Статистика

Конспект практических занятий

Мат-Мех, ПМИ, СМ–СМ, конспект семестров 5–8
Лекции Голяндиной Н.Э.

Первоначальный набор текста сделан Дмитрием Зотиковым
(NEG: UNDER CONSTRUCTION 01.11.2020)

СтатМод

Оглавление

[1] ANalysis Of VAriance

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.

Определение. *Характеристика* распределения — функционал от распределения:

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

Чаще всего, $D = \mathbb{R}$.

Определение. *Оценка* $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ *характеристики* θ — функция от выборки, не зависящая от этой характеристики.

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)**.**
Вопрос: $\mathsf{E}\xi =$? Ответ $\widehat{\mathsf{E}}\xi = \bar{x}$.
Вопрос: насколько точна оценка? Ответ: $\mathsf{D}\bar{x} = \frac{\mathsf{D}\xi}{n}$.
Вопрос: а как по выборке оценить эту точность оценки? Ответ: $\widehat{\mathsf{D}\bar{x}} = \frac{s^2}{n}$.
Получили оценку дисперсии оценки математического ожидания.

Мера разброса для качественных признаков.

Определение (Энтропия)**.** Для дискретного распределения

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

энтропией (мера 'размазанности' распределения) вычисляется как :

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

Замечание. Минимальная мера разброса, если .... В случае равномерного дискретного распределения энтропия максимальна.

SE и SD

Определение. *Выборочное стандартное отклонение* есть

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

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

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

Определение. *Стандартная ошибка* оценки есть

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

Это показатель разброса оценки случайной величины.

Замечание. В частном случае $\theta = \mathsf{E}\xi$, $\hat{\theta} = \bar{x}$ 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»[1])

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

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’.

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.

[1] «Skewness».

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 $\xi_n = \hat{\xi}_n$ — 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

Полезное свойство для любой случайной величины ζ:

$$\mathsf{E}(\zeta - a)^2 = \mathsf{D}\zeta + (\mathsf{E}\zeta - a)^2. \tag{6.1}$$

Доказывается легко: $\mathsf{E}(\zeta - a)^2 = \mathsf{E}((\zeta - \mathsf{E}\zeta) + (\mathsf{E}\zeta - a))^2 = ...$

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. В (6.1) берем $\zeta = \hat{\theta}_n$, $a = \theta$ и получим

$$\underbrace{\mathsf{E}(\hat{\theta}_n - \theta)^2}_{\text{MSE}} = \mathsf{D}\hat{\theta}_n + \underbrace{(\mathsf{E}(\hat{\theta}_n - \theta))^2}_{\text{bias}^2}. \tag{6.2}$$

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 $\theta = \mathsf{E}\xi$, $\hat{\theta}_n = \hat{\mathsf{E}}\xi_n = \bar{x}$. 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. В (6.1) берем $\zeta = \hat{\xi}_n$, $a = \mathsf{E}\xi$ и выразим дисперсию; получим для $s^2 = \hat{\mathsf{D}}\xi_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$:

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

.

Берем мат.ож. и получаем (так как $\mathsf{E}\bar{x} = \mathsf{E}\xi$):

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

□

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

Определение. *Исправленная* дисперсия:

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

Очевидно, исправленная дисперсия — несмещенная оценка дисперсии.

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 (1.1), 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 Чебышёва,

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

□

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

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:*

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

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 $\operatorname{med}\xi = \mathsf{E}\xi$ и имеет смысл сравнить две оценки одной и той же характеристики, выборочное среднее и выборочную медиану.

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

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

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

откуда

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

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

Замечание. В то же время, $\widehat{\operatorname{med}\xi}$ более устойчива к аутлаерам, чем $\bar{x}$, и этим лучше. Это легко увидеть, устремив одно из значений к бесконечности. Выборочное среднее тоже устремится к бесконечности, а выборочная медиана либо не изменится, либо почти не изменится.

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 ($\eta \sim \chi^2(m)$), если

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

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

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

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

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

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

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

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

(Вообще, справедливость $\mathbf{BC} = 0$ не удивительна в случае, когда $\mathbf{B}$ — матрица проектора на линейное подпространство, а $\mathbf{C} = \mathbf{I}_n - \mathbf{B}$ — матрица проектора на ортогональное дополнение. В данном примере, это проектор на подпространство, натянутое на вектор из единиц.)

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).

Проблема: случайно может произойти что угодно, т.е. безошибочных решений практически не бывает. Приходится задавать максимальный уровень вероятности ошибки, на который можно согласиться при принятии решения.

Задаем уровень значимости (significance level) $0 < \alpha < 1$.

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 $\alpha_I = \mathsf{P}_{H_0}(\mathbf{x} \in \mathcal{A}_\alpha^{(\text{крит})}) = \alpha$ (здесь P_{H_0} — вероятность (мера), 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, а 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 $A_{\text{крит}}^{(\alpha)}$ выбирается так, чтобы $\alpha_I = \mathsf{P}_{H_0}(t \in A_{\text{крит}}^{(\alpha)}) = \alpha$. Заметим, что мы поменяли обозначение для областей, потому что это разбиение значений статистики критерия, а ранее это было разбиение множества значений выборки.

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. Для $H_0 :, \mathsf{E}\xi = a_0$ в модели $\xi \sim N(a, \sigma^2)$ с известной дисперсией σ^2 удобно использовать статистику критерия $t = \sqrt{n}(\bar{\mathbf{x}} - a_0)/\sigma \sim N(0, 1)$.

     Пример. Еще примеры см. в разделе 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 (см. про нее в след.разделе) 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 $A^{(\text{крит})}_{\alpha}$ «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 $\alpha_{\text{II}} = \alpha_{\text{II}}(\alpha, n, H_1)$. 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, $\mathsf{E}\bar{\mathbf{x}} = 1/n \cdot \sum_{i=1}^{n} \mathsf{E}x_i = n/n \cdot a_1$ 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).

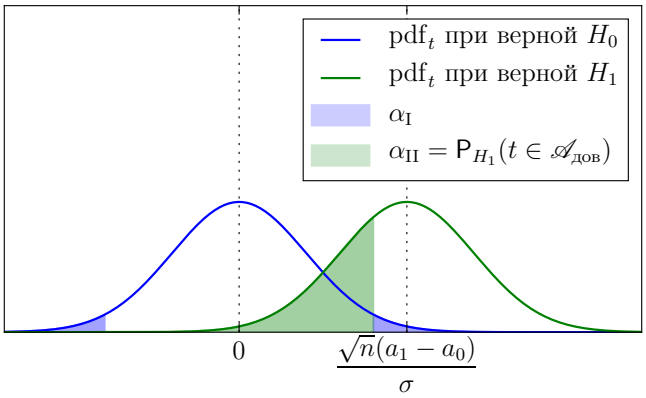


Рис. 1.1.: Плотности распределения t (неоптимальное разбиение)

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

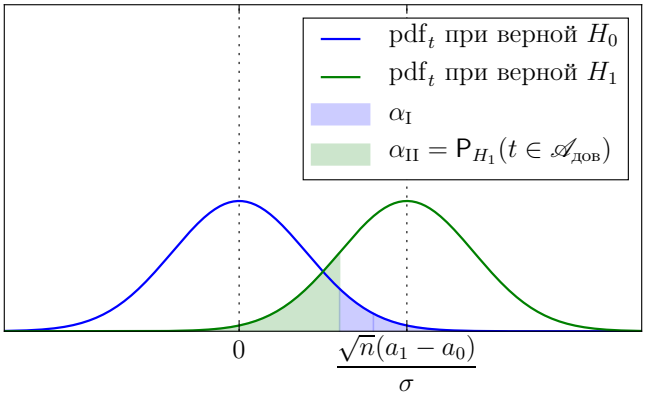


Рис. 1.2.: Плотности распределения t (оптимальное разбиение)

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^2/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:

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

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

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

то

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

⌟

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 $n = 4040$, $\#\text{H} = 2048$, $\#\text{T} = 1992$. Проверим $H_0 : \mathcal{P} = \text{Ber}(0.5)$ с $\alpha = 0.1$. Условия критерия выполняются, поэтому посчитаем

$$T = \frac{(2048 - 2020)^2}{2020} + \frac{(1992 - 2020)^2}{2020} = \frac{28^2 + 28^2}{2020} \approx 0.78,$$

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 с $p = 0.5$ для альтернирующей (и явно не случайной) последовательности $\mathbf{x} = (0, 1, 0, 1, \ldots)$, когда $T = 0$. 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: $H_1 : \xi \not\sim \mathcal{P}_0$; $A^{(\alpha)}_{\text{крит}} = (\mathrm{qnt}_{\text{K-S}}(1-\alpha), \infty)$.

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/

7. Гипотеза о равенстве распределений

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

Возможно рассматривать два случая:

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

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

или

ξ	x_1	$\ldots$	x_{n_1}	y_1	$\ldots$	y_{n_2}

то есть одному признаку сопоставлено $n_1 + n_2$ индивидов.

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

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

или

ξ	x_1	$\ldots$	x_n
η	y_1	$\ldots$	y_n

то есть по строчкам стоят признаки, по столбцам — индивиды.

Такие тесты называются парными.

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

Замечание. Непараметрические критерии хороши тем, что основаны на рангах, значит устойчивы к аутлаерам; плохи тем, что не используют всю информацию о значении — только порядок, из-за чего обладают меньшей мощностью.
Также непараметрические аналоги параметрических тестов могут проверять несколько другую гипотезу (точнее — быть мощными против других альтернатив).

8. Равенство математических ожиданий для независимых выборок

8.1. Двухвыборочный t-критерий

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

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

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

Пусть выборки *независимы*[1], $(x_1, \ldots, x_{n_1}), (y_1, \ldots, y_{n_2})$, $n = n_1 + n_2$ (на самом деле, нужно говорить про независимость ξ_1 и ξ_2). Значит $\mathsf{D}(\bar{x} - \bar{y}) = \mathsf{D}\bar{x} + \mathsf{D}\bar{y}$.

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

Предложение. *Если дисперсия известна,*

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

откуда

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

Если данные нормальные, то

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

Предложение. *Если дисперсия неизвестна,*

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

Если данные нормальные, то

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

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

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

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

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

[1] Случай зависимой выборки рассматривается в другом параграфе.

Разбиение

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

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

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

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

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

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

Если данные нормальные, то

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

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

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

Замечание. Точное распределение неизвестно, примерно равно t с дробным числом степеней свободы (что вычисляется интерполяцией по соседним степеням). Всегда ожидается, что если данные нормальны, то распределение известно. Это противоречие носит название *проблемы Беренса-Фишера*[2].

Разбиение

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

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

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

8.2. Непараметрический t-критерий

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

Пусть, как и прежде, дана выборка $(\mathbf{x}, \mathbf{y})$. Следующие два критерия — Wilcoxon и Mann-Whitney — проверяют гипотезу $H_0 : \mathsf{P}(\xi_1 > \xi_2) = \mathsf{P}(\xi_1 < \xi_2)$ или, альтернативно, $H_0 : \mathcal{P}_{\xi_1} = \mathcal{P}_{\xi_2}$ против $H_1 : \mathcal{P}_{\xi_1} \neq \mathcal{P}_{\xi_2}$ (что выборки получены из одной «генеральной совокупности») в случае абсолютно непрерывных распределений.

[2] Behrens-Fisher problem.

8.3. Критерии суммы рангов Wilcoxon

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

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

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

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

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

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

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

Иногда в качестве статистики берут количество инверсий в объединенной выборке.

8.4. Критерий Mann-Whitney (U test)

Используется статистика

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

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

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

Асимптотически,

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

но для малых объемов выборки можно посчитать и точные распределения.

Замечание. Критерий состоятельный против альтернативы

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

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

Замечание. Критерии Манна-Уитни и Вилкоксона *эквивалентны* — в том смысле, что выделяют один и тот же p-value.

8.5. Критерий серий (runs)

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

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

8.6. Двухвыборочный тест Колмогорова–Смирнова

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

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

9. Равенство математических ожиданий для парных (зависимых) выборок

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

9.1. t-критерий

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

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

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

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

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

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

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

9.2. Непараметрический тест знаков (Sign test)

$H_0 : \mathsf{P}(\xi_1 < \xi_2) = \mathsf{P}(\xi_1 > \xi_2)$. Используется статистика
$$W = \sum_{i=1}^n \psi_i, \quad \psi_i = \begin{cases} 1 & x_i > y_i \\ 0 & x_i < y_i. \end{cases}$$

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

Пусть после удаления всех пар, таких, что $x_i = y_i$, объем выборки стал равен m. Тогда

$$W \sim \mathrm{Bin}(m, 0.5)$$

и для построения разбиения можно пользоваться $\mathrm{qnt}_{\mathrm{Bin}(m,0.5)}$.

Замечание. Критерий применим к порядковым признакам.

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

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

Увеличить мощность предыдущего критерия можно, учтя больше информации:

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

Для симметрии можно рассмотреть статистику

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

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

Замечание. Критерий уже не применим к порядковым признакам.

10. Равенство дисперсии для двух распределений

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

10.1. Критерий Фишера

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

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

Замечание. При отклонении от нормальности не становится асимптотическим.

10.2. Критерий Левена (Levene's test)

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

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

10.3. Критерий Brown–Forsythe

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

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

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.

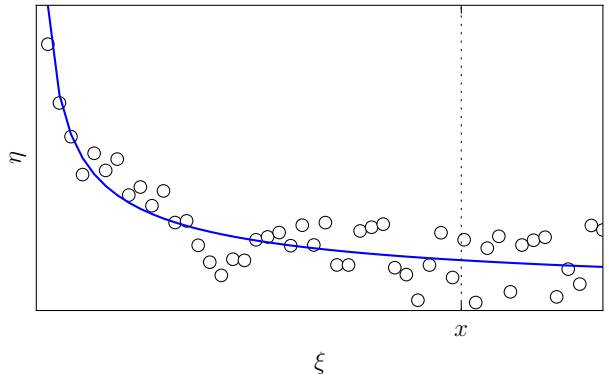


Рис. 2.1.: Нелинейная зависимость

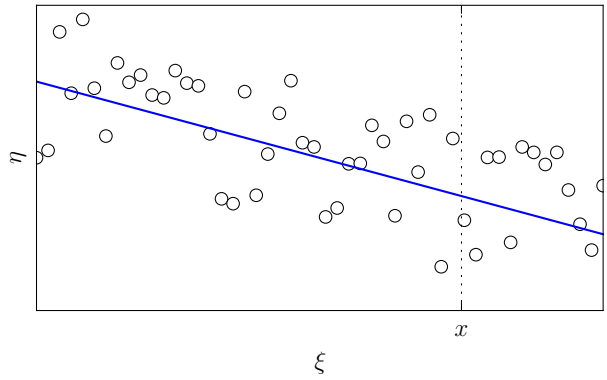


Рис. 2.2.: Линейная зависимость

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

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

□

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:

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

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

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

Этот критерий асимптотический даже в нормальной модели.

