Прикладной статистический анализ данных

Проверка гипотез. Параметрические и непараметрические критерии

Чупраков Д. В.

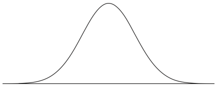
Проверка гипотез

Нулевая гипотеза H_0

Конкурирующая гипотеза H_1

Проверка гипотез

Основные понятия



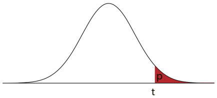
выборка: $X^n = (X_1, \ldots, X_n)\,,\ X \sim \mathbf{P} \in \Omega$

нулевая гипотеза: $H_0\colon\ \mathbf{P} \in \omega,\quad \omega \in \Omega$

альтернативная гипотеза $H_1\colon\ \mathbf{P} \notin \omega$

статистика: $T(X^n)\,,\quad T(X^n) \sim F(x)$ при $\mathbf{P} \in \omega$

$T(X^n) \nsim F(x)$ при $\mathbf{P} \notin \omega$



Реализация выборки: $x^n = (x_1, \ldots, x_n)$

Наблюдаемая статистика: $t = T(x^n)$

Достигаемый уровень значимости: $p(x^n)$ — вероятность при H_0 получить $T(X^n) = t$ или более экстремальное

$$p(x^n) = \mathbf{P}(T \geqslant t \,|\, H_0)$$

Гипотеза отвергается при $p(x^n) \leqslant \alpha,\ \ \alpha$ — уровень значимости

Ошибки I и II рода

матрица ошибок

	H_0 верна	H_0 неверна
H_0 принимается	 верно принята	 Ошибка II рода (False negative)
H_0 отвергается	 Ошибка I рода (False positive)	 H_0 верно отвергнута

Ошибки I и II рода

Свойства ошибок

Задача проверки гипотез несимметрична относительно пары (H_0, H_1): вероятность ошибки первого рода ограничивается сверху величиной α, а второго рода — минимизируется путём выбора критерия.

- Корректный критерий: $\mathbf{P}\left(p\left(T\right) \leqslant \alpha \,|H_0\right) \leqslant \alpha \;\forall\, \mathbf{P} \in \Omega$.
- Мощность: $\mathrm{pow} = \mathbf{P}\left(p\left(T\right) \leqslant \alpha \,|H_1\right)$.
- Состоятельный критерий: $\mathrm{pow} \to 1$ для всех альтернатив H_1 при $n \to \infty$.
- T_1 — равномерно наиболее мощный критерий, если $\forall T_2$

$$
\begin{aligned}
&\mathbf{P}\left(p\left(T_1\right) \leqslant \alpha \,|H_1\right) \geqslant \mathbf{P}\left(p\left(T_2\right) \leqslant \alpha \,|H_1\right) \quad \forall H_1 \neq H_0,\\
&\mathbf{P}\left(p\left(T_1\right) \leqslant \alpha \,|H_0\right) = \mathbf{P}\left(p\left(T_2\right) \leqslant \alpha \,|H_0\right),
\end{aligned}
$$

причём хотя бы для одной H_1 неравенство строгое.

Интерпретация результата Absence of evidence $\not\Rightarrow$ evidence of absence.

- Если величина p достаточно мала, то данные свидетельствуют против нулевой гипотезы в пользу альтернативы Гипотезу H_0 можно отвергнуть
- Если величина p недостаточно мала, то данные не свидетельствуют против нулевой гипотезы в пользу альтернативы. Нет причин отвергать гипотезу H_0

При помощи инструмента проверки гипотез нельзя доказать справедливость нулевой гипотезы!

Статистическая и практическая значимость

- Выбранная статистика может отражать не всю информацию, содержащуюся в выборке:
$$H_0\colon X \sim N\left(\mu,\sigma^2\right), \qquad H_1\colon \not\sim N\left(\mu,\sigma^2\right)$$
$$T\left(X^n\right) = \operatorname{skew}(X^n).$$
Все симметричные распределения будут признаны нормальными!
- Вероятность отвергнуть нулевую гипотезу зависит не только от того, насколько она отличается от истины, но и от размера выборки.
- По мере увеличения n могут выявиться более тонкие несоответствия выборки гипотезе H_0, и она будет отвергнута.

При любой проверке гипотез нужно:

- оценивать размер эффекта — степень отличия нулевой гипотезы отличается от истины,
- оценивать его практическую значимость.

Статистическая и практическая значимость
Примеры

- За 3 года женщины, упражнявшиеся не меньше 1 часа в день, набрали значимо меньше веса, чем женщины, упражнявшиеся меньше 20 минут в день ($p < 0.001$). (*Lee et al, 2010*)
  - Разница в набранном весе составила 150 г.
  - Практическая значимость такого эффекта сомнительна.
- В 2002 году клинические испытания гормонального препарата Премарин были досрочно прерваны. (*Ellis, 2010, гл. 2*)
  - обнаружено, что его приём ведёт к значимому увеличению риска развития рака груди на 0.08%, риска инсульта на 0.08% и инфаркта на 0.07%.
- если при испытании гипотетического лекарства, позволяющего замедлить прогресс ослабления интеллекта больных Альцгеймером, оказывается, что разница в IQ контрольной и тестовой групп составляет 13 пунктов, возможно, изучение лекарства стоит продолжить, даже если эта разница статистически незначима. (*Kirk, 1996*)

Гипотезы о числовых характеристиках

Гипотезы вида $H_0\colon \theta = \theta_0$ можно проверять при помощи доверительных интервалов для θ:

- если θ_0 не попадает в $(1-\alpha)$ доверительный интервал для θ, то H_0 отвергается на уровне значимости α;
- p-value — максимальное α, при котором θ_0 попадает в соответствующий доверительный интервал.

Пример: Shaken, not stirred

Джеймс Бонд говорит, что предпочитает мартини взболтанным, но не смешанным.

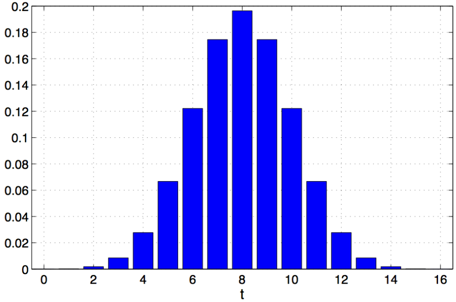
- Проведём слепой тест: n раз предложим ему пару напитков и выясним, какой из двух он предпочитает.
- Выборка: бинарный вектор длины n, 1 — Джеймс Бонд предпочёт взболтанный, 0 — смешанный.
- Нулевая гипотеза: Джеймс Бонд не различает два вида мартини.

- Статистика T — число единиц в выборке.

Нулевое распределение

Если нулевая гипотеза справедлива, то равновероятны все выборки длины n из нулей и единиц.

