Статистика

Конспект практических занятий

Мат-Мех, ПМИ, СМ–СМ

$Revision: ... (NEG: UNDER CONSTRUCTION) $

СтатМод

Оглавление

[1] ANalysis Of VAriation

Часть I.

Оценки характеристик и параметров распределения

1. Выборка и эмпирическая случайная величина

Пусть $\xi : \big(\Omega, \mathcal{F}, \mathsf{P}\big) \to \big(V, \mathfrak{A}\big)$ — случайная величина с распределением $\mathcal{P}$ (пишем $\xi \sim \mathcal{P}$).

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

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

независимых в совокупности одинаково распределенных случайных величин с распределением $\mathcal{P}$. (Знак $\perp\!\!\!\perp$ иногда используется для обозначения независимости случайных величин.)

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

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

Замечание. Подходящее определение выбирается по контексту.

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

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

Если ξ имеет дискретное распределение, то выборку можно *сгруппировать*; тогда распределение случайной величины $\hat{\xi}_n$ запишется как

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

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

Следствие. *По ЗБЧ (теореме Бернулли),*

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

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

Выше используется сходимость по вероятности (P от слова probability). Обозначение $\zeta_n \xrightarrow{\mathsf{P}} \zeta$ означает, что $\forall \varepsilon > 0$, $\lim\limits_{n\to\infty} \mathsf{P}(|\zeta_n - \zeta| > \varepsilon) = 0$.

2. Виды признаков

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

Количественные признаки: $V \subset \mathbb{R}$, заданы операции с вещественными числами.

По типу операций:

- Аддитивные: заданы, т.е. имеют смысл в контексте данного признака, операции $+, -$. Разница между значениями характеризуется разностью значений.
- Мультипликативные: заданы операции $\times, /$; признак принимает не отрицательные значения. Разница между значениями измеряется в процентах (определяется делением).

По типу данных:

- Непрерывные
- Дискретные

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

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

Пример. Цвет глаз, имена, пол.

3. Характеристики распределений и метод подстановки

Определение. *Статистика* — измеримая функция от выборки.

Обобщением статистики является понятие характеристики.

Определение. *Характеристика* распределения — функционал от распределения:

$$T : \{\mathcal{P}\} \to D;$$

Чаще всего, $D = \mathbb{R}$.

Определение. *Оценка* — функция от выборки, не зависящая от генеральной характеристики θ.

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

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

Здесь используется обозначение cdf, сокращение от cumulative distribution function, т.е. от названия функции распределения. Мы здесь оставим это обозначение, однако часто на занятиях будем обозначать функцию распределения просто $F(x)$ или $F_\xi(x)$, чтобы подчеркнуть, какой случайной величины это функция распределения.

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

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

(сходимость a.s. означает сходимость almost surely, почти наверно, почти всегда; она вводится для случайных величин, заданных на вероятностном пространстве и означает, что сходимость имеет место для почти всех элементарных событий кроме, может быть, событий меры ноль).

Может возникнуть вопрос, почему слева от стрелки случайная величина. Дело в том, что в этом утверждении(как и в любом теоретическом утверждении в математической статистике, выборка понимается как «до эксперимента», т.е. все x_i — случайные величины, поэтому эмпирическая функция распределения в точке x, равная числу x_i, меньших x, тоже является случайной.

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

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

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

Выше используется другой тип сходимости, более слабый, по распределению (d от слова distribution). Эта сходимость означает, что функция распределения случайной величины слева сходится в функции распределения, указанного справа, во всех точках ее непрерывности.

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

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

Следствие. *Значит, при достаточно больших n, в качестве интересующей характеристики $\theta = f(\xi)$ распределения $\mathcal{P}_\xi$ можем брать ее оценку $\hat{\theta} = \hat{\theta}_n = f(\hat{\xi}_n)$ — аналогичную характеристику $\hat{\mathcal{P}}_n$. Этот метод называется методом подстановки.*

4. Характеристики распределений и их оценки

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

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

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

4.1. Характеристики положения

В качестве характеристики положения для количественных признаков выделяется 1-й момент — математическое ожидание и его оценка *выборочное среднее*:

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

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

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

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

Квартиль есть квантиль уровня, кратного 1/4; дециль — 1/10; перцентиль — 1/100. Эти характеристики определены для порядковых признаков (и, следовательно, для количественных тоже).

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

Определение. *Медиана* есть 0.5-квантиль:

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

Определение. Мода ($\mathrm{mode}\,\xi$) есть точка локального максимума плотности или состояние с максимальной вероятностью для качественных признаков.

По методу подстановки можем получить аналогичные выборочные характеристики.

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

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

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

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

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

4.2. Характеристики разброса

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

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

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

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

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

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

откуда

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

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

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

$$H = \log_2 m$$

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

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

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

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

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

Определение. *Выборочное стандартное отклонение* есть

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

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

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

SD позволяет оценивать стандартное отклонение распределения ξ.

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

Определение. *Стандартная ошибка* оценки есть

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

Это показатель разброса оценки случайной величины.

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

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

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

Пусть $c_\gamma = \mathrm{qnt}_{\mathrm{N}(0,1)}\,\gamma$. $\mathrm{N}(\mu, \sigma^2$ обозначает нормальное распределение с математическим ожиданием μ и дисперсией σ^2.

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

$$\left(\bar{\mathbf{x}} \pm \mathrm{SD} \cdot c_\gamma\right).$$

SE как показатель разброса выборочного среднего использовать по смыслу нельзя.

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

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

Подробности, включая определение c_γ см. в разделе посвященном доверительным интервалам (Глава 3).

4.3. Характеристики формы распределения

Для удобства, обозначим $\sigma^2 = \mathsf{m}_2^{(0)} = \mathsf{D}\xi$.

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

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

Замечание. Не зависит от линейных преобразований.

Замечание. Старое определение скошенности было $\frac{\mathsf{E}\xi - \mathrm{med}\,\xi}{\sigma}$.

Замечание. Типичный случай соответствует тому, что при положительном коэффициенте асимметрии ‘хвост вправо’.

[2] «Skewness».

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

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

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

Замечание. Положительный коэффициент эксцесса соответствует медленному убыванию на концах отрезка. Причём, так как распределение стандартизуется, имеется в виду убывание на хвостах, которое медленнее по порядку (!), чем убывание на хвостах у нормального распределения. Например, сравните e^{-x^2}, $e^{-x^2/10}$ и e^{-10x}. Часто говорят об островершинности при положительном эксцессе, но это просто вторая сторона скорости убывания на хвостах. Медленное убывание на хвостах означает на практике, что далекие от среднего значения встречаются необычно часто.

4.4. Характеристики зависимости

Определение. Пусть $(\xi_1, \xi_2)^{\mathrm{T}} \sim \mathcal{P}$ и $(x_1, y_1)^{\mathrm{T}}, \ldots, (x_n, y_n)^{\mathrm{T}} \sim \mathcal{P}$ — выборка из этого распределения. Тогда можно записать две другие важные характеристики: *ковариацию* и *коэффициент корреляции*:

$$\operatorname{cov}(\xi_1, \xi_2) = \mathsf{E}(\xi_1 - \mathsf{E}\xi_1)(\xi_2 - \mathsf{E}\xi_2) = \mathsf{E}\xi_1\xi_2 - \mathsf{E}\xi_1\mathsf{E}\xi_2 \qquad \widehat{\operatorname{cov}}(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

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

5. Точечная оценка параметров распределения

5.1. Метод подстановки

Метод подстановки заключается в подстановке вместо неизвестного теоретического распределения известного эмпирического распределения. Например, вас интересует некоторая характеристика $f(\xi)$, а вы в качестве оценки предлагаете $\widehat{f(\xi)} = f(\xi_n)$, где $\hat{\xi}_n = \xi_n$ — эмпирическая случайная величина.

5.2. Метод моментов

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

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

Примем

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

Часто, $g_i(\xi) = \xi^i$. Или, еще чаще, $g_1(\xi) = \xi$ и $g_i(\xi) = (\xi - \mathsf{E}\xi)^i$, $i > 1$, так как для таких моментов обычно известны формулы.