2.3. Метод наименьших квадратов (Ordinary Least Squares)

Пусть $\eta, \xi \in L^2(\mathcal{F}, \mathsf{P})$ пространству $\mathcal{F}$-измеримых по мере P функций с нулевым мат.ожиданием, конечным вторым моментом и скалярным произведением $(\eta,\xi) = \mathsf{E}\eta\xi$. По свойству УМО, случайная величина

$$\hat\eta^* = \mathsf{E}\left(\eta \mid \xi\right)$$

будет ортогональной проекцией η на K, где $K = \{\phi(\xi), \phi \text{ измерима}\}$, т.е. $(\eta - \hat\eta^*, \hat\eta) = 0\ \forall\hat\eta \in K$. Значит, она минимизирует квадрат нормы расстояния от η до K:

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

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

2.4. Происхождение и сравнение мер зависимости разного типа

Если K — линейное пространство, то теорема Пифагора принимает вид

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

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

Откуда можно записать меру аппроксимации линейным пространством K как

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

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

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

ρ — коэффициент корреляции Пирсона.

(Это будет доказано позже, в теме про парную регрессию.)

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

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

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 (в смысле метода наименьших квадратов).

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

2.4.2. 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

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

Пусть ξ — дискретная случайная величина со значениями $(x_1^*, \dots, x_k^*)$. Переобозначим элементы выборки:

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

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, $r_{\eta|\xi}^2 = \mathsf{DE}(\eta \mid \xi)/\mathsf{D}\eta$,

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

2.5. Множественный коэффициент корреляции

Введенные выше меры обобщаются естественным образом и на случай зависимости не от одной случайной величины, а сразу от нескольких.

Определение. *Множественный коэффициент корреляции* определяется на основе метода МНК с $K = \left\{\sum_{i=1}^{k} b_i\xi_i + b_0\right\}$.

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

Замечание. $R^2 \geq \rho^2$; если же $R^2(\eta, \xi_1, \dots, \xi_k) = \rho^2(\eta, \xi_1)$, то либо через ξ_1 линейно выражаются все остальные ξ_i (но это вырожденный случай), либо η и ξ_i, $i = 2, \dots, k$, независимы.

2.6. Приложение. Свойства условного математического ожидания

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

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

3. Частная корреляция

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

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

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.

Частная корреляция есть, по сути, корреляция между остатками от линейной регрессии (т.е. вычитается наилучшее линейное приближение по МНК).

3.1. Примеры, когда ξ имеет два состояния

Рассмотрим пример, когда вычитается признак с двумя значениями. В этом случае функцию, задающую наилучшее приближение по МНК, всегда можно считать линейной функцией (так как через две точки всегда можно провести прямую). Поэтому вычитание наилучшего линейного приближения можно заменить на вычитание условного математического ожидания.

Пример. Это пример, в котором обнаружилась отрицательная корреляция между ростом и длиной волос.

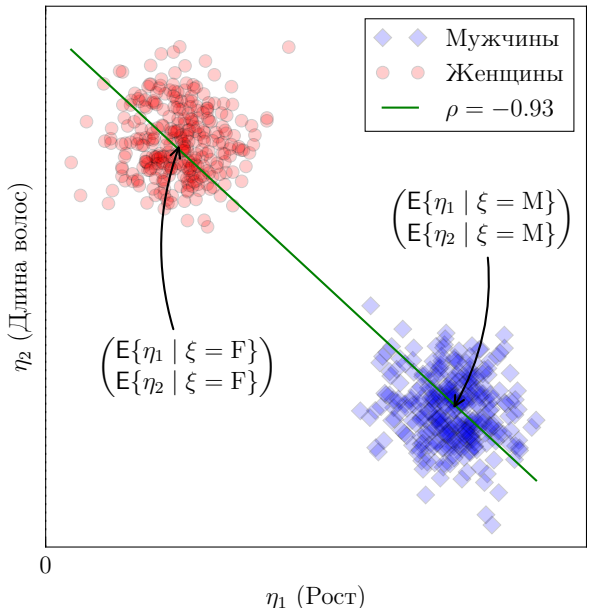


Рис. 3.1.: Исходные данные

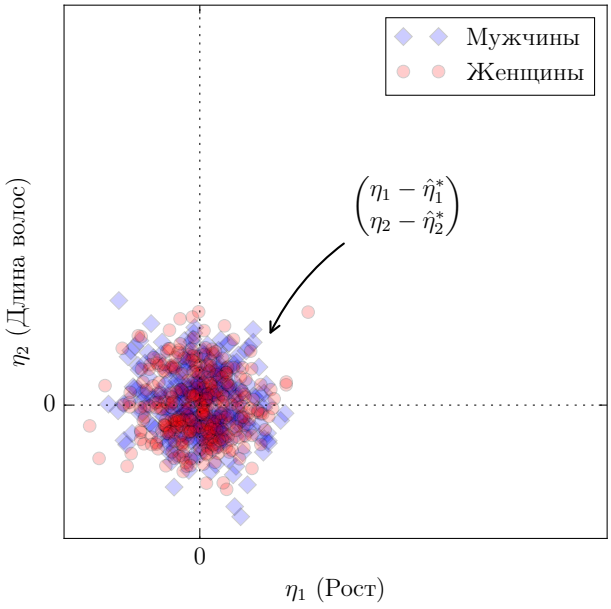


Рис. 3.2.: Данные после вычитания условного матем.ожидания; примерно нулевая корреляция

Пример. Возможна и ситуация как на (3.3), где, скорее всего (а если раздвинуть облака, то определенно) $\rho(\eta_1, \eta_2) > 0$, но $\rho(\eta_1, \eta_2 \mid \xi) < 0$.

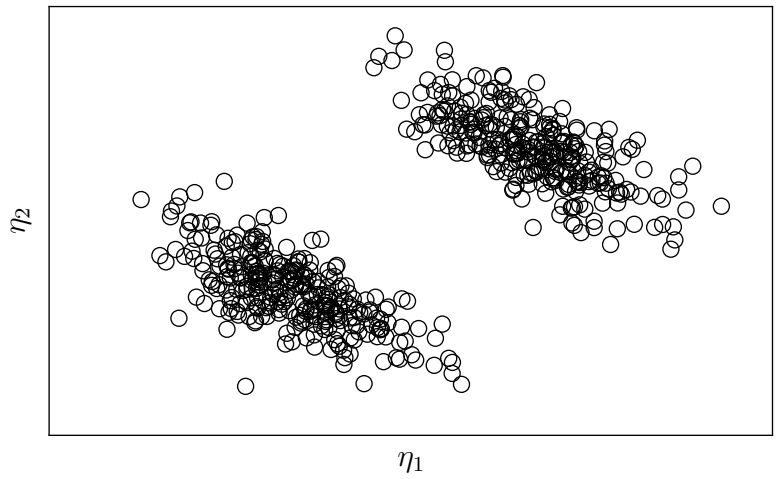


Рис. 3.3.: $\rho(\eta_1, \eta_2) > 0$, но $\rho(\eta_1, \eta_2 \mid \xi) < 0$

Заметим, что если корреляции внутри групп разные, то результат будет бессмысленным.

3.2. Более двух значений у ξ

Для числа групп, большего, чем 2, все происходит примерно так же, но со след. замечанием. Здесь уже наилучшее функциональное приближение, вообще говоря, нелинейное. Поэтому вычитание условных мат.ожиданий может соответствовать частному корреляционному отношению, но не частной корреляции.

Если же центры групп лежат примерно на одной прямой, то подход через вычитание условного мат.ожидания (центров групп) по-прежнему работает. Заметим, что группы могут пересекаться, сливаться и пр., т.е. не обязаны быть раздельными, как в предыдущем примере.

Однако, наиболее общий случай, когда ξ — количественный признак и тогда никаких групп, вообще говоря, нет. В этом случае можно мысленно разбить значения ξ на градации и все равно рассуждать в терминах групп.

Интерпретация Интерпретируем частную корреляцию мы так: какая была бы зависимость, если бы рассмотреть только объекты с зафиксированным значением признака ξ. Результат интерпретации будет осмысленным, если это действие (фиксирование значения ξ и одновременное изменение η_i) осмысленно.

3.3. Пример анализа данных CARDATA

Давайте ответим на вопрос, почему цена не зависит от других характеристик

1. Смотрим на график pairs.
2. Видим много распределений с хвостом вправо. Логарифмируем данные.
   a) Длинные хвосты (вправо) укорачиваются
   b) Можно больше доверять визуальному восприятию выбросов

c) Если признак lognormal, то log(lognormal) = normal и нелинейные зависимости становятся линейными.

d) Ряд критериев становятся точными (те, что строятся в нормальной модели)

3. Убираем аутлаеров в логданных. (Мерседесы?)

4. Ищем неоднородности в данных. По качественному признаку (origin) нельзя считать частные корреляции, поэтому строим скаттерплоты, раскрашенные согласно значению origin. Видим, что американские данные отличаются от европейских + японских. Можем построить correlation matrix отдельно для каждого значения origin. Там видно, что для японских машин корреляции значительно увеличились. Однако, для американских этого не произошло.

5. Смотрим раскрашенные скаттерплоты по году (например, цена против мощности) и, эврика, видим, что год размывает отчетливую тенденцию. Все облако горизонтальное, с нулевой корреляцией, но состоит из сдвинутых в противоположном направлении облаков с положительной корреляцией.

6. Это означает в точности то, что частная корреляция за вычетом влияния года будет положительной, как и должно быть по смыслу. Напомним, что интерпретация частной корреляции как раз такая и есть — какая будет корреляция внутри выборки, если фиксируем значение того признака(ов), влияние которых вычитаем.

7. Можно попробовать раскрасить скаттерплот одновременно по году и origin, но там мало индивидов будет в группах, не знаю, что там останется.

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)

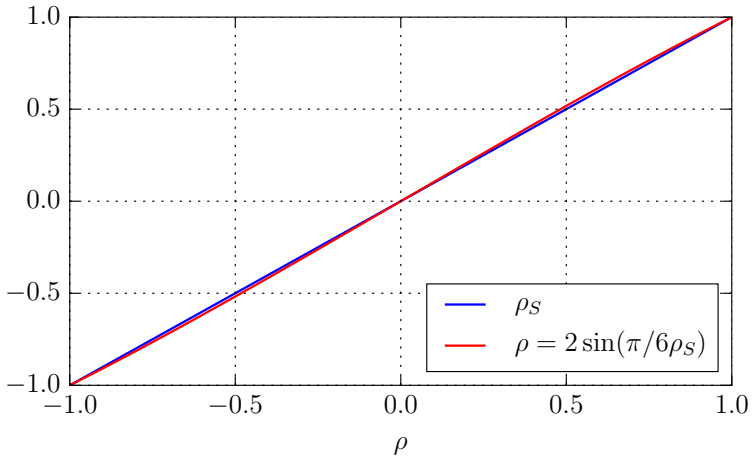


Рис. 4.1.: $\rho \approx \rho_S$

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.

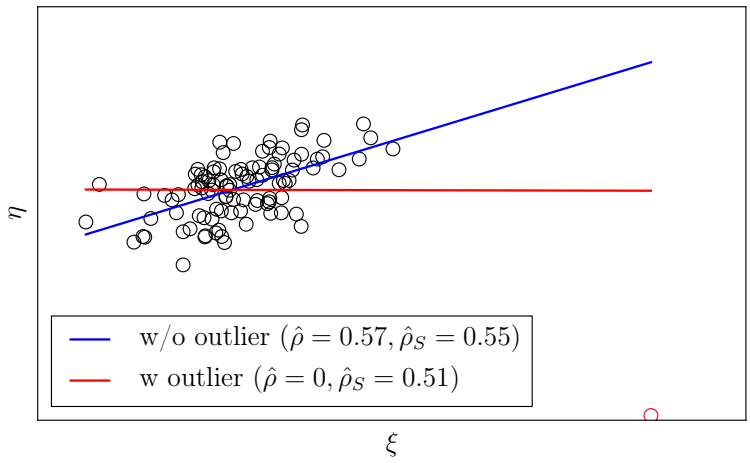


Рис. 4.2.: $\hat{\rho}$ до и после добавления аутлаера

- 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.

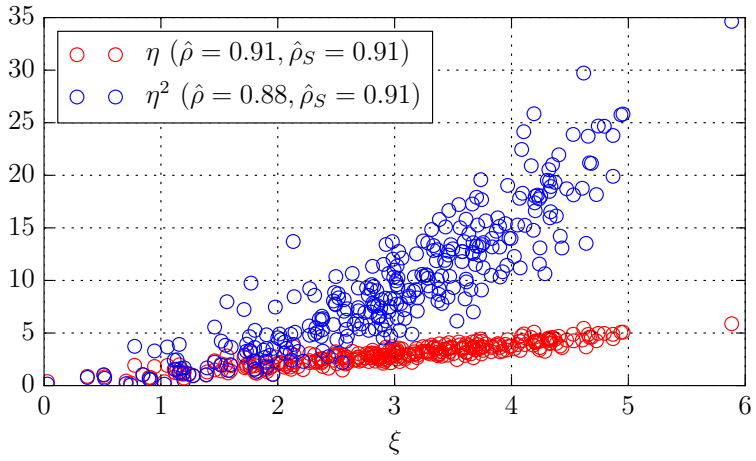


Рис. 4.3.: Монотонное преобразование нормальных данных

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).

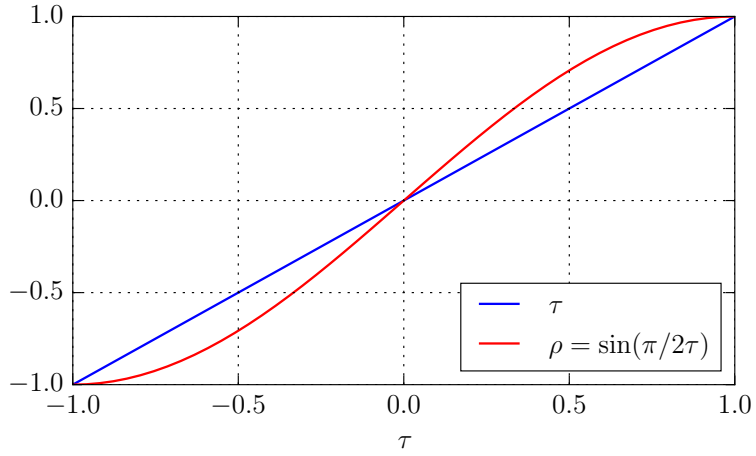


Рис. 4.4.: ρ и τ

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.

