Прикладной статистический анализ данных. 2. Параметрическая проверка гипотез.

Рябенко Евгений
riabenko.e@gmail.com

I/2015

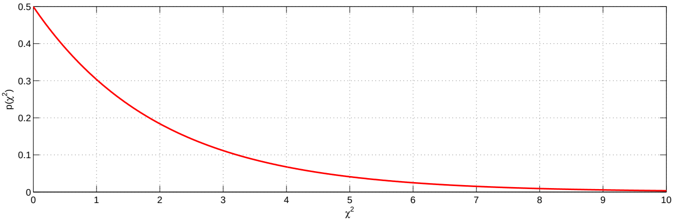
О нормальном распределении

Благодаря центральной предельной теореме и удобству вывода критериев для нормально распределённых выборок методы, основанные на предположении о нормальности данных, наиболее широко распространены.

- Перед использованием методов, предполагающих нормальность, стоит проверить нормальность.
- Если принять предположение о нормальности, то можно применять более мощные критерии. Зачастую они также чувствительны к небольшим отклонениям от нормальности.
- Если гипотеза нормальности отвергается, следует использовать непараметрические методы.

Критерий Харке-Бера

выборка: $X^n = (X_1, \ldots, X_n)$;
нулевая гипотеза: $H_0\colon X \sim N(\mu, \sigma^2)$;
альтернатива: $H_1\colon H_0$ неверна;
статистика: $\chi^2(X^n) = \frac{n}{6}\left(g_1^2 + \frac{1}{4}g_2^2\right)$;
$\chi^2(X^n) \sim \chi^2_2$ при H_0;



достигаемый уровень значимости:

$$p\left(\chi^2\right) = 1 - pchisq(\chi^2, 2).$$

Критерий Шапиро-Уилка

выборка: $X^n = (X_1, \ldots, X_n)$;

нулевая гипотеза: $H_0\colon X \sim N\left(\mu, \sigma^2\right)$;

альтернатива: $H_1\colon H_0$ неверна;

статистика: $$W\left(X^n\right) = \frac{\left(\sum\limits_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum\limits_{i=1}^{n}\left(X_i - \bar{X}\right)^2},$$

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{\left(m^T V^{-1} V^{-1} m\right)},$$

$m = (m_1, \ldots, m_n)^T$ — матожидания порядковых статистик $N(0,1)$, V — их ковариационная матрица; $W\left(X^n\right)$ при H_0 имеет табличное распределение.

Значения a_i также табулированы.

Критерий согласия Пирсона (хи-квадрат)

выборка: $X^n = (X_1, \dots, X_n)$;

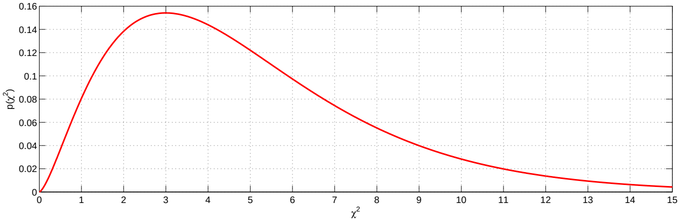
нулевая гипотеза: $H_0\colon X \sim N\left(\mu, \sigma^2\right)$;

альтернатива: $H_1\colon H_0$ неверна;

статистика: $\chi^2\left(X^n\right) = \sum\limits_{i=1}^{K} \frac{(n_i - np_i)^2}{np_i}$;

$$\chi^2\left(X^n\right) \sim \begin{cases} \chi^2_{K-1}, & \mu, \sigma \text{ заданы}, \\ \chi^2_{K-3}, & \mu, \sigma \text{ оцениваются} \end{cases} \text{ при } H_0;$$

$[a_i, a_{i+1}], i = 1, \dots, K$ — интервалы гистограммы,
n_i — число элементов выборки в $[a_i, a_{i+1}]$,
$p_i = F\left(a_{i+1}\right) - F\left(a_i\right)$ — вероятность попадания в i-й интервал при H_0.



Недостатки:

- разбиение на интервалы неоднозначно;
- требует больших выборок ($np_i > 5$ в 80% ячеек).

Критерии, основанные на эмпирической функции распределения

Ряд критериев согласия основаны на различиях между $F(x)$ и $F_n(x)$:

- Джини:

$$\int |F_n(x) - F(x)|\, dx;$$

- Крамера-фон Мизеса:

$$\int (F_n(x) - F(x))^2\, dx;$$

- Колмогорова (одновыборочный Колмогорова-Смирнова):

$$\sup_{-\infty<x<\infty} |F_n(x) - F(x)|\,;$$

- Смирнова-Крамера-фон Мизеса:

$$\int (F_n(x) - F(x))^2\, dF(x)\,;$$

Критерии, основанные на эмпирической функции распределения

- Андерсона-Дарлинга:

$$\int \frac{\left(F_n\left(x\right) - F\left(x\right)\right)^2}{F\left(x\right)\left(1 - F\left(x\right)\right)} dF\left(x\right);$$

- Купера:

$$\sup_{-\infty<x<\infty} \left(F_n\left(x\right) - F\left(x\right)\right) + \sup_{-\infty<x<\infty} \left(F\left(x\right) - F_n\left(x\right)\right);$$

- Ватсона:

$$\int \left(F_n\left(x\right) - F\left(x\right) - \int \left(F_n\left(x\right) - F\left(x\right)\right) dF\left(x\right)\right) dF\left(x\right);$$

- Фроцини:

$$\int \left|F_n\left(x\right) - F\left(x\right)\right| dF\left(x\right).$$

Предполагается, что $F\left(x\right)$ известна с точностью до параметров (если они оцениваются по выборке, нулевое распределение корректируется).