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

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

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

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

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

Пример 5.2 ($r = 1$)**.** Пусть $\xi \sim \mathrm{Exp}(\lambda)$. Тогда $\mathsf{E}\xi = \lambda$ и $\bar{x} = \lambda$.

5.3. Метод оценки максимального правдоподобия

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

Определение. Пусть

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

Тогда *функция правдоподобия* определяется как значение распределения выборки (плотности в непрерывном случае и вероятности значений в дискретном) с подстановкой выборки вместо аргумента:

$$\mathsf{L}(\boldsymbol{\theta} \mid \mathbf{x}) = \mathsf{P}(\mathbf{x} \mid \boldsymbol{\theta}).$$

Пример 5.3. Пусть $\xi \sim \mathrm{N}(\mu, \sigma^2)$. По независимости x_i, $p_{\boldsymbol{\theta}}(\mathbf{x})$ распадается в произведение:

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

Пример 5.4. $\xi \sim \mathrm{Pois}(\lambda)$,

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

Утверждение. Пусть $\mathbf{x}$ — выборка. В качестве оценки максимального правдоподобия[1] $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ следует взять

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

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

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

откуда

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

Утверждение. В условиях регулярности:

1. Существует один глобальный максимум, так что

$$\left.\frac{\partial \ln \mathsf{L}(\lambda \mid \mathbf{x})}{\partial \lambda}\right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\mathrm{MLE}}} = 0.$$

2. $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ обладает всеми свойствами (про определение этих свойст написано в следующих разделах):
   a) Состоятельность;
   b) Асимптотическая несмещенность;
   c) Асимптотическая нормальность;
   d) Асимптотическая эффективность.

[1] Maximum likelihood estimate (MLE).

6. Свойства оценок

6.1. Несмещенность

Определение. *Смещение*[1] есть

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

Определение. *Среднеквадратичная ошибка*[2] есть

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

Замечание. Поскольку

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

то

$$\underbrace{\mathsf{E}(\hat{\theta}_n - \theta)^2}_{\text{MSE}} = \mathsf{D}\hat{\theta}_n + \underbrace{(\mathsf{E}(\hat{\theta}_n - \theta))^2}_{\text{bias}^2}. \tag{6.1}$$

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

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

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

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

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

□

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

Доказательство. Для $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ имеем:

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

□

Определение. *Исправленная* дисперсия:

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

Очевидно, исправленная дисперсия — несмещенная оценка дисперсии.

[1] Bias.
[2] Mean squared error (MSE).

6.2. Состоятельность

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

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

Как следует из равенства (6.1), для асимптотически несмещенных оценок состоятельность в средне-квадратическом следует из сходимости дисперсии оценки к нулю.

Определение. Оценка называется *состоятельной*, если

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

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

Доказательство. В самом деле, по неравенству Чебышева,

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

□

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

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

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

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

Замечание. Состоятельность выполняется и для центральных моментов $\mathsf{m}_k^{(0)}$, так как они выражаются через нецентральные, а свойство состоятельности сохраняется для линейной комбинации.

В частности, $\bar{x}$ — состоятельная оценка $\mathsf{E}\xi$ и s^2 — состоятельная оценка $\mathsf{D}\xi$.

6.3. Асимптотическая нормальность

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

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

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

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

Доказательство. По ЦПТ,

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

□

Мы обсуждали, что асимптотическую нормальность можно определять и в слабом смысле — как сходимость по распределению к нормальному распределению $N(0,1)$ стандартизированной случайной величины.

6.4. Эффективность

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

$$\operatorname{MSE}\hat\theta^{(1)} \le \operatorname{MSE}\hat\theta^{(2)}.$$

Замечание. Для несмещенных оценок определение эквивалентно, конечно,

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

Определение. В классе несмещенных оценок оценка называется эффективной (в средне-квадратическом), если ее дисперсия минимальна. В классе асимптотически несмещенных оценок оценка $\hat\theta = \hat\theta_n$ называется асимптотически эффективной, если для любой другой оценки $\hat\theta^*$ выполнено $\lim_{n\to\infty} \mathsf{D}\hat\theta_n/\mathsf{D}\hat\theta_n^* \le 1$.

6.4.1. Эффективность и неравенство Рао-Крамера

Пусть $\mathcal{P}_\xi(\boldsymbol\theta),\ \boldsymbol\theta = (\theta_1,\dots,\theta_r)^\mathsf{T}$ — параметрическая модель. Пусть $r = 1$.

Определение. *Информанта n-го порядка*:

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

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

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

Замечание.

$$\ln \mathsf{L}(\theta \mid \mathbf{x}) = \sum_{i=1}^n \ln p_\theta(x_i) \implies S_2 = \frac{\mathrm{d}^2 \ln \mathsf{L}(\theta \mid \mathbf{x})}{\mathrm{d}\theta^2} = \sum_{i=1}^n (\ln p_\theta(x_i))'',$$

откуда, для повторной независимой выборки,

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

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

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

Определение. Условие регулярности: имеют отношение к независимости носителя распределения от параметра, а также к существованию и ограниченности производных функции логправдоподобия по параметру до определённого порядка дифференцирования.

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

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

$$\mathsf{D}\hat\theta_n \ge \frac{1}{I_n(\theta)}.$$

Для смещенных оценок,

$$\mathsf{D}\hat\theta_n \ge \frac{(1+\operatorname{bias}'(\theta))^2}{I_n(\theta)}.$$

Следствие. *Несмещенная оценка является эффективной, если:*

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

Следствие. *Асимптотически несмещенная оценка является асимптотически эффективной, если:*

$$\mathsf{D}\hat{\theta}_n \cdot I_n(\theta) \to 1 \ \textit{as}\ n \to \infty.$$

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

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

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

$$\begin{aligned} \mathsf{D}\hat{\lambda}_n &= \mathsf{D}\bar{x} = \mathsf{E}\xi/n = \lambda/n \\ I_n(\lambda) &= n/\lambda, \end{aligned}$$

то $\hat{\lambda}_n$ — эффективная оценка (по свойствам $\hat{\theta}_{\mathrm{MLE}}$, гарантировано, что она асимптотически эффективная).

6.5. Устойчивость оценок

Так как в реальных данных часто бывают те или иные ошибки, часто жертвуют точностью для увеличения устойчивости (робастности) к выбросам. Устойчивые аналоги оценок часто строятся на основе рангов (номеров по порядку в упорядоченной выборке). Приведем пример.

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

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

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

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

откуда

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

значит $\bar{x}$ эффективнее $\widehat{\mathrm{med}\,\xi}$.

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

Часть II.

Некоторые распределения, связанные с нормальным

1. Распределение $N(a, \sigma^2)$

Свойства хорошо известны. В частности, плотность имеет вид

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

математическое ожидание равно a, дисперсия σ^2, асимметрия и эксцесс равны 0.

Рассмотрим вопрос про измерение расстояния в сигмах. Будет говорить, что точка далеко от мат.ожидания, если это и более далекие значения маловероятны.

Формально, пусть $\xi \sim N(a, \sigma^2)$. Рассмотрим $\mathsf{P}(|\xi - a| > k\sigma)$. Эта вероятность не зависит от σ и равна $2(1 - \Phi(k))$, где $\Phi(x)$ — функция стандартного нормального распределения $\mathrm{N}(0, 1)$.

Значения $\mathsf{P}(|\xi - a| > k\sigma)$:

k	Вероятность
1	0.317
1.64	0.101
1.96	0.050
2	0.046
3	2.70E-03
6	1.97E-09

Отсюда правило двух сигма (вероятность быть на расстоянии от мат.ож. больше двух сигм примерно равна 0.05), правило трех сигм, правило шести сигм.

2. Распределение $\chi^2(m)$

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

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

Свойства[1] $\chi^2(m)$

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

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

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

Пример. $m = 50,\ \eta_m = 80$. Тогда

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

и

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

Предложение. $\chi^2(m)/m \xrightarrow[m\to\infty]{} 1$.