Часть V.

Дисперсионный анализ

1. Однофакторный дисперсионный анализ (One-way ANOVA[1])

Задача может быть поставлена двумя эквивалентными образами:

1. Пусть $\eta_i \sim \mathcal{P}_i,\ i \in 1:k$. Проверить гипотезу, что все распределения равны:
$$H_0 : \mathcal{P}_1 = \cdots = \mathcal{P}_k.$$
2. Пусть дан двумерный вектор $\begin{pmatrix}\xi & \eta\end{pmatrix}^\mathsf{T}$, причем ξ («фактор») принимает k значений $A_1, \ldots, A_k$. Рассмотрим $\eta_i \sim \mathcal{P}_i = \mathcal{P}_{\eta|\xi=A_i}$. Проверить гипотезу
$$H_0 : \mathcal{P}_{\eta|\xi=A_1} = \cdots = \mathcal{P}_{\eta|\xi=A_k}.$$

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

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

Для построения критерия, вспомним разложение дисперсии на выборочном языке (раздел 2.4.2, y_{ij} — j-й элемент из i-й группы):

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

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

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

и

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

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

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

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

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

[1] ANalysis Of VAriance

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

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

Пример. Пусть дана выборка вида $\{(\xi = \text{пол}, \eta = \text{вес})\}$. Выдвинем H_0 : вес не зависит от пола. Очевидно, что ξ — категориальная случайная величина, а η — количественная. Значения ξ разобьют всю выборку на две группы. Тогда проверка гипотезы сведется к проверке равенства распределений в двух группах, $\mathcal{P}_{\eta|\xi=g_1} = \mathcal{P}_{\eta|\xi=g_2}$. В предположении, что $(\eta \mid \xi = g_i) \sim \mathrm{N}(\mu_i, \sigma^2)$, равенство распределений будет следовать из равенства математических ожиданий.

2. Множественные сравнения

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

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

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

	Retain H_0 (отличие от H_0 не значимо)	Reject H_0 (отличие от H_0 значимо)
True H_0	# True Negative	# False Positive (False Discovery)
False H_0	# False Negative	# True Positive (True Discovery)

Используя обозначения таблички,

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

Определение. Family-wise error rate (FWER):

$$\mathrm{FWER} = \mathsf{P}(\text{хотя бы один раз отвергнута верная гипотеза}).$$

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

Требуется контролировать FWER на предзаданном уровне α, т.е. чтобы $\mathrm{FWER} \sim \alpha$, где $\sim \in \{=, \leq, \to\}$.

Выше мы предполагали, что все гипотезы верны и мы везде ошиблись. В *сильном* смысле контроль FWER на уровне α должен гарантироваться для *любой* конфигурации верных и не верных $H_0^{(j)}$.

Определение. Пусть $I := \left\{ i : H_0^{(i)} \text{ верна} \right\}$. Тогда

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

Определение.

$$\text{strong}\,\mathrm{FWER} = \max_{I : I \subset \{1,\dots,m\}} \mathrm{FWER}_I.$$

Это осуществляется двумя процедурами:

- Single
- Stepdown

2.1. Single

Каждая $H^{(i)}$ проверяется отдельно с уровнем значимости α_1. Задача сводится к тому, чтобы найти такое α_1, что $\mathrm{FWER} \leq \alpha$ для какого-то нужного предзаданного α. Пусть $T = 1 : m$, т.е. будто все тесты верны; тогда

$$\mathrm{FWER}_{\{1,\dots,m\}} = \mathsf{P}\left(\bigcup_{i=1}^{m}\{H_0^{(i)} \text{ отв}\}\right) \leq \sum_{i=1}^{m}\mathsf{P}(H_0^{(i)} \text{ отв}) = m\alpha_1 = \alpha \implies \alpha_1 := \frac{\alpha}{m}.$$

Замечание. Из-за неравенства тест консервативный, т.е. $\mathrm{FWER} \leq \alpha$. Значит не максимально мощный.

Вопрос: когда тест максимально консервативный?

Утверждение. $\operatorname{strong}\mathrm{FWER} \leq \alpha$

Доказательство.

$$\begin{aligned}\operatorname{strong}\mathrm{FWER} &= \max_{I\subset\{1,\dots,m\}}\mathsf{P}(H_0^{(i)} \text{ отвергается},\ i \in I)\\ &\leq \sum_{i\in I}\mathsf{P}(H_0^{(i)} \text{ отвергается}) = |\{i : i \in I\}|\,\alpha_1 \leq \alpha.\end{aligned}$$

□

Определение. Поправка Бонферрони

$$\alpha_1 = \frac{\alpha}{m}.$$

Тест нужно проверять не с α_1, а с α/m. Так критерий будет консервативным (иначе — радикальным, что хуже).

Определение. Поправка Бонферрони для p-value:

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

2.2. Stepdown (Holm's algorithm)

Алгоритм Хольма/Холма (Holm algorithm, или Holm-Bonferroni algorithm):

1. Пусть $H_1, ..., H_m$ — семейство нулевых гипотез и $p_1, ..., p_m$ — соответствующие вероятностные уровни (p-values).
2. Отсортируем их по возрастанию: $p_{(1)} \leq \ldots \leq p_{(m)}$, соответствующие отсортированным уровням нулевые гипотезы переобозначим как $H_{(1)} \ldots H_{(m)}$.
3. Для заданного уровня значимости α пусть k — минимальный индекс такой, что $p_{(k)} > \frac{\alpha}{m+1-k}$.
4. Отвергаем все нулевые гипотезы $H_{(1)} \ldots H_{(k-1)}$ и не отвергаем $H_{(k)} \ldots H_{(m)}$
5. Если $k = 1$, то никакая нулевая гипотеза не отвергается, а если нет такого k, то все нулевые гипотезы отвергаются.

Утверждение. При множественном тестировании с помощью алгоритма Хольма $\operatorname{strong}\mathrm{FWER} \leq \alpha$.

Доказательство. Обозначим I_0 множество индексов, соответствующих верным нулевым гипотезам (неизвестно, каким), $m_0 = |I_0|$ — число верных нулевых гипотез.

Пусть h — номер первой (имеем в виду порядок гипотез, введенный в алгоритме Хольма) отвергнутой нулевой гипотезы $H_{(h)}$ среди всех верных гипотез. Так как если одна гипотеза не отверглась, то и следующие гипотезы не отвергаются, то все предыдущие гипотезы $H_{(1)}, \ldots, H_{(h-1)}$ тоже отвергались, но были неверными. Поэтому m_0 верных нулевых гипотез должны поместиться между h и m; отсюда получаем $m_0 \leq m - h + 1$ и поэтому $\dfrac{1}{m-h+1} \leq \dfrac{1}{m_0}$.

По определению, $FWER_{I_0} = \Pr(\exists i \in I_0 : H_i \text{ отвергается})$. Так как события отвержения гипотез являются вложенными, т.е. не может отвергнуться гипотеза с бОльшим номером (i), если не отверглась с меньшим, получаем $FWER_{I_0} = \Pr(H_{(h)} \text{ отверглась})$. ($\Pr(A) = \Pr(AB)$, если $B \subset A$.)

В силу неравенства для m_0,

$$FWER_{I_0} = \Pr(p_{(h)} \leq \frac{\alpha}{m-h+1}) \leq \Pr(p_{(h)} \leq \frac{\alpha}{m_0}).$$

Так как для h может быть не более m_0 вариантов,

$$\Pr(p_{(h)} \leq \frac{\alpha}{m_0}) \leq \Pr(p_i \leq \frac{\alpha}{m_0} \text{ for } i \in I_0) \leq \sum_{i \in I_0} \Pr(p_i \leq \frac{\alpha}{m_0}) = \alpha.$$

Таким образом, $FWER_{I_0} \leq \alpha$ для любого I_0. Поэтому strong FWER $\leq \alpha$. □

Замечание. Тест по алгоритму Хольма более мощный, чем с поправкой Боферрони.

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

2.2.1. Частный случай

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

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

FWER без поправки ($\alpha_1 = \alpha$):

1	2	5	10	20
α_1	α_2	α_5	α_{10}	α_{20}
0.01	0.02	0.05	0.10	0.20
0.05	0.10	0.23	0.40	0.64
0.10	0.19	0.41	0.65	0.88

Определение. *Поправка Šidák'a*:

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

Вопрос: как выглядит поправка Šidák'a для p-value?

Вопрос: что происходит с тестом, если поправку Šidák'a заменить на поправку Бонферрони?

3. ANOVA Post-Hoc Comparison

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

3.1. Least Significant Difference (LSD)

LSD test — это просто попарный t-test:

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

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

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

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

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

3.2. Распределение размаха

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

Определение. *Размах* есть случайная величина

$$w_d = \xi_{(d)} - \xi_{(1)}$$

с функцией распределения

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

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

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

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

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

Определение. Пусть натуральное число ℓ и случайная величина η такие, что $\ell\eta^2/\sigma^2 \sim \chi^2(\ell)$; тогда

$$\frac{w_d}{\eta} \sim q(d, \ell),$$

т.е. имеет распределение стьюдентизированного размаха с параметрами d и ℓ. Это распределение затабулировано.

Пример (Проверка выборки на outliers)**.** В нормальной модели, H_0 : нет outliers. Статистика при $d = n$

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

потому что, естественно,

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

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

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

3.3. Tukey's Honest Significant Difference (HSD) Test

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

По определению стьюдентизированного размаха (с $\xi_i = \bar{y}_{(i)}$, $l = n - k$),

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

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

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

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

Замечание. Если $k = 2$, то HSD и LSD — эквивалентные тесты, но статистики критерия отличаются в $\sqrt{2}$ раз и наличием/отсутствием модуля: в t-test

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\tilde{s}_{1,2}\sqrt{2/n_0}},$$

Предложение. $\mathrm{FWER}_{\{1:m\}} = \alpha$.

Доказательство. Действительно

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

□

3.4. Другие критерии

Newman-Keals stepdown вариант HSD.

Tukey-Cramer HSD вариант Tukey для несбалансированного дизайна

Dunnett сравнивает все группы с контрольной

3.5. Scheffé's Method

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

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

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

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

Полученную сумму следует отнормировать и получить статистику

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

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

Пусть $\mathbf{c}^{(1)}, \dots, \mathbf{c}^{(d)},\ d \le k-1$ — наборы ортогональных контрастов. Тогда для любого вектора

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

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

Сколько бы ни захотелось проверить контрастов, хочется уверенности, что $\mathrm{FWER} \le \alpha$. Статистика

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

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

3.6. Сравнение мощностей

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

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

Чем больше критическое значение, тем ниже мощность, конечно.

Часть VI.

Регрессионный анализ

1. Регрессия

Определение. *Регрессией* η на $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_p)^\mathsf{T}$ называется $\mathsf{E}\{\eta \mid \boldsymbol{\xi}\}$.

Замечание. Таким образом осуществляется предсказание η по ξ с минимальной среднеквадратичной ошибкой. Детали можно посмотреть в разделе 2.4.

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

Замечание. f находится по МНК: $f(\boldsymbol{\xi}) = \operatorname{argmin}_{\hat{\eta} \in K} \mathsf{E}(\eta - \hat{\eta})^2$ для $K = \{\psi(\boldsymbol{\xi}) : \psi \text{ — измеримая}\}$.

Если класс K отличается от $\{\psi(\boldsymbol{\xi}) : \psi \text{ — измеримая}\}$, то результат тоже называют регрессией, добавляя, например, линейная регрессия с $K = \left\{A^\mathsf{T}\boldsymbol{\xi} + b\right\}$, полиномиальная регрессия и пр.

Если зависимость линейная, то линейная регрессия является действительно регрессией (условным мат.ожиданием), но в общем случае линейная регрессия — наилучшее линейное приближение по МНК.

Виды регрессий

- Парные (предсказывая величину по одной случайной величине) и множественные (по многим).

Неслучайные регрессоры Есть другая постановка задачи, когда регрессоры не случайные, т.е. есть неслучайная матрица $\mathbf{X}$, состоящая из n индивидов $\mathbf{x} = (x_1, \ldots, x_p)^\mathsf{T}$, а наблюдается только случайный вектор η с компонентами $\eta_i = \psi(\mathbf{x}_i) + \varepsilon_i$, $i = 1, \ldots, n$, где $\mathsf{E}\varepsilon_i = 0$. В этой постановке функция ψ называется функцией регрессии, $\mathsf{E}\eta_i = \psi(\mathbf{x}_i)$ (просто математическое ожидание, не условное). В этом случае ищется функция ψ (или ее параметры), на которой достигается минимум $\min_{\hat{y}_i = \psi(\mathbf{x}_i), \psi \in K} \mathsf{E} \sum_{i=1}^n (\eta_i - \hat{y}_i)^2$. При такой постановке задачи возможна только параметрическая модель регрессии, когда ψ не просто измеримая функция, а задана числом параметров меньше n.

В этом случае линейная модель соответствует линейной функции ψ. Стандартные дополнительные предположения — ошибки ε_i независимы и одинаково распределены с дисперсией σ^2. Еще одно необязательное предположение — это нормальное распределение ошибок; оно нужно только для дополнительных свойств оценок коэффициентов регрессии.

2. Парная линейная регрессия

Когда используют название «линейная регрессия» или «линия регрессии», обычно даже не предполагают, что полученная прямая линия является регрессией (условным математическим ожиданием). Имеется в виду просто приближение линейной функцией по методу наименьших квадратов.

Определение. Пусть $\xi, \eta \in L^2$ (с конечной дисперсии). *Парной линейной регрессией* η *на* ξ называется наилучшее среднеквадратичное приближение $h_{b_1^*,b_0^*}(\xi) = b_1^*\xi + b_0^*$ в классе линейных по ξ функций $K = \mathcal{L} = \{b_1\xi + b_0\}$. Иными словами,

$$h_{b_1^*,b_0^*}(\xi) = \operatorname*{argmin}_{b_1,b_0} \|\eta - h_{b_1,b_0}(\xi)\|^2 = \operatorname*{argmin}_{b_1,b_0} \underbrace{\mathsf{E}(\eta - (b_1\xi + b_0))^2}_{\phi(b_1,b_0)} = b_1^*\xi + \beta_0^*.$$

Замечание. Найти минимум ϕ можно, как обычно, решив систему $\partial\phi/\partial\beta_i = 0$[1].

