{d}\bar{P}_n = \sum_{x_i : x_i \le x} \frac{1}{n} = \frac{\left|\{x_i \in \mathbf{x} : x_i < x\}\right|}{n}.$$

Утверждение. Пусть $\bar{F}_n$ — эмпирическая функция распределения, F_ξ — функция распределения ξ. Тогда, по теореме Гливенко-Кантелли,

$$\sup_x \left|\bar{F}_n(x) - F_\xi(x)\right| \xrightarrow{\text{a.s.}} 0.$$

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

Определение (Моменты и центральные моменты)**.** Генеральные и соответствующие им выборочные характеристики *k-го момента* и *k-го центрального момента*:

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

$$m_k^{(0)} = \int_{\mathbb{R}} (x - m_1)^k \,\mathrm{d}\mathcal{P} \qquad \bar{m}_k^{(0)} = \int_{\mathbb{R}} (x - \bar{m}_1)^k \,\mathrm{d}\bar{P}_n = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{m}_1)^k.$$

2.1.2. Характеристики положения

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

$$m_1 = \mathsf{E}\xi, \qquad \bar{m}_1 =: \bar{\mathbf{x}} = \widehat{\mathsf{E}\xi} = \mathsf{E}\bar{\xi}_n.$$

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

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

$$z_p = \sup\left\{z : F(z) \le p\right\}.$$

Определение. *Квартиль* есть 1/4- или 3/4-квантиль.

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

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 $\bar{z}_p$, что она больше по значению $|\mathbf{x}| \cdot p = np$ точек из выборки:

$$\bar{z}_p = \sup\left\{z : \bar{F}_n(z) \le p\right\} = x_{\lfloor np \rfloor + 1}.$$

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

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

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

2.1.3. Характеристики разброса

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

$$m_2^{(0)} = \mathsf{D}\xi \qquad \bar{m}_2^{(0)} =: s^2 = \widehat{\mathsf{D}\xi} = \mathsf{D}\bar{\xi}_n = \begin{cases} \mathsf{E}\left(\bar{\xi}_n - \mathsf{E}\bar{\xi}_n\right)^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2 \\ \mathsf{E}\bar{\xi}_n^2 - \left(\mathsf{E}\bar{\xi}_n\right)^2 = \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^2\right) - \bar{\mathbf{x}}^2. \end{cases}$$

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

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

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

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

откуда

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

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

Определение. *Энтропия*

2.1.4. Другие характеристики

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

$$\gamma_3 = \mathsf{A}\xi = \frac{m_3^{(0)}}{\sigma^3}.$$

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

$$\gamma_4 = \mathsf{K}\xi = \frac{m_4^{(0)}}{\sigma^4} - 3.$$

Замечание. Величина $m_4^{(0)}/\sigma^4 = 3$ соответствует стандартному нормальному распределению.

Замечание. При замене $z := (\xi - \mathsf{E}\xi)/\sigma$ величину $m_4^{(0)}/\sigma^4 = \mathsf{E}(z^4)$ 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.

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

[2] «Skewness».

Замечание. Справедлива оценка
$$\gamma_3^2 + 4 \le \gamma_4 + 3 \le \infty,$$
где минимум достигается $\mathrm{Ber}(1/2)$.

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

$$\mathrm{cov}(\xi_1, \xi_2) = \iint_{\mathbb{R}^2} (u - m_1(u))(v - m_1(v))\mathcal{P}(du \times dv) \qquad \overline{\mathrm{cov}}(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})(y_i - \bar{\mathbf{y}})$$
$$\mathrm{cor}(\xi_1, \xi_2) = \frac{\mathrm{cov}(\xi_1, \xi_2)}{\sigma_{\xi_1}\sigma_{\xi_2}} \qquad \overline{\mathrm{cor}}(\mathbf{x}, \mathbf{y}) = \frac{\overline{\mathrm{cov}}(\mathbf{x}, \mathbf{y})}{s(\mathbf{x})s(\mathbf{y})}.$$

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

Замечание (Проблема моментов). Для заданной последовательности моментов $m_1, m_2, \dots$ не обязано существовать подходящее распределение. Помимо требований $m_{2k} \ge 0$ и взаимосвязи между соседними моментами по неравенству Гёльдера, существенно, что ряд Тейлора по m_ℓ, в который, как известно, раскладывается характеристическая функция, должен сходиться равномерно.

2.2. Свойства оценок

2.2.1. Определения

Определение. *Смещение*[3] есть
$$\mathrm{bias}\,\hat{\theta}_n := \mathsf{E}\hat{\theta}_n - \theta \quad \forall \theta \in \Theta.$$

Определение. *Среднеквадратичная ошибка*[4] есть
$$\mathrm{MSE}\,\hat{\theta}_n := \mathsf{E}(\hat{\theta}_n - \theta)^2.$$

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

Выделяются следующие свойства оценок:

Несмещенность $\mathrm{bias}\,\hat{\theta}_n = 0$, т.е.
$$\mathsf{E}\hat{\theta}_n = \theta.$$

Состоятельность в среднеквадратичном смысле
$$\mathrm{MSE}\,\hat{\theta}_n \xrightarrow[n\to\infty]{} 0.$$

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

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

[3] Bias.
[4] Mean square error (MSE).

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

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

Доказательство. В самом деле, по неравенству Чебышева,

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

$\square$

Предложение. *$\bar{m}_k$ является состоятельной оценкой m_k.*

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

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^n x_i}{n} \xrightarrow{\mathsf{P}} m_1(\mathcal{P}).$$

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

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

Предложение. *$\bar{m}_k^{(0)}$ является состоятельной оценкой $m_k^{(0)}$.*

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 $(\bar{m}_2, \bar{m}_1) \xrightarrow{\mathsf{P}} (m_2, m_1)$. Тогда

$$f(\bar{m}_2, \bar{m}_1) = \bar{m}_2 - \bar{m}_1^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{\mathbf{x}}^2 = s^2 \xrightarrow{\mathsf{P}} f(m_2, m_1) = \mathsf{D}\xi.$$

Для $m_k^{(0)}$ доказывается аналогично. $\square$

2.2.2. $\bar{m}_1$

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

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

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

$\square$

Пример (Сравнение оценок мат. ожидания симметричного распределения)**.** Пусть $\mathcal{P}$ *симметрично* — в этом случае $\widehat{\mathrm{med}\,\xi} = \bar{\mathbf{x}}$ и имеет смысл сравнить две этих характеристики. Найдем среднеквадратичную ошибку $\widehat{\mathrm{med}\,\xi}$ и $\bar{\mathbf{x}}$. Поскольку обе этих оценки несмещенные, $\mathrm{MSE} = \mathsf{D}$ и

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

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

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

откуда

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

значит $\widehat{\operatorname{med}\xi}$ хуже, чем $\bar{\mathbf{x}}$.