Доказательство. По ЗБЧ. □

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

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

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

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

Свойства $\mathrm{t}(m)$

- При $m = 1$ это распределение Коши, у него не существует математического ожидания.
- При $m > 1$, $\mathsf{E}\xi = 0$ по симметричности.
- При $m > 2$, $\mathsf{D}\xi = m/(m-2)$.
- При $m > 3$, $\mathsf{A}\xi = 0$ по симметричности.
- При $m > 4$, $\mathsf{K}\xi = 6/(m-4)$.

Предложение. *Распределение Стьюдента сходится к стандартному нормальному:*

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

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

4. Распределение Фишера

Определение. Распределение Фишера имеет вид

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

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

5. Квадратичные формы от нормально распределенных случайных величин

(Это на след. семестр, сейчас можно не вникать.)
Пусть $\boldsymbol{\xi} = (\xi_1, \dots, \xi_p)^\mathsf{T} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_p)$, $\mathbf{B}$ — симметричная, неотрицательно определенная матрица. Найдем распределение $\boldsymbol{\xi}^\mathsf{T}\mathbf{B}\boldsymbol{\xi}$.

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

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

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

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

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

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

так как

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

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

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

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

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

Пример. Покажем, что

$$ns^2/\sigma^2 \sim \chi^2(p-1).$$

Воспользуемся представлением из предыдущего примера: $ps^2 = \mathbf{x}^\mathsf{T}\mathbf{C}\mathbf{x}$. 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 $\boldsymbol{\xi} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_p)$, $\boldsymbol{\xi}^\mathsf{T}\boldsymbol{\xi} = \sum_i Q_i$, где Q_i — квадратичная форма, заданная $\mathbf{B}_i$, $\operatorname{rk}\mathbf{B}_i = r_i$. Тогда следующие утверждения эквивалентны:

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

6. Распределение важных статистик

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

Предложение. $t = \sqrt{n}\frac{(\bar{x}-\mathsf{E}\xi)}{\sigma}$ *имеет стандартное нормальное распределение.*

Доказательство.

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

□

Определим $s_a^2 = \sum_{i=1}^n (x_i - a)^2/n$.

Предложение. $ns_a^2/\sigma^2 \sim \chi^2(n)$.

Доказательство.

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

□

Предложение. $ns^2/\sigma^2 = (n-1)\tilde{s}^2/\sigma^2 \sim \chi^2(n-1)$.

Доказательство. См. раздел 5). □

Альтернативное доказательство. По определению запишем

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

Домножив обе части на n/σ^2, получим

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

□

Замечание. Для строгого доказательства, нужно использовать независимость $\bar{x}^2$ и s^2 (см. раздел 5).

Предложение. *Следующая статистика имеет распределение Стьюдента:*

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

Предложение. $t = \sqrt{n-1}\,\frac{\bar{x}-a}{s} = \sqrt{n}\,\frac{\bar{x}-a}{\tilde{s}} \sim \mathrm{t}(n-1)$.

Доказательство.

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

поскольку

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

и они независимы (также пока без доказательства — используется техника квадратичных форм или можно доказать через разложение дисперсии). □

Часть III.

Проверка гипотез и доверительные интервалы

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

1. Построение критерия

1.1. Общие сведения

Пусть H_0 — это гипотеза, т.е. некоторое предположение о случайной величине ξ, которое мы хотим проверить (модель — это предположение, которое считается верным без проверки).

Задаем уровень значимости $0 < \alpha < 1$.

Тогда критерий — это разбиение множества V^n всевозможных значений выборки $\mathbf{x}$ на две области, критическую $\mathcal{A}_\alpha^{(\text{крит})}$ и доверительную $\mathcal{A}_\alpha^{(\text{дов})}$ так, что $\alpha_I = \mathsf{P}_{H_0}(\mathbf{x} \in \mathcal{A}_\alpha^{(\text{крит})}) = \alpha$ (P_{H_0} — вероятность, соответствующая предположению, что H_0 верна).

При проверки гипотезы (уже после эксперимента), если выборка попала в критическую область $\mathcal{A}_\alpha^{(\text{крит})}$, то нулевая гипотеза отвергается, а если в доверительную $\mathcal{A}_\alpha^{(\text{дов})}$, то не отвергается (важно, что именно нет основания отвергнуть, но и принять нельзя).

Обычно разбиение строят с помощью статистики критерия $t = t(\mathbf{x})$. В этом случае на доверительную и критическую область нужно делать область значений статистики критерия, а это подмножество вещественных чисел. Поэтому критическая область $A_\alpha^{(\text{крит})}$ выбирается так, чтобы $\alpha_I = \mathsf{P}_{H_0}(t \in A_\alpha^{(\text{крит})}) = \alpha$.

Допустимо строить разбиение так, чтобы выполнялось $\alpha_I \leq \alpha$ (тогда критерий называется консервативным).

Часто удается построить только асимптотический критерий, когда $\alpha_I \to \alpha$ при $n \to \infty$. В этом случае критерий можно применять при достаточно (для критерия) большом объеме выборки, где допустимый объем выборки зависит от скорости сходимости.

Ниже более подробно.

1.2. Схема построение критерия на основе статистики критерия

1. Строим статистику критерия t так, что:
   - Статистика критерия t должна измерять то, насколько выборка соответствует гипотезе. В этом случае мы получаем значение статистики критерия для «идеального соответствия».

     Например, если гипотеза про математическое ожидание, то $t = \bar{x} - \mathsf{E}\xi$ подходит под это требование. Если гипотеза про дисперсию, то соответствие правильнее измерять отношением и поэтому подошло бы $t = s^2/\mathsf{D}\xi$.
     Пример. Пусть $H_0 : \mathsf{E}\xi = a_0$; тогда $t = \bar{\mathbf{x}} - a_0$ и «идеальное значение» $t = 0$.
   - Распределение t при верной H_0 должно быть известно хотя бы асимптотически. Из-за этого часто преобразовывают вариант меры несоответствия, приведенный выше.
     Пример. См. раздел 2.1.
2. Строим разбиение области значений статистики критерия t так, что:
   - $\mathsf{P}(t \in A_\alpha^{\text{крит}}) = \alpha$.
   - Если альтернативная гипотеза H_1 не конкретизирована, то $A_\alpha^{(\text{крит})}$ следует выбрать так, чтобы она располагалась как можно дальше от идеального значения.
     Пример. В случае $t \sim \mathrm{N}(0,1)$ при идеальном значении 0, разумно определить $A_\alpha^{(\text{крит})}$ «на хвостах» графика $\mathrm{pdf}_{\mathrm{N}(0,1)}$ симметрично по обе стороны от 0 так, что для $A_\alpha^{(\text{крит})} =$

$(-\infty, -t_\alpha) \cup (t_\alpha, \infty)$

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

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 $A_\alpha^{(\text{крит})}$ выбирается так, чтобы максимизировать мощность критерия против альтернативы H_1, определения см. ниже.

1.3. Ошибки первого и второго рода

Определение (Ошибки I-го и II-го родов)**.** Пусть мы проверяем нулевую гипотезу H_0. Зафиксируем альтернативную гипотезу H_1 — такое отклонение от H_0, что его обнаружение важно для нас. Тогда

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

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

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

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

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

Определение. *Мощность* критерия против альтернативы есть

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

Иными словами, это способность критерия отличать H_1 от H_0.

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 могущей быть довольно большой. Поэтому гипотезу H_0 можно только отвергать или не отвергать, так как мы контролируем ошибку неправильного решения (отвергнуть). Принимать гипотезу нельзя, так как мы не контролируем, вообще говоря, ошибку неправильного решения в случае принятия гипотезы.

Замечание. В связи с введенным понятием мощности можно описать две проблемы:

- проблема маленьких объемов выборки состоит в том, что в этом случае мощность маленькая и критерий не заметит отличие от H_1 от H_0, т.е. с большой вероятностью не отвергнет H_0, хотя будет верна H_1;

- как ни странно, но есть также проблема слишком больших объемов выборки, когда мощность слишком большая, т.е. критерий может отвергнуть H_0 с большой вероятностью, даже если она «почти» верна (например, из-за ошибок округления).