Утверждение. b_1^*, b_0^* таковы, что для $y = h_{b_1^*,b_0^*}(x)$

$$\frac{y - \mathsf{E}\eta}{\sqrt{\mathsf{D}\eta}} = \rho\frac{x - \mathsf{E}\xi}{\sqrt{\mathsf{D}\xi}}.$$

Это уравнение задает линию регрессии. Иными словами,

$$y = \underbrace{\rho\frac{\sqrt{\mathsf{D}\eta}}{\sqrt{\mathsf{D}\xi}}}_{b_1^*} x + \underbrace{\mathsf{E}\eta - \rho\frac{\sqrt{\mathsf{D}\eta}}{\sqrt{\mathsf{D}\xi}}\mathsf{E}\xi}_{b_0^*}.$$

Отсюда можно получить соотношение между коэффициентом линейной регрессии b_1^* (наклоном регрессионной прямой) и коэффициентом корреляции:

$$b_1^* = \rho\frac{\sigma_\eta}{\sigma_\xi}.$$

Замечание. Подстановкой проверятся, что

$$\phi(b_0^*, b_1^*) = \min_{\hat\eta \in K} \mathsf{E}\,(\eta - \hat\eta)^2 = \mathsf{D}\eta(1 - \rho^2),$$

откуда можно найти уже известное выражение для коэффициента корреляции Пирсона

$$\rho^2(\eta, \xi) = 1 - \frac{\phi(b_1^*, b_0^*)}{\mathsf{D}\eta} = 1 - \frac{\min_{\hat\eta\in\mathcal{L}} \mathsf{E}\,(\eta - \hat\eta)^2}{\mathsf{D}\eta}, \quad \hat\eta := h(\xi).$$

Определение. Линейная регрессия *значима*, если $b_1^* \neq 0 \implies \rho \neq 0$. Значимость регрессии эквивалентна значимости (осмысленности) предсказания по ней. В случае отсутствия угла наклона предсказание не имеет смысла, так как равно $\mathsf{E}\eta$ и не зависит от значений ξ.

В случае, когда наблюдения происходят в фиксированных, а не случайных точках, линия регрессии находится из условия

$$\sum_{i=1}^{n}(\eta_i - (b_1 x_i + b_0))^2.$$

Заметим, что здесь n относится не к объему выборки, а к числу измерений.

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

2.1. Переход на выборочный язык

Если модель была двумерная $(\xi, \eta)^\mathsf{T}$, то выборка состоит из n реализаций $(x_i, y_i)^\mathsf{T}$ вектора $(\xi, \eta)^\mathsf{T}$.

Если измерения происходят в фиксированных точках x_i, то выборка является одной реализацией $(y_1, \dots, y_n)$ вектора η.

Методом подстановки получаем формулу для линии регрессии (ее оценки), причем для формулы неважно, фиксированные иксы или случайные:

$$\frac{y - \bar{x}}{s_y} = \hat{\rho}\frac{x - \bar{x}}{s_x};$$

по этой формуле несложно выписать оценки коэффициентов линии регрессии.

Обозначим $\hat{y}^i = \hat{b}_1 x_i + \hat{b}_0$. Тогда оценкой дисперсии ошибки методом подстановки будет $\sum_{i=1}^n (y_i - \hat{y}_i)^2/n = \mathrm{SSE}/n$. Но рассматривают исправленную оценку, чтобы получить несмещенную оценку дисперсии ошибки

$$\hat{\sigma^2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2/(n-2) = \mathrm{SSE}/(n-2).$$

Таким образом, у нас есть оценки всех параметров модели, включая мешающий параметр σ^2.

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

2.2. Доверительные интервалы для параметров регрессии

Рассмотрим модель с фиксированными неслучайными иксами $\eta_i = b_1 x_i + b_0 + \varepsilon_i$, $i = 1, \dots, n$, где $\mathsf{E}\varepsilon_i = 0$, $\mathsf{D}\varepsilon_i = \sigma^2$.

Как обычно, помимо точечной оценки $\hat{b}_1$ и $\hat{b}_0$, интересуемся диапазоном значений, которые может принимать оценка с заданной вероятностью, т.е. доверительными интервалами.

Удобнее заменить параметр b_0 на другой, переписав уравнение регрессии как $y = b_1(x_i - \bar{x}) + \tilde{b}_0$, где $\tilde{b}_0 = b_0 + b_1\bar{x}$. Соответственно поменяются и оценки параметров. Для них можно получить следующие формулы для дисперсии (это частный случай множественной регрессии, поэтому без доказательства):

$$\mathsf{D}\hat{b}_1 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{n s_x^2}, \qquad \mathsf{D}\hat{\tilde{b}}_0 = \frac{\sigma^2}{n}.$$

Более того, если распределение ошибок нормальное, что оценки $\hat{b}_1$ и $\hat{\tilde{b}}_0$ являются независимыми.

Значимость регрессии определяется с помощью проверки гипотезы $H_0 : \rho = 0$. Если гипотезы отвергается, то регрессия называется значимой (предсказание имеет смысл в том смысле, что зависит от значения x).

2.3. Предсказание по линейной регрессии

На основе дисперсий оценок параметров, можно построить доверительный интервал для среднего предсказания $y = b_1 x_i + b_0$ в точке x с уровнем доверия γ:

$$\left((\hat{b}_1 x_i + \hat{b}_0) \pm c_\gamma \sigma \sqrt{1/n + \frac{(x - \bar{x})^2}{n s_x^2}} \right),$$

где $c_\gamma = \Phi^{-1}((1+\gamma)/2)$ (если ошибки не имеют нормальное распределение или дисперсия ошибки неизвестна и вместо нее стоит ее оценка $\hat{\sigma}^2$, то доверительный интервал асимптотический).

Из формулы ясно видно, что чем дальше x от среднего $\bar{x}$, тем больше доверительный интервал (менее точное предсказание).

Замечание. На картинке доверительные интервалы изображаются в виде «рукавов» вокруг графика линейной регрессии — т.е. область всевозможных положений прямой при варьировании $b_1, \tilde{b}_0$ в заданных интервалах.

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

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

Если мы хотим построить предсказательный интервал для $y = b_1 x_i + b_0 + \varepsilon_i$, то получим

$$\left((\hat{b}_1 x_i + \hat{b}_0) \pm c_\gamma \sigma \sqrt{1 + 1/n + \frac{(x - \bar{x})^2}{n s_x^2}} \right).$$

Доверительный интервал строится на основе SE оценки среднего предсказания и показывает, что в среднем может быть при таком x. Его размер уменьшается с ростом n (с улучшением точности оценивания параметров регрессии). Предсказательный интервал предсказывает, что вообще может случиться.

3. Множественная линейная регрессия

Будем рассматривать случай не случайных регрессоров,

3.1. Псевдообратные матрицы

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

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

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}, \quad \mathbf{A}^-\mathbf{A}\mathbf{A}^- = \mathbf{A}^-.$$

2. *(Псевдообратная по Муру-Пенроузу)* если по аналогии с $\mathbf{A}^{-1} = \mathbf{A}^\mathsf{T} \implies (\mathbf{A}^{-1}\mathbf{A})^\mathsf{T} = \mathbf{A}^\mathsf{T}\left(\mathbf{A}^{-1}\right)^\mathsf{T} = \mathbf{A}^{-1}\mathbf{A}$, дополнительно выполняется

$$\mathbf{A}^-\mathbf{A} = (\mathbf{A}^-\mathbf{A})^\mathsf{T}, \quad \mathbf{A}\mathbf{A}^- = (\mathbf{A}\mathbf{A}^-)^\mathsf{T}.$$

Свойства

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

$$\mathbf{A}^- = (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}.$$

2. Пусть ищут решение $\mathbf{X}\mathbf{b} = \mathbf{y}$ относительно $\mathbf{b}$

   a) Если уравнение не имеет решений, то на $\mathbf{b} = \mathbf{X}^-\mathbf{y}$ достигается минимум невязки между левой и правой частями:

$$\mathbf{b}^* = \mathbf{X}^-\mathbf{y} = \underset{\mathbf{b}}{\operatorname{argmin}} \|\mathbf{X}\mathbf{b} - \mathbf{y}\|^2.$$

   b) Если решение не единственно, то $\mathbf{b} = \mathbf{X}^-\mathbf{y}$ есть решение с минимальной нормой.

3.2. Проекторы на подпространства

Пусть $\mathscr{L}_d \subset \mathbb{R}^m$ — линейное подпространство размерности d с базисом $\{\mathbf{p}_1, \ldots, \mathbf{p}_d\}$, $\mathbf{P} = [\mathbf{p}_1 : \ldots : \mathbf{p}_d]$. Тогда проектор на $\mathscr{L}_d$ будет задан как

$$\underset{\mathscr{L}_d}{\operatorname{proj}} = \mathbf{\Pi} = \mathbf{P}(\mathbf{P}^\mathsf{T}\mathbf{P})^{-1}\mathbf{P}^\mathsf{T} = \mathbf{P}\mathbf{P}^-.$$

Если $\{\mathbf{p}_i\}_{i=1}^d$ — ортонормированный базис, то

$$\mathbf{\Pi} = \mathbf{P}\mathbf{I}\mathbf{P}^\mathsf{T} = \mathbf{P}\mathbf{P}^\mathsf{T}.$$

Кроме того,

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

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

Свойства

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

3.3. Ordinary and Total Least Squares

Пусть

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

матрица данных с n индивидами [1] по столбцам, каждый из которых описывается k признаками;

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

вектор наблюдений[2];

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

вектор неизвестных коэффициентов.

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

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

найти

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

Откуда регрессией будет[3]

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

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

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

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

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

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

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

Дальше рассматривается OLS.

[1] Также «predictors», «regressors», «controlled variables», «explanatory variables», «features», «inputs».
[2] Также «regressands», «response», «explaining variables», «outcome», «experimental variables».
[3] $\mathbf{H}$ — «hat matrix».

3.4. Свободный член

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

1. Ввести фиктивный столбец из единиц:
$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \in \mathbb{R}^{n \times m}, \quad m = k + 1.$$
2. Центрировать признаки.

Предложение. *Оба способы эквиваленты.*

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

Доказательство. Без доказательства. □

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

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

3.5. Стандартизованные признаки

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

Определение. Чтобы стандартизировать наблюдения, следует поделить центрированные столбцы на нормы каждого столбца, получится $\mathbf{X}^{(s)} \in \mathbb{R}^{n \times k}$; $\mathbf{y}^{(s)} = \mathbf{y}^{(c)} / \left\|\mathbf{y}^{(c)}\right\|$. Тогда
$$\hat{\mathbf{b}}^{(s)} = (\mathbf{X}^{(s)})^-\mathbf{y}^{(s)} = \left(\left(\mathbf{X}^{(s)}\right)^\mathsf{T}\mathbf{X}^{(s)}\right)^{-1}\left(\mathbf{X}^{(s)}\right)^\mathsf{T}\mathbf{y}^{(s)} = \hat{\boldsymbol{\beta}}, \quad \hat{\beta}_i = \frac{\left\|\mathbf{x}_i^{(c)}\right\|}{\left\|\mathbf{y}^{(c)}\right\|}\hat{b}_i.$$

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

3.6. Свойства оценки $\hat{\mathrm{b}}$

1. *Несмещенность* (по $\mathsf{E}\boldsymbol{\epsilon} = \mathbf{0}$):
$$\mathsf{E}\hat{\mathbf{b}} = \mathsf{E}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathsf{E}\mathbf{y} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathsf{E}\left(\mathbf{Xb} + \boldsymbol{\epsilon}\right) = \mathbf{b}.$$
2. Ковариационная матрица:
$$\operatorname{cov}\hat{\mathbf{b}} = \operatorname{cov}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\operatorname{cov}\boldsymbol{\epsilon}\,\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}.$$

Заметим, что никакие асимптотические свойства невозможны, так как $\mathbf{X}$ фиксирована.

3.7. Свойства $\hat{\mathrm{b}}^{(c)}$ и $\hat{\mathrm{b}}^{(s)}$

1. $\left(\mathbf{X}^{(c)}\right)^{\mathsf{T}}\mathbf{X}^{(c)}/n = \mathbf{S_{xx}}$, $\left(\mathbf{X}^{(c)}\right)^{\mathsf{T}}\mathbf{y}^{(c)}/n = \mathbf{S_{xy}}$ суть выборочные ковариационные матрицы (это можно вручную расписать и убедиться); тогда в их терминах

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

2. Ковариационная матрица:

$$\operatorname{cov}\hat{\mathbf{b}}^{(c)} = \sigma^2(\left(\mathbf{X}^{(c)}\right)^{\mathsf{T}}\mathbf{X}^{(c)})^{-1} = \frac{\sigma^2}{n}\cdot\mathbf{S_{xx}^{-1}} \xrightarrow[n\to\infty]{} 0$$

3. Аналогично,

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

и

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

3.8. Сравнение оценок

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

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

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

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

Теорема (Гаусс-Марков)**.** *В условиях* $\mathsf{E}\epsilon_i = 0,\ \mathsf{D}\epsilon_i = \sigma^2,\ \epsilon_i \perp\!\!\!\perp \epsilon_j$, $\hat{\mathbf{b}}_{\mathrm{OLS}}$ *является «BLUE»: «best linear unbiased estimate». То есть* $\hat{\mathbf{b}}_{\mathrm{OLS}}$ *обладает наименьшей ковариационной матрицей среди всех линейных (линейно зависящих от* $\mathbf{y}$*) несмещенных оценок.*

Замечание. Наименьшая ковариационная матрица гарантирует, что дисперсия оценки каждого коэффициента минимальна (следует из предыдущего замечания).

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

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

Доказательство. Так как $\mathbf{y} \sim \mathrm{N}(\mathbf{Xb}, \sigma^2\mathbf{I})$, то

$$\begin{aligned}
\hat{\mathbf{b}}_{\mathrm{MLE}} = \operatorname*{argmax}_{\mathbf{b}} \mathsf{P}(\mathbf{y} \mid \mathbf{b}) &= \operatorname*{argmax}_{\mathbf{b}} \frac{1}{(2\pi)^{n/2}\sqrt{\det\sigma^2\mathbf{I}}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{Xb})\,\sigma^{-2}\mathbf{I}(\mathbf{y}-\mathbf{Xb})^{\mathsf{T}}\right\} \\
&= \operatorname*{argmax}_{\mathbf{b}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^{\mathsf{T}}\right\} \\
&= \operatorname*{argmax}_{\mathbf{b}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{Xb}-\mathbf{y}\|^2\right\} \\
&= \operatorname*{argmin}_{\mathbf{b}} \|\mathbf{Xb}-\mathbf{y}\|^2 = \hat{\mathbf{b}}_{\mathrm{OLS}}
\end{aligned}$$

□

3.9. Разложение суммы квадратов и оценка σ^2

Разложение суммы квадратов в случае модели линейной регрессии имеет вид (даже если модель не верна):

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2.$$

Обозначим $\mathrm{SSE}=\mathrm{SSError}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$. Пусть модель верна и ошибки имеют нормальное распределение $N(0,\sigma^2)$. Тогда, с помощью теореме Cochran можно получить (без док-ва):

$$\frac{\mathrm{SSE}}{\sigma^2}\sim\chi^2(\underbrace{n-m}_{n-k-1})$$

и оценкой методом подстановки для σ^2 будет SSE/n; несмещенной оценкой (с поправкой на число степеней свободы) будет

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

3.10. Проверка значимости коэффициентов линейной регрессии и доверительные интервалы

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

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

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

Заметим, что в $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ нумерация идет от 0, а в $\mathbf{S}_\mathbf{x}$ от 1. Подставляя оценку σ, получают

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

3.10.1. Расстояние Махаланобиса

Если на прямой разброс удобно измерять стандартных отклонениях σ, то в многомерном пространстве аналогом такой характеристики является расстояние Махаланобиса (Mahalanobis distance).