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

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

Доказательство. Либо по (2.2.1) для $k = 1$, либо из того факта, что $\operatorname{bias}\bar{\mathbf{x}} = 0$, значит

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

и по (2.2.1) получаем утверждение. □

2.2.3. $\bar{m}_2^{(0)}$

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

Доказательство. В самом деле,

$$\mathsf{E}s^2 = \mathsf{E}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2 = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}\left(x_i - \bar{\mathbf{x}}\right)^2.$$

По несмещенности $\bar{\mathbf{x}}$, $\mathsf{E}\bar{\mathbf{x}} = \mathsf{E}\xi$, откуда

$$\mathsf{D}(x_i - \bar{\mathbf{x}}) = \mathsf{E}(x_i - \bar{\mathbf{x}})^2 - (\mathsf{E}(x_i - \bar{\mathbf{x}}))^2 = \mathsf{E}(x_i - \bar{\mathbf{x}})^2 - (\mathsf{E}\xi - \mathsf{E}\xi)^2 = \mathsf{E}(x_i - \bar{\mathbf{x}})^2,$$

значит

$$\mathsf{E}(x_i - \bar{\mathbf{x}})^2 = \mathsf{D}\xi - \mathsf{D}\bar{\mathbf{x}} = \mathsf{D}\xi - \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{D}x_i = \mathsf{D}\xi - \frac{\mathsf{D}\xi}{n}.$$

Продолжив равенство, получим

$$\frac{1}{n}\sum_{i=1}^{n}\mathsf{E}\left(x_i - \bar{\mathbf{x}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\mathsf{D}\xi - \frac{\mathsf{D}\xi}{n} = \mathsf{D}\xi - \frac{\mathsf{D}\xi}{n} = \frac{n-1}{n}\mathsf{D}\xi \xrightarrow[n\to\infty]{} \mathsf{D}\xi.$$

□

Альтернативное доказательство. Распишем

$$\begin{aligned}
\mathsf{E}s^2 &= \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \mathsf{E}\bar{\mathbf{x}}^2 = \mathsf{E}\xi^2 - \mathsf{E}\bar{\mathbf{x}}^2 = \mathsf{E}\xi^2 - \frac{\mathsf{E}\left(x_1 + \dots + x_n\right)\left(x_1 + \dots + x_n\right)}{n^2} \\
&= \mathsf{E}\xi^2 - \frac{1}{n^2}\left(n\mathsf{E}\xi^2 + n(n-1)\left(\mathsf{E}\xi\right)^2\right) = \mathsf{E}\xi^2 - \frac{\mathsf{E}\xi^2}{n} - \frac{(n-1)\left(\mathsf{E}\xi\right)^2}{n} = \frac{n-1}{n}\left(\mathsf{E}\xi^2 - (\mathsf{E}\xi)^2\right) \\
&= \frac{n-1}{n}\mathsf{D}\xi \xrightarrow[n\to\infty]{} \mathsf{D}\xi.
\end{aligned}$$

□

Определение. *Исправленная* дисперсия:

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

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

Доказательство. По (2.2.1) с $k = 2$. □

2.2.4. $\bar{F}_n$

Предложение. *$\bar{F}_n$ — состоятельная оценка F в каждой точке.*

Доказательство. Введем

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

Тогда по теореме Хинчина о ЗБЧ,

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

Независимость y_i очевидна, если расписать $F_{y_i}(x)F_{y_j}(y) = F_{y_i,y_j}(x,y)$. □

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

2.2.5. $\bar{z}_p$

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

2.3. Метод моментов

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

$$\begin{cases} m_1 = \phi_1(\theta_1, \ldots, \theta_r) \\ \quad\vdots \\ m_r = \phi_r(\theta_1, \ldots, \theta_r) \end{cases} \Longrightarrow \begin{cases} \theta_1 = f_1(m_1, \ldots, m_r) \\ \quad\vdots \\ \theta_r = f_r(m_1, \ldots, m_r). \end{cases}$$

Примем

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

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

Как правило, эти оценки смещенные.

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

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

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

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

2.4. Метод оценки максимального правдоподобия

2.4.1. Функция правдоподобия и метод

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

Определение. *Функция правдоподобия:*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

откуда

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

[5] Maximum likelihood estimate (MLE).

Утверждение. В условиях регулярности[6]:

1. Существует один глобальный максимум, так что
$$\left.\frac{\mathrm{d}\mathcal{L}(\boldsymbol{\theta}\mid\mathbf{x})}{\mathrm{d}\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}}=0.$$

2. $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ обладает всеми свойствами:
    a) Состоятельность;
    b) Асимптотическая несмещенность;
    c) Асимптотическая нормальность;
    d) *Эффективность.*

2.4.2. Информационное количество Фишера и эффективность оценки MLE

Пусть $r=1$.

Определение. *Информанта n-го порядка*:
$$S_n(\mathbf{x},\theta)=\frac{\mathrm{d}^n\ln\mathcal{L}(\theta\mid\mathbf{x})}{\mathrm{d}\theta^n}.$$

Определение. *Информационное количество Фишера*:
$$I_n(\theta):=-\mathsf{E}S_2(\mathbf{x},\theta).$$

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

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

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

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

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

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. $\xi\sim\mathrm{Pois}(\lambda)$,
$$\mathsf{D}\bar{\mathbf{x}}=\frac{\lambda}{n}=I_n(\lambda).$$

[6] Область, где $x=0$ не должна зависеть от $\boldsymbol{\theta}$ + условия на производные (существование, ограниченность, ...).

3. Некоторые распределения, связанные с нормальным

3.1. Распределение $\chi^2(m)$

Определение (Распределение $\chi^2(m)$)**.** η имеет распределение χ^2 с m степенями свободны:

$$\eta \sim \chi^2(m) \iff \eta = \sum_{i=1}^{m} \beta_i^2, \quad \beta_i \sim \mathrm{N}(0,1),\ \beta_i \text{ независимы.}$$

Свойства[7] $\chi^2(m)$

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

Утверждение. Пусть $\eta_m \sim \chi^2(m)$. Тогда, по ЦПТ,

$$\frac{\eta_m - \mathsf{E}\eta_m}{\sqrt{\mathsf{D}\eta_m}} = \frac{\eta_m - m}{\sqrt{2m}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1).$$

Пример. $m = 50,\ \eta_m = 80$. Тогда

$$\frac{80-50}{10} = 3$$

и можно посчитать, к примеру, $\Phi(3) \approx 0.9986$.