Критерий Колмогорова (Лиллиефорса)

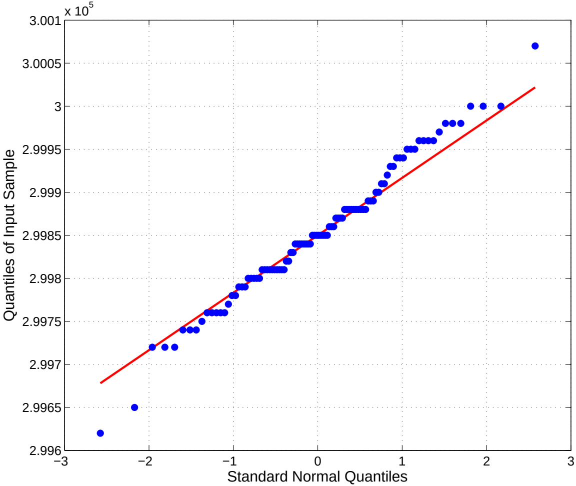
$$
\begin{aligned}
\text{выборка:} &\quad X^n = (X_1, \ldots, X_n)\,; \\
\text{нулевая гипотеза:} &\quad H_0 \colon X \sim N\left(\mu, \sigma^2\right); \\
\text{альтернатива:} &\quad H_1 \colon H_0 \text{ неверна}; \\
\text{статистика:} &\quad D\left(X^n\right) = \sup_{-\infty < x < \infty} \left|F_n(x) - \Phi(x)\right|; \\
&\quad D\left(X^n\right) \text{ при } H_0 \text{ имеет табличное распределение.}
\end{aligned}
$$

Недостатки:

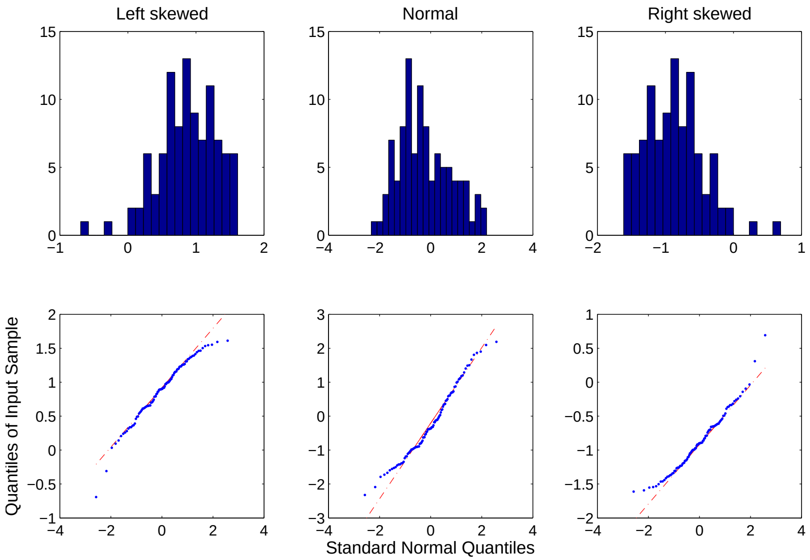
- имеет низкую мощность;
- не чувствителен к различиям на хвостах распределений.

Q-Q plot

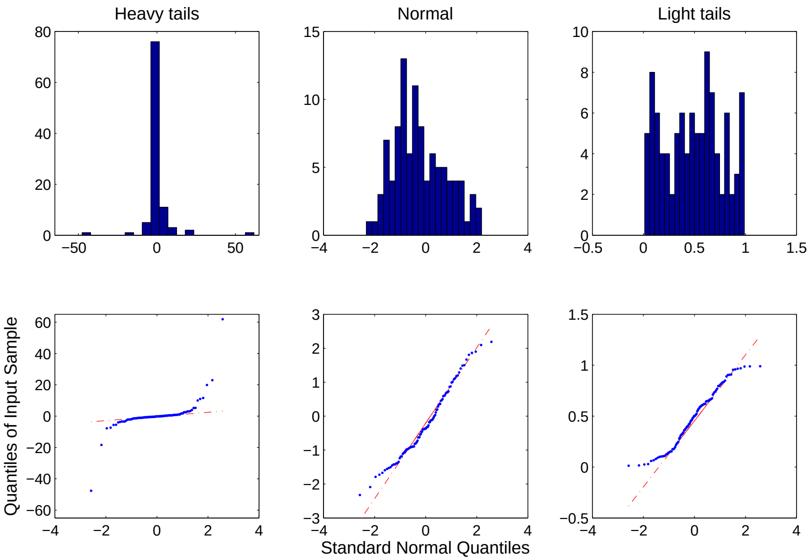
Визуальный метод проверки согласия выборки и распределения — q-q plot (для нормального распределения называется также normal probability plot)



Q-Q plot


Left skewed
Normal
Right skewed
Quantiles of Input Sample
Standard Normal Quantiles

Q-Q plot


Heavy tails
Normal
Light tails
Quantiles of Input Sample
Standard Normal Quantiles

Пример

Измерения скорости света:
https://yadi.sk/d/RWmzyOd3egbLe

Итого о проверке нормальности

- **очень маленькие выборки**: любой критерий может пропустить отклонения от нормальности, графические методы бесполезны;
- **очень большие выборки**: любой критерий может выявлять небольшие статистически, но не практически значимые отклонения от нормальности; значительная часть методов, предполагающих нормальность, демонстрируют устойчивость к отклонениям;
- **выбросы**: сильно влияют на выборочные коэффициенты асимметрии и эксцесса;
- **критерий Лиллиефорса**: представляет только исторический интерес;
- **критерий хи-квадрат**: слишком общий, не самый мощный, потеря информации из-за разбиения на интервалы.

Итого о проверке нормальности

Сравнение критериев проверки нормальности распределения случайных величин

Наименование критерия (раздел)	Характер альтернативного распределения					Ранг
	асимметричное		симметричное		≈ нормальное	
	$\alpha_4 < 3$	$\alpha_4 > 3$	$\alpha_4 < 3$	$\alpha_4 > 3$	$\alpha_4 \approx 3$	
Критерий Шапиро–Уилка (3.2.2.1)	1	1	3	2	2	1
Критерий K^2 (3.2.2.16)	7	8	10	6	4	2
Критерий Дарбина (3.1.2.7)	11	7	7	15	1	3
Критерий Д'Агостино (3.2.2.14)	12	9	4	5	12	4
Критерий α_4 (3.2.2.16)	14	5	2	4	18	5
Критерий Васичека (3.2.2.2)	2	14	8	10	10	6
Критерий Дэвида–Хартли–Пирсона (3.2.2.10)	21	2	1	9	1	7
Критерий χ^2 (3.1.1.1)	9	20	9	8	3	8
Критерий Андерсона–Дарлинга (3.1.2.4)	18	3	5	18	7	9
Критерий Филлибена (3.2.2.5)	3	12	18	1	9	10
Критерий Колмогорова–Смирнова (3.1.2.1)	16	10	6	16	5	11
Критерий Мартинеса–Иглевича (3.2.2.14)	10	16	13	3	15	12
Критерий Лина–Мудхолкара (3.2.2.13)	4	15	12	12	16	13
Критерий α_3 (3.2.2.16)	8	6	21	7	19	14
Критерий Шпигельхальтера (3.2.2.11)	19	13	11	11	8	15
Критерий Саркади (3.2.2.12)	5	18	15	14	13	16
Критерий Смирнова–Крамера–фон Мизеса (3.1.2.2)	17	11	20	17	6	17
Критерий Локка–Спурье (3.2.2.7)	13	4	19	21	17	18
Критерий Оя (3.2.2.8)	20	17	14	13	14	19
Критерий Хегази–Грина (3.2.2.3)	6	19	16	19	21	20
Критерий Муроты–Такеучи (3.2.2.17)	15	21	17	20	20	21

Кобзарь, 3.2.2.19, табл. 80.