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

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

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

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

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

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

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

как сумма квадратов центрированных и нормированных нормальных случайных величин.

Действительно,

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

3.10.2. Доверительный эллипсоид

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

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

Т.е. как такое множество значений, что расстояние их от среднего с учетом стандартного отклонения меньше квантиля уровня γ. В случае оценки среднего μ_0, например, получают стандартное с $\mathrm{SE} = \sqrt{\mathsf{D}\bar{x}}$

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

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

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

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

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

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

Значит,

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

3.11. Значимость регрессии

$H_0 : \mathbf{b}^{(c)} = \mathbf{0}$. Напомним что $\mathbf{b}^{(c)} = (b_1, \ldots, b_k)^{\mathrm{T}}$ и равенство его нулю означает то, что предсказание равно константе и не зависит от значений иксов. Эту гипотезу можно проверить тремя способами:

1. Аналогично парной регрессии: критерий

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

   а именно,

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

   Неизвестный σ^2 следует оценить как

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

   тогда

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

2. Через ANOVA (разложение дисперсии): Разложение дисперсии

$$\mathsf{D}\eta = \mathsf{E}(\eta - \mathsf{E}\eta)^2 = \mathsf{E}(\hat{\eta}^* - \mathsf{E}\eta)^2 + \mathsf{E}(\eta - \hat{\eta}^*)^2,$$

где $\hat{\eta}^*$ — наилучшее линейное приближение от $\boldsymbol{\xi}$, на выборочном языке будет иметь вид

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

В случае, когда регрессоры не случайны (есть неслучайная матрица данных $\mathbf{X}$ и случайный отклик $\mathbf{y}$, как у нас сейчас), то же самое разложение имеет место.

Замечание. Иногда также пишут

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

что ведет к неиллюзорной путанице!

Пусть $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Тогда, с помощью теореме Cochran можно получить (без док-ва):

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

и SSE $\perp\!\!\!\perp$ SSR.

Замечание. Утверждение про распределение SSE справедливо всегда при нормальном распределении ошибок; про SST и SSR это верно только если $\mathbf{b}^{(c)} = \mathbf{0}$. Именно поэтому применяется F-критерий для проверки значимости регрессии.

Таким образом, в качестве статистики F-критерия можно взять, как и в дисперсионном анализе,

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

Критическая область, очевидно, справа, так как 'идеальное значение' — 0.

Замечание. У этой статистики с предыдущей совпадает также и числитель, хотя, чтобы в этом убедиться, надо провести некоторые выкладки, так это не очевидно.

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

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

на выборочном языке для множественной линейной регрессии получают

$$R^2 \quad = \quad \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}. \tag{3.1}$$

Если матрица регрессоров $\mathbf{X}$ фиксирована (т.е. не является выборкой из распределения $\boldsymbol{\xi}$), то R^2, вычисленный по той же формуле (3.1), называется коэффициентом детерминации.

Замечание. При удалении даже незначимого признака R^2 уменьшится; однако adjusted R^2

$$\text{adjusted}\, R^2 \quad = \quad 1 - \frac{\text{SSE}/(n-(k+1))}{\text{SST}/(n-1)} \xrightarrow[n\to\infty]{} R^2$$

не обязательно в силу поправки $n-(k+1)$, действующей как штраф за количество переменных.

Несложные манипуляции позволяют выписать статистику критерия ANOVA через коэффициент детерминации:

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

3.12. Анализ оценок коэффициентов

Для анализа оценок коэффициентов можно посмотреть на попарные срезы доверительного эллипсоида; точнее, на двумерные эллипсоиды. Для пары коэффициентов β_i, β_j его можно нарисовать (самостоятельно), в качестве центра взяв точку $(\hat\beta_i, \hat\beta_j)^\mathsf{T}$, наклон главной оси и вытянутость определив по величине $\mathrm{corr}(\hat\beta_i, \hat\beta_j)$.

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

3.12.1. Корреляция между оценками коэффициентов в двумерном случае

При возрастании корреляции признаков:

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

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

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

Тогда

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

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

Комментарий по поводу того, почему естественно, что знак корреляции между оценками коэффициентов регрессии обратен к знаку корреляции между признаками.

Пусть $\rho(\xi_1, \xi_2) = 1$, например, $\xi_1 = \xi$, $\xi_2 = \xi$. Пусть $\eta = 3\xi$. Тогда возможны варианты:

$$\begin{aligned} \eta &= 2\cdot\xi_1 &&+1\cdot\xi_2 \\ \eta &= 1\cdot\xi_1 &&+2\cdot\xi_2 \\ \eta &= 0\cdot\xi_1 &&+3\cdot\xi_2 \\ \eta &= -1\cdot\xi_1 &&+4\cdot\xi_2 \\ &\ldots \end{aligned}$$

Видно, что между коэффициентами регрессии явная отрицательная линейная зависимость.

3.12.2. Супрессоры

TODO

3.12.3. Избыточность (redundancy) и ручное удаление признаков

С этой проблемой можно бороться, удаляя подходящие признаки из анализа[4] по следующим критериям:

1. Множественный коэффициент корреляции
$$R^2(\xi_i; \{\xi_j,\ j \neq i\}).$$
Чем он больше, тем скорее i-й признак нужно удалить.

2. Допустимость i-го признака:
$$\text{tolerance}_i = 1 - R^2(\xi_i; \{\xi_j,\ j \neq i\}).$$
Чем он меньше, тем скорее i-й признак нужно удалить. Помимо предыдущего соотношения справедливо
$$\mathsf{D}\hat{b}_i = \frac{\sigma^2}{\sum_{\ell=1}^n (x_\ell - \bar{\mathbf{x}}_i)^2} \frac{1}{\text{tolerance}_i}, \quad \frac{1}{\text{tolerance}_i} \text{ — Variance Inflation Factor,}$$
так что при маленькой допустимости дисперсия велика.

3. Частные корреляции
$$\rho(\xi_i, \eta \mid \{\xi_j,\ j \neq i\}) = \rho(\xi_i - \hat{\xi}_i, \eta - \hat{\eta})$$
Чем i-я частная корреляция больше, тем больше вклад признака в регрессию (тем менее он предпочтителен для удаления).

4. Полу-частные корреляции
$$\rho(\xi_i - \hat{\xi}_i, \eta).$$

3.12.4. Проверка гипотезы о том, что набор признаков избыточен

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

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

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

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

[4] Нет признака — нет проблемы.

3.12.5. Stepwise автоматическое удаление/добавление признаков

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

$$R^2_{1,i+1} - R^2_{1,i};$$

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

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

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

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

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

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

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

В Statistica есть критерий автоматической остановке для stepwise отбора признаков. F to enter в forward варианте — это пороговое значение для F, если F $<$ этого числа, то останов. В backward варианте есть F to remove: если F $>$ F to remove, то STOP. Только F to remove должно быть больше F to enter (на самом деле, на каждом шаге проверяется, можно ли добавить признак, а потом какой-то другой удалить и неравенство нужно, чтобы процедура не зациклилась.) Имеет смысл установить такие пороги, чтобы критерий остановки не сработал, а потом посмотреть на таблицу Stepwise summary.

Можно нарисовать, как ведет себя коэффициент детерминации в варианте forward и backward. Это монотонные функции, но необязательно вторая производная одного знака. Отсюда можно увидеть, что критерий остановке в варианте backward более безопасен.

3.12.6. Выбор модели на основе информационных критериев AIC и BIC

См. отдельный файл на wiki курса. Для информационных критериев нужна параметрическая модель, так как они строятся на основе функции правдоподобия и вводят штраф за число параметров.

Отдельно отметим, что информационные критерии показывают, какая модель лучше, но не утверждают, что лучшая модель является правильной.

Заметим, что в случае нормального распределения для всех случайных величин справедливо

$$\eta = \mathsf{E}(\eta \mid \xi_1, \ldots, \xi_i) + (\eta - \mathsf{E}(\eta \mid \xi_1, \ldots, \xi_i)),$$

так как имеем ортогональность второго слагаемого первому и линейность ошибок. Значит все модели «верны» и можно среди них выбрать наилучшую.

3.12.7. О множественном коэффициенте корреляции и саппрессорах

Известно, что $\rho(\eta, \xi)$ есть косинус угла между η и ξ в соответствующем пространстве. Аналогично можно думать, что R^2 есть косинус между η и линейным пространством, натянутым на $\xi_1, \ldots, \xi_k$:

$$R^2 = \cos^2(\eta, \mathcal{L}(\xi_1, \ldots, \xi_k)).$$

Для коэффициента детерминации то же самое, только вместо случайных величин стоят вектора-признаки и косинус — это обычный косинус угла между векторами.

Возможна ситуация, когда $\cos^2(\eta, \mathcal{L}(\xi_1, \xi_2)) = 1 = R^2(\eta; \xi_1, \xi_2)$ — т.е. η лежит в $\mathcal{L}(\xi_1, \xi_2)$ (и предсказание абсолютно точно), но, тем не менее, $\mathrm{cor}(\xi_1, \eta) \approx 0$, $\mathrm{cor}(\xi_2, \eta) \approx 0$. Это возможно, если $\mathrm{cor}(\xi_1, \xi_2) \approx \pm 1$ (почти коллинеарны). ξ_1 называется «саппрессором» (suppressor) по отношению к ξ_2 (или наоборот). Подробнее, см https://stats.stackexchange.com/a/73876.

3.12.8. Как понять, что все хорошо

Если stepwise регрессия вперед и назад дает примерно одинаковые результаты и вторая производная одного знака (отрицат.), если нет супрессоров, нет плохих доверительных эллипсоидов, нет коэффициентов регрессии (перед стандартизованными признаками) больше 1.

Но самый хороший вариант, конечно, если регрессоры (почти) независимы. Если удается найти набор слабо зависимых признаков, которые предсказывают лишь немного хуже, чем полный набор, то это удача.

Также, если данные (регрессоры) должны собираться, то добавляются еще неформальные характеристики признаков — признаки должны легко и дешево собираться и не иметь много пропусков.

3.12.9. Заполнение пропусков

К стандартным вариантам casewise и pairwise добавляет вариант заполнения пропусков средним значением. Здесь такая опасность: если большое количество пропусков заполнить средними, то искусственно уменьшится дисперсия признаков и, тем самым, ширина доверительных интервалов для оценок параметров (увеличится значимость).

Есть еще варианты заполнения пропусков по регрессии на признаки с малым числом пропусков, но это нужно делать в ручном режиме.

3.13. Анализ аутлаеров

Выделяющиеся наблюдения всегда выделяются по отношению к какой-то закономерности. Одним из стандартных методов определения выделяющегося наблюдения по отношению к конкретному методу является сравнение результатов методв, которые получены с участием индивида и без его участия.

3.13.1. Matrix plot

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

3.13.2. Deleted residuals

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

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

между $\hat{y}_i^{(i)}$ — предсказание по регрессии на i-м индивиде с помощью коэффициентов, оцененных без этого индивида, и y_i — наблюдении на i-м индивиде. $r_i^{(i)}$ будет «большой» также по сравнению с $r_i = y_i - \hat{y}_i$. Напротив, если i-й индивид аутлаером не является, то будет справедливо приближенное равенство $r_i^{(i)} \approx r_i$. Deleted residuals ('удаленные остатки') всегда не больше (по модулю) residuals, поэтому прямая $y = x$ для точек $(r_i, r_i^{(i)})$ не получится.

3.13.3. Studentized residuals

А как просто посмотреть, остатки слишком большие или не слишком? Для этого нужно знать распределение остатков. Оно отличается от распределения ошибок с ковар.матрицей $\sigma^2 \mathrm{I}_n$

Справедливо

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

откуда

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

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

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

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

Определение. h_{ii} — рычаг[5].

Чем больше i-й рычаг, тем меньше ошибка на i-м индивиде, так как он перетягивает на себя. Получаем в нормальной модели $\frac{r_i}{\sqrt{\mathsf{D}r_i}} = \frac{r_i}{\sigma\sqrt{1-h_{ii}}} \sim \mathrm{N}(0,1)$.

Определение. Стандартизированные (также называют internally studentized) остатки:

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

Можно рассмотреть $\hat{\sigma}^{(i)}$ — оценку дисперсии без i-го индивида (т.е. на основе суммы квадратов не всех остатков, без i-го индивида); тогда, при нормально распределенных ошибках наблюдения,

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

(«-1» потому что меньше на одного индивида). Такие остатки называют стьюдентизированными (или externally studentized).

Замечание. Полученную величину рычага можно сравнивать со «средним» значением рычага

$$\frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{1}{n}\operatorname{tr}\mathbf{H} = \frac{1}{n}\operatorname{rk}\mathbf{H} = \frac{k+1}{n}.$$

(как след идемпотентной матрицы, равный её рангу[6]: след есть сумма собственных чисел, однако у идемпотента два возможных собственных числа: 0 и 1, а кратность 1 в точности равна рангу).

3.13.4. Расстояние по Куку и расстояние Махаланобиса

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

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

так что расстояние по Куку определяется как

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

[5] «Leverage».
[6] http://math.stackexchange.com/a/101515

Расстояние по Куку показывает выбросы по отношению к регрессии (outliers всегда по отношению к чему-то, какой-то закономерности).

Можно еще рассмотреть выбросы по отношению к распределению регрессоров (зависимая переменная тут не участвует). Это делается стандартным способом, через расстояние Махаланобиса в пространстве независимых признаков (регрессоров): если x_i — i-й индивид, $\bar{\mathbf{x}}$ — вектор средних, то аутлаером можно назвать индивида, для которого велико

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

Правда, тут мы незаметно перешли к пониманию матрицы $\mathbf{X}$ как выборки из многомерного распределения. Если $\mathbf{X}$ — детерминированная матрица, то в этом смысле выбросов быть не может (так как нет закономерности).