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

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

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

Свойства $\mathrm{t}(m)$

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

Предложение. *Распределение Стьюдента сходится к стандартному нормальному:*

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

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

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

3.3. Квадратичные формы от нормально распределенных случайных величин

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

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

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

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

тогда

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

В то же время, пусть

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

тогда

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

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

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

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

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

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

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

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

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

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

4. Построение гипотез

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

4.1. Построение критерия

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

Определение. *Гипотеза* — утверждение о $\mathcal{P}$, требующее проверки.

При построении эксперимента, фиксируется *нулевая гипотеза* H_0, обыкновенно состоящая в том, что эффект, который хотят выявить *не присутствует*. Задача статистического эксперимента тогда — опровергнуть H_0.

Пример (Презумпция невиновности)**.** H_0: подсудимый невиновен. Работа обвиняющей стороны сводится к опровержению H_0.

Поскольку выборка конечного объема позволяет делать только вероятностные заключения, со статистическим критерием ассоциирована величина α_{I} — вероятность ошибочно отвергнуть H_0 («вероятность ошибки I-го рода»[8]). До эксперимента фиксируется величина $\alpha : \alpha \geq \alpha_{\mathrm{I}}$ называемая *уровнем значимости* критерия. Неформально, α обратно пропорциональна «строгости» критерия, выбираемой экспериментатором.

Определение (Критерий)**.** Для выборки

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

критерий есть разбиение

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

такое, что H_0 отвергается, если $\mathbf{x} \in \mathscr{A}^{(\alpha)}_{\text{крит}}$ и не отвергается, если $\mathbf{x} \in \mathscr{A}^{(\alpha)}_{\text{дов}}$.

Определение. *Уровень значимости* $\alpha \in [0, 1]$ есть вероятность попадания в критическую область при верной нулевой гипотезе:

$$\alpha_{\mathrm{I}} = \mathsf{P}_{H_0}(\mathbf{x} \in \mathscr{A}^{(\alpha)}_{\text{крит}}) \sim \alpha, \quad \text{где } \sim \in \{=, \leq, \to\}\,.$$

Замечание. Если $\alpha_{\mathrm{I}} = \alpha$, то критерий называется *точным*; если $\alpha_{\mathrm{I}} > \alpha$, то *радикальным*, если $\alpha_I < \alpha$, то *консервативным*.

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

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

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

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

Среди всех таких c следует выбрать наименьшее, чтобы минимизировать ошибку второго рода (см. далее).

[8] Size.

Замечание. Стандартный уровень значимости $\alpha = 0.05$ или $\alpha = 0.01$.

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

Определение (Ошибки I-го и II-го родов)**.** Ошибка

- *I-го рода* есть отвержение H_0, при верной H_0; соответствующая вероятность есть

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

  Замечание. Для точного критерия вероятность α_{I} совпадает с уровнем значимости α.

- *II-го рода* есть не отвержение H_0 при верной H_1; соответствующая вероятность есть

$$\alpha_{\mathrm{II}} := \mathsf{P}_{H_1}(\mathbf{x} \in \mathscr{A}^{(\alpha)}_{\text{дов}}).$$

  Замечание. Если H_1 — отрицание H_0, то её распределение может считаться сложно.

Определение. *Мощность* критерия против альтернативы это

$$\beta := 1 - \alpha_{\mathrm{II}} = 1 - \mathsf{P}_{H_1}(\mathbf{x} \in \mathscr{A}^{(\alpha)}_{\text{дов}}) = \mathsf{P}_{H_1}(\mathbf{x} \in \mathscr{A}^{(\alpha)}_{\text{крит}}),$$

т.е. вероятность отвергнуть H_0 при верной H_1 («справедливо»).

Иными словами, это способность критерия отличать H_0 от H_1.

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

Замечание. Утверждать об *отвержении* гипотезы можно с вероятностью ошибки α (достаточно малой); утверждать о *принятии* гипотезы можно с вероятностью ошибки α_{II} — не контролируемой и потенциально довольно большой. Поэтому безопасно гипотезу можно только отвергать или не отвергать.

Замечание. При высокой вероятности ошибки II-го рода возможна ситуация не отвержении заведомо ложной гипотезы. Это, в свою очередь, может произойти из-за маленького объема выборки (критерий не находит разницу, см. 6). 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.

4.2. Статистика критерия

Определение. *Статистика критерия* есть отображение

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

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

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

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

Схема построения критерия с помощью статистики критерия

1. Выдвигают H_0 (и H_1).

2. Фиксируют предположение относительно данных (например, независимость).

3. Выбирают подходящий критерий и статистику T. T должна измерять то, насколько выборка соответствует гипотезе. В этом случае должно быть известно «идеальное соответствие» T.

   Пример. Пусть $H_0 : \mathsf{E}\xi = a_0$; тогда $T = \bar{\mathbf{x}} - a_0$ и идеальное значение $T = 0$.

4. Распределение T при верной H_0 должно быть известно хотя бы асимптотически.

   Пример. См. 4.3.1.

5. Фиксируют уровень значимости α.

6. Строят разбиение $\operatorname{im} T$ так:

   - Если H_1 неизвестна, то $\mathscr{A}_{\text{крит}}$ следует выбрать так, чтобы она располагалась как можно дальше от идеального значения, т.е. разбиение должно максимизировать мощность.
     Пример. В случае $T \sim \mathrm{N}(0,1)$, разумно определить $\mathscr{A}_{\text{крит}}$ «на хвостах» графика $\mathrm{pdf}_{\mathrm{N}(0,1)}$ симметрично по обе стороны от 0 так, что для $\mathscr{A}_{\text{крит}} = (-\infty, T_0) \cup (T_1, \infty)$

     $$\alpha/2 = \int_{-\infty}^{T_0} \mathrm{pdf}_{\mathrm{N}(0,1)}(y)\,\mathrm{d}y = \int_{T_1}^{+\infty} \mathrm{pdf}_{\mathrm{N}(0,1)}(y)\,\mathrm{d}y.$$

     Иными словами,

     $$\alpha/2 = 1 - \mathrm{cdf}_{\mathrm{N}(0,1)}(T_1) \implies T_1 = \mathrm{cdf}^{-1}_{\mathrm{N}(0,1)}(1-\alpha/2)$$

     и аналогично для T_0.

   - Если H_1 известна, то $\mathscr{A}_{\text{крит}}$ выбирается так, чтобы минимизировать α_{II}.
     Пример. Пусть $\xi \sim \mathrm{N}(a, \sigma^2)$, σ^2 известна, $H_0 : \mathsf{E}\xi = a_0$, $H_1 : \mathsf{E}\xi = a_1$. Тогда по 4.3.1,

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

     В то же время, поскольку при верной H_1, $\mathsf{E}\bar{\mathbf{x}} = 1/n \cdot \sum_{i=1}^n \xi_i = n/n \cdot a_1$ то

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

     (дисперсия, конечно, не меняется при сдвиге).