Пусть $n = 16$, тогда существует $2^{16} = 65536$ равновероятных варианта. Статистика T принимает значения от 0 до 16:

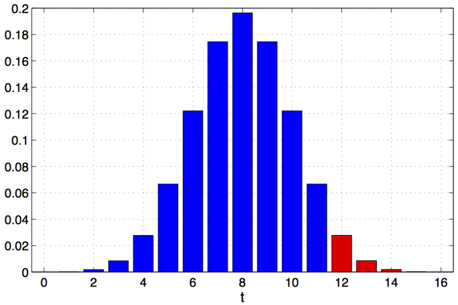


Односторонняя альтернатива

H_1: Джеймс Бонд предпочитает взболтанный мартини.

При справедливости такой альтернативы более вероятны большие значения T (большие значения T свидетельствуют против H_0 в пользу H_1).

Вероятность того, что Джеймс Бонд предпочтёт взболтанный мартини в 12-и или более случаях из 16 при справедливости H_0, равна $\frac{2517}{65536} \approx 0.0384$.



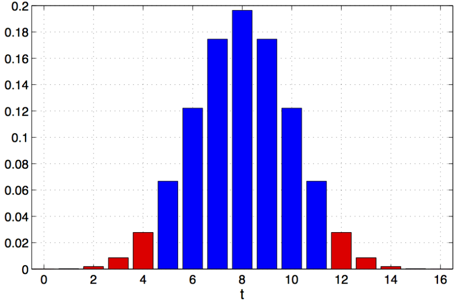
0.0384 — достигаемый уровень значимости при реализации $t = 12$.

Двусторонняя альтернатива

H_1: Джеймс Бонд предпочитает какой-то определённый вид мартини.

При справедливости такой альтернативы и большие, и маленькие значения T свидетельствуют против H_0 в пользу H_1).

Вероятность того, что Джеймс Бонд предпочтёт взболтанный мартини в $\geqslant 12$ случаях из 16 при справедливости H_0, равна $\frac{5034}{65536} \approx 0.0768$.



0.0768 — достигаемый уровень значимости при реализации $t = 12$.

Достигаемый уровень значимости

Чем ниже достигаемый уровень значимости, тем сильнее данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

0.0384 — вероятность реализации $t \geqslant 12$ при условии, что нулевая гипотеза справедлива, т. е. Джеймс Бонд выбирает мартини наугад.

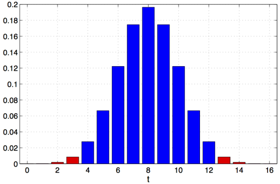
Пример: Допустим Джеймс Бонд выбирает взболтанный мартини в 51% случаев. Однако, по итогам 100 испытаний взболтанный мартини был выбран 49 раз.

Достигаемый уровень значимости против односторонней альтернативы — $p \approx 0.6178$.

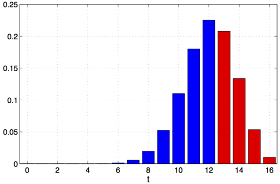
Нулевая гипотеза не отвергается.

Мощность

Проверяя нулевую гипотезу против двусторонней альтернативы, мы отвергаем H_0 при $t \geqslant 13$ или $t \leqslant 3$, что обеспечивает достигаемый уровень значимости $p = 0.0213 \leq \alpha = 0.05$.



Пусть Джеймс Бонд выбирает взболтанный мартини в 75% случаев.



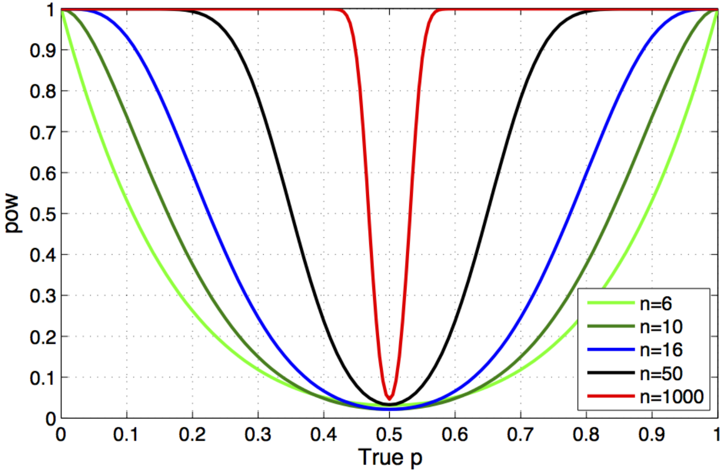
$\mathrm{pow} \approx 0.6202$, т.е. при многократном повторении эксперимента гипотеза будет

Мощность

Мощность критерия зависит от следующих факторов:

- размер выборки
- размер отклонения от нулевой гипотезы
- чувствительность статистики критерия
- тип альтернативы

Мощность



Размер выборки

Особенности прикладной задачи: 1 порция мартини содержит 55 мл джина и 15 мл вермута — суммарно около 25 мл спирта. Смертельная доза алкоголя при массе тела 80 кг составляет от 320 до 960 мл спирта в зависимости от толерантности (от 13 до 38 мартини).

Обеспечение требуемой мощности: размеры выборки подбирается так, чтобы при размере отклонения от нулевой гипотезы не меньше заданного (например, вероятность выбора взболтанного мартини не меньше 0.75) мощность была не меньше заданной.

Параметрические критерии проверки гипотез

Гипотезы о параметрах нормальных выборок

Одна выборка:

$X^n \sim N(\mu, \sigma^2)$

- $H_0\colon \sigma = \sigma_0$ 2
- $H_0\colon \mu = \mu_0$
  - σ неизвестна 3
  - σ известна 1

Две выборки:

$X_1^{n_1} \sim N(\mu_1, \sigma_1^2),$
$X_2^{n_2} \sim N(\mu_2, \sigma_2^2)$

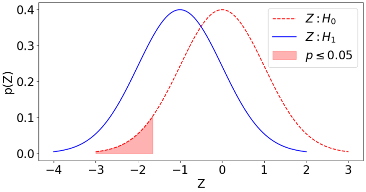
- $H_0\colon \sigma_1 = \sigma_0$ 7
- $H_0\colon \mu_1 = \mu_2$
  - X_1, X_2 зависимые 6
  - X_1, X_2 независимые
    - σ_1, σ_2 неизвестна 5
    - σ_1, σ_2 известны 4

Z-критерий Стьюдента: $\mu = \mu_0$, σ известна

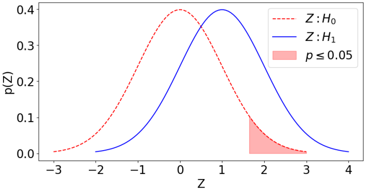
- Выборка $X \sim N\left(\mu, \sigma^2\right)$
- $H_0\colon \mu = \mu_0,\qquad H_1\colon \mu \neq \mu_0 \qquad H_1\colon \mu < \mu_0 \qquad H_1\colon \mu > \mu_0$
- статистика: $Z\left(X^n\right) = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0{,}1)$
- достигаемый уровень значимости:

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

$H_1\colon \mu < \mu_0$



$H_1\colon \mu > \mu_0$



Пример $\mu = \mu_0$, σ известна

Пример

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.