Заметим, что вероятность ошибки первого рода α_I фиксирована (как минимум, она ограничена сверху выбранным значением α, в то время как $\alpha_{\mathrm{II}} = \alpha_{\mathrm{II}}(\alpha, n, H_1)$. Ниже продемонстрируем эту зависимость на примере.

Пример. Пусть $\xi \sim \mathrm{N}(a, \sigma^2)$, σ^2 известна — это модель. Гипотезы имеют вид $H_0 : a = a_0,\ H_1 : a = a_1$. Тогда

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

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

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

(дисперсия, конечно, не меняется при сдвиге).

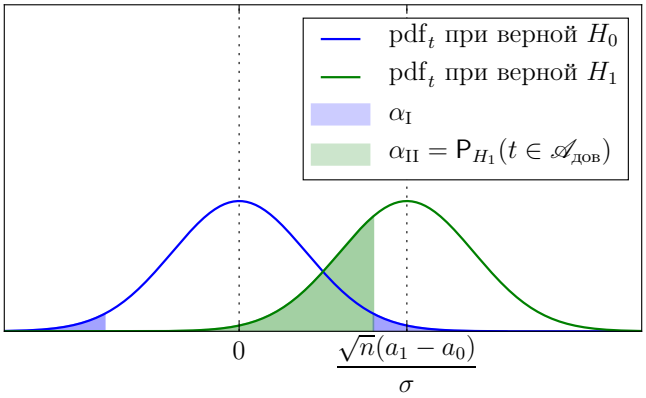

Рис. 1.1.: Плотности распределения t (неоптимальное разбиение)

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

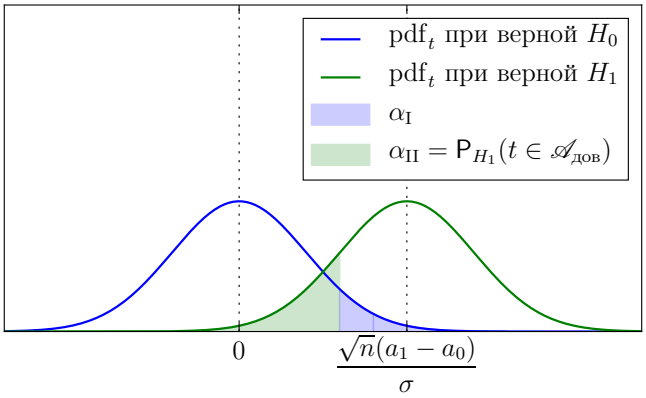

Рис. 1.2.: Плотности распределения t (оптимальное разбиение)

Таким образом, мы увидели, что если известна альтернатива, то можно выбирать критерий (= разбиение на доверительную и критическую области; статистика критерия — лишь удобной средство для этого), более мощный против именно этой альтернативы. И вообще, разные критерии для одной и той же основной гипотезы сравниваются по мощности против интересующей нас альтернативы.

Помимо этого, по рисунку видно, что ошибка второго рода α_{II} уменьшается при (1) увеличении ошибки первого рода (уровня значимости α) (поэтому мы ее выбираем максимальной из всех допустимых); (2) увеличении объема выборки (если α_{II} стремится к нулю при увеличении объема выборки, то критерий называется состоятельным после данной альтернативы) и (3) увеличении разницы между H_0 и H_1 в том же смысле, в каком статистика критерия измеряет разницу между

выборкой и нулевой гипотезой. Напомним, что мощность критерия против некоторой альтернативы (чего-то отличного от основной гипотезы) говорит нам о том, насколько хорошо критерий обнаруживает это отличие.

1.4. Понятие вероятностного уровня p-value.

Определение. *p-value* есть такой значение, что при значениях уровня значимости α, больших p-value, H_0 отвергается (по причине попадания t в $A_\alpha^{\text{крит}}$), а при меньших — не отвергается.

p-value - не вероятность, это пороговое значение. Неформально его можно интерпретировать как меру согласованности H_0 и выборки. Например, при больших значениях p-value практически при всех разумных уровнях значимости гипотеза не отвергается. При близких к нулю значениях p-value, наоборот, гипотеза будет отвергаться.

p-value — максимальное значение уровня значимости, при котором гипотеза не отвергается (значение статистики критерия попадает в доверит. область). Или, что эквивалентно, минимальное значение уровня значимости, при котором гипотеза отвергается.

Заметим, что для вычисления значения p-value даже для одной и той же статистики критерия нет единой формулы.

Неформальный комментарий, как находить p-value. Нужно сначала 1) нарисовать плотность статистики критерия, если верна H_0, и на ней сделать разметку — где там критическая область, где доверительная для некоторого уровня значимости α (α равняется вероятности попасть в критическую область, т.е. равно площади под соотв.частью графика плотности). Разбиение на доверительную и критическую область обычно делают так, чтобы в критическую область попадали значения, наиболее далекие от того значения статистики критерия, которое соответствует идеальному соотношению выборки и гипотезы. 2) нарисовать еще раз эту плотность и нанести значение статистики критерия.

Глядя на обе каринки, мысленно или рисуя, меняйте α и двигайте границу между доверительной и критической областями. Как упоминали выше, p-value — максимальное значение уровня значимости, при котором гипотеза не отвергается (значение статистики критерия попадает в доверит. область). Или, что эквивалентно, минимальное значение уровня значимости, при котором гипотеза отвергается. Замечу, что на основе этого определения сразу понятно, как, получив значение p-value, определить, при каких уровнях значимости (можно написать неравенство) гипотеза отвергается, а при каких — нет. Именно в таком виде и надо выдавать ответ.

На основе этих манипуляций должно стать понятно, что нужно для вычисления p-value — считать функцию распределения $F(t)$ или считать $1-F(t)$, умножать потом на два или нет. После того как поймете, можно уже находить p-value с помощью R или Python. Например, в R есть pnorm — функция распределения нормального распределения, pt — распределения Стъюдента, pf — распределение Фишера.

На всякий случай, замечание про уровень значимости α: уровень значимости задается заранее, еще до проверки гипотезы. Его смысл — максимальная вероятность ошибки (отвергнуть H0 неправильно), на которую согласен тот, кто будет отвечать за последствия ошибки. Но проблема в том, что те, кто делают обработку данных (в том числе, компьютеры) — это не те, которые отвечают за последствия, т.е. не те, кто устанавливают уровень значимости. Поэтому надо дать ответ в общем виде — в таком диапазоне уровней значимости гипотеза отвергается, а в таком - нет. Чтобы это сделать, все статистические пакеты выдают p-value. На его основе уже можно сформулировать общий ответ: при таких-то уровнях значимости гипотеза отвергается, при таких-то не отвергается. Иногда, когда некого спросить об уровне значимости, используются некоторые стандартные значения типа 0.05.

2. Проверка гипотезы о значении параметра (характеристики)

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

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

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

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

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

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

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

Доказательство.

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

□

2.1.2. $\mathsf{D}\xi$ неизвестна

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

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

Сходимость к нормальному распределению следует из модифицированной теоремы Леви (модифицированная ЦПТ), которая позволяет заменять дисперсию на ее состоятельную оценку с сохранением сходимости к тому же нормальному распределению.