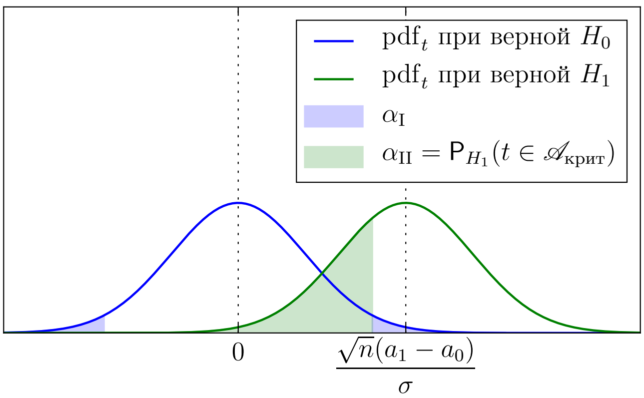


Рис. 1: Плотности распределения z (неоптимальное разбиение)

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

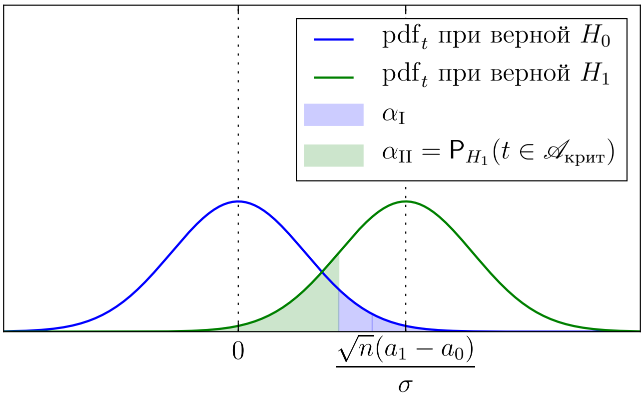


Рис. 2: Плотности распределения z (оптимальное разбиение)

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

Аналогично, чем a_1 дальше от a_0, тем α_{II} меньше. Стоит отметить, что H_1 не выбирается, но берется из смысла задачи.

7. Считают значение статистики и принимают решение об отвержении H_0 в зависимости от того, попадает ли оно в $\mathscr{A}_{\text{крит}}$ или $\mathscr{A}_{\text{дов}}$. Альтернативно, можно посчитать p-value, на основании чего сделать вывод.

Пример (С гранатами)**.**

Определение. *p-value* значения статистики T есть вероятность получить такое же или большее по модулю значение T при верной H_0.
p-value обратно пропорционален «существенности» результата.

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

Пример (Средняя температура в холодильнике)**.** Хотят купить холодильник, такой, чтобы температура не опускалась ниже 0 — иначе продукты померзнут. Известно количество измерений $n = 25$ и $\bar{\mathbf{x}} = 0.7$.

1. Выдвинута $H_0 : \mathsf{E}\xi = 0$ — если гипотеза опровергнется, то холодильник купят.

2. Пусть $\xi \sim \mathrm{N}(a, \sigma^2),\ \sigma^2 = 4$.

3. Поскольку модель нормальная и известная σ^2, выберем статистику 4.3.1 («z-test»):

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

   Идеальное значение статистики — 0.

4. Зафиксируем два уровня значимости: 0.2 (храним петрушку) и 0.01 (храним дорогую красную икру). Поскольку идеально значение статистики 0, обозначим на хвостах $\mathrm{pdf}_{\mathrm{N}(0,1)}$.

5. Посчитаем

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

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

6. Вычислим критическое значение:

   - Для $\alpha = 0.2$,
   $$z_r = \mathrm{cdf}^{-1}_{\mathrm{N}(0,1)}(1-\alpha/2) = \mathrm{cdf}^{-1}_{\mathrm{N}(0,1)}(0.9) \approx 1.28.$$
   Но $z(\mathbf{x}) = 1.75$. Таким образом, $z \in \mathscr{A}_{\text{крит}}$, H_0 отвергнется и холодильник купят.
   - Для $\alpha = 0.01$, $z_r \approx 2.57$, $z(\mathbf{x}) = 1.75 \in \mathscr{A}_{\text{дов}}$ и H_0 не отвергается; значит холодильник, быть может, не купят.

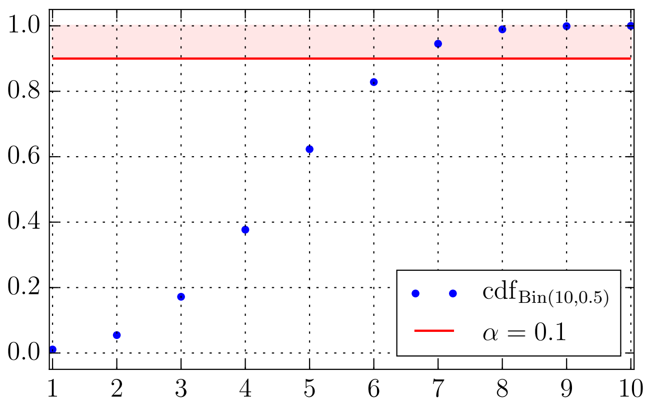
7. Можно посчитать p-value:
   $$2 \cdot (1 - \mathrm{cdf}_{\mathrm{N}(0,1)}(1.75)) \approx 0.08.$$
   Поэтому при уровне значимости $\alpha = 0.2 > 0.08$ H_0 не отвергается, а при $\alpha = 0.01 < 0.08$ отвергается.

Пример (С мышой)**.** В одном из рукавов T-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, $H_0 : p = p_0 = 0.5$. По ЦПТ,
$$\frac{\sum_{i=1}^n \xi_i - \sum_{i=1}^n \mathsf{E}\xi_i}{\sqrt{\sum_{i=1}^n \mathsf{D}\xi_i}} = \frac{n\bar{\mathbf{x}} - n\mathsf{E}\xi}{\sqrt{n}\sqrt{\mathsf{D}\xi}} = \frac{\sqrt{n}\,(\bar{\mathbf{x}} - p_0)}{\sqrt{p_0(1-p_0)}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1).$$
Пусть это будет статистикой критерия с идеальным значением 0. Тогда
$$T = \frac{\sqrt{n}\,(\bar{\mathbf{x}} - p_0)}{\sqrt{p_0(1-p_0)}} = \frac{\sqrt{10}\cdot 0.2}{0.5} \approx 1.2649 \implies p\text{-value} = 2\cdot(1 - \mathrm{cdf}_{\mathrm{N}(0,1)}(1.2649)) \approx 0.2.$$
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 6, 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.

4.3. t-критерий

4.3.1. $\widehat{\mathsf{E}\xi}$

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

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

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

  по ЦПТ.

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

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

  Доказательство.