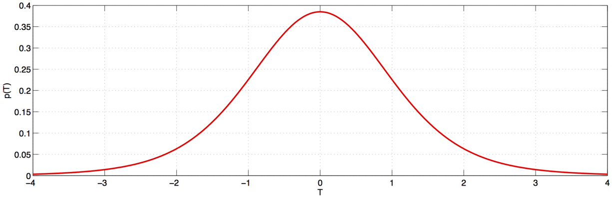
- H_0: средний вес пудры в упаковке соответствует норме.
- H_1: средний вес пудры в упаковке не соответствует норме.
  $\Rightarrow p = 0.0719$, 95% доверительный интервал для среднего веса — $[3.95, 5.25]$ г.
  H_1: средний вес пудры в упаковке превышает норму $\Rightarrow p = 0.0359$, нижний 95% доверительный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

Одностороннюю альтернативу можно использовать, если знак изменения среднего известен заранее.

3 t-критерий Стьюдента: $\mu = \mu_0$, σ неизвестна

- Выборка $X \sim N\left(\mu, \sigma^2\right)$
- $H_0\colon \mu = \mu_0,\qquad H_1\colon \mu \neq \mu_0 \qquad H_1\colon \mu < \mu_0 \qquad H_1\colon \mu > \mu_0$
- статистика: $\quad t\left(X^n\right) = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim St(n-1)$
- достигаемый уровень значимости:

$$p\left(t\right) = \begin{cases} 1 - F_{St(n-1)}(t), & H_1\colon \mu > \mu_0, \\ F_{st(n-1)}(t), & H_1\colon \mu < \mu_0, \\ 2\left(1 - F_{St(n-1)}(|t|)\right), & H_1\colon \mu \neq \mu_0. \end{cases}$$



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

Пример: $\mu = \mu_0$, σ неизвестна I

Средний вес детей при рождении составляет 3300 г. В то же время, если мать ребёнка живёт за чертой бедности, то средний вес таких детей — 2800 г.
С целью увеличить вес тех детей, чьи матери живут за чертой бедности, разработана программа ведения беременности.
Чтобы проверить ее эффективность, проводится эксперимент. В нём принимают участие 25 женщин, живущих за чертой бедности. У всех них рождаются дети, и их средний вес составляет 3075 г, выборочное СКО — 500 г.
Эффективна ли программа? Уровень значимости: $\alpha = 0.05$

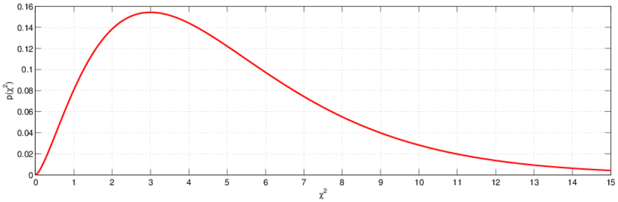
Пример: $\mu = \mu_0$, σ неизвестна II

- Вопрос 1: Влияет ли программа не вес детей?
  - $H_0\colon \mu = 2800$ — программа не влияет на вес детей.
  - $H_1\colon \mu \neq 2800$ программа как-то влияет на вес детей.
  - $t(X^n) = \frac{3075-2800}{500/\sqrt{25}} = 2.75$
  - $F_{St(24)}(2.75) \approx 0.9944$ `scipy.stats.t(24).cdf(t)`
  - $p(t) = 2\big(1 - F_{St(24)}(2.75)\big) \approx 0.0111$
  - $p(t) < \alpha = 0.05$, поэтому влияние существенно.
  - 95% доверительный интервал для изменения веса — $[68.6, 481.4]$ г.
- Вопрос 2: увеличивается ли вес ребенка в следствие программы?
  - $H_0\colon \mu = 2800$ — программа не влияет на вес детей.
  - $H_1\colon \mu > 2800$ программа увеличивает вес детей
  - $p(t) = 2\big(1 - F_{St(24)}(2.75)\big) \approx 0.0111$ $p(t) = 1 - F_{St(24)}(2.75) \approx 0.0056$
  - 95% нижний доверительный предел для увеличения веса — 103.9 г.

2 Критерий хи-квадрат: $\sigma = \sigma_0$

- Выборка $X^n = (X_1, \dots, X_n)\,, X \sim N\left(\mu, \sigma^2\right)$
- $H_0\colon \sigma = \sigma_0, \qquad H_1\colon \sigma \neq \sigma_0 \qquad H_1\colon \sigma < \sigma_0 \qquad H_1\colon \sigma > \sigma_0$
- статистика: $\chi^2\left(X^n\right) = \dfrac{(n-1)S^2}{\sigma_0^2}$
- достигаемый уровень значимости:

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



Пример: $\sigma = \sigma_0$ I

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 выборочная дисперсия объёма жидкости составляет 12 кв. мл.

- Вопрос 1: Существенно ли отклонение дисперсиипри уровне значимости $\alpha = 0.05$?
  - $H_0 : \sigma = 9$ дисперсия объёма жидкости соответствует стандарту.
  - $H_1 : \sigma \neq 9$ дисперсия объёма жидкости не соответствует стандарту
  - $\chi^2(X^n) = \frac{(25-1)\cdot 12}{9} = 32$
  - $F_{\chi^2_{25-1}}(32) \approx 0.8730$ `scipy.stats.chi2(24).cdf(32)`
  - $p(\chi^2) = 2\min\{(1 - F_{\chi^2_{24}}(32), F_{\chi^2_{24}}(32))\} \approx 0.254$
  - $p(\chi^2) > 0.05$ нет причин считать дисперсию не соответствующей стандарту
  - 95% доверительный интервал для дисперсии — $[7.3, 23.2]$ кв. мл.

Пример: $\sigma = \sigma_0$ II

- Вопрос 2: Действительно ли дисперсия объёма жидкости превышает допустимое значениё
  - $H_1 : \sigma > 9$ дисперсия объёма жидкости превышает допустимое значение
  - $p(\chi^2) = 1 - F_{\chi^2_{25-1}}(32) \approx 0.127$
  - $p(\chi^2) > 0.05$ нет причин считать дисперсию не соответствующей стандарту
  - односторонний нижний 95% доверительный предел — 7.9 кв. мл.

Выбор альтернативной гипотезы

По умолчанию используется двухсторонняя гипотеза

Одностороннюю альтернативу можно использовать, если знак изменения среднего известен заранее.

Альтернатива должна выбираться до получения данных!