Замечание. Если индивид аутлаер по Махаланобису, то $\mathbf{S_{xx}}$ оценивается неправильно (если понимать ее как выборочную ковариационную матрицу) и все значимости/доверительные интервалы становятся неправильными.

	Аутлаер по Куку	Не аутлаер по Куку
Аутлаер по Махаланобису	Далеко от линии регрессии, далеко от $\bar{\mathbf{x}}$	Далеко от $\bar{\mathbf{x}}$, на линии регрессии
Не аутлаер по Махаланобису	Далеко от линии регрессии, недалеко от $\bar{\mathbf{x}}$	Недалеко от $\bar{\mathbf{x}}$, на линии регрессии

3.14. Проверка правильности и выбор модели

- Если известно, что ошибки нормально распределены (например, в случае измерений прибора), то если остатки не имеют нормального распределения, то модель не является правильной.
  Вообще, на нормальность остатков имеет смысл смотреть, если верно то, что написано выше, но это особый специфичный случай, а также (более общий вариант) — для того, чтобы понимать, как относиться к результатам критериев (точные они или асимптотические). Поэтому всегда имеет смысл посмотреть на гистограмму остатков и/или их normal probability plot.

- Если исходные данные имеют нелинейную зависимость, то и расположение остатков по линейной регрессии на графике будет отражать характер этой зависимости. Имеет смысл рассмотреть график, где по оси X откладывают предсказанные значения predicted $\hat{y}_i$, а по оси Y — остатки residuals $r_i = y_i - \hat{y}_i$. Обращаю внимание, что остатки всегда (!) ортогональны предсказанным значениям, по построению. Поэтому по наклону линии парной регрессии на графике residuals vs predicted невозможно определить, правильная ли была модель регрессии. Однако, по облаку точек это можно сделать, так как если зависимость нелинейная, то в ϵ войдет кусочек ξ и независимости residuals и predicted не будет.
  Упражнение Нарисуйте, как будет выглядеть это график residuals vs predicted, если регрессия квадратичная типа $y = x^2$, $x > 0$.

  Замечание. Есть только один вариант, когда линия парной регрессии на графике residuals vs predicted может быть не горизонтальной — когда выборка неправильная, что получается в случае пропущенных наблюдений и варианте pairwise.

3.15. Доверительные интервалы для среднего предсказания и предсказательные интервалы

Предваряя следующие рассуждения, сразу скажу, что первое в названии строится на основе SE, а второе — на основе SD. Смысл тот же.

Пусть $\mathbf{x} = (1, \mathbf{z})^\mathsf{T} \in \mathbb{R}^{k+1}$; тогда среднее предсказание (мат.ож., mean prediction) в модели $y = \mathbf{x}^\mathsf{T}\mathbf{b} + \varepsilon$ имеет вид

$$\bar{y} = \begin{pmatrix}1\\ \mathbf{z}\end{pmatrix}^\mathsf{T} \mathbf{b},$$

а его оценка —

$$\hat{\mathbf{y}} = \begin{pmatrix}1\\ \mathbf{z}\end{pmatrix}^\mathsf{T} \hat{\mathbf{b}}.$$

Эта оценка несмещенная, $\mathsf{E}\hat{y} = \bar{y}$. Несложно увидеть, что её дисперсия есть

$$\mathsf{D}\hat{y} = \sigma^2 \begin{pmatrix}1\\ \mathbf{z}\end{pmatrix}^\mathsf{T} (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \begin{pmatrix}1\\ \mathbf{z}\end{pmatrix}.$$

Так как у матрицы $\mathbf{X}$ первый столбец состоит из единиц, технические выкладки дают более интерпретируемое выражение

$$\mathsf{D}\hat{y} = \frac{\sigma^2}{n} + \frac{\sigma^2}{n}(\mathbf{z} - \bar{z})^\mathsf{T}\mathbf{S}_{\mathbf{xx}}^{-1}(\mathbf{z} - \bar{z}),$$

частный случай чего выписывался в случае парной регрессии как

$$\mathsf{D}\hat{y} = \frac{\sigma^2}{n} + \frac{\sigma^2(z - \bar{z})^2}{\sum_{i=1}^n (z_i - \bar{z})^2}.$$

Доверительным интервалом для $\bar{y}$ будет

$$(\hat{y} \pm c_\gamma \,\mathrm{SE}) = \left(\bar{y} \pm c_\gamma \sqrt{\widehat{\mathsf{D}\hat{y}}}\right) = \left(\hat{y} \pm c_\gamma \hat{\sigma} \sqrt{\begin{pmatrix}1\\ \mathbf{z}\end{pmatrix}^\mathsf{T} (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \begin{pmatrix}1\\ \mathbf{z}\end{pmatrix}}\right),$$

где c_γ находится из распределения $\mathrm{t}(n-m)$ стандартным способом. Если же мы хотит предсказать не среднее значение, а построить диапазон значений, который может быть, т.е. предсказательный интервал (prediction interval), то получим, добавляя дисперсию ε:

$$\left(\hat{\mathbf{y}} \pm c_\gamma \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{1}{n}(\mathbf{z} - \bar{z})^\mathsf{T}\mathbf{S}_{\mathbf{xx}}^{-1}(\mathbf{z} - \bar{z})}\right).$$

3.16. Сведение нелинейной модели к линейной

Тут мы смешаем случайные и неслучайные регрессоры. Если регрессоры случайны, то стандартно под регрессией понимаем условное математическое ожидание.

Существует три базовых модели, в которых функция регрессии линейная:

1. $\eta, \xi_1, \ldots, \xi_k$ имеют нормальное распределение.
2. $\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon}$, $\mathsf{E}\boldsymbol{\epsilon} = \mathbf{0}$, $\operatorname{cov}\boldsymbol{\epsilon} = \sigma^2\mathbf{I}$. Если регрессоры случайные, то модель имеет вид измерений со случайными ошибками: $\eta = \boldsymbol{\xi}^\mathsf{T}\mathbf{b} + \epsilon$, где $\boldsymbol{\xi}$ и ϵ независимы.
3. одномерный случай $\eta = \phi(\xi) + \epsilon$, ξ принимает всего два значения (возможно, как качественный признак).

Теперь рассмотрим случаи, когда регрессию можно свести к линейной. Пусть

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

и ϕ — нелинейная функция.

- ϕ — многочлен. Можно свести к линейной, добавляя признаки $\xi, \xi^2, \ldots$ и для этих признаков строить модель (для неслучайных регрессоров аналогично).
  Опасность: может появиться сильная зависимость построенных регрессоров.
- ξ — качественный признак, $A_1, \ldots, A_k$ — его градации. Можно ввести $k-1$ штук[7] фиктивных признаков со значениями $\{0, 1\}$, где 1 стоит на месте A_i, $i = 1, \ldots, k-1$, и для них строить модель.

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

4. Модификации линейной регрессии.

4.1. Взвешенная регрессия (Weighted Least Squares)

Пусть $\mathbf{W}$ — симметричная, положительно определенная матрица, тогда

$$\hat{\mathbf{b}}_W = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{y}$$

есть «взвешенная» оценка. При $\mathbf{W} = \mathbf{I}$, $\hat{\mathbf{b}}_W = \hat{\mathbf{b}}$, конечно.

Если $\mathsf{E}\boldsymbol{\epsilon} = \mathbf{0}$, то

$$\mathsf{E}\hat{\mathbf{b}}_W = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathsf{E}\mathbf{y} = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}\mathbf{b} = \mathbf{b}$$

и оценка несмещенная.

- Если $\operatorname{cov}\boldsymbol{\epsilon} = \sigma^2\mathrm{I}$, то $\hat{\mathbf{b}}$ — BLUE и $\hat{\mathbf{b}}_W$ уже не лучшая.
- Если $\operatorname{cov}\boldsymbol{\epsilon} = \mathbf{C}$ (то есть шум не белый) то нужно подобрать $\mathbf{W}$ такую, что $\hat{\mathbf{b}}_W$ — BLUE. Это делается операцией отбеливания: пусть всё центрированное; тогда $\mathbf{C}^{-1/2}\boldsymbol{\epsilon}$ — центрированный и нормированный белый шум, и

  $$\underbrace{\mathbf{C}^{-1/2}\mathbf{y}}_{\tilde{\mathbf{y}}} = \underbrace{\mathbf{C}^{-1/2}\mathbf{X}}_{\tilde{\mathbf{X}}}\mathbf{b} + \mathbf{C}^{-1/2}\boldsymbol{\epsilon}$$

  откуда

  $$\operatorname{cov}(\mathbf{C}^{-1/2}\mathbf{y}) = \left(\mathbf{C}^{-1/2}\right)^\mathsf{T}\operatorname{cov}(\mathbf{X}\mathbf{b} + \boldsymbol{\epsilon})\mathbf{C}^{-1/2} = \mathbf{I}.$$

  Так как теперь шум белый, следующая оценка будет BLUE:

  $$\begin{aligned}\hat{\tilde{\mathbf{b}}} &= \tilde{\mathbf{X}}^-\tilde{\mathbf{y}} = ((\mathbf{C}^{-1/2}\mathbf{X})^\mathsf{T}(\mathbf{C}^{-1/2}\mathbf{X}))^{-1}\left(\mathbf{C}^{-1/2}\mathbf{X}\right)^\mathsf{T}\mathbf{C}^{-1/2}\mathbf{y} \\ &= (\mathbf{X}^\mathsf{T}\left(\mathbf{C}^{-1/2}\right)^\mathsf{T}\mathbf{C}^{-1/2}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\left(\mathbf{C}^{-1/2}\right)^\mathsf{T}\mathbf{C}^{-1/2}\mathbf{y}\end{aligned}$$

  Значит, следует положить $\mathbf{W} = \mathbf{C}^{-1}$.

  Для $\mathbf{W}$ итеративный процесс: берем начальное значение, находим коэффициент, оцениваем $\mathbf{C}$ и т.д.

Пример. Стандартный случай — измерения с разной точностью, откуда

$$\mathbf{C} = \begin{pmatrix}\sigma_1^2 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sigma_n^2\end{pmatrix}.$$

Наблюдениям, таким образом, придается разный вес — чем меньше точность наблюдения, тем больше σ_i^2 и меньший, соответственно, вес.

Замечание. $\mathbf{W}$ можно также назначить и руками.

4.2. Гребневая (Ridge) регрессия

Чтобы бороться с вырожденностью $\mathbf{R_{xx}}$ в оценке $\hat{\boldsymbol{\beta}} = \mathbf{R_{xx}^{-1}R_{xy}}$ рассматривают

$$\hat{\boldsymbol{\beta}} = (\mathbf{R_{xx}} + \lambda\mathbf{I})^{-1}\mathbf{R_{xy}}.$$

Получается смещенная оценка, но с меньшей дисперсией, что может привести к уменьшению MSE (например, если регрессоры линейно зависимы, то точно будет лучше, так как конечная дисперсия меньше бесконечной). Напомним, что MSE равно сумме дисперсии и квадрата смещения. Для поиска λ используют кросс-валидацию.

Эта процедура эквивалентна тому, как если бы мы к регрессорам (после стандартизации) добавили искусственный шум с дисперсией λ.

Часть VII.

Материалы для курса ‘Вероятностные и статистические модели’ (магистры 1 курса)

1. Робастные оценки, критерии, ...

Робастными (robust) оценками и критериями называются те, которые слабо зависят от предположений о виде распределения и/или выбросов (outliers). Поэтому, вообще говоря, нужно уточнять, относительно чего роюастность рассматривается. Часто, по умолчанию, предполагается робастность по отношению к выбросам.

Очевидные робастные кандидаты — оценки/критерии, основанные на рангах, когда вместо значений подставляются ранги индивидов (номера по порядку в упорядоченной выборке, ранг 1 у минимального значения, ранг n у максимального). Так как ранги не меняются при монотонном преобразовании, такие оценки/критерии вообще не зависят от вида распределения (и называются непараметрическими). Также, так как выброс меняет каждый ранг не больше, чем на 1, то характеристики, основанные на рангах, устойчивы к выбросам.

Другой класс оценок/критериев, который относительно устойчив по отношению к виду распределения, это те, которые основаны на центральной предельной теореме (ЦПТ). Конечно, устойчивость зависит от распределения (от скорости сходимости в ЦПТ для случайных величин с таким распределением). Самый общий пример — t-test для проверки гипотезы о математическом ожидании на основе выборочного среднего. Если случайная величины не имеет нормальное распределение, распределение стандартизованного выборочного среднего по ЦПТ все равно сходится к нормальному распределению.

Отдельное замечание по поводу того, что такое выброс. Если это просто ошибка в данных (неправильно набрано число, к примеру), то проблем нет, это значение не рассматривается или заменяется на пропуск или заменяется на среднее по признаку. Но, вообще, к выбросам нужно относиться внимательно. Например, вы изучаете состояние организма человек и делаете о нем какие-то выводы. При анализе обнаружился выброс, его удалили. Получили выводы, применили к женщине (например, чтобы предсказать, здорова она или нет), и ошиблись (!). Оказалось, что удаленный выброс был женщиной, а остальные были мужчинами. Соответственно, метод годился только для мужчин, а был применен к женщине. Этот пример случая, когда выбросы — не ошибка, а просто данные из другого распределения (другой закономерности). Если бы таких точек было много, мы бы распознали неоднородность данных, а так просто выбросили и не заметили, что это точки из другой закономерности.

Также, не бывает просто выброса (outlier), нужно обязательно сказать, по отношению к чему, к какой закономерности, это выброс.

1.1. Непараметрические оценки и критерии

1.1.1. Оценки

Например, вместо выборочного среднего рассматривать выборочную медиану. Получаем более устойчивую оценку (не реагирует на выброс, не реагирует на монот.преобразование). Простейший пример (10 3 6 4 7) и (10 3 6 4 777). Но (!) того ли это оценка? Если интересует мат.ож., но распределение несимметричное. Тогда не того. А если интересует характеристика положения у логнормального распределения – то ровно того, что нужно (а выб.среднее – как раз не того, что нужно). За устойчивость оценки чаще всего платят точностью. Как видели на прошлом занятии, даже не всегда. ($\pi/2$ для n, но потом извлечь корень).

1.1.2. Критерии

Проверяем гипотезу, что два распределения равны.

t-Test, тест Манна-Уитни (Mann-Whitney or Wilcoxon, or тест суммы рангов). Та же история – вопрос, то же самое ли проверяют. (Переформулируем – мощный ли против той альтернативы, которая нам важна.) Здесь за устойчивость можем заплатить мощностью. 5% для норм.распр. с одинаковой дисп.