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

  □

- Пусть $\mathsf{D}\xi$ неизвестна; тогда используется следующая статистика

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

  Замечание. При условии нормальности данных,

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

Доказательство.

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

поскольку

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

□

4.3.2. $\widehat{\mathsf{D}\xi}$

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

- Пусть $\mathsf{E}\xi = a$; При условии нормальности данных используется следующая статистика («chi-squared test for variance»):

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

Доказательство.

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

□

- Пусть $\mathsf{E}\xi$ неизвестно. При условии нормальности данных используется следующая статистика («chi-squared test for variance»):

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

Доказательство. См. (3.3). □

Альтернативное доказательство. По определению запишем

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

Но

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

откуда

$$s^2 = s_a^2 - (\bar{\mathbf{x}} - a)^2.$$

Домножив обе части на n/σ_0^2, получим

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

□

Замечание.

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

Замечание. Для строгого доказательства, нужно использовать независимость $\bar{\mathbf{x}}^2$ и s^2 (см. 3.3).

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:

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

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

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

⌟

4.4. Гипотеза согласия с видом распределения

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

4.4.1. Критерий χ^2

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

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 n_i; тогда np_i — *ожидаемая* абсолютная частота. Построим статистику критерия:

- Введем

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

  с идеальным значением 0.

- $T \xrightarrow{\mathrm{d}} \chi^2(k-1)$.
- Поскольку pdf_{χ^2} не симметрична относительно 0, зададим $\mathscr{A}_{\text{крит}}$ на хвосте графика.

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 $np_i = 100 \cdot 1/4 = 25$,

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

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

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

⌟

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

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

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

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

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

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

Найдем

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

Так как $\mathrm{pdf}_{\mathrm{Geom}(p)}(k) = (1-p)^k p$,

$$\begin{aligned} \log \mathcal{L}(\mathbf{x}; p) &= \log \prod_{k=1}^{n} (1-p)^k p = \log (1-p)^{n\bar{x}} p^n = n\bar{\mathbf{x}} \log(1-p) + n \log p \\ &= n(\bar{\mathbf{x}} \log(1-p) + \log p) \end{aligned}$$

откуда

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

Учитывая

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

найдем

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

Посчитаем статистику χ^2, найдя соответствующие p_i:

$$p_0 = \mathsf{P}_{\mathrm{Geom}(0.55)}(0) = 0.55, \quad p_1 \approx 0.26, \quad p_2 \approx 0.11, \quad p_3 \approx 0.09.$$

Тогда

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

Наконец, поскольку $\chi^2 \xrightarrow[n\to\infty]{} \chi^2(k-r-1)$,

$$p\text{-value} = 1 - \mathrm{cdf}_{\chi^2(2)}(17.77) \approx 0.00014.$$

⌟

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 $T = 0$.

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.

4.4.2. Критерий Колмогорова-Смирнова

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

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

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

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

Замечание. $\sup_x \sqrt{n}\,|\mathrm{ecdf}_n(x) - \mathrm{cdf}_0(x)| \xrightarrow{\mathrm{d}} \mathcal{P}_{\text{K.S.}}$, 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.

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

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

Критерий Лиллиефорса[9] для проверки $H_0 : \mathcal{P}_\xi = \mathrm{N}(a, \sigma^2)$ считается статистикой D_n с $\mathrm{cdf}_0(x) = \mathrm{cdf}_{\mathrm{N}(\bar{\mathbf{x}}, s^2)}(x)$, сходящейся к распределению Лиллиефорса (Колмогорова-Смирнова с учетом подстановки оценок).

Критерий Шапиро-Уилка есть $T \approx \rho^2$ для $H_0 : \mathcal{P}_\xi = \mathrm{N}(a, \sigma^2)$.

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

Пример. В R:

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

4.4.3. Критерий типа ω^2

Статистика

$$Q = n \int (F_n(x) - F_0(x))^2 \, \psi(x) \, \mathrm{d}F_0(x)$$

может быть проинтерпретирована как площадь разницы между соответствующими функциями распределения.

Cramer von Mises Q с $\psi \equiv 1$.

Anderson-Darling Q с

$$\psi(x) = \frac{1}{F_0(x)(1 - F_0(x))}.$$

Все эти критерии точны.