4 Сравнение средних выборок. Z-критерий Стьюдента

выборки не связаны, σ_1, σ_2 известны

- Выборки:
  $X_1^{n_1} = (X_{11}, \ldots, X_{1n_1}), X_1 \sim N(\mu_1, \sigma_1^2)$
  $X_2^{n_2} = (X_{21}, \ldots, X_{2n_2}), X_2 \sim N(\mu_2, \sigma_2^2)$
- σ_1, σ_2 известны
- $H_0\colon \mu_1 = \mu_2, \qquad H_1\colon \mu_1 \neq \mu_2,\ H_1\colon \mu_1 < \mu_2,\ H_1\colon \mu_1 > \mu_2.$
- Статистика: $Z(X_1^{n_1}, X_2^{n_2}) = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$
- достигаемый уровень значимости:

$$p(Z) = \begin{cases} 1 - F_{N(0,1)}(Z), & H_1\colon \mu_1 > \mu_2, \\ F_{N(0,1)}(Z), & H_1\colon \mu_1 < \mu_2, \\ 2\left(1 - F_{N(0,1)}(|Z|)\right), & H_1\colon \mu_1 \neq \mu_2. \end{cases}$$

4 Сравнение средних выборок. t-критерий Стьюдента

выборки не связаны, σ_1, σ_2 неизвестны

- Выборки:
  $X_1^{n_1} = (X_{11}, \ldots, X_{1n_1}), X_1 \sim N(\mu_1, \sigma_1^2)$
  $X_2^{n_2} = (X_{21}, \ldots, X_{2n_2}), X_2 \sim N(\mu_2, \sigma_2^2)$
- σ_1, σ_2 неизвестны
- $H_0\colon \mu_1 = \mu_2, \qquad H_1\colon \mu_1 \neq \mu_2,\ H_1\colon \mu_1 < \mu_2,\ H_1\colon \mu_1 > \mu_2.$
- Статистика: $T(X_1^{n_1}, X_2^{n_2}) = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx St(\nu)$, $\qquad \nu = \dfrac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}$
- достигаемый уровень значимости:

$$p(t) = \begin{cases} 1 - F_{St(n-1)}(t), & H_1\colon \mu > \mu_0, \\ F_{st(n-1)}(t), & H_1\colon \mu < \mu_0, \\ 2\left(1 - F_{St(n-1)}(|t|)\right), & H_1\colon \mu \neq \mu_0. \end{cases}$$

Приближение достаточно точно при $n_1 = n_2$ или $[n_1 > n_2] = [\sigma_1 > \sigma_2]$.

6 Сравнение средних связанных выборок. t-критерий Стьюдента

- Выборки:
  $X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), X_1 \sim N(\mu_1, \sigma_1^2)$
  $X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), X_2 \sim N(\mu_2, \sigma_2^2)$
- выборки связаны
- $H_0\colon \mu_1 = \mu_2, \qquad H_1\colon \mu_1 \neq \mu_2,\ H_1\colon \mu_1 < \mu_2,\ H_1\colon \mu_1 > \mu_2.$
- Статистика: $t\left(X_1^{n_1}, X_2^{n_2}\right) = \dfrac{\bar{X}_1 - \bar{X}_2}{S/\sqrt{n}} \sim St(n-1), \qquad S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(D_i - \bar{D}\right)^2},$

  $D_i = X_{1i} - X_{2i}, \quad \bar{D} = \frac{1}{n}\sum_i D_i$
- достигаемый уровень значимости:

$$p(t) = \begin{cases} 1 - F_{St(n-1)}(t), & H_1\colon \mu > \mu_0, \\ F_{st(n-1)}(t), & H_1\colon \mu < \mu_0, \\ 2\left(1 - F_{St(n-1)}(|t|)\right), & H_1\colon \mu \neq \mu_0. \end{cases}$$

Пример: сравнение средних 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
- $p = 0.916$;
- 95% доверительный интервал для разности средних $[-2.1, 0.9]$.

Пример

Пусть имеются следующие связные выборки:

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

требуется оценить разность $\Delta = \mathbb{E}X_1 - \mathbb{E}X_2$.

1. 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 $n \approx 200$.

2. Если же **попарные соответствия неизвестны**, лучшая оценка — $\hat{\Delta}_i = \bar{X}_1 - \bar{X}_2$; её дисперсия:

F-критерий Фишера. Гипотеза о дисперсиях

- Выборки:
  $X_1^{n_1} = \left(X_{11}, \ldots, X_{1n_1}\right), X_1 \sim N\left(\mu_1, \sigma_1^2\right)$
  $X_2^{n_2} = \left(X_{21}, \ldots, X_{2n_2}\right), X_2 \sim N\left(\mu_2, \sigma_2^2\right)$
- $H_0\colon \sigma_1 = \sigma_2, \qquad H_1\colon \sigma_1 \neq \sigma_2,\ H_1\colon \sigma_1 < \sigma_2,\ H_1\colon \sigma_1 > \sigma_2.$
- Статистика: $F\left(X_1^{n_1}, X_2^{n_2}\right) = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$
- достигаемый уровень значимости:

$$p\left(F\right) = \begin{cases} \mathcal{F}_{F_{n_1-1,n_2-1}}(1/F), & H_1\colon \sigma_1 > \sigma_2, \\ \mathcal{F}_{F_{n_1-1,n_2-1}}(F), & H_1\colon \sigma_1 < \sigma_2, \\ 2\min\left\{\mathcal{F}_{F_{n_1-1,n_2-1}}(F), \mathcal{F}_{F_{n_1-1,n_2-1}}(1/F)\right\}, & H_1\colon \sigma_1 \neq \sigma_2 \end{cases}$$

Критерий Фишера неустойчив к отклонениям от нормальности даже асимптотически.

7 F-критерий Фишера. Гипотеза о дисперсиях

Требования к выборкам

1. Нормальность распределения: Каждая из групп, между которыми проводится сравнение, должна иметь нормальное распределение данных. Это условие особенно важно, когда размеры выборок малы (Тест Шапиро).
2. Гомогенность дисперсий: Дисперсии внутри каждой из групп должны быть приблизительно равны. (тест Левена или тест Бартлетта).
3. Независимость выборок: Выборки в каждой из групп должны быть независимыми друг от друга.