Виды задач: одновыборочные

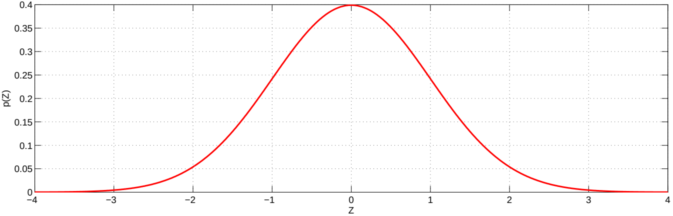
- $X^n \sim N(\mu, \sigma^2)$
  - $H_0\colon \mu = \mu_0$
    - σ известна 1
    - σ неизвестна 3
  - $H_0\colon \sigma = \sigma_0$ 2

Виды задач: двухвыборочные

- $X_1^{n_1} \sim N(\mu_1, \sigma_1^2), \quad X_2^{n_2} \sim N(\mu_2, \sigma_2^2)$
  - $H_0\colon \mu_1 = \mu_2$
    - X_1, X_2 независимые
      - $\sigma_1 = \sigma_2 = \sigma$
        - σ известна 4
        - σ неизвестна 5
      - $\sigma_1 \neq \sigma_2$
        - σ_1, σ_2 известны 6
        - σ_1, σ_2 неизвестны 7
    - X_1, X_2 связанные 8
  - $H_0\colon \sigma_1 = \sigma_2$ 9

Z-критерий

выборка: $X^n = (X_1, \dots, X_n)\,, X \sim N\left(\mu, \sigma^2\right)$,
σ известна;
нулевая гипотеза: $H_0\colon \mu = \mu_0$;
альтернатива: $H_1\colon \mu <\neq> \mu_0$;
статистика: $Z\left(X^n\right) = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$;
$Z\left(X^n\right) \sim N(0,1)$ при H_0;



достигаемый уровень значимости:

$$p\left(z\right) = \begin{cases} 1 - pnorm(z,0,1), & H_1\colon \mu > \mu_0, \\ pnorm(z,0,1), & H_1\colon \mu < \mu_0, \\ 2\left(1 - pnorm(|z|,0,1)\right), & H_1\colon \mu \neq \mu_0. \end{cases}$$

1 Z-критерий

Пример (Kanji, критерий 1): линия по производству пудры должна обеспечивать средний вес пудры в упаковке 4 грамма, заявленное стандартное отклонение — 1 грамм.
В ходе инспекции выбрано 9 упаковок, средний вес продукта в них составляет 4.6 грамма.

$H_0:$ средний вес пудры в упаковке соответствует норме.
$H_1:$ средний вес пудры в упаковке не соответствует норме $\Rightarrow p = 0.0719$, 95% доверительный интервал для среднего веса — $[3.95, 5.25]$ г.
$H_1:$ средний вес пудры в упаковке превышает норму $\Rightarrow p = 0.0359$, односторонний нижний 95% доверительный предел для среднего веса — 4.05 г.

Критерий хи-квадрат

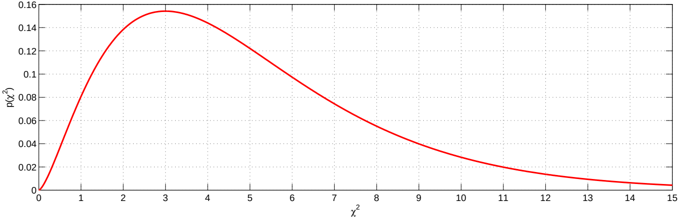
выборка: $X^n = (X_1, \ldots, X_n), X \sim N(\mu, \sigma^2);$

нулевая гипотеза: $H_0\colon \sigma = \sigma_0;$

альтернатива: $H_1\colon \sigma <\neq> \sigma_0;$

статистика: $\chi^2(X^n) = \frac{(n-1)S^2}{\sigma_0^2};$

$\chi^2(X^n) \sim \chi^2_{n-1}$ при $H_0;$



достигаемый уровень значимости:

$$p(\chi^2) = \begin{cases} 1 - pchisq(\chi^2, n-1), & H_1\colon \sigma > \sigma_0, \\ pchisq(\chi^2, n-1), & H_1\colon \sigma < \sigma_0, \\ 2\min\left(1 - pchisq(\chi^2, n-1), pchisq(\chi^2, n-1)\right), & H_1\colon \sigma \neq \sigma_0. \end{cases}$$

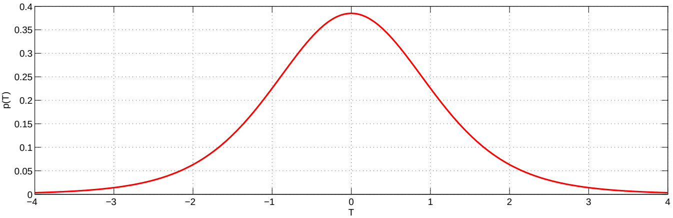
2 Критерий хи-квадрат

Пример (Kanji, критерий 15): при производстве микрогидравлической системы делается инъекция жидкости. Дисперсия объёма жидкости — критически важный параметр, установленный стандартом на уровне 9 кв. мл. В выборке из 25 микрогидравлических систем дисперсия объёма жидкости составляет 12 кв. мл.

$H_0:$ дисперсия объёма жидкости соответствует стандарту.
$H_1:$ дисперсия объёма жидкости не соответствует стандарту $\Rightarrow p = 0.254$, 95% доверительный интервал для дисперсии — $[7.3, 23.2]$ кв. мл.
$H_1:$ дисперсия объёма жидкости превышает допустимое значение $\Rightarrow p = 0.127$, односторонний нижний 95% доверительный предел — 7.9 кв. мл.