Предложение. *При условии нормальности данных,*

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

2.1.3. Проверка гипотезы о мат.ож. в модели с одним параметром

Разница с общим случаем состоит в том, что в параметрической модели с одним параметром не нужно оценивать дисперсию. Так как все выражается через этот параметр, то имеем формулу для дисперсии через значение параметра, предполагаемое в нулевой гипотезе.

z-критерий для пропорции в модели Бернулли Пусть $\xi \sim \mathrm{Ber}(p)$. Поскольку $\mathsf{E}\xi = p$, можно воспользоваться только что введенной статистикой; учитывая $\mathsf{D}\xi = p(1-p)$, получаем статистику критерия для гипотезы $H_0 : p = p_0$:

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

z-критерий для интенсивности потока в модели Пуассона Пусть $\xi \sim \mathrm{Pois}(\lambda)$. Поскольку $\mathsf{E}\xi = \lambda$, можно воспользоваться только что введенной статистикой; учитывая $\mathsf{D}\xi = \lambda$, получаем статистику критерия для гипотезы $H_0 : \lambda = \lambda_0$:

$$t = \sqrt{n}\frac{\bar{x} - \lambda_0}{\sqrt{\lambda_0}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1).$$

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

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

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

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

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

2.2.2. $\mathsf{E}\xi$ неизвестно

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

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

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

Решение. Воспользуемся статистикой

$$\chi^2 = \frac{ns^2}{\sigma_0^2} = 101 \cdot 0.96 = 96.96.$$

«Идеальные» значения близки к $\mathsf{E}\xi_{\chi^2(100)} = 100$, так что определим критическую область на концах плотности:

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

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

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

⌟

2.3. Асимптотический критерий для гипотезы о значении параметра на основе MLE

Если умеем находить $\hat{\theta}_{\mathrm{MLE}}$, то 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,

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

и, учитывая асимптотическую эффективность ($\mathsf{D}\hat{\theta}_{\text{MLE}} I_n(\theta) \xrightarrow[n\to\infty]{} 1$), запишем статистику для $H_0 : \theta = \theta_0$:

$$t = \left(\hat{\theta}_{\text{MLE}} - \theta_0\right) \sqrt{I_n(\theta_0)} \xrightarrow{\text{d}} \text{N}(0,1).$$

Задание: построить критерий для гипотезы о значении параметра для распределений Бернулли и Пуассона.

3. Доверительные интервалы

3.1. Мотивация и определение

Точечные оценки не дают информации о том, насколько (количественно) настоящее значение далеко от оценки.

Определение. $[b_1, b_2]$ — *доверительный интервал* для параметра θ с уровнем доверия $\gamma \in [0,1]$, если $\forall \theta$

$$\mathsf{P}(\theta \in [b_1, b_2]) = \gamma, \quad \text{где } b_1 = b_1(\mathbf{x}), b_2 = b_2(\mathbf{x}),$$

т.е. границы доверительного интервала — это статистики (функции от выборки, случайные величины «до эксперимента»).

Замечание. Если выборка из дискретного распределения, то b_1, b_2 — тоже дискретны. Поэтому наперед заданную точность получить может не получиться; в таких случаях знак «=» заменяют «$\geq$». Аналогично с заменой на «$\xrightarrow[n\to\infty]{}$» для асимптотических доверительных интервалов, когда точные получить невозможно или трудоемко.

3.2. Доверительный интервал для проверки гипотезы о значении параметра

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

$$\mathsf{P}(\theta \in [b_1(\mathbf{x}), b_2(\mathbf{x})]) = \gamma = 1 - \alpha \implies \alpha = 1 - \mathsf{P}(\theta \in [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})]) = \mathsf{P}(\theta \notin [a_\gamma(\mathbf{x}), b_\gamma(\mathbf{x})]).$$

Соответственно,

$$\begin{cases} \text{отвергаем } H_0, \text{ если} & \theta_0 \notin [b_1(\mathbf{x}), b_2(\mathbf{x})] \\ \text{не отвергаем } H_0, \text{ если} & \theta_0 \in [b_1(\mathbf{x}), b_2(\mathbf{x})] \end{cases}.$$

Вероятность ошибки первого рода равна α, что соответствует определению критерия. Заметим, что здесь мы пользуемся общим определением критерия, а не частным случаем, когда критерий строится через статистику критерия.

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

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

3.3.1. Доверительный интервал для a

- Пусть σ^2 известно. Свяжем a с выборкой через статистику критерия $t = \sqrt{n}\frac{(\bar{x}-a)}{\sigma} \sim \mathrm{N}(0,1)$:

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

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

$$\mathsf{P}\left(a \in \left(\bar{x} \pm \frac{\sigma}{\sqrt{n}} c_\gamma\right)\right) = \gamma.$$

- Пусть σ^2 неизвестно. По аналогии,

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

и

$$\mathsf{P}\left(a \in \left(\bar{x} \pm \frac{\tilde{s}}{\sqrt{n}} c_\gamma\right)\right) = \gamma.$$

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

Решение.

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

⌟

3.3.2. 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

$$\mathsf{P}\left(c_1 < \frac{n s_a^2}{\sigma^2} < c_2\right) = \gamma \iff \mathsf{P}\left(\sigma^2 \in \left(\frac{n s_a^2}{c_2}, \frac{n s_a^2}{c_1}\right)\right) = \gamma.$$

- Пусть a неизвестно. Тогда аналогично

$$\mathsf{P}\left(\sigma^2 \in \left(\frac{n s^2}{c_2}, \frac{n s^2}{c_1}\right)\right) = \gamma,$$

где

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

3.4. 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 $\mathsf{E}\xi = a$, не задавая параметрическую модель. Пусть $\{x_i\}$ i.i.d., тогда

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

Если заменить σ на ее состоятельную оценку (s), то по модифицированной теореме Леви (будет у Владимира Викторовича в след.году) сходимость не испортится. Тогда

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

Все это так же, как было в разделе 2.1.2.

Аналогично разделу 2.1.3, доверительные интервалы можно улучшить (сделать большее точными, т.е. вероятность попадания в них ближе к γ при фиксированном n) в параметрической модели, если вместо независимой оценки дисперсии использовать оценку, полученную на основе оценок параметров. Разница в том, что там можно было использовать значение параметра, взятое из гипотезы, а в доверительных интервалах придется подставлять оценки (или решать нелинейные неравенства, см. следующий раздел).

3.5. Асимптотический доверительный интервал для параметра на основе MLE

В точности, как было при проверке гипотез,

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

Чтобы по аналогии с предыдущим выразить θ в $\mathsf{P}(c_1 < T < c_2) = \mathsf{P}(|T| < c_\gamma) = \gamma$, необходимо знать зависимость $I_n(\theta)$ от θ. Для Pois и Ber разрешение неравенства относительно θ эквивалентно решению неравенства для квадратичного полинома.

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

$$\mathsf{P}\left(-c_\gamma < \left(\hat\theta_{\text{MLE}} - \theta\right)\sqrt{I_n(\hat\theta_{\text{MLE}})} < c_\gamma\right) \to \gamma \iff \tag{3.1}$$

$$\mathsf{P}\left(\theta \in \left(\hat\theta_{\text{MLE}} \pm \frac{c_\gamma}{\sqrt{I_n(\hat\theta_{\text{MLE}})}}\right)\right) \to \gamma, \tag{3.2}$$

где

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

Пример. $\xi \sim \text{Pois}(\lambda)$. В разделе 5.3 получали: $\hat\lambda_{\text{MLE}} = \bar x$ и $I_n(\lambda) = n/\lambda$ и $I_n(\hat\lambda) = n/\bar x$, откуда

$$\mathsf{P}\left(\lambda \in \left(\bar x \pm c_\gamma \frac{\sqrt{\bar x}}{\sqrt{n}}\right)\right) \xrightarrow[n\to\infty]{} \gamma.$$

Замечание. Этот доверительный интервал не очень хорош, потому что может включать значения меньше 0.

Пример. $\xi \sim \text{Ber}(p)$. $p = \mathsf{E}\xi$. $\hat p = \bar x$, 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 обязательно принадлежит $[0,1]$.

Задание: Построить хорошие доверительные интервалы в моделях Бернулли и Пуассона, решив неравенства с квадратными уравнениями.

Задание: Построить доверительный интервал для параметра λ экспоненциального распределения $\text{Exp}(\lambda)$.

3.6. Использование SE для проверки гипотез и построения доверительных интервалов

Пусть оценка $\hat{\theta}_n$ имеет (асимптотически) нормальное распределение и является асимптотически несмещенной. Тогда доверительный интервал уровня γ для θ (т.е. такой интервал, в котором лежит γ всех значений величины) задается как

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

где c_γ — $(\gamma+1)/2$-квантиль стандартного нормального распределения. К примеру, для $\mathrm{N}(0,1)$ и 95%-квантили это был бы интервал$(-1.96; 1.96)$, а так нужно передвинуть его на среднее и растянуть на корень из дисперсии.