Параметрические критерии проверки гипотез о законах распределения

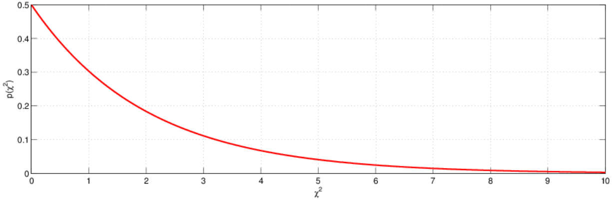
Критерий Харке—Бера

- Выборка: $X^n = (X_1, \ldots, X_n)$
- $H_0\colon X \sim N(\mu, \sigma^2)$, $\qquad H_1\colon \neg H_0$
- Статистика: $\chi^2(X^n) = \frac{n}{6}\left(\gamma_1^2 + \frac{1}{4}\gamma_2^2\right) \sim \chi_2^2$

$$\gamma_1 = \mathbb{E}\left(\frac{X - \mathbb{E}X}{\sqrt{\mathbb{D}X}}\right)^3, \qquad \gamma_2 = \frac{\mathbb{E}(X - \mathbb{E}X)^4}{(\mathbb{D}X)^2} - 3$$

- достигаемый уровень значимости:

$$p\left(\chi^2\right) = 1 - F_{\chi_2^2}\left(\chi^2\right)$$



Критерий согласия Пирсона

- Выборка: $X^n = (X_1, \ldots, X_n)$
- $H_0\colon X \sim N\left(\mu, \sigma^2\right), \qquad H_1\colon \neg H_0$
- Статистика: $\chi^2\left(X^n\right) = \sum_{i=1}^K \frac{(n_i - np_i)^2}{np_i} \sim \begin{cases} \chi^2_{K-1}, & \text{если } \mu, \sigma \text{ заданы} \\ \chi^2_{K-3}, & \text{если } \mu, \sigma \text{ оцениваются} \end{cases}$
- достигаемый уровень значимости: $p\left(\chi^2\right) = 1 - F_{\chi^2_{K-k-1}}\left(\chi^2\right)$
- в Python `scipy.stats.shapiro(x)`

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:
$$\int \left| F_n(x) - F(x) \right| dx$$
- Крамера-фон Мизеса:
$$\int \left(F_n(x) - F(x) \right)^2 dx$$
- Колмогорова (одновыборочный Колмогорова-Смирнова):
$$\sup_{-\infty < x < \infty} \left| F_n(x) - F(x) \right|$$
- Смирнова-Крамера-фон Мизеса:
$$\int \left(F_n(x) - F(x) \right)^2 dF(x)$$

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 Шапиро-Уилка

- выборка: $X^n = (X_1, \ldots, X_n)$
- $H_0\colon X \sim N\left(\mu, \sigma^2\right)$, $\qquad H_1\colon \neg H_0$
- Статистика:

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

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{\left(m^T V^{-1} V^{-1} m\right)^{1/2}}$$

$m = (m_1, \ldots, m_n)^T$ математические ожидания порядковых статистик $N(0,1)$,

V — их ковариационная матрица

- нулевое распределение: табличное
- в Python `scipy.stats.shapiro(x)`

О проверке нормальности

Что не так в критериях?

- **выбросы**: сильно влияют на выборочные коэффициенты асимметрии и эксцесса
- **критерий Колмогорова**: представляет только исторический интерес
- **критерий хи-квадрат**: слишком общий, не самый мощный, потеря информации из-за разбиения на интервалы

О проверке нормальности

Какой критерий лучше?

Сравнение критериев проверки нормальности распределения случайных величин

Наименование критерия (раздел)	Характер альтернативного распределения					Ранг
	асимметричное		симметричное		$\approx$ 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.

О проверке нормальности

А нужно ли вообще?

- **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 от неё

О проверке нормальности

так что же делать?

- если *данные явно ненормальны* (например, бинарны или дискретны), нужно выбрать метод, специфичный для такого распределения

- если *на ку-ку графике не видно существенных отклонений от нормальности*, можно сразу использовать методы, устойчивые к небольшим отклонениям (например, критерии Стьюдента)

- если метод *чувствителен к отклонениям от нормальности* (например, критерий Фишера), проверять её рекомендуется критерием Шапиро-Уилка

- если *нормальность отвергается*, чувствительные методы, предполагающие нормальность, использовать нельзя!

Непараметрические критерии проверки гипотез

Виды задач

- Одновыборочные X^n:
  - среднее выборки равно заданному числу 1 3 8
- Двухвыборочные $X_1^{n_1}, X_2^{n_2}$:
  - среднее выборок равны
    - X_1, X_2 связаны 2 4
    - X_1, X_2 независимы 5 10
  - дисперсии выборок равны 6 11

Варианты двухвыборочных гипотез

О положении:

$$
\begin{aligned}
&H_0\colon \mathbb{E}X_1 = \mathbb{E}X_2, && H_1\colon \mathbb{E}X_1 <\neq> \mathbb{E}X_2;\\
&H_0\colon \operatorname{med} X_1 = \operatorname{med} X_2, && H_1\colon \operatorname{med} X_1 <\neq> \operatorname{med} X_2;\\
&H_0\colon \mathbf{P}\left(X_1 > X_2\right) = 0.5, && H_1\colon \mathbf{P}\left(X_1 > X_2\right) <\neq> 0.5;\\
&H_0\colon F_{X_1}(x) = F_{X_2}(x), && H_1\colon F_{X_1}(x) = F_{X_2}(x+\Delta), \Delta <\neq> 0;\\
&H_0\colon F_{X_1}(x) = F_{X_2}(x), && H_1\colon F_{X_1}(x) <\neq> F_{X_2}(x).
\end{aligned}
$$

О рассеянии:

$$
\begin{aligned}
&H_0\colon \mathbb{D}X_1 = \mathbb{D}X_2, && H_1\colon \mathbb{D}X_1 <\neq> \mathbb{D}X_2;\\
&H_0\colon F_{X_1}(x) = F_{X_2}(x+\Delta), && H_1\colon F_{X_1}(x) = F_{X_2}(\sigma x+\Delta), \sigma <\neq> 1.
\end{aligned}
$$

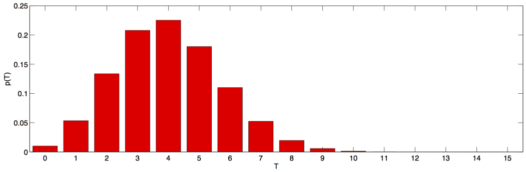
Биномиальный критерий

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

нулевая гипотеза: $H_0 \colon p = p_0$

альтернатива: $H_1 \colon p \gtrless p_0$

статистика: $T\,(X^n) = \sum\limits_{i=1}^{n} X_i \sim Bin(n, p_0)$



достигаемый уровень значимости:

$$p\,(T) = \begin{cases} 1 - F_{Bin(n,p_0)}(T-1), & H_1 \colon p > p_0, \\ F_{Bin(n,p_0)}(T), & H_1 \colon p < p_0, \end{cases}$$

(1) Одновыборочный критерий знаков

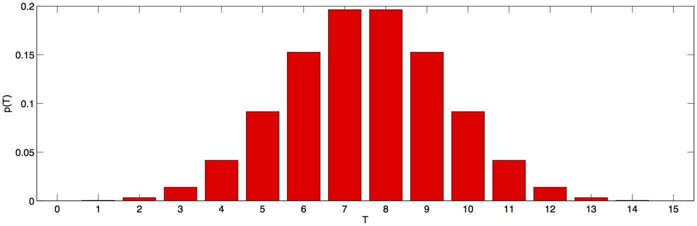
выборка: $X^n = (X_1, \ldots, X_n)\,, X_i \neq m_0$

нулевая гипотеза: $H_0\colon \operatorname{med} X = m_0$

альтернатива: $H_1\colon \operatorname{med} X \gtrless m_0$

статистика: $T(X^n) = \sum\limits_{i=1}^{n} [X_i > m_0]$

нулевое распределение: $Bin(n, \frac{1}{2})$



Пример: одновыборочный критерий знаков

Dinse, 1982

Выживаемость пациентов с лимфоцитарной лимфомой (в неделях):

$$49, 58, 75, 110, 112, 132, 151, 276, 281, 362^*$$

Исследование длилось 7 лет, поэтому для пациентов, проживших дольше, выживаемость неизвестна (выборка цензурирована сверху).
Превышает ли среднее время дожития 200 недель?