3 t-критерий Стьюдента

выборка: $X^n = (X_1, \ldots, X_n)\,, X \sim N\left(\mu, \sigma^2\right),$
σ неизвестна;
нулевая гипотеза: $H_0\colon \mu = \mu_0;$
альтернатива: $H_1\colon \mu <\neq> \mu_0;$
статистика: $T\left(X^n\right) = \frac{\bar{X} - \mu_0}{S/\sqrt{n}};$
$T\left(X^n\right) \sim St(n-1)$ при H_0.



С ростом объёма выборки разница между t- и z-критериями уменьшается.

t-критерий Стьюдента

Пример: в 1975 году с помощью лазерного интерферометра была получена оценка скорости света 299792458 м/с.
Насколько этому значению соответствуют данные эксперимента Майкельсона?

$H_0:$ оценки Майкельсона являются несмещёнными.
$H_1:$ оценки Майкельсона смещены $\Rightarrow p = 1.8 \times 10^{-11}$, 95% доверительный интервал для смещения — $[44.2, 75.6]$ км/с.
$H_1:$ оценки Майкельсона завышены $\Rightarrow p = 9.1 \times 10^{-12}$, односторонний нижний 95% доверительный предел для смещения — 46.8 км/с.

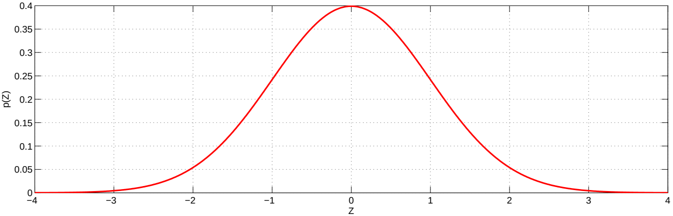
Z-критерий

выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), X_1 \sim N(\mu_1, \sigma^2)$,
$X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), X_2 \sim N(\mu_2, \sigma^2)$,
σ известна;

нулевая гипотеза: $H_0\colon \mu_1 = \mu_2$;

альтернатива: $H_1\colon \mu_1 <\neq> \mu_2$;

статистика: $Z(X_1^{n_1}, X_2^{n_2}) = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$;
$Z(X_1^{n_1}, X_2^{n_2}) \sim N(0,1)$ при H_0.



4 Z-критерий

Пример (Kanji, критерий 2): два отдела сбыта сравниваются по коэффициенту результативности при выполнении схожих операций. В первом отделе на 9 операциях среднее значение коэффициента результативности составило 1.2, во втором на 16 операциях — 1.7. Дисперсии коэффициента результативности в обоих отделах равны 2.075.

H_0: средняя результативность в обоих отделах одинакова.
H_1: средняя результативность в двух отделах различается $\Rightarrow p = 0.405$, 95% доверительный интервал для разности — $[-1.7, 0.7]$.

5 t-критерий Стьюдента

выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1})$, $X_1 \sim N(\mu_1, \sigma^2)$,
$X_2^{n_2} = (X_{21}, \dots, X_{2n_2})$, $X_2 \sim N(\mu_2, \sigma^2)$,
σ неизвестна;

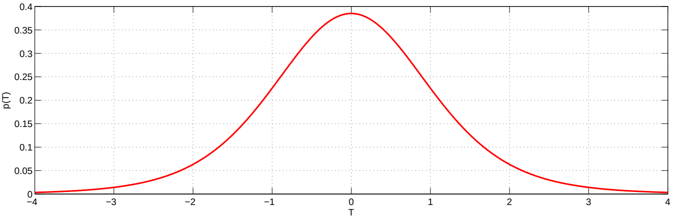
нулевая гипотеза: $H_0\colon \mu_1 = \mu_2$;

альтернатива: $H_1\colon \mu_1 <\neq> \mu_2$;

статистика: $T(X_1^{n_1}, X_2^{n_2}) = \dfrac{\bar{X}_1 - \bar{X}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$,

$S = \sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$;

$T(X_1^{n_1}, X_2^{n_2}) \sim St(n_1 + n_2 - 2)$ при H_0.



5 t-критерий Стьюдента

Пример (Kanji, критерий 8): чипсы продаются в тридцатиграммовых пакетах двух разновидностей. В выборке из 12 пачек каждого вида средние веса равны 31.75 г и 28.67 г, выборочные дисперсии — 112.25 г^2 и 66.64 г^2.

H_0: количество чипсов в упаковках двух разновидностей совпадает.
H_1: количество чипсов в упаковках двух разновидностей различается
$\Rightarrow p = 0.433$, 95% доверительный интервал для разности — $[-4.9, 11.1]$.

6 Z-критерий

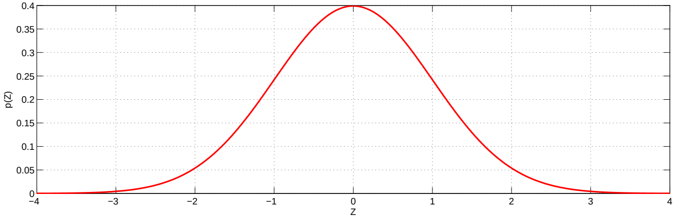
выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), X_1 \sim N(\mu_1, \sigma_1^2)$,
$X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), X_2 \sim N(\mu_2, \sigma_2^2)$,
σ_1, σ_2 известны;

нулевая гипотеза: $H_0\colon \mu_1 = \mu_2$;

альтернатива: $H_1\colon \mu_1 <\neq> \mu_2$;

статистика: $Z(X_1^{n_1}, X_2^{n_2}) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$;

$Z(X_1^{n_1}, X_2^{n_2}) \sim N(0, 1)$ при H_0.



6 Z-критерий

Пример (Kanji, критерий 3): известно, что одна из линий по расфасовке чипсов даёт упаковки с более вариабельным весом продукта, чем вторая. Дисперсии равны 0.000576 г^2 и 0.001089 г^2 соответственно, средние значения веса в выборках из 13 и 8 элементов — 80.02 г и 79.98 г.

$H_0:$ средний вес продукта в упаковках, произведённых на двух линиях, совпадает.
$H_1:$ средние веса продукта в упаковках, произведённых на двух линиях, различаются $\Rightarrow p = 0.001$, 95% доверительный интервал для разности — $[0.039, 0.041]$.