Но стандартное отклонение $\sqrt{\mathsf{D}\hat{\theta}_n}$ распределения $\hat{\theta}_n$ можно оценить как SE (standard error, стандартная ошибка). Значит, доверительный интервал будет иметь вид

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

Аналогично, статистика критерия $H_0 : \theta = \theta_0$ будет иметь вид

$$t = (\hat{\theta}_n - \theta_0)/\,\mathrm{SE}(\hat{\theta}_n),$$

которая имеет асимптотически нормальное распределение $\mathrm{N}(0,1)$ с ‘идеальным’ значение в нуле.

Заметим, что SE играет роль ‘сигмы’ распределения оценки.

4. Критерии проверки гипотезы о согласии с видом распределения

4.1. Критерий χ^2

По выборке возможно проверить гипотезу о виде распределения случайной величины, реализацией которой является выборка. Для проверки гипотезы согласия с видом произвольного *дискретного* распределения используется асимптотический *критерий* χ^2 *(«chi-squared test for goodness of fit»).*

4.1.1. Распределение с известными параметрами

Пусть

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

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

В качестве меры расхождения между эмпирическим и генеральным распределением рассматривается величина

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

откуда записывается статистика критерия

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

с ‘идеальным’ значением 0 (следовательно, критическая область только справа).

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

Определение. Критерий *применим*, если $\alpha_{\text{I}} = \alpha$ или $\alpha_{\text{I}} \approx \alpha$ с достаточной степенью точности.

Замечание. Поскольку критерий асимптотический, с достаточной (тому, кто дает такие рекомендации) 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.

Замечание. Почему бы не подстраховаться и не объединить состояния так, чтобы было >10? Ответ: теряем в мощности.
Задание Привести пример, демонстрирующий потерю мощности.

Пример (С монеткой)**.** Пусть $n = 4040$, $\#\text{H} = 2048$, $\#\text{T} = 1092$. Проверим $H_0 : \mathcal{P} = \text{Ber}(0.5)$ с $\alpha = 0.1$. Условия критерия выполняются, поэтому посчитаем

$$T = \frac{(2048 - 2020)^2}{2020} + \frac{(1092 - 2020)^2}{2020} = \frac{28^2 + 28^2}{2020} \approx 0.78,$$

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. Если нужно подстраховаться от подгонки (искусственно составленных под гипотезу выборок), то критическую область можно выбрать с двух сторон, слева и справа. Например, чтобы отверглась гипотеза о $p = 0.5$ для альтернирующей (и явно не случайной) последовательности $\mathbf{x} = (0, 1, 0, 1, \ldots)$ имеет $T = 0$. Однако, если мы не подозреваем данные в обмане, то так не делают.

4.1.2. Распределение с неизвестными параметрами

В случае сложной гипотезы $\mathcal{P} \in \{\mathcal{P}(\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\Theta}$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^\mathsf{T}$, следует найти оценку $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ 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,

$$T = \sum_{i=1}^{k} \frac{(\nu_i - np_i(\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}))^2}{np_i} \xrightarrow{\mathrm{d}} \chi^2(k - r - 1).$$

Важно: параметр можно считать известным, только если его значение выбрано без знания, какая получилась выборка.

Оценки по методу минимума хи-квадрат Предельное распределение статистики критерия не поменяется, если вместо оценки максимального правдоподобия подставить любую другую оценка с тем же предельным распределением. Рассмотрим оценки по минимуму хи-квадрат:

$$\boldsymbol{\theta}_{\mathrm{minChiSq}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{k} \frac{(\nu_i - np_i(\boldsymbol{\theta}))^2}{np_i} \xrightarrow{\mathrm{d}} \chi^2(k - r - 1).$$

Утверждение (без доказательства) заключается в том, что в условиях регулярности оценки, полученные по методу максимального правдоподобия, эквивалентны оценкам, полученным по методу минимума хи-квадрат. Таким образом, в критерии хи-квадрат можно использовать статистику в виде

$$T = \min_{\boldsymbol{\theta}} \sum_{i=1}^{k} \frac{(\nu_i - np_i(\boldsymbol{\theta}))^2}{np_i} \xrightarrow{\mathrm{d}} \chi^2(k - r - 1).$$

5. Критерий Колмогорова-Смирнова согласия с видом распределения

5.1. Произвольное абсолютно непрерывное распределение

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

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

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

где cdf_0 — функция распределения $\mathcal{P}_0$ нулевой гипотезы. Распределение D_n, оно разное для разных n, но не зависит от распределения.

Распределение не зависит от cdf_0, так как любое распределение можно привести, например, к равномерному, монотонным преобразованием: для любой cdf_0 верно $\mathrm{cdf}_0^{-1}(\xi) \sim \mathrm{U}(0,1)$, как будто, мы проверяем гипотезу $H_0 : \mathrm{cdf}_0^{-1}(\xi) \sim \mathrm{U}(0,1)$.

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

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

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

6. Визуальное определение согласия с распределением

6.1. P-P plot

Определение. *P-P plot* есть график

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

Пример. В R:

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

6.2. Q-Q plot

Определение. *Q-Q plot* есть график

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

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

Пример. В R:

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

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

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/

Часть IV.

Корреляционный анализ

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

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

1. Вероятностная независимость

1.1. Визуальное определение независимости

- Поскольку при $p_\eta(y_0) \neq 0$

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

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

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

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

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

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

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

  Пример. Таблица сопряженности похожей на независимую выборки:

$$\begin{array}{ccc} 1 & 3 & 2 \\ 2 & 5 & 3 \\ 9 & 20 & 11 \end{array}$$

1.2. Критерий независимости χ^2

По определению, для двумерных дискретных распределений, независимость есть

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

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

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

Следовательно,

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

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

Запишем статистику

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

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

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

Решение. Сводится к гипотезе о независимости, так как независимость эквивалентна равенству условных распределений. ⌟

Замечание. На маленьких выборках ($n < 50$ или $np_{ij} < 5$) возникают проблемы со сходимостью, потому что можно объединять только столбцы / строки и каждый раз терять сразу $S - 1$ ($K - 1$) степень свободы. В этих случаях используют критерием с перестановкой[1] или, в случае таблиц сопряженности 2×2, точным критерием Фишера.

Замечание. Критерий верен для количественных, порядковых и качественных признаков, потому что нигде не участвуют значения из выборки. Однако есть требование дискретности (конечного числа значений).

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

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

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

или коэффициент сопряженности Пирсона

$$\hat{p}^2 := \frac{\chi^2}{\chi^2 + n} = \frac{\hat{r}^2}{\hat{r}^2 + 1}.$$

(тогда 1 никогда не достигается).

Заметим, что $\hat{r}^2 := \frac{\chi^2}{n}$ является оценкой следующей меры зависимости (меры сопряженности) для двумерного дискретного распределения, задаваемого набором p_{ij}:

$$r^2 = \sum_{i=1}^{K}\sum_{j=1}^{S} \frac{(p_{ij} - p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}}$$

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

2. Линейная / нелинейная зависимость

2.1. Определение вида зависимости

Пусть теперь ξ, η — количественные признаки.

Напомним, условное математическое ожидание $\mathsf{E}(\eta \mid \xi)$ является такой функцией от ξ, на которой достигается минимум $\min_{\hat{\eta} \in \{\varphi(\xi)\}} \mathsf{E}(\eta - \hat{\eta})^2$.