- H_0: медиана времени дожития не отличается от 200 недель.
- H_1: медиана времени дожития больше 200 недель.
- $T(X^n) = \sum\limits_{i=1}^{n} [X_i > 200] = 3$
- $F_{Bin(10,1/2)}(3) = C_{10}^{0}\frac{1}{2^{10}} + C_{10}^{1}\frac{1}{2^{10}} + C_{10}^{2}\frac{1}{2^{10}} \sum() \approx 0.0547$

  `scipy.stats.binom(10,1/2).cdf(2)`
- Критерий знаков: $p = 1 - 0.0547 \approx 0.9453$.

(2) Двухвыборочный критерий знаков

выборки: $X_1^n = (X_{11}, \ldots, X_{1n})$

$X_2^n = (X_{21}, \ldots, X_{2n})$, $X_{1i} \neq X_{2i}$

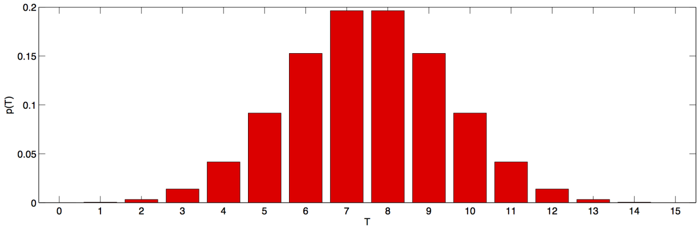
выборки связанные

нулевая гипотеза: $H_0\colon \mathbf{P}(X_1 > X_2) = \frac{1}{2}$

альтернатива: $H_1\colon \mathbf{P}(X_1 > X_2) <\neq> \frac{1}{2}$

статистика: $T(X_1^n, X_2^n) = \sum\limits_{i=1}^{n} [X_{1i} > X_{2i}]$

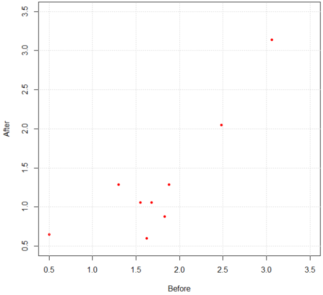
нулевое распределение: $Bin(n, \frac{1}{2})$



(2) Двухвыборочный критерий знаков

Пример, Hollander & Wolfie, 29f

Депрессивность 9 пациентов была измерена по шкале Гамильтона до и после первого приёма транквилизатора. Подействовал ли транквилизатор?



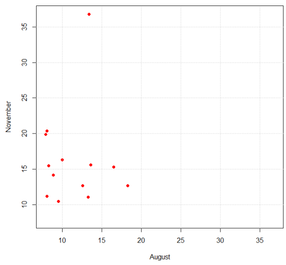
H_0: уровень депрессивности не изменился.
H_1: уровень депрессивности снизился.

Критерий знаков: $p = 0.09$, 95% нижний доверительный предел для медианы изменения — -0.041.

(2) # Двухвыборочный критерий знаков

Пример, Laureysens et al., 2004

Для 13 разновидностей тополей, растущих в зоне интенсивного загрязнения, в августе и ноябре измерялась средняя концентрация алюминия в микрограммах на грамм древесины.



$H_0:$ концентрация алюминия не менялась.
$H_1:$ концентрация алюминия изменилась.
Для тополей 10 из 13 разновидностей концентрация алюминия увеличилась.
Критерий знаков: $p = 0.0923$,
95% доверительный интервал для медианы изменения: $[-0.687, 10.107]$.

Причины использовать критерий знаков

- Точные разности Δx_i неизвестны, известны только их знаки (сравнение агрессивности комаров).
- Разности Δx_i при H_1 могут быть небольшими по модулю, но иметь систематический характер по знаку (пример с мышами).
- Разности Δx_i при H_0 могут быть большими по модулю, но случайными но знаку (влияние меди на число личинок комаров).

Вариационный ряд, ранги, связки

$$X_1, \ldots, X_n \quad \Rightarrow \quad X_{(1)} \leq \ldots < \underbrace{X_{(k_1)} = \ldots = X_{(k_2)}}_{\text{связка размера } k_2 - k_1 + 1} < \ldots \leqslant X_{(n)}$$

Ранг наблюдения X_i:

- если X_i не в связке, то $\mathrm{rank}\,(X_i) = r \colon X_i = X_{(r)}$,
- если X_i в связке $X_{(k_1)}, \ldots, X_{(k_2)}$, то $\mathrm{rank}\,(X_i) = \frac{k_1 + k_2}{2}$.

(3) Одновыборочный критерий знаковых рангов Уилкоксона

выборка: $X^n = (X_1, \ldots, X_n)\,, X_i \neq m_0$

$F(X)$ симметрично относительно медианы

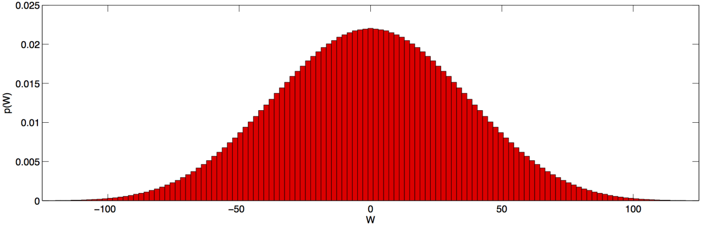
нулевая гипотеза: $H_0\colon\ \mathrm{med}\, X = m_0$

альтернатива: $H_1\colon\ \mathrm{med}\, X <\neq> m_0$

статистика: $W(X^n) = \sum\limits_{i=1}^{n} \mathrm{rank}\,(|X_i - m_0|) \cdot \mathrm{sign}\,(X_i - m_0)$