7 t-критерий Стьюдента / Аспина-Уэлша (проблема Беренца-Фишера)

выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), X_1 \sim N(\mu_1, \sigma_1^2)$,
$X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), X_2 \sim N(\mu_2, \sigma_2^2)$,
σ_1, σ_2 неизвестны;

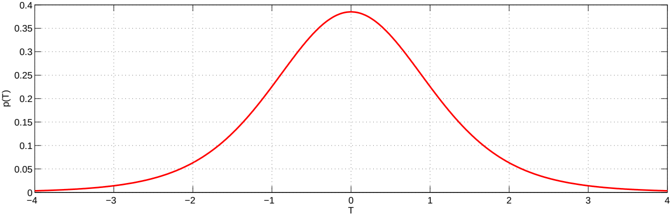
нулевая гипотеза: $H_0\colon \mu_1 = \mu_2$;

альтернатива: $H_1\colon \mu_1 <\neq> \mu_2$;

статистика: $T(X_1^{n_1}, X_2^{n_2}) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$,

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}};$$

$T(X_1^{n_1}, X_2^{n_2}) \approx\sim St(\nu)$ при H_0.



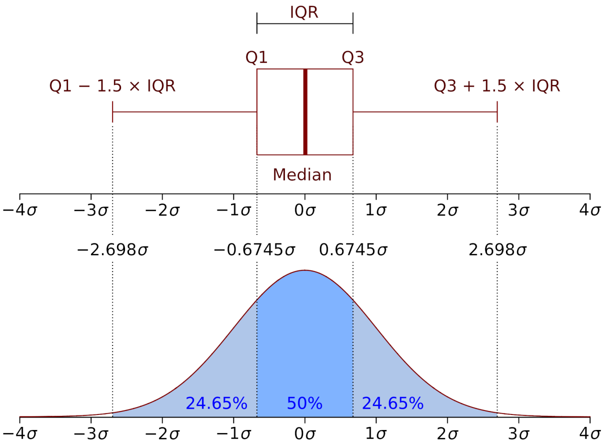
7 t-критерий Стьюдента / Аспина-Уэлша (проблема Беренца-Фишера)

Пример (Kanji, критерий 9): в связи со слиянием двух финансовых организаций решается вопрос о ликвидации отделов, выполняющих дублирующиеся функции. Рассматриваются две команды, занимающиеся сбытом похожих продуктов; первая продаёт 4 продукта, вторая — 9. Для каждого из продуктов рассчитывается уровень принесённой прибыли на одного работника за две недели, средние значения составляют 3166.0 и 2240.4, дисперсии — 6328.27 и 221661.3.

H_0: эффективность работы двух команд одинакова.
H_1: эффективность работы двух команд различна $\Rightarrow p = 1.342 \times 10^{-4}$,
95% доверительный интервал для разности — $[559.1245, 1292.075]$.

Boxplot

Ящик с усами — способ визуализации основных характеристик распределения:



Иногда дополняется доверительным интервалом для медианы и выбросами.

Пример

Продолжительность жизни крыс:
https://yadi.sk/d/Cp9yoQtRegfyy

8 t-критерий Стьюдента для связанных выборок

выборки: $X_1^n = (X_{11}, \dots, X_{1n})\,, X_1 \sim N\left(\mu_1, \sigma_1^2\right),$
$X_2^n = (X_{21}, \dots, X_{2n})\,, X_2 \sim N\left(\mu_2, \sigma_2^2\right),$
выборки связанные;

нулевая гипотеза: $H_0\colon \mu_1 = \mu_2;$

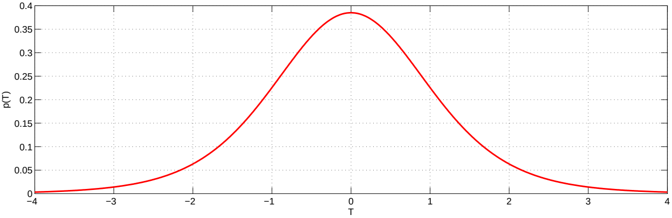
альтернатива: $H_1\colon \mu_1 <\neq> \mu_2;$

статистика: $T\left(X_1^n, X_2^n\right) = \frac{\bar{X}_1 - \bar{X}_2}{S/\sqrt{n}},$

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(D_i - \bar{D}\right)^2},$$

$D_i = X_{1i} - X_{2i};$

$T\left(X_1^n, X_2^n\right) \sim St(n-1)$ при H_0.



8 t-критерий Стьюдента для связанных выборок

Пример (Kanji, критерий 10): на 10 испытуемых сравниваются два лекарства против респираторного заболевания. Каждый из испытуемых вдыхает первое лекарство с помощью ингалятора, после чего проходит упражнение беговой дорожке. Измеряется время достижения максимальной нагрузки. Затем после периода восстановления эксперимент повторяется со вторым лекарством.

$H_0:$ время достижения максимальной нагрузки не отличается для исследуемых лекарств.

$H_1:$ время достижения максимальной нагрузки для исследуемых лекарств отличается $\Rightarrow p = 0.916$; 95% доверительный интервал для разницы — $[-2.1, 0.9]$.

Пример

Пусть имеются следующие связные выборки:

$$\begin{aligned}&X_1^n, \;\; X_1 \sim N(0,1),\\ &X_2^n, \;\; X_2 = X_1 + \varepsilon, \;\; \varepsilon \sim N(0.1, 0.25) \Rightarrow X_2 \sim N(0.1, 1.25);\end{aligned}$$

требуется оценить разность $\Delta = \mathbb{E}X_1 - \mathbb{E}X_2$.
Если попарные соответствия элементов известны, лучшая оценка $\hat{\Delta}_p = \frac{1}{n}\sum\limits_{i=1}^{n}(X_{1i} - X_{2i})$ имеет дисперсию

$$\mathbb{D}\hat{\Delta}_p = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{D}\left(X_{1i} - X_{2i}\right) = \frac{1}{n}\mathbb{D}\varepsilon = \frac{1}{2n};$$

мощность 0.8 достигается при $n \approx 200$.
Если же попарные соответствия неизвестны, лучшая оценка — $\hat{\Delta}_i = \bar{X}_1 - \bar{X}_2$; её дисперсия:

$$\mathbb{D}\hat{\Delta}_i = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{D}X_{1i} + \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{D}X_{2i} = \frac{1}{n} + \frac{5}{4n} = \frac{9}{4n}$$

— в 4.5 раза больше; мощность 0.8 достигается при $n \approx 1900$.