4.4.4. Визуальное определение согласия с распределением

P-P plot

Определение. *P-P plot* есть график

$$\{(\mathrm{cdf}_0(x_i), \mathrm{ecdf}_n(x_i))\}_{i=1}^n .$$

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)
}
```

Q-Q plot

Определение. *Q-Q plot* есть график

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

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

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

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

4.5. Гипотеза о равенстве распределений

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

Возможно рассматривать два случая:

1. Для независимых выборок: признак один, но измерен на разных индивидах. Пусть $\zeta \in \{1, 2\}$ — номер индивида, ξ — признак. Тогда $\xi_1 \sim \mathcal{P}_{\xi|\zeta=1}$, $\xi_2 \sim \mathcal{P}_{\xi|\zeta=2}$ и $\xi_1 \not\perp\!\!\!\perp \xi_2$.

2. Для зависимых выборок: две характеристики на одних и тех же индивидах (либо же «до» и «после»); $\xi = (\xi_1, \xi_2)^{\mathrm{T}}$.

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

4.5.1. Проверка гипотезы о равенстве математических ожиданий

$H_0 : \mathsf{E}\xi_1 = \mathsf{E}\xi_2$. Используется t-критерий

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

Как обычно, если данные нормальные, то $t \sim \mathrm{N}(0,1)$. Можно считать и по оценке дисперсии $\widehat{\mathsf{D}(\bar{\mathbf{x}} - \bar{\mathbf{y}})}$.

Пусть выборка независима, $(x_1, \ldots, x_{n_1}), (y_1, \ldots, y_{n_2})$, $n = n_1 + n_2$. Значит $\mathsf{D}(\bar{\mathbf{x}} - \bar{\mathbf{y}}) = \mathsf{D}\bar{\mathbf{x}} + \mathsf{D}\bar{\mathbf{y}}$. Если дисперсия неизвестна, $\widehat{\mathsf{D}(\bar{\mathbf{x}} - \bar{\mathbf{y}})} = s_1^2/n_1 + s_2^2/n_2$, откуда

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

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

В случае, если дисперсии одинаковы, то

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

Оценку можно найти по объединенной и центрированной выборке:

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

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

Итого, когда σ^2 известна,

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

причем, если данные нормальные, то $t \sim \mathrm{t}(n_1 + n_2 - 2)$.

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

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 $n_1 \approx n_2$.

5. Доверительные интервалы

5.1. Мотивация и определение

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

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

все-равно будет, конечно, состоятельной.

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

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

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

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

---

[10] Behrens-Fisher problem.

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

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

5.2.1. Доверительный интервал для a

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

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

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

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

  Наконец,

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

- Пусть σ^2 неизвестно. По аналогии,

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

  и

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

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

Решение.

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

⌟

5.2.2. Доверительный интервал для σ^2

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

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

  Тогда

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

- Пусть a неизвестно. Тогда аналогично

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

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

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

1. Её распределение («центральное распределение») не зависит от распределения θ.
2. G_n (функция распределения центрального распределения) непрерывна.
3. $\forall z_1, z_2$ и $\mathcal{P}_\theta$-почти всюду
$$z_1 < g(x_1, \ldots, x_n, \theta) < z_2$$
монотонно разрешимо относительно θ, т.е.
$$\exists f_1, f_n : f_1(x_1, \ldots, x_n, \theta, z_1, z_2) < \theta < f_2(x_1, \ldots, x_n, \theta, z_1, z_2).$$

Рассмотрим всегда разрешимое

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

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

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

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

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

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

Альтернативно замену σ на s можно обосновать по теореме Слуцкого.

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

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

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

5.4. Асимптотический доверительный интервал для параметра на основе MLE

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

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

по асимптотической несмещенности,

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

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

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

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

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

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

где

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

Пример. $\xi \sim \mathrm{Pois}(\lambda)$. По 2.4.1, $\hat{\lambda}_{\text{MLE}} = \bar{\mathbf{x}}$, по 2.4.2 $I_n(\lambda) = n/\lambda = n/\bar{\mathbf{x}}$ откуда

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

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

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

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

5.5. Доверительный интервал для проверки гипотезы о значении параметра