1.1.3. Корреляции

Обычный коэффициент корреляции Пирсона и ранговый Спирмена. Разное-одинаковое? Если двумерное распределение нормальное, то $\rho = 2\sin(\pi\rho_S/6)$, т.е. коэффициенты примерно равны (синус вблизи нуля хорошо приближается линейной функцией $y = x$. А в других случаях даже если зависимость (условное мат.ож.) линейная, коэффициенты не обязаны быть равны даже примерно.

1.2. M-оценки

M-оценки являются обобщением оценок максимального правдоподобия (ОМП, MLE) в том смысле, что оценка ищется путем максимизации некоторого функционала, а точнее, ищется экстремум некоторой суммы по наблюдениям:
$\sum_{i=1}^n d(x_i, \theta) \to \min_\theta$.

В случае ОМП это ...
В случае оценок по методу наименьших квадратов это ...
В случае оценок по методу наименьших абсолютных отклонений это ...

ОМП: $\sum_{i=1}^n (-\ln p_\xi(x_i; \theta))$

МНК: $\sum_{i=1}^n (y_i - \phi(\mathbf{x}_i; \theta))^2$ (ϕ — модель).

Мат.ож. Рассмотрим два последних случая: обсуждали, что в случае МНК решение — это среднее арифметическое, а в случае абсолютных отклонений - это выборочная медиана.

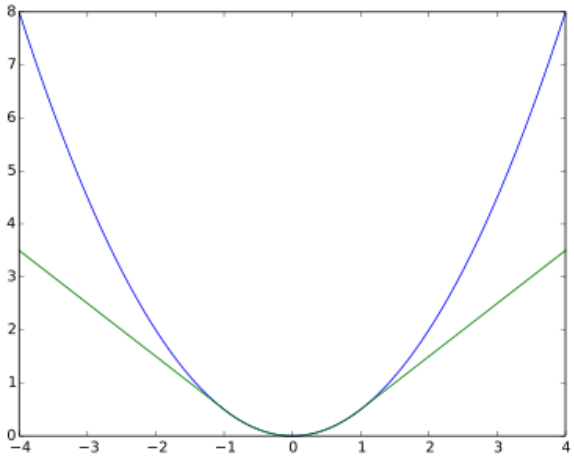
Поэтому методы, основанные на $d(x_i, \theta) = (x_i - \theta)^2$, не являются робастными, а методы, основанные на $d(x_i, \theta) = |x_i - \theta|$, являются робастными.

В случае нормального распределения, рассматривая робастные оценки, теряем в точности.

Идея — сделать оценки, где сочетается точность одного и устойчивость другого (для симметричного распределения). При этом предполагается, что выбросы тоже расположены симметрично. Например, исходная модель $N(a, 1)$, модель выбросов $N(a, 10)$, данные имеют вид смеси $0.95 \cdot N(a, 1) + 0.05 \cdot N(a, 10)$ или $0.95 \cdot N(a, 1) + 0.05 \cdot Lapl(a, 10)$.

Пусть $d(x_i, \theta) = f(x_i - \theta)$.
Huber's оценки:

$$f(x) = \begin{cases} 0.5x^2 \text{ if } |x| < \delta \\ \delta(|x| - 0.5\delta) \text{ otherwise.} \end{cases}$$



Для оценивания мат.ож. симметричного распределения используются еще варианты trimmed или windsorized. trimmed — это когда какое-то количество крайних значений (например, больше трех сигм) удаляется перевычислением среднего арифметического, windsorized — когда соотв.значения не удаляются, а устанавливаются равными крайним неудаленным.

Задание — какой функции f это соответствует?

Задать порог — непросто, так как нужно оценивать дисперсию для стандартизацию. Часто особые правила (например, trimmed) применяют к заданному проценту крайних точек.

Регрессия Есть и M-оценки для параметров регрессии, те же идеи.

Взвешенная регрессия, когда уменьшается вклад далеких элементов. В задачах типа регрессии, выброс часто соотносят с ошибкой в регрессии ($y_i = ax_i + b + e_i$), т.е. все ошибки подчиняются $N(a, s^2)$, а выброс нет (модель выброса — у него больше дисперсия). Тогда, как мы обсуждали (говоря про), можно выбросам дать меньший вес. Как это определить? Итеративная процедура. iteratively reweighted least-squares, IRLS

Общая идея для построения робастных оценок — замена в оптимизационных (!) задачах L_2-нормы на L_1-норму. *Не путайте с L_1- и L_2-регуляризацией.*

1.3. Не про робастность (!), но про увеличение точности оценки

1.3.1. Variance-bias trade-off

Определение. *Среднеквадратичная ошибка* (mean squared error, MSE) есть

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

Замечание. Поскольку

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

то

$$\underbrace{\mathsf{E}(\hat{\theta}_n - \theta)^2}_{\mathrm{MSE}} = \mathsf{D}\hat{\theta}_n + \underbrace{(\mathsf{E}(\hat{\theta}_n - \theta))^2}_{\mathrm{bias}^2}. \tag{1.1}$$

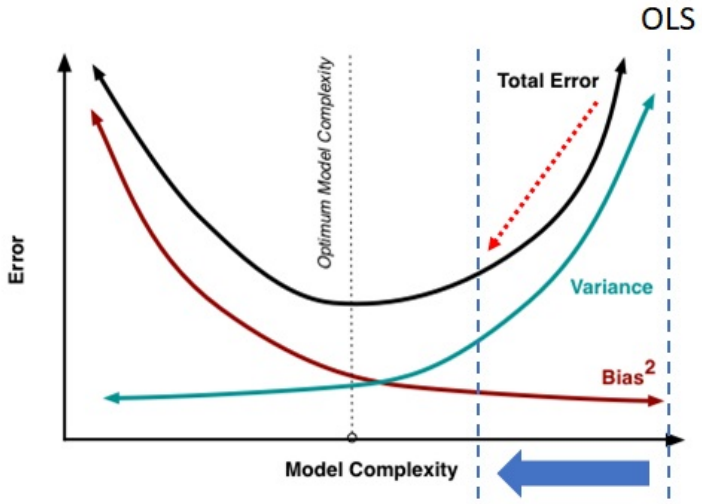
1.3.2. L_1- и L_2-регуляризация

Модель регрессии:
$y_i = \phi(\mathbf{x}_i;\theta) + \varepsilon_i$, $\mathsf{E}\varepsilon = 0$, $\mathsf{D}\varepsilon = \sigma^2$.

МНК: $\sum_{i=1}^{n}(y_i - \phi(\mathbf{x}_i;\theta))^2 \to \min_\theta$.

Для линейной регрессии (функция ϕ линейная функция) известно, то оценки параметров регрессии по методу наименьших квадратов являются BLUE — best linear unbiased estimate.

Идея: разрешить оценке иметь смещение, но уменьшить MSE.



Испортим оптимизируемую функцию:
$\sum_{i=1}^{n}(y_i - \phi(\mathbf{x}_i;\theta))^2 + \lambda f(\theta) \to \min_\theta$, $\lambda \geq 0$.

$f(\mathrm{t}) = \sum_i t_i^2$ — ridge regression;
$f(\mathrm{t}) = \sum_i |t_i|$ — Lasso regression.

Почему работает:
$\|\hat{\theta}\|^2 = \mathsf{D}\hat{\theta} + \|\theta\|^2$.

Чем больше портим (больше λ), тем меньше дисперсия оценки получается, но больше смещение. И, наоборот, при нулевом λ смещение нулевое, но дисперсия побольше.

Хорошая ссылка:
https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net

2. Доверительные интервалы

2.1. Мотивация и определение доверительных интервалов

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.

2.2. 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 (критическая область — область значений выборки, вероятность которой равна α), а не частным случаем, когда критерий строится через статистику критерия.

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

Если модель неизвестна, но известно, что $\mathsf{D}\xi < \infty$, можно построить доверительный интервал для $\mathsf{E}\xi = a$, 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 сходимость не испортится. Тогда

$$\mathsf{P}\left(a \in \left(\bar{x} \pm c_\gamma \frac{s}{\sqrt{n}}\right)\right) \xrightarrow[n\to\infty]{} \gamma, \quad c_\gamma = \mathrm{cdf}^{-1}_{\mathrm{N}(0,1)}\left(\frac{1+\gamma}{2}\right).$$

Чтобы в случае, когда модель нормальная, доверительный интервал становился точным, рассматривают его модификацию:

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

2.4. Доверительные интервалы для пропорций

Модель для распределения ξ — испытания Бернулли с вероятностью успеха p ($\mathsf{E}\xi = p$, $\mathsf{D}\xi = p(1-p)$). Задача — оценить p. Так как $p = \mathsf{E}\xi$, то ответ очевиден: $\hat{p} = \bar{x}$. То же самое получается, если искать ОМП.

Что такое $\bar{x}$, если выборка состоит из результатов испытаний Бернулли?

Итак, нас интересует процент успехов.

Замечание. Хотя речь о модели Бернулли и повторной независимой выборке объема n, часто говорят, что данные подчиняются биномиальной модели. В этом случае у нас модель $Bin(n, p)$ и всего одно испытание (выборка объема 1) из числа успехов. Если разделим на n, будет доля успехов.

Если с оценкой все очевидно, то с доверительными интервалами не так очевидно.

Первая идея — воспользоваться асимптотическим дов. интервалом для мат.ож.

Задание: Построить доверительный интервал для p с помощью аналогичной идеи, как для математического ожидания.

$$\mathsf{P}\left(p \in \left(\hat{p} \pm c_\gamma \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right)\right) \xrightarrow[n\to\infty]{} \gamma, \quad c_\gamma = \mathrm{cdf}^{-1}_{\mathrm{N}(0,1)}\left(\frac{1+\gamma}{2}\right).$$

Разница с общим случаем:

- можно получить более точную оценку дисперсии, так как всего один параметр ($\mathsf{D}\xi = p(1-p)$), и тем самым улучшить сходимость (asymptotic c.i.).
- можно получить более точный (все еще асимптотический) несимметричный доверительный интервал (он будет всегда внутри $[0, 1]$), называется Wilson's confidence interval for propotions.
- можно получить точный доверительный интервал (но не для всех доверительных уровней) (exact).

https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval

Wilson's confidence interval for proportions:

$$\left(\frac{1}{1+\frac{c_\gamma^2}{n}}\left(\hat{p} + \frac{c_\gamma^2}{2n}\right) \pm \frac{c_\gamma}{1+\frac{c_\gamma^2}{n}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{c_\gamma^2}{4n^2}}\right)$$

or the equivalent

$$\left(\frac{n_S + \frac{1}{2}c_\gamma^2}{n + c_\gamma^2} \pm \frac{c_\gamma}{n + c_\gamma^2}\sqrt{\frac{n_S\, n_F}{n} + \frac{c_\gamma^2}{4}}\right),$$

где n_S — число успехов в выборке, n_F — число неудач.

Построение точного доверительного интервала основано на том, что распределение $n\bar{x}$ известно. Какое оно?

Это биномиальное распределение с мат.ож. np. Находим границы: $P(np - c_1 < n\bar{x} < np + c_2) \geq \gamma$ (интервал покороче), отсюда получаем интервал для p. Распределение дискретное, поэтому точные границы не найти.

Примеры:

(1) Интуитивно, сколько раз может выпасть орел на правильной монетке из 10 бросаний с вер. 0.95? Из 100?
Если $p \approx 0.5$, то $\sqrt{p(1-p)} \approx 0.5$. Получаем примерный доверительный интервал $(\hat{p} \pm 0.5c_\gamma/\sqrt{n})$, а для 95% доверительного интервала примерно $(\hat{p} \pm 1/\sqrt{n})$.

(2) Про американскую вакцину было сказано, что доля случаев, когда она помогает — 95%. А про другие — 90 и 92%. Позволяет ли точность эксперимента говорить, что какая-то из вакцин лучше? Там было 95 заболевших, из них 90 из группы с плацебо и 5 из группы с вакциной, группы равные. (Если группы плацебо-вакцина делятся 1 к k, обычно плацебо дают меньшей по размеру группе, то для случаев, когда помогает, оценивается как
(больные среди плацебо $* k$)/(больные среди плацебо $* k +$ больные с вакциной). Например, $= 16 * 3/(16 * 3 + 4) \approx 0.92$, т.е. у нас из 20 человек 4 человека заболели после вакцины.

(3) Каков доверительный интервал, если получили ноль успехов? Правило трех: 95% доверительный интервал имеет примерный вид $[0, 3/n]$ (при больших n, $3 \approx -\ln 0.05$) — здесь имеется в виду односторонний доверительный интервал, так как с одной стороны возможные значения ограничены нулем. Был пример — 10 человек перешли по мостику и не упали. Кажется, что переход по мостику безопасный. Но (!) с вероятностью 0.95 вероятность упасть может иметь значение из $[0, 0.3]$, т.е. может иметь значение вплоть до 0.3, что немало.

(4) Подряд 10 дней случились практически одинаковые значения числа успехов (не долей, а именно числа). Так могло быть? Построим доверительный интервал для числа успехов (при маленьком p и большом n). Оказывается, он не зависит от n и p по отдельности!

Воспользуемся тем, что $p(1-p) \approx p$. Получим:

$$\mathsf{P}\left(np \in \left(n\hat{p} \pm c_\gamma\sqrt{n\hat{p}}\right)\right) \xrightarrow[n\to\infty]{} \gamma, \quad c_\gamma = \mathrm{cdf}^{-1}_{\mathrm{N}(0,1)}\left(\frac{1+\gamma}{2}\right),$$

здесь np — теоретическое число успехов, а $n\hat{p}$ — эмпирическое число успехов.

Если построить, к примеру, 50% доверительный интервал, то вероятность того, что 10 дней подряд число успехов будет лежать в нем, равно $0.5^{10} \approx 0.001$ (а если внутри 20-процентного, то 10^{-7}).

3. Множественные тесты

Когда мы в тестах говорим об ошибке первого рода (ложно отвергнуть H_0), то мы выбираем эту ошибку так, как будто в единичном эксперименте эта ошибка не может произойти (примерно так можно представлять себе допустимую ошибку).

Однако, если наши эксперименты повторяются, то это не так. Если проводить эксперименты долго, то ошибка произойдет неминуемо. Пусть α_1 — ошибка в одном эксперименте. Нас интересует так называемая групповая ошибка (FWER, family-wise error rate)

$$\text{FWER} = \alpha_m = \mathsf{P}(\text{хотя бы 1 раз из } m \text{ экспериментов произойдет ошибка}).$$

3.1. Независимые тесты

Пусть у нас тесты проводятся независимо (как с леденцами). (Чему там равно m?)