Пример

Метилфенидат и способность к подавлению импульсивных поведенческих реакций:
https://yadi.sk/d/vH-hmNNmehR6Z

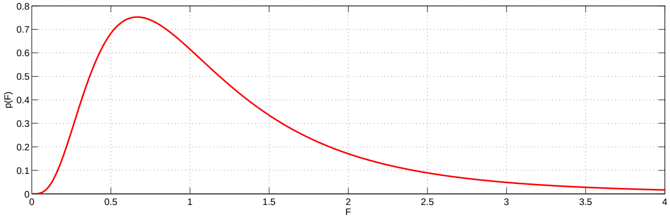
9 F-критерий Фишера

выборки: $X_1^{n_1} = \left(X_{11}, \ldots, X_{1n_1}\right), X_1 \sim N\left(\mu_1, \sigma_1^2\right),$
$X_2^{n_2} = \left(X_{21}, \ldots, X_{2n_2}\right), X_2 \sim N\left(\mu_2, \sigma_2^2\right);$

нулевая гипотеза: $H_0\colon \sigma_1 = \sigma_2;$

альтернатива: $H_1\colon \sigma_1 <\neq> \sigma_2;$

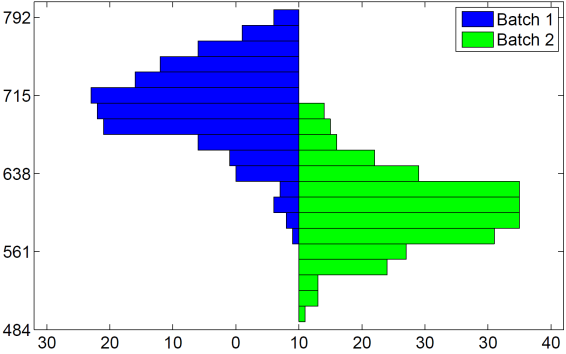
статистика: $F\left(X_1^{n_1}, X_2^{n_2}\right) = \frac{S_1^2}{S_2^2};$
$F\left(X_1^{n_1}, X_2^{n_2}\right) \sim F(n_1 - 1, n_2 - 1)$ при H_0.



Критерий Фишера неустойчив к отклонениям от нормальности даже асимптотически.

F-критерий Фишера

Пример (NIST/industry ceramics consortium for strength optimization of ceramic, 1996): собраны данные о прочности материала 440 керамических изделий из двух партий по 220 в каждой.
Одинакова ли дисперсия прочности в разных партиях?



Гипотезы нормальности не отклоняются критерием Шапиро-Уилка ($p_1 = 0.2062, p_2 = 0.7028$).
Критерий Фишера: $p = 0.1721$, $[C_L, C_U] = [0.9225, 1.5690]$.

Теоретическая база

$$X^n = (X_1, \ldots, X_n), \; X \sim Ber(p), \; y \equiv \sum_{i=1}^{n} X_i.$$

Критерии для проверки гипотез о p основаны на ММП:

$$L(p) = y \ln p + (n - y) \ln (1 - p),$$

$$l(p) \equiv -\mathbb{E} \frac{\partial^2 L}{\partial p^2} = \frac{n}{p(1-p)},$$

$$\hat{p} = \frac{y}{n}, \; \mathbb{E}\hat{p} = p, \; \mathbb{D}\hat{p} = \sqrt{\frac{p(1-p)}{n}}.$$

ММП порождает критерии МП, Вальда и множителей Лагранжа:

$$Z_{MLE} = -2\left(L(p_0) - L(\hat{p})\right) \sim \chi_1^2;$$

$$Z_W = \frac{\hat{p} - p_0}{\sqrt{1/l(\hat{p})}} = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1);$$

$$Z_{LM} = \frac{\left.\frac{\partial L}{\partial p}\right|_{p=p_0}}{\sqrt{l(p_0)}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1).$$

Z-критерий для доли

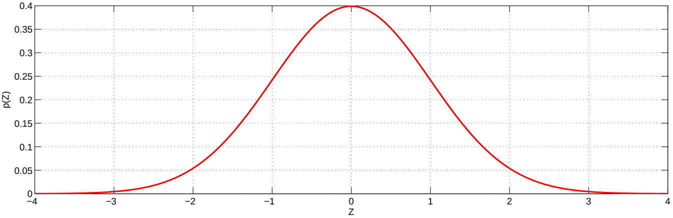
выборка: $X^n = (X_1, \dots, X_n), X \sim Ber(p);$

нулевая гипотеза: $H_0\colon p = p_0;$

альтернатива: $H_1\colon p <\neq> p_0;$

статистика: $Z(X^n) = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$

$Z(X^n) \sim N(0,1)$ при H_0.



Выборочное распределение статистики критерия множителей Лагранжа ближе к стандартному нормальному, чем критерия Вальда.

Z-критерий для доли

Пример 1 (Королёв, задача 7.2.2): Бенджамин Спок, знаменитый педиатр и автор большого количества книг по воспитанию детей, был арестован за участие в антивоенной демонстрации в Бостоне. Его дело должен был рассматривать суд присяжных. Присяжные назначаются с помощью многоступенчатой процедуры, на очередном этапе которой было отобрано 300 человек. Однако среди них оказалось только 90 женщин. Адвокаты доктора Спока подали протест на предвзятость отбора.

$H_0:$ процедура отбора была беспристрастной, женщины попадали в выборку с вероятностью $1/2$.
$H_1:$ предпочтение отдавалось кандидатам-мужчинам $\Rightarrow p = 2.3 \times 10^{-12}$.

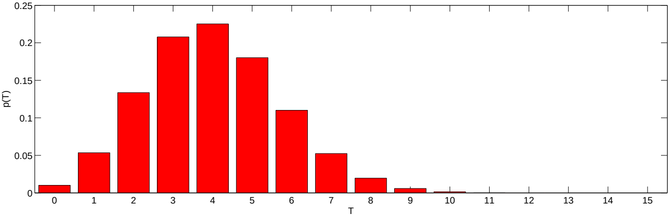
Z-критерий для доли

Пример 2 (Кобзарь, задача 227): нормируемый уровень дефектных изделий в партии $p_0 = 0.05$. Из партии извлечена выборка $n = 20$ изделий, в которой обнаружены при проверке $t = 2$ дефектных.

$H_0:$ доля дефектных изделий в партии не превосходит нормируемого значения.
$H_1:$ доля дефектных изделий в партии превышает нормируемое значение $\Rightarrow p = 0.15$.