Определение. Определим функцию условного математического ожидания

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

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

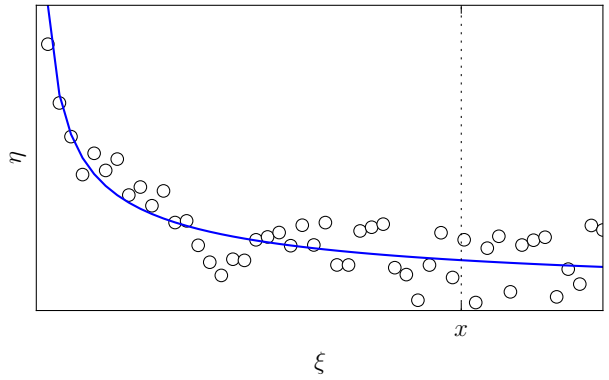

Рис. 2.1.: Нелинейная зависимость

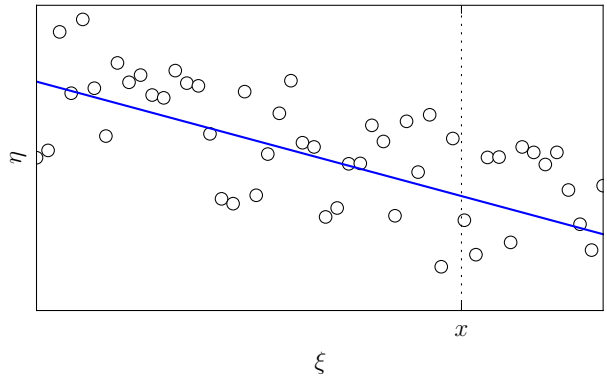

Рис. 2.2.: Линейная зависимость

2.2. Коэффициент корреляции Пирсона

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

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

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

Замечание (Важное).

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

Предложение. *Для линейно зависимых данных, конечно,* $\rho = \operatorname{sign} b$.

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

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

□

Предложение.

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

2.2.1. Оценка коэффициента корреляции

Оценка коэффициента корреляции строится стандартным методом подстановки в формулу для корреляции двумерного эмпирического распределения, в котором каждая пара значений $(x_i, y_i)^{\mathrm{T}}$, $i = 1\ldots, n$, имеет вероятность $1/n$.

В знаменателе стоят дисперсии, поэтому они в оценке просто заменяются на выборочные дисперсии.

Для оценки ковариации $\operatorname{cov}(\xi,\eta) = \mathsf{E}(\xi - \mathsf{E}\xi)(\eta - \mathsf{E}\eta) = \mathsf{E}\xi\eta - \mathsf{E}\xi\mathsf{E}\eta$ можно использовать два варианта (одной и той же) оценки, поэтому получим:

$$\hat\rho(\xi,\eta) = \frac{\sum_{i=1}^n (x_i-\bar x)(y_i-\bar y)}{s_x s_y} = \frac{\sum_{i=1}^n x_i y_i - \bar x\bar y}{s_x s_y}.$$

2.2.2. Значимость коэффициента корреляции

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

Пусть $(\xi,\eta)^{\mathsf{T}} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Тогда при $H_0 : \rho = 0$ статистика критерия имеет вид и распределение:

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

Идеальное значение — 0, критическая область двухсторонняя. Если предположения о нормальности $(\xi,\eta)^{\mathsf{T}}$ нет, а гипотеза не о некоррелированности, а о независимости, то критерий становится асимптотическим (т.е. им все равно можно пользоваться).

Проверим теперь гипотезу $H_0 : \rho = \rho_0$ (чаще проверяют $H_0 : \rho > \rho_0$). Тогда применяется z-преобразование Фишера

$$z = \frac12 \ln\frac{1+\rho}{1-\rho}, \quad z_0 = \frac12 \ln\frac{1+\rho_0}{1-\rho_0}.$$

Если $(\xi,\eta)^{\mathsf{T}} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

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

2.3. Происхождение и сравнение других мер зависимости

Если K — линейное пространство, то теорема Пифагора принимает вид

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

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

Откуда можно записать меру аппроксимации линейным пространством K как

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

Если $K = \mathcal{L} = \{a\xi + b\}$, то полученная величина является квадратом коэффициентом корреляции ρ^2:

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

ρ — коэффициент корреляции Пирсона.

Определение. *Множественный коэффициент корреляции* есть полученная величина для МНК с $K = \mathcal{M} = \left\{\sum_{i=1}^{k} b_i\xi_i + b_0\right\}$.

$$R^2(\eta, \xi_1, \ldots, \xi_k) := 1 - \frac{\min_{\hat\eta\in\mathcal{M}} \mathsf{E}\,(\eta - \hat\eta)^2}{\mathsf{D}\eta}.$$

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

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

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

Сравнивая полученные формулы, мы получаем, что коэффициент корреляции измеряет, насколько хорошо случайную величину η можно приблизить линейной функцией от ξ, а корреляционное отношение — произвольной (измеримой) функцией от ξ.

2.3.1. Свойства корреляционного отношения

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

2.3.2. Выборочное корреляционное отношение

По разложению дисперсии,

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

Перейдем на выборочный язык. Пусть дана выборка

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

Сгруппируем её:

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

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

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

на выборочном языке получаем (домножив на n):

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}_{\text{total sum of squares}} = \underbrace{\sum_{i=1}^{k} n_i\,(\bar{y}_i-\bar{y})^2}_{\text{межгрупповой разброс}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}_{\text{внутригрупповой разброс}}$$

$$ns_y^2 = ns_{y|x}^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2.$$

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

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

3. 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.

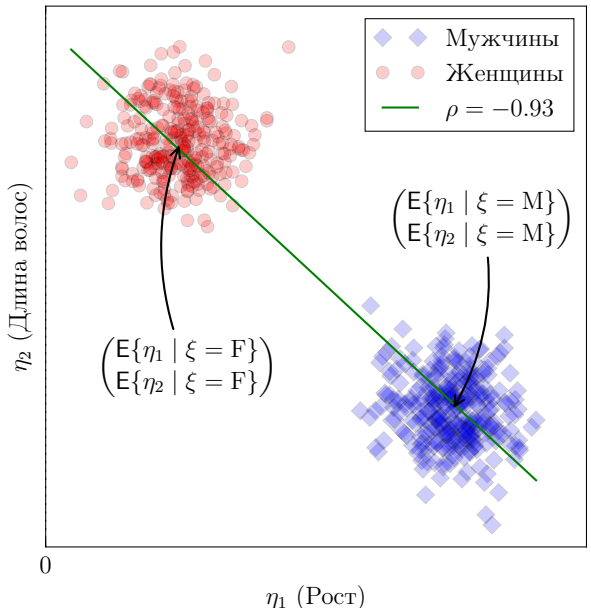

Рис. 3.1.: Исходные данные (бимодальность)

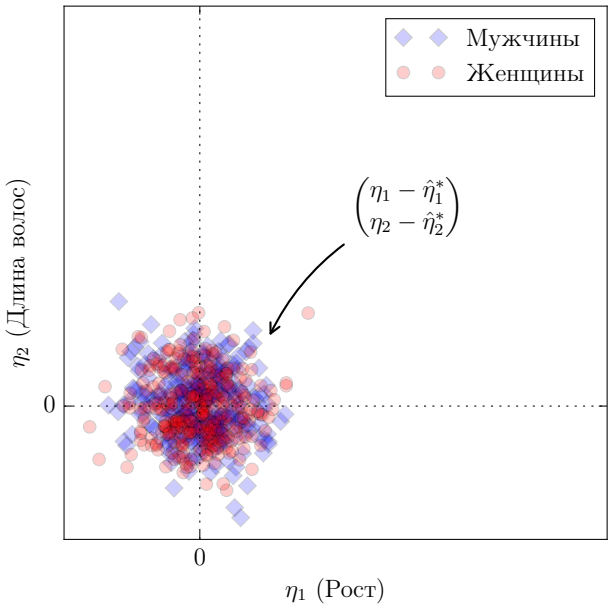

Рис. 3.2.: Центрированные данные

Пример. Возможна и ситуация как на (3.3), где определенно $\rho(\eta_1, \eta_2) > 0$, но $\rho(\eta_1, \eta_2 \mid \xi) < 0$.