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

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

причем

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

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

6. Зависимость и корреляция

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

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

6.1. Вероятностная независимость

6.1.1. Визуальное определение независимости

- Поскольку при $p_\eta(y_0) \ne 0$

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

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

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

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

Тогда выборка с большей чем случайной вероятностью будет независима при пропорциональных строчках / столбцах.

Пример. Таблица сопряженности похожей не независимую выборки:

1	3	2
2	5	3
9	20	11

6.1.2. Критерий независимости χ^2

По определению, для двумерных дискретных распределений, независимость есть

$$\xi \not\parallel \eta \iff \underbrace{\mathsf{P}(\xi = i, \eta = j)}_{p_{ij}} = \underbrace{\mathsf{P}(\xi = i)}_{p_i}\underbrace{\mathsf{P}(\eta = j)}_{p_j} = \underbrace{\sum_{k=1}^{K} \mathsf{P}(\xi = i, \eta = k)}_{p_{i\cdot}} \cdot \underbrace{\sum_{s=1}^{S} \mathsf{P}(\xi = s, \eta = j)}_{p_{\cdot j}}.$$

Проверим $H_0 : \xi \not\parallel \eta$.

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

По утверждению, $\hat{p}_{ij} = \hat{p}_{i\cdot}\hat{p}_{\cdot j} = n_{i\cdot}/n \cdot n_{\cdot j}/n$. Запишем статистику

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

Количество параметров таково, потому что если $\xi \parallel \eta$, то всего $ks - 1$ параметров (-1 потому что $\sum_{ij} p_{ij} = 1$); если $\xi \not\parallel \eta$, 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 возникают проблемы со сходимостью, потому что может объединять только столбцы / строки и каждый раз терять сразу $S-1$ ($K-1$) степень свободы.

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

Замечание. Критерий не удовлетворяет 1-му пункту определения меры зависимости ($\chi^2 \notin [-1, 1]$). Это обычно исправляют так: рассматривают среднеквадратичную сопряженность χ^2/n или коэффициент сопряженности Пирсона $\chi^2/(\chi^2 + n)$ (тогда 1 никогда не достигается). Могли бы работать с $1 - p$-value, но так почему-то никогда не делают.

6.2. Линейная / полиномиальная зависимость

Определим $\phi(x) := \mathsf{E}(\eta \mid \xi = x)$. Тогда назовем зависимость *линейной*, если $\phi(x)$ — линейная функция, *квадратичной* — если квадратичная и т.д.

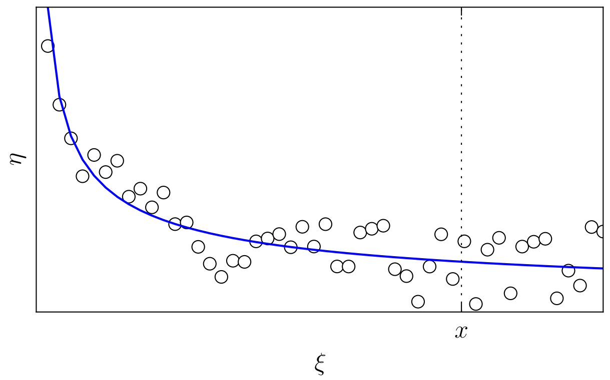


Рис. 3: Нелинейная зависимость

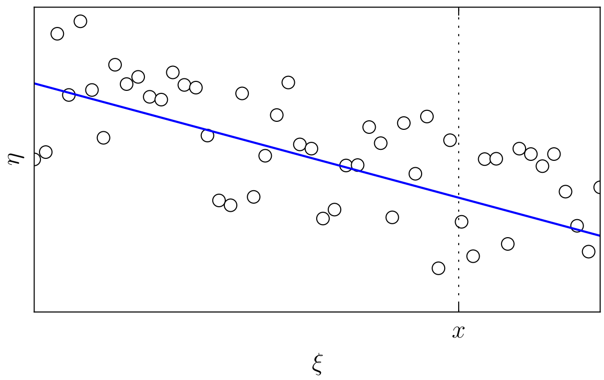


Рис. 4: Линейная зависимость

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

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

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

Замечание. $\xi \not\perp\!\!\!\perp \eta \implies \rho = 0$ и $\xi, \eta \sim \mathrm{N}(a, \sigma^2)$, $\xi \not\perp\!\!\!\perp \eta \iff \rho = 0$.

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

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

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

□

О соотношении ρ и коэффициента линейной регрессии По (6.8), 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 $k = \rho = \operatorname{cov}(\xi,\eta)$.

6.3. Корреляционная зависимость

По свойству УМО(2.2.1), для заданного вектора η из $L_2 = \left\{\eta : \mathsf{E}\eta^2 < \infty\right\}$ со скалярным произведением $(\eta, \xi) = \mathsf{E}\eta\xi$ и подпространства $K = \{\phi(\xi)\} = \{\hat\eta : \sigma(\phi(\xi))\text{-измерима}\}$ вектор $\hat\eta^* = \mathsf{E}\{\eta \mid \phi(\xi)\}$ 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:

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

Если $K = \{a\xi + b\}$ — линейное пространство, то теорема Пифагора принимает вид

$$\mathsf{E}\,(\eta - \mathsf{E}\eta)^2 = \underbrace{\mathsf{E}\,(\hat\eta^* - \mathsf{E}\eta)^2}_{\text{объяснённая доля аппроксимации}} + \underbrace{\mathsf{E}\,(\eta - \hat\eta^*)^2}_{\text{ошибка аппроксимации}}.$$

Откуда можно записать меру аппроксимации

$$\frac{\mathsf{E}(\hat\eta^* - \mathsf{E}\eta)^2}{\mathsf{D}\eta} = 1 - \frac{\mathsf{E}\,(\eta - \hat\eta^*)^2}{\mathsf{D}\eta} = 1 - \frac{\min_{\hat\eta \in K} \mathsf{E}\,(\eta - \hat\eta)^2}{\mathsf{D}\eta}.$$

$\hat\eta^*$ — предсказание η по ξ по МНК.

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

и можно выразить коэффициент корреляции Пирсона

$$\rho^2(\eta, \xi) = 1 - \frac{\min_{\hat\eta \in K} \mathsf{E}\,(\eta - \hat\eta)^2}{\mathsf{D}\eta}.$$

На случай нескольких случайных величин обобщается естественно («*множественный коэффициент корреляции*»):

$$R^2(\eta, \xi_1, \ldots, \xi_k) = 1 - \frac{\min_{\hat\eta \in \left\{\sum_{i=1}^k b_i\xi_i + b_0\right\}} \mathsf{E}\,(\eta - \hat\eta)^2}{\mathsf{D}\eta}.$$

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

В общем случае, («*корреляционное отношение*»)

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

6.3.1. Свойства корреляционного отношения

1. $r^2_{\eta|\xi} \in [0, 1]$.
2. $\eta \not\perp\!\!\!\perp \xi \implies r^2_{\eta|\xi} = 0$.
3. $\eta = \phi(\xi) \iff r^2_{\eta|\xi} = 1$.
4. Вообще говоря, $r^2_{\eta|\xi} \neq r^2_{\xi|\eta}$. К примеру, для любой не монотонной функции (так, чтобы не существовала обратная).
5. $r^2_{\eta|\xi} \geq \rho^2(\eta, \xi)$ (потому что минимум по всем функциям меньше, чем лишь по линейным, значит $1 - \min$ больше).
6. $(\xi, \eta)^{\mathrm{T}} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies r^2_{\eta|\xi} = \rho^2(\eta, \xi)$.

6.4. Частная корреляция

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

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

Если регрессия линейна, то

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

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

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

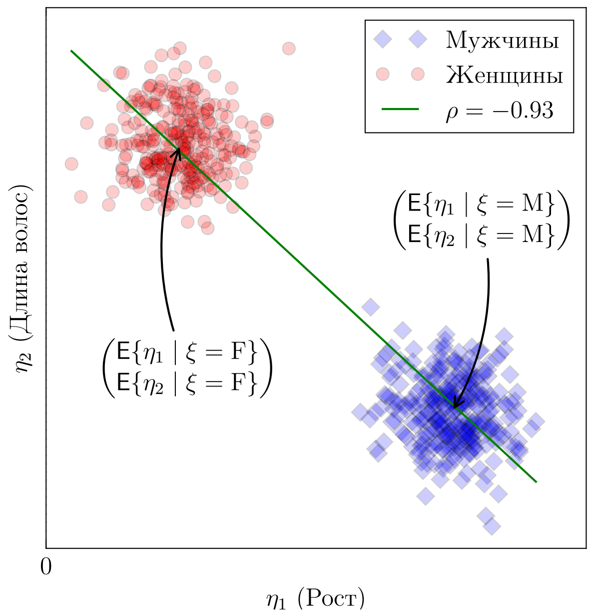


Рис. 5: Исходные данные (бимодальность)

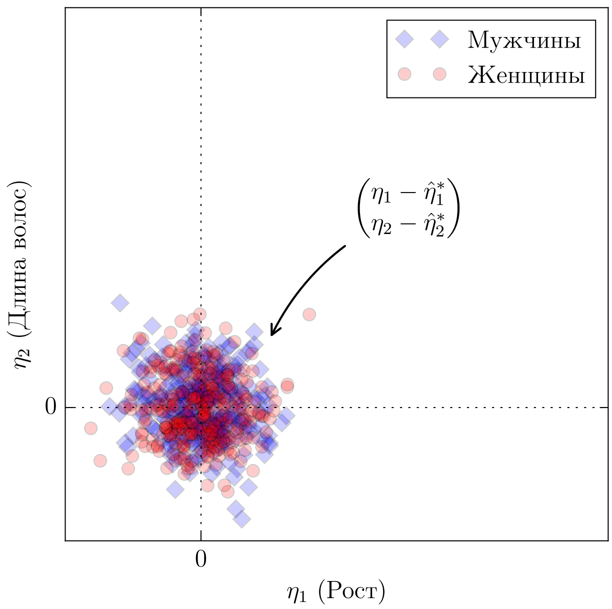