нулевое распределение: табличное

в Python: `scipy.stats.wilcoxon(d)`



(3) Одновыборочный критерий знаковых рангов Уилкоксона

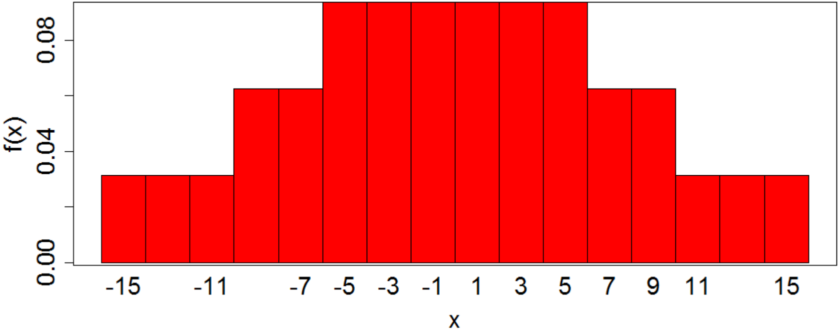
Откуда берётся табличное распределениё

1	2	3	4	5	W
$-$	$-$	$-$	$-$	$-$	-15
$+$	$-$	$-$	$-$	$-$	-13
$-$	$+$	$-$	$-$	$-$	-11
$+$	$+$	$-$	$-$	$-$	-9
$-$	$-$	$+$	$-$	$-$	-9
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$+$	$+$	$-$	$+$	$+$	9
$-$	$-$	$+$	$+$	$+$	9
$+$	$-$	$+$	$+$	$+$	11
$-$	$+$	$+$	$+$	$+$	13
$+$	$+$	$+$	$+$	$+$	15

Всего 2^n вариантов.

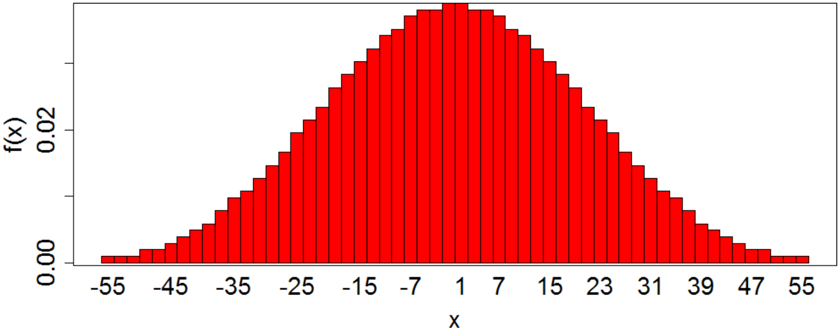
(3) Одновыборочный критерий знаковых рангов Уилкоксона

$n = 5$:



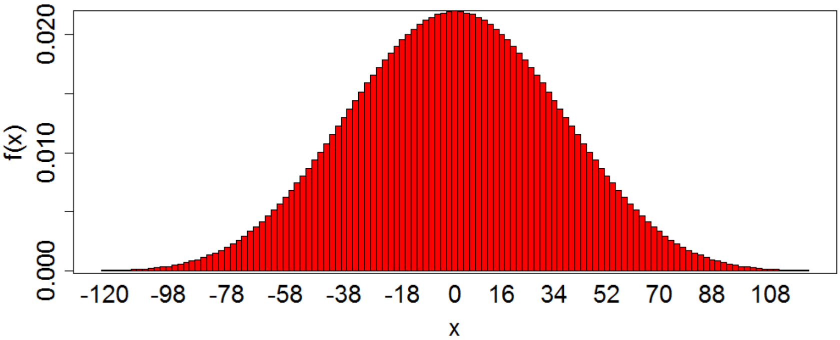
(3) Одновыборочный критерий знаковых рангов Уилкоксона

$n = 10$:



(3) Одновыборочный критерий знаковых рангов Уилкоксона

$n = 15$:

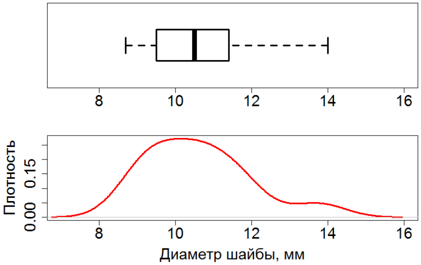


Аппроксимация для $n > 20$:

$$W \approx\sim N\left(0, \frac{n(n+1)(2n+1)}{6}\right).$$

(3) Одновыборочный критерий знаковых рангов Уилкоксона

Пример 1 (Bonnini, табл. 1.4): диаметры шайб на производстве ($n = 24$):



Соответствуют ли шайбы стандартному размеру 10 мм?

$H_0:$ средний диаметр шайбы — 10 мм, $\mathrm{med}\, X = 10$.
$H_1:$ средний диаметр шайбы не соответствует стандарту, $\mathrm{med}\, X \neq 10$.

Критерий знаковых рангов: $p = 0.0673$, выборочная медиана диаметра — 10.5 мм (95% доверительный интервал — $[9.95, 11.15]$ мм).

(4) Критерий знаковых рангов Уилкоксона для связанных выборок

выборки: $X_1^n = (X_{11}, \dots, X_{1n})$

$X_2^n = (X_{21}, \dots, X_{2n})\,, X_{1i} \neq X_{2i}$

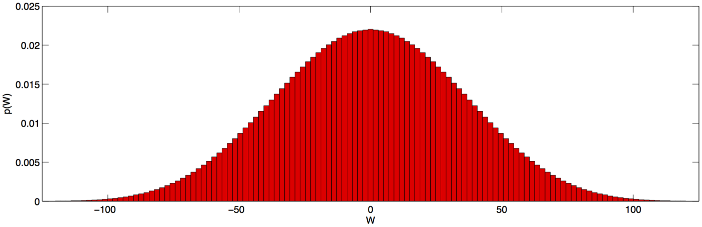
выборки связанные, разность выборок симметрична относите

нулевая гипотеза: $H_0\colon \operatorname{med}(X_1 - X_2) = 0$

альтернатива: $H_1\colon \operatorname{med}(X_1 - X_2) <\neq> 0$

статистика: $W(X_1^n, X_2^n) = \sum\limits_{i=1}^{n} \operatorname{rank}(|X_{1i} - X_{2i}|) \cdot \operatorname{sign}(X_{1i} - X_{2i})$

нулевое распределение: табличное



(4) Критерий знаковых рангов Уилкоксона для связанных выборок

Пример, Kanji, критерий 48

Управляемый вручную станок на каждом шаге процесса производит пару пружин. Для 14 пар измерена прочность:

$$X_1: \{1.38, 0.39, 1.42, 0.54, 5.94, 0.59, 2.67, 2.44, 0.56, 0.69, 0.71, 0.95, 0.50, 9.69\},$$
$$X_2: \{1.42, 0.39, 1.46, 0.55, 6.15, 0.61, 2.69, 2.68, 0.53, 0.72, 0.72, 0.93, 0.53, 10.37\}.$$

Одинакова ли прочность пружин в парё

H_0: средние значение прочности пружин в паре равны.
H_1: средние значение прочности пружин в паре не равны $\Rightarrow p = 0.0142$, 95% доверительный интервал для медианной разности — $[0.005, 0.14]$.