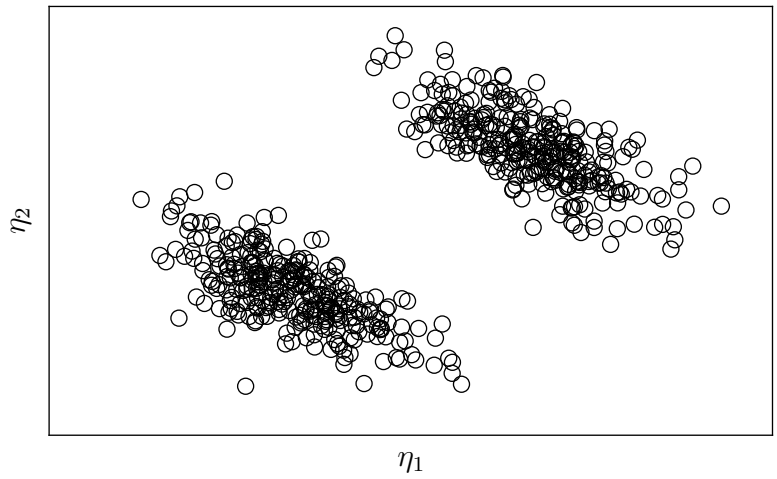

Рис. 3.3.: $\rho(\eta_1, \eta_2) > 0$, но $\rho(\eta_1, \eta_2 \mid \xi) < 0$

4. Зависимость между порядковыми признаками

Пусть признаки порядковые, т.е. их значения можно только сравнивать. Это означает, что нельзя читать, например, математическое ожидание, плотность и пр., но понятие функции распределения, основанное на сравнении, корректно.

Оценки значений функции распределения строятся по выборке на рангах случайных величин. Поэтому в случае порядковых признаков рассматриваются оценки, основанные на рангах. В частности, ранговые коэффициенты корреляции. Заметим, что ранговые характеристики хорошо работают на выборках без совпадающих наблюдений.

4.1. Ранговый коэффициент Спирмана

Определение. *Ранговый коэффициент Спирмана* есть

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

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

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

$$\mathrm{rk}\, x_{(i)} = i.$$

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

Можем ввести эмпирическое распределение

$$\mathrm{cdf}_{\xi_n}(x_i + 0) = \frac{\mathrm{rk}\, x_i}{n}, \quad \mathrm{cdf}_{\eta_n}(y_i + 0) = \frac{\mathrm{rk}\, y_i}{n} = \frac{T_i}{n}.$$

Тогда будет справедливо следующее

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

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

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 $|\rho_S|$, тем более явно выражена зависимость y от x в виде некоторой монотонной функции.

4.1.1. Согласованость ρ и ρ_S

Для количественных признаков, мера монотонной зависимости ρ_S не согласована с мерой линейной зависимости ρ в том же смысле, что ρ и мера функциональной зависимости $r_{\xi|\eta}$, где, в частности, $\rho \leq r_{\xi|\eta}$, а равенство достигается в случае линейной зависимости (линейного условного математического ожидания).

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

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

- С точностью до погрешности, по значению, ρ и ρ_S — это одно и то же (см. 4.1)

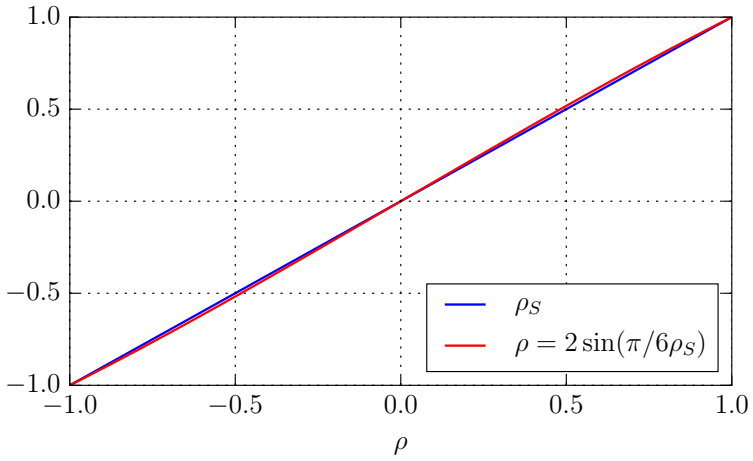

Рис. 4.1.: $\rho \approx \rho_S$

Значит, в случае нормального распределения можем сравнить оценки $\hat{\rho}$ и $\hat{\rho}_S$ между собой.

- Выборочную дисперсию оценок сравнить довольно сложно. Тем не менее, можем заметить, что $\hat{\rho}_S$ более устойчив к аутлаерам (см. 4.2). 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.

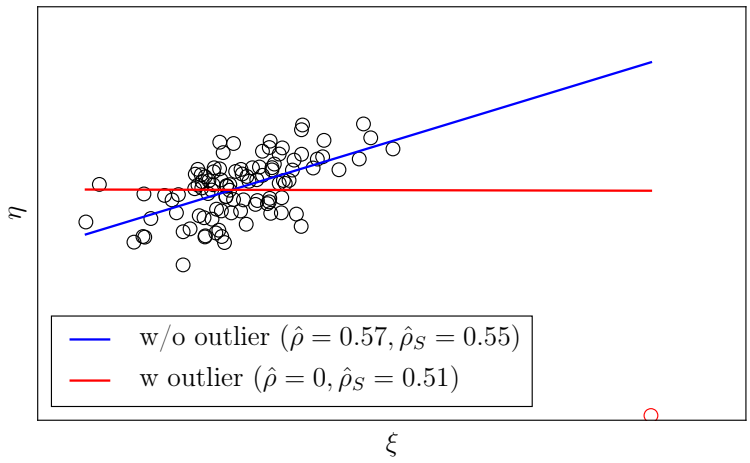

Рис. 4.2.: $\hat{\rho}$ до и после добавления аутлаера

- Монотонным преобразованием можем всегда сделать так, чтобы ρ изменился (например, возведя в квадрат); при монотонном преобразовании, однако, не меняется ρ_S (см. 4.3).

Значит, чтобы узнать ρ исходных (нормальных) данных, можно не выполнять обратного преобразования, а сразу посчитать ρ_S.

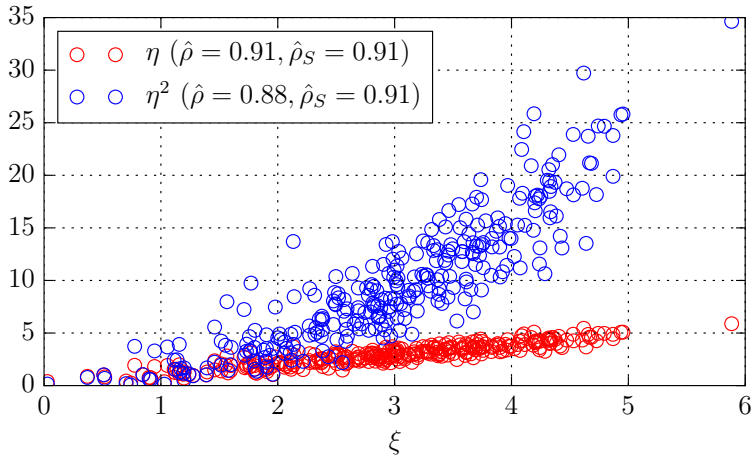

Рис. 4.3.: Монотонное преобразование нормальных данных

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

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

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

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

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

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

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

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

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

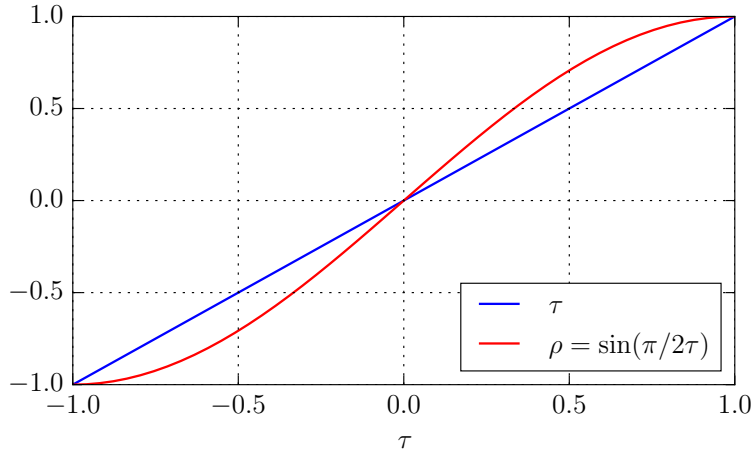

Рис. 4.4.: ρ и τ

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