Точный критерий для доли

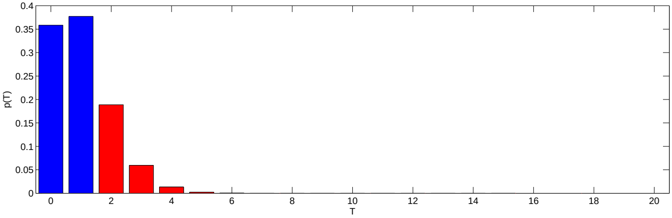
выборка: $X^n = (X_1, \dots, X_n)\,, X \sim Ber\,(p)\,;$
нулевая гипотеза: $H_0\colon p = p_0;$
альтернатива: $H_1\colon p <\neq> p_0;$
статистика: $T\,(X^n) = y;$
$T\,(X^n) \sim Bin(n, p_0)$ при H_0.



Поскольку нулевое распределение дискретно, нельзя добиться, чтобы вероятность ошибки первого рода была равна в точности α.
Критерий консервативен — истинная вероятность ошибки первого рода ограничена уровнем значимости сверху.

Точный критерий для доли

Пример 2:



$H_0:$ доля дефектных изделий в партии не превосходит нормируемого значения.

$H_1:$ доля дефектных изделий в партии превышает нормируемое значение $\Rightarrow p = 0.26$.

Доверительные интервалы для доли

$100\,(1-\alpha)\%$ доверительный интервал Вальда:

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

В примере 1 95% доверительный интервал Вальда — $[0.2481, 0.3519]$.
В примере 2 — $[-0.0315, 0.2315]$.

Недостатки:

- доверительные пределы могут выходить за границы $[0,1]$ (вообще, при $\hat{p} \in (0,1)$ нежелательно даже $C_L = 0$ и $C_U = 1$);
- при $\hat{p} = 0$ и $\hat{p} = 1$ вырождается в точку;
- антиконсервативен — накрывает p реже, чем в $100\,(1-\alpha)\%$ случаев.

Доверительные интервалы для доли

Более точный доверительный интервал Уилсона (score confidence interval):

$$\left[\hat{p}\left(\frac{n}{n+z}\right)+\frac{1}{2}\left(\frac{z}{n+z}\right)\right]\pm z\sqrt{\frac{1}{n+z}\left[\hat{p}\left(1-\hat{p}\right)\left(\frac{n}{n+z}\right)+\frac{1}{2}\frac{1}{2}\left(\frac{z}{n+z}\right)\right]},$$

$z \equiv z^2_{1-\frac{\alpha}{2}}$.
Центр интервала — $\frac{y+z/2}{n+z}$.

В примере 1 95% доверительный интервал Уилсона — $[0.2509, 0.3541]$.
В примере 2 — $[0.0279, 0.3010]$.

Z-критерий для разности двух долей, независимые выборки

выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), X_1 \sim Ber(p_1)$;
$X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), X_2 \sim Ber(p_2)$, выборки независимы;

Исход \ Выборка	$X_1^{n_1}$	$X_2^{n_2}$
1	a	b
0	c	d
$\sum$	n_1	n_2

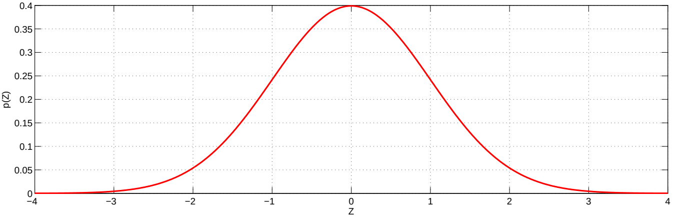
$$p_1 = \frac{\mathbb{E}A}{n_1}, \quad \hat{p}_1 = \frac{a}{n_1}, \quad p_2 = \frac{\mathbb{E}B}{n_2}, \quad \hat{p}_2 = \frac{b}{n_2};$$

нулевая гипотеза: $H_0\colon p_1 = p_2$;

альтернатива: $H_1\colon p_1 <\neq> p_2$;

статистика: $Z\left(X_1^{n_1}, X_2^{n_2}\right) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad P = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2};$

$Z\left(X_1^{n_1}, X_2^{n_2}\right) \sim N(0,1)$ при H_0.



Z-критерий для разности двух долей, независимые выборки

Пример (Кобзарь, задача 226): в двух партиях объёмами $n_1 = 100$ шт. и $n_2 = 200$ шт. обнаружено соответственно $t_1 = 3$ и $t_2 = 5$ дефектных приборов. Необходимо проверить гипотезу о равенстве долей дефектных приборов в партиях.

Наличие дефекта \ Номер партии	1	2
Есть	$a = 3$	$b = 5$
Нет	$c = 97$	$d = 195$
Всего	$n_1 = 100$	$n_2 = 200$

H_0: доли дефектных изделий в партиях равны.
H_1: доли дефектных изделий в партиях различаются $\Rightarrow p = 0.8$.
H_1: доля дефектных изделий в первой партии выше $\Rightarrow p = 0.4$.
H_1: доля дефектных изделий в первой партии ниже $\Rightarrow p = 0.6$.

Доверительный интервал для разности двух долей

Доверительный интервал Уилсона:

$$[C_L, C_U] = \left[\hat{p}_1 - \hat{p}_2 - \delta, \hat{p}_1 - \hat{p}_2 + \varepsilon\right],$$
$$\delta = z_{1-\frac{\alpha}{2}} \sqrt{\frac{l_1(1-l_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}},$$
$$\varepsilon = z_{1-\frac{\alpha}{2}} \sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{l_2(1-l_2)}{n_2}},$$

l_1, u_1 — корни уравнения $|x - \hat{p}_1| = z_{1-\frac{\alpha}{2}} \sqrt{\frac{x(1-x)}{n_1}}$,
l_2, u_2 — корни уравнения $|x - \hat{p}_2| = z_{1-\frac{\alpha}{2}} \sqrt{\frac{x(1-x)}{n_2}}$.

В примере 95% доверительный интервал — $[-0.0331, 0.0616]$, минимальное значение α, при котором интервал не содержит нуля — 0.8003.