Рис. 6: Центрированные данные

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

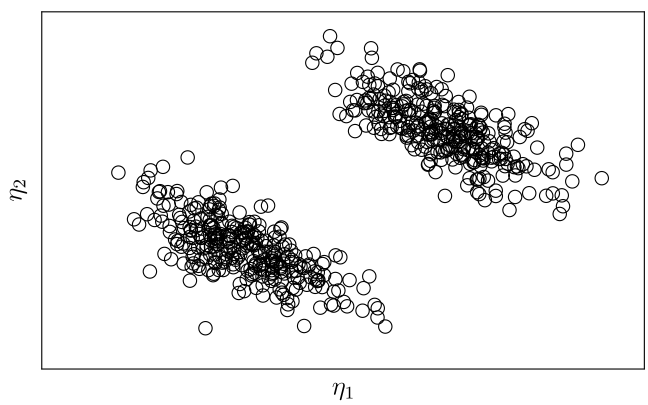


Рис. 7: $\rho(\eta_1, \eta_2) > 0$, но $\rho(\eta_1, \eta_2 \mid \xi) < 0$

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

6.5. Значимость коэффициента корреляции

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

Статистика для проверки значимости:

$$T = \frac{\sqrt{n-2}\hat{\rho}_n}{\sqrt{1-\hat{\rho}_n^2}} \sim \mathrm{t}(n-2) \quad \text{если } \begin{pmatrix}\xi\\ \eta\end{pmatrix} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Идеальное значение — 0, два хвоста.

6.6. 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).

6.6.1. Ранговый коэффициент Спирмана

Определение. *Ранговый коэффициент Спирмана* есть

$$\rho_S = \rho(F_\xi(\xi), F_\eta(\eta)).$$

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

Определение. *Ранг* элемента из выборки x_i есть его порядковый номер:

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

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

Замечание. Можем ввести эмпирическое распределение

$$\begin{pmatrix}\begin{pmatrix}R_1/n\\ T_1/n\end{pmatrix} & \cdots & \begin{pmatrix}R_n/n\\ T_n/n\end{pmatrix}\\ 1/n & \cdots & 1/n\end{pmatrix}.$$

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

$$\hat{\rho}_S = \frac{1/n \cdot \sum_{i=1}^n R_i T_i - \bar{R}\bar{T}}{\sqrt{1/n \cdot \sum_{i=1}^n \left(R_i - \bar{R}\right)^2}\sqrt{1/n \cdot \sum_{i=1}^n \left(T_i - \bar{T}\right)^2}}.$$

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. Справедлива формула

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

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

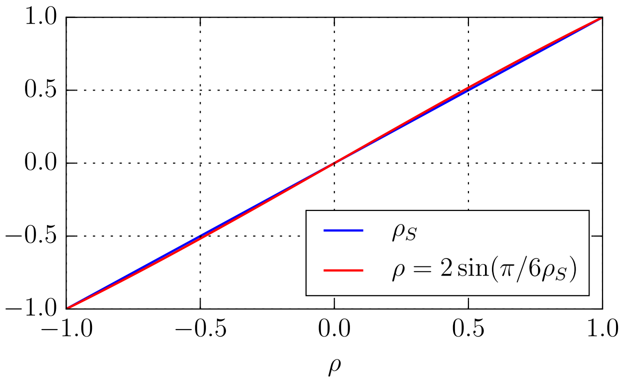


Рис. 8: $\hat{\rho} \approx \hat{\rho}_S$

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

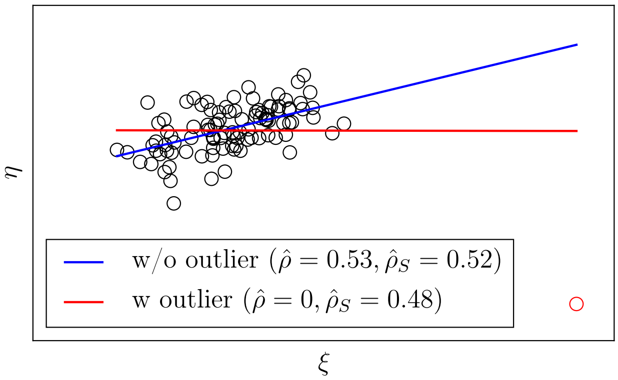


Рис. 9: $\hat{\rho}$ до и после добавления аутлаера

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

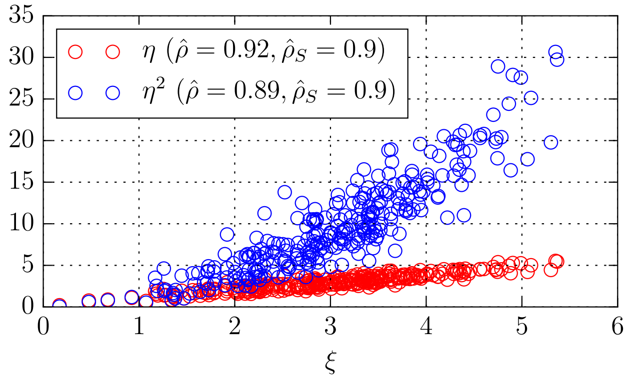


Рис. 10: Монотонное преобразование нормальных данных

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

Определение. Пусть $\begin{pmatrix}\xi_1\\ \eta_1\end{pmatrix} \not\!\perp\!\!\!\perp \begin{pmatrix}\xi_2\\ \eta_2\end{pmatrix} \sim \mathcal{P}_{\xi,\eta}$; тогда *ранговым коэффициентом Кэндалла* называется

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

На выборочном языке, пусть $(x_1, y_1), \ldots, (x_n, y_n) = \mathbf{\bar{x}}$, 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 $\mathrm{sign}(x_i - x_j) = \mathrm{sign}(y_i - y_j)$, а $\#(\text{комбинаций пар}) = C_n^2 = n(n-1)/2$.

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

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

Из утверждения следует, что τ все время меньше ρ и ρ_S .

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

6.7. 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 $\rho_i : \rho_0 < \rho_1 < \ldots$. Тогда сначала следует составить плеяду по ρ_0, затем внутри полученного по ρ_1 и т.д.

6.8. Парная линейная регрессия

Определение. Пусть $\xi, \eta \in L^2(\mathcal{F}, \mathsf{P})$ пространству $\mathcal{F}$-измеримых по мере P функций с конечным вторым моментом, причем $\xi \in H \subset L^2$ — пространству линейных функций вида $h(\xi) = \beta_1\xi + \beta_2$. *Линейной регрессией* будет называться такая линейная функция $h_{\beta_1,\beta_2}(\xi) = \beta_1\xi + \beta_2$, что она минимизирует расстояние до η по L^2-метрике:

$$\begin{aligned} h_{\beta_1,\beta_2}(\xi) &= \operatorname*{argmin}_{\beta_1,\beta_2} \|\eta - h_{\beta_1,\beta_2}(\xi)\|^2 = \operatorname*{argmin}_{\beta_1,\beta_2} \int_\Omega (\eta - h_{\beta_1,\beta_2}(\xi))^2 \, \mathrm{d}\mathsf{P} \\ &= \operatorname*{argmin}_{\beta_1,\beta_2} \underbrace{\mathsf{E}(\eta - (\beta_1\xi + \beta_2))^2}_{\phi(\beta_1,\beta_2)} = \beta_1^*\xi + \beta_2^*. \end{aligned}$$

Замечание. Найти минимум ϕ можно, как обычно, решив систему $\partial\phi/\partial\beta_i = 0$[11].

Утверждение. β_1^*, β_2^* таковы, что

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

Иными словами,

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

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

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

Замечание. Подстановкой проверятся, что

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

откуда можно найти уже известное выражение для

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

Определение. Величина sum of squares residual есть

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

Модель линейной регрессии Можно описать выборку как

$$y_i = \beta_1 x_i + \beta_2 + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2),\ \epsilon_i \not\perp\!\!\!\perp \epsilon_j.$$

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,

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

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

Значимость линейной регрессии Значимость регрессии эквивалентна значимости предсказания по ней. Пусть $H_0 : \rho = 0$. Если H_0 отвергается, то линейная регрессия значима.

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[12], что

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

Кроме того,

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

Предложение. *Статистика*

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

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

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

Тогда

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

□

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

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

Аналогично, для β_2:

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

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.

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

A. Другие полезные распределения случайных величин

A.1. Пуассона

A.2. Логнормальное

B. Свойства условного математического ожидания

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

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