Давайте получим формулу для FWER в случае, когда все гипотезы справедливы. Это упражнение по теорверу. Решается довольно просто — нас интересует событие, дополнение к которому ‘событие не произойдет ни разу’.

$FWER = \alpha_m = 1 - (1 - \alpha_1)^m$. При $m \to \infty$ FWER стремится к 1.

1	2	5	10	20
α_1	α_2	α_5	α_{10}	α_{20}
0.01	0.02	0.05	0.10	0.20
0.05	0.10	0.23	0.40	0.64
0.10	0.19	0.41	0.65	0.88

Но мы хотим ограничивать именно групповую ошибку. Что делать?

Идея — поменять α_1, т.е., уровень значимости для единичного эксперимента.

Напишем формулу, как задать α_1, чтобы получить $FWER = \alpha$.

Очевидно, ответ:

$$FWER = 1 - (1 - \alpha_1)^m = \alpha.$$

Следовательно,

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

Мало кто любит извлекать корень степени m, особенно это было сложно раньше, когда вычисления делались вручную.

Есть выход. Для малых x приближение для корня такое: $\sqrt[m]{1 - x} \approx 1 - x/m$. Тогда $\alpha_1 \approx \alpha/m$.

Поправка уровня значимости Šidak’a: $\alpha_1 = 1 - \sqrt[m]{1 - \alpha}$.
Поправка уровня значимости Бонферрони (Bonferroni): $\alpha_1 = \alpha/m$.

Однако неудобно менять уровень значимости. Гораздо удобнее менять p-value.

Одна гипотеза отвергается, если $p < \alpha_1$.

Получим эквивалентное неравенство ... $< \alpha$.

Поправка p-value Šidak’a: $1 - (1 - p)^m$ вместо p.
Поправка p-value Бонферрони (Bonferroni): mp вместо p.

Поправка Бонферрони очень удобна в использовании. По-прежнему, если ’поправленный p-value’ в гипотезе меньше заданного уровня значимости α, эта гипотеза отвергается.

3.2. Зависимые тесты, общий случай

Докажем, что поправка Бонферрони для p-value в каждом отдельном тесте приводит к тому, что групповая ошибка ограничена α.

Каждая $H^{(i)}$ проверяется отдельно с уровнем значимости α_1. Задача сводится к тому, чтобы найти такое α_1, что $\mathrm{FWER} \le \alpha$ для выбранного группового уровня значимости α. Имеем:

$$\mathrm{FWER} = \mathsf{P}\left(\bigcup_{i=1}^{m}\{H_0^{(i)} \text{ отв}\}\right) \le \sum_{i=1}^{m}\mathsf{P}(H^{(i)} \text{ отв}) = m\alpha_1 = \alpha \implies \alpha_1 := \frac{\alpha}{m}.$$

Таким образом, деля уровень значимости на число тестов или, наоборот, умножая p-value на число тестов, мы получаем консервативный тест.

Замечание. Из-за неравенства тест консервативный, т.е. может быть, что $\mathrm{FWER} \ll \alpha$. Значит, тест может быть маломощным.

Вопрос. В каком случае поправка Бонферрони приведет к максимально консервативному тесту? Если все тесты полностью зависимы (выдают одинаковые p-values).

Замечание. Есть множественный тест с гарантированной групповой ошибкой, у которого мощность больше, чем у теста с поправками Боферрони. Это тест Хольма (Holm). Увеличение мощности удается получить за счет того, что гипотезы проверяются не параллельно, а в определенном порядке, причем поправка на каждом шаге разная.

3.3. Множественное тестирование переходом к многомерному случаю

3.3.1. Одна группа, много признаков

До сих пор мы обсуждали, как проверить гипотезу, что $\mathsf{E}\xi = a_0$. Для этого использовался t-test в случае, если предполагали, что ξ имеет нормальное распределение, и его асимптотические вариант в общем случае. Статистика критерия

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

Доверительный интервал:

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

Предположим, что нам нужно проверить одновременно равенство каким-то значениям (например, нулю) сразу нескольких математических ожиданий для зависимых признаков. И пусть нас интересует именно групповая ошибка.

Есть обобщение t-test, который называется тестом Хотеллинга для гипотезы:

$$H_0 : \begin{pmatrix} \mathsf{E}\xi_1 \\ \mathsf{E}\xi_2 \\ \dots \\ \mathsf{E}\xi_p \end{pmatrix} = \mathbf{a}_0.$$

статистика имеет вид

$$T^2 = (\bar{\mathbf{x}} - \mathbf{a}_0)^{\mathrm{T}}\Big(\frac{\tilde{\mathbb{S}}}{n}\Big)^{-1}(\bar{\mathbf{x}} - \mathbf{a}_0) \xrightarrow[n\to\infty]{\sim} \chi^2(p),$$

где $\bar{\mathbf{x}}$ — вектор из выборочных средних, $\mathbb{S}$ — выборочная ковариационные матрицы.

Задание Где критическая область?

3.3.1.1. Доверительные интервалы/области

Этот подход распространяется на проверку гипотез о параметрах, например,

$$H_0 : \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \theta_1^{(0)} \\ \theta_2^{(0)} \end{pmatrix}.$$

В свою очередь, мы знаем, что проверку гипотез про параметры можно проводить на основе доверительных интервалов, попадает туда значение, предполагаемое в нулевой гипотезе, или нет.

Но также у нас речь шла о проблеме с множественным тестирование. Наверняка, как-то эта проблема должна отразиться и на доверительных интервалах (если параметр не одномерный, то говорят о доверительных областях).

Определение доверительной области:

$$\mathsf{P}\left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \in A(\mathbf{x})\right) = (\text{или } \geq)\ \gamma.$$

Пусть мы умеем строить доверительный интервал отдельно для θ_1 и отдельно для θ_2.

$$\mathsf{P}(\theta_i \in \left[b_{\text{low}}^{(i)}(\mathbf{x};\gamma), b_{\text{up}}^{(i)}(\mathbf{x};\gamma)\right]) = \gamma \text{ для любого } \gamma.$$

Задание Построить доверительную область с уровнем доверия γ для вектора из параметров

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}.$$

Сначала рассмотрите случай, когда доверительные интервалы независимые (их границы независимые для разных параметров случайные величины). А потом, если получится, попробуйте использовать поправки Бонферрони для общего случая.

В общем случае доверительный интервал на основе поправок Бонферрони получается слишком большим (вероятность для параметра попасть туда больше заданного уровня доверия). Опять же, в конкретных случаях можно строить (асимптотически) точные доверительные области. Например, в случае проверки гипотезы о значении вектора из математических ожиданий это будет эллипсоид.

3.3.2. Две группы, много признаков

Предположим, что нам нужно проверить одновременно равенство сразу нескольких математических ожиданий для зависимых признаков. Т.е. у нас есть две группы и нужно проверить гипотезу о том, что они в среднем не различаются.

Пример: есть две группы, кто занимается физическими упражнениями (например, фитнесом) и кто нет. Признаки — давление, пульс, вес, мешки под глазами... Гипотеза: все равно, заниматься или нет.

Пример: есть 2 группы людей. Снимаются показания по росту, длине ног, длине рук и т.д. Необходимо сравнить так называемые “средние размеры”.

Мы упоминали постановку задачи классификации. Чтобы строить правило, по которому группы различаются, нужно, чтобы они различались (иначе можно ооооочень долго строить такое правило).

Здесь нас интересует именно групповая ошибка, так как гипотеза, что все мат.ож. равны, а альтернатива — что хотя бы по одному из признаков мат.ож. не равны.

Статистика асимптотического варианта критерия в случае p признаков (многомерный аналог t-критерия, называется тестом Хотеллинга) получается переходом к гипотезе, что разность значений равна нулю:

$$T^2 = (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^{\mathrm{T}} \Big(\frac{\tilde{\mathbb{S}}_1}{n_1} + \frac{\tilde{\mathbb{S}}_2}{n_2} \Big)^{-1} (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) \xrightarrow[n_1, n_2 \to \infty]{\sim} \chi^2(p),$$

где $\bar{\mathbf{x}}^{(1)}$ и $\bar{\mathbf{x}}^{(1)}$ — вектора из выборочных средних в двум группам, $\mathbb{S}_1$ и $\mathbb{S}_2$ — выборочные ковариационные матрицы.

Замечание. Есть групп больше двух (пусть их k), то у теста если обобщение, которое называется one-way (M)ANOVA, ANalysis Of VAriance). В этом случае, если ковариационные матрицы в группах одинаковые, то можно проверять гипотезу, что вектора из мат.ож. равны для всех k групп.

Задание Пусть есть две группы и два признака. Поэтому мы можем выборку нарисовать на плоскости. Нарисуйте два случая (группы разными значками) — ковариационные матрицы одинаковые и ковариационные матрицы разные.

3.4. Post-hoc сравнения, много групп, один признак

В ANOVA рассматривается модель, когда $\mathcal{P}(\eta_i) = \mathcal{N}(\mu_i, \sigma^2)$, $i = 1, \ldots, k$. В рамках этой модели гипотеза о равенстве k распределений равносильна следующему:

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_k. \tag{3.1}$$

Это просто одна гипотеза, здесь нет множественных сравнений. Но пусть эта гипотеза отвергнулась. Какой вопрос сразу возникает?

Возникает вопрос, между какими группами есть разница.

Вопрос: А это сколько нужно сделать сравнений, чему равно число тестов, если групп k?

Если не обращать внимание на то, что проверяем сразу много тестов, а просто сравнивать каждое среднее с каждым по обычному t-test, то групповая ошибка будет большой. Такой тест называется **LSD (Least significant difference)**.

В этом случае тест будет радикальным. Шанс, что гипотеза про равенство средних для групп, где минимальное и максимальное выборочное среднее, отвергнется, гораздо больше, чем для двух случайно взятых групп.

Есть специальные тесты, с помощью которых можно построить точный тест (т.е. $FWER = \alpha$). Для равных объемов выборки в разных группах это **HSD (Honestly significant difference)** тест Тьюки (Tukey).

Какой бы честный тест не был, но его мощность будет падать при увеличении числа групп. Поэтому всегда стараются уменьшить число одновременно проверяемых тестов. Это можно сделать путем **плановых сравнений**. Для этого заранее (!), до эксперимента, определяются важные сравнения. Например, есть контрольная группа с плацебо и группы людей, которые принимают разные лекарства. Можно принять решения, что будут сравниваться только с группой плацебо.

Общее число сравнений — $k(k-1)/2$, а если сравнивать только с контрольной группой, то сравнений $k - 1$.

Пример-задание. Ниже представлены результаты тестов LSD и HSD (p-values).

multiple comparisons

```
## Warning: package 'agricolae' was built under R version 4.0.3

## [1] "means by groups"

##       yield      std r       LCL      UCL  Min  Max   Q25  Q50   Q75
## cc 24.40000 3.609709 3 18.086268 30.71373 21.7 28.5 22.35 23.0 25.75
## fc 12.86667 2.159475 3  6.552935 19.18040 10.6 14.9 11.85 13.1 14.00
## ff 36.33333 7.333030 3 30.019601 42.64707 28.0 41.8 33.60 39.2 40.50
## oo 36.90000 4.300000 3 30.586268 43.21373 32.1 40.4 35.15 38.2 39.30

## [1] "LSD test"

##          difference pvalue signif.        LCL        UCL
## cc - fc  11.5333333 0.0176       *   2.604368  20.462299
## cc - ff -11.9333333 0.0151       * -20.862299  -3.004368
## cc - oo -12.5000000 0.0121       * -21.428965  -3.571035
## fc - ff -23.4666667 0.0003     *** -32.395632 -14.537701
## fc - oo -24.0333333 0.0003     *** -32.962299 -15.104368
## ff - oo  -0.5666667 0.8873          -9.495632   8.362299

## [1] "HSD test"

##          difference pvalue signif.         LCL         UCL
## cc - fc  11.5333333 0.0686       .  -0.8663365  23.9330031
## cc - ff -11.9333333 0.0592       . -24.3330031   0.4663365
## cc - oo -12.5000000 0.0482       * -24.8996698  -0.1003302
## fc - ff -23.4666667 0.0014      ** -35.8663365 -11.0669969
## fc - oo -24.0333333 0.0012      ** -36.4330031 -11.6336635
## ff - oo  -0.5666667 0.9988         -12.9663365  11.8330031
```

На их основе напишите, что (какие сравнения) будет отвергнуто и что не отвергнуто при групповом уровне значимости $\alpha = 0.05$ и как относиться к результатам тестов (a) LSD, (b) HSD, (c) Все сравнения, с поправками Бонферрони (d) Плановые сравнения всех групп с контрольной, с поправками Бонферрони.

Задача — исследовать влияние картофельного вируса на вес выращенной картошки. FF — один вирус, CC — другой, FC — оба вируса сразу, OO — здоровый картофель (с ней и надо сравнивать в пункте (d)).

3.5. Одна группа, p признаков

Часто хочется проверить гипотезу о равенстве средних всех p признаков. Конечно, для этого нужно, чтобы признаки были на одну тему. Часто это измерение одного и того же в разные моменты времени.

Например, люди сидят на модной диете (когда можно есть сколько угодно, но не всё) и измеряют свой вес раз в месяц. Гипотеза: такая диета не влияет на вес:

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_p, \tag{3.2}$$

где $\mu_i = \mathsf{E}\xi_i$. Выглядит так же, как в случае ANOVA, но там один признак и много групп, а здесь одна группа и много признаков (называется Repeated Measures ANOVA).

Идея: свести к тому, что мы уже умеем делать (проверять гипотезу про многомерный вектор из мат.ожиданий критерием типа тест Хотеллинга, обобщение t-test.)

Эквивалентная гипотеза:

$$H_0 : \mathbf{C} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \ldots \\ \mu_p \end{pmatrix} = 0$$

где требуется, чтобы матрица контрастов $\mathbf{C}$ размера $p-1$ на p была полного ранга и сумма по строкам была равна нулю. Мы проверяем $p-1$ гипотез про то, что разные линейные комбинации (строка матрицы $\mathbf{C}$ состоит из коэффициентов одной линейной комбинации) равны нулю, причем сумма коэффициентов равна нулю.

Например, если мы хотим проверить, что $\mu_2 - \mu_1 = 0$, то коэффициенты (строка матрицы контрастов) будут выглядеть как $(-1, 1, 0, 0, \ldots, 0)$.

Разные матрицы контраста — разные критерии. Критерии отличаются чем? Они более мощные против разных альтернатив (хотя состоятельные против любого неравенства).

Задание: Напишите, какой матрице соответствует случай, когда (a) нам важно сравнить всё с первым моментом времени (b) важно сделать акцент на динамике, т.е. сравнивать последовательные моменты времени. Пусть число признаков $p = 4$.