Z-критерий для разности двух долей, связанные выборки

выборки: $X_1^n = (X_{11}, \dots, X_{1n}), X_{1i} \sim Ber(p_1);$
$X_2^n = (X_{21}, \dots, X_{2n}), X_{2i} \sim Ber(p_2)$, выборки связанные;

$X_1^n \backslash X_2^n$	1	0	$\sum$
1	e	g	$e+g$
0	f	h	$f+h$
$\sum$	$e+f$	$g+h$	n

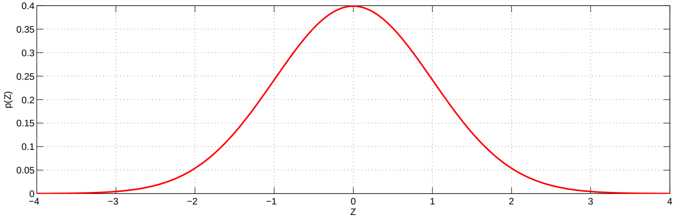
$$p_1 = \frac{\mathbb{E}(E+F)}{n}, \quad \hat{p}_1 = \frac{e+f}{n}, \quad p_2 = \frac{\mathbb{E}(E+G)}{n}, \quad \hat{p}_2 = \frac{e+g}{n};$$

нулевая гипотеза: $H_0\colon p_1 = p_2;$

альтернатива: $H_1\colon p_1 <\neq> p_2;$

статистика: $Z\left(X_1^n, X_2^n\right) = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{f+g-(f-g)^2}{n^3}}} = \dfrac{f-g}{\sqrt{f+g-\frac{(f-g)^2}{n}}};$

$Z\left(X_1^n, X_2^n\right) \sim N(0,1)$ при H_0.



Z-критерий для разности двух долей, связанные выборки

Пример (Agresti, табл. 10.1): из опрошенных 1600 граждан Великобритании, имеющих право голоса, 944 высказали одобрение деятельности премьер-министра. Через 6 месяцев эти же люди были опрошены снова, на этот раз одобрение высказали только 880 опрошенных.

I \ II	+	-	$\sum$
+	$e = 794$	$g = 150$	944
-	$f = 86$	$h = 570$	656
$\sum$	880	720	1600

H_0: рейтинг премьер-министра не изменился.
H_1: рейтинг премьер-министра изменился $\Rightarrow p = 2.8 \times 10^{-5}$.
H_1: рейтинг премьер-министра снизился $\Rightarrow p = 1.4 \times 10^{-5}$.
H_1: рейтинг премьер-министра повысился $\Rightarrow p = 0.999986$.

Z-критерий для разности двух долей, связанные выборки

Без учёта информации о связи между выборками:

Результат \ Опрос	I	II
+	$a = 994$	$b = 880$
-	$c = 606$	$d = 720$
$\sum$	$n_1 = 1600$	$n_2 = 1600$

H_0: рейтинг премьер-министра не изменился.
H_1: рейтинг премьер-министра изменился $\Rightarrow p = 4.3 \times 10^{-5}$.
H_1: рейтинг премьер-министра снизился $\Rightarrow p = 2.1 \times 10^{-5}$.
H_1: рейтинг премьер-министра повысился $\Rightarrow p = 0.999979$.

Доверительный интервал для разности двух долей

Доверительный интервал Уилсона:

$$
\begin{aligned}
[C_L, C_U] &= \left[\hat{p}_1 - \hat{p}_2 - \delta, \hat{p}_1 - \hat{p}_2 + \varepsilon\right], \\
\delta &= \sqrt{dl_1^2 - 2\hat{\phi} dl_1 du_2 + du_2^2}, \\
\varepsilon &= \sqrt{du_1^2 - 2\hat{\phi} du_1 dl_2 + dl_2^2}, \\
\hat{\phi} &= \begin{cases} \frac{eh-fg}{(e+f)(g+h)(e+h)(f+h)}, & \text{если знаменатель не равен нулю,} \\ 0, & \text{иначе;} \end{cases} \\
dl_1 &= \hat{p}_1 - l_1, \\
du_1 &= u_1 - \hat{p}_1, \\
dl_2 &= \hat{p}_2 - l_2, \\
du_2 &= u_2 - \hat{p}_2,
\end{aligned}
$$

l_1, u_1 — корни уравнения $\left|x - \hat{p}_1\right| = z_{1-\frac{\alpha}{2}}\sqrt{\frac{x(1-x)}{n}}$,

l_2, u_2 — корни уравнения $\left|x - \hat{p}_2\right| = z_{1-\frac{\alpha}{2}}\sqrt{\frac{x(1-x)}{n}}$.

В примере 95% доверительный интервал — $[0.0214, 0.0590]$, минимальное значение α, при котором интервал не содержит нуля — 3.1×10^{-5}.

Примеры

«Разрушители легенд» и обратная сторона руки:
https://yadi.sk/d/-chv8FwqehNVm

Качество бинарных классификаторов (для самостоятельной работы):
https://yadi.sk/i/lRYBvUcyejVbT

Литература

Критерии нормальности:

- Харке-Бера (Jarque–Bera) — Кобзарь, 3.2.2.16;
- Шапиро-Уилка (Shapiro-Wilk) — Кобзарь, 3.2.2.1;
- хи-квадрат (chi-square) — Кобзарь, 3.1.1.1, 3.2.1.1;
- согласия (goodness-of-fit), основанные на эмпирической функции распределения — Кобзарь, 3.1.2, 3.2.1.2.

Для нормальных распределений:

- Z-критерии (Z-tests) — Kanji, №№ 1, 2, 3;
- t-критерии Стьюдента (t-tests) — Kanji, №№ 7, 8, 9;
- критерий хи-квадрат (chi-square test) — Kanji, №15;
- критерий Фишера (F-test) — Kanji, №16.

Литература

Для распределения Бернулли:

- всё про одновыборочную задачу — Agresti, 1.3, 1.4;
- Z-критерии (Z-tests) — Kanji, №№ 4, 5;
- точный критерий (exact binomial test) — McDonald, `http://www.biostathandbook.com/exactgof.html`;
- доверительные интервалы Уилсона (score confidence intervals) — Newcombe, 1998a, 1998b, 1998c.

Кобзарь А.И. *Прикладная математическая статистика*. — М.: Физматлит, 2006.
Kanji G.K. *100 statistical tests*. — London: SAGE Publications, 2006.
Agresti A. *Categorical Data Analysis*. — Hoboken: Wiley, 2013.
McDonald J.H. *Handbook of Biological Statistics*. — Baltimore: Sparky House Publishing, 2008.
Newcombe R.G. (1998). *Two-sided confidence intervals for the single proportion: comparison of seven methods*. Statistics in Medicine, 17, 857–872.
Newcombe R.G. (1998). *Improved confidence intervals for the difference between binomial proportions based on paired data*. Statistics in Medicine, 17, 2635–2650.
Newcombe R.G. (1998). *Interval estimation for the difference between independent proportions: comparison of eleven methods*. Statistics in Medicine, 17, 873–890.

Литература

Королёв В.Ю. *Теория вероятностей и математическая статистика*. — М.: Проспект, 2008.

NIST/SEMATECH. *e-Handbook of Statistical Methods*.

`http://www.itl.nist.gov/div898/handbook/`