(5) Критерий Манна-Уитни-Уилкоксона

выборки: $X_1^{n_1} = (X_{11}, \ldots, X_{1n_1})$
$X_2^{n_2} = (X_{21}, \ldots, X_{2n_2})$
выборки независимые

нулевая гипотеза: $H_0\colon F_{X_1}(x) = F_{X_2}(x)$

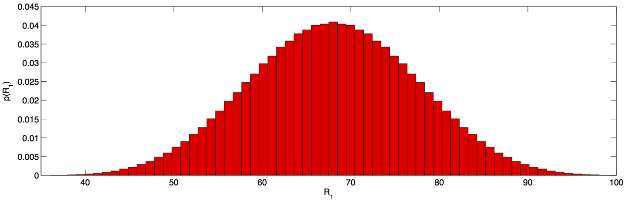
альтернатива: $H_1\colon F_{X_1}(x) = F_{X_2}(x + \Delta), \Delta <\neq> 0$

статистика: $X_{(1)} \leq \ldots \leqslant X_{(n_1+n_2)}$ — вариационный ряд объединённой выборки $X = X_1^{n_1} \bigcup X_2^{n_2}$

$$R_1\left(X_1^{n_1}, X_2^{n_2}\right) = \sum_{i=1}^{n_1} \operatorname{rank}(X_{1i})$$

нулевое распределение: табличное

в Python: `scipy.stats.mannwhitneyu(x1, x2)`



(5) Критерий Манна-Уитни-Уилкоксона

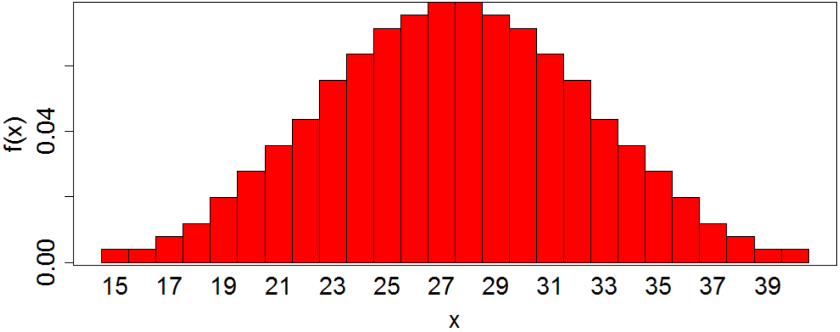
Откуда берётся табличное распределение?

X_1	X_2	R_1
{1,2,3}	{4,5,6,7}	6
{1,2,4}	{3,5,6,7}	7
{1,2,5}	{3,4,6,7}	8
{1,2,6}	{3,4,5,7}	9
{1,2,7}	{3,4,5,6}	10
{1,3,4}	{2,5,6,7}	8
...	...	...
{3,5,7}	{1,2,4,6}	15
{3,6,7}	{1,2,4,5}	16
{4,5,6}	{1,2,3,7}	15
{4,5,7}	{1,2,3,6}	16
{4,6,7}	{1,2,3,5}	17
{5,6,7}	{1,2,3,4}	18

Всего $C_{n_1+n_2}^{n_1}$ вариантов.

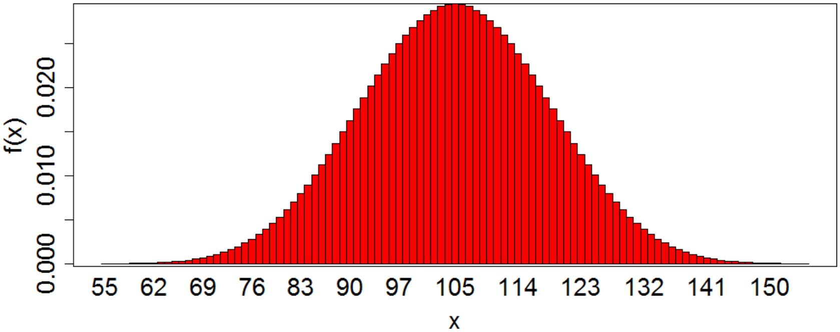
Критерий Манна-Уитни-Уилкоксона

$n_1 = n_2 = 5$:



(5) Критерий Манна-Уитни-Уилкоксона

$n_1 = n_2 = 10$:



Аппроксимация для $n_1, n_2 > 10$:

$$R_1 \sim N\left(\frac{n_1(n_1+n_2+1)}{2}, \frac{n_1 n_2 (n_1+n_2+1)}{12}\right).$$

Критерий Манна-Уитни-Уилкоксона

Пример, Kanji, критерий 52

Сотрудник налоговой службы хочет сравнить средние значения в двух выборках заявленных трат на компенсацию командировочных расходов в одной и той же компании в двух разных периодах (расходы скорректированы на инфляцию).

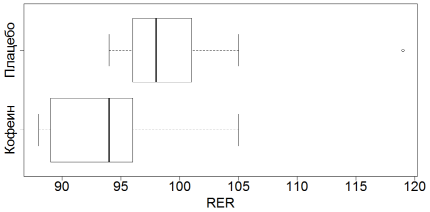
$$X_1\colon \{50.5, 37.5, 49.8, 56.0, 42.0, 56.0, 50.0, 54.0, 48.0\},$$
$$X_2\colon \{57.0, 52.0, 51.0, 44.2, 55.0, 62.0, 59.0, 45.2, 53.5, 44.4\}.$$

Равны ли средние расходы?

$H_0\colon$ средние расходы равны.
$H_1\colon$ средние расходы не равны $\Rightarrow p = 0.3072$, 95% доверительный интервал для медианной разности — $[-9, 4]$.

(5) Критерий Манна-Уитни-Уилкоксона

RER — соотношение числа молекул CO_2 и O_2 в выдыхаемом воздухе.
В эксперименте измерялся респираторный обмен 18 испытуемых в процессе физических упражнений. За час до этого 9 из них получили таблетку кофеина, 9 — плацебо.



Повлиял ли кофеин на значение RER?

H_0: среднее значение показателя респираторного обмена не отличается в двух группах.
H_1: среднее значение показателя респираторного обмена отличается в двух группах.

(5) Критерий Манна-Уитни-Уилкоксона

Ранг	Наблюдение	Номер наблюдения	Наблюдение	Ранг
16.5	105	1	96	9
18	119	2	99	13
14	100	3	94	5.5
11	97	4	89	3
9	96	5	96	9
15	101	6	93	4
5.5	94	7	88	1.5
7	95	8	105	16.5
12	98	9	88	1.5

Статистика R_1 — сумма рангов в одной из групп.

$p = 0.0521$, сдвиг между средними — 6 пунктов, (95% доверительный интервал — $[-0.00005, 12]$ пт).

(6) Критерий Ансари-Брэдли

выборки: $X_1^{n_1} = (X_{11}, \ldots, X_{1n_1})$
$X_2^{n_2} = (X_{21}, \ldots, X_{2n_2})$
выборки независимые, $\mathrm{med}(X_1) = \mathrm{med}(X_2)$

нулевая гипотеза: $H_0\colon \mathbb{D}X_1 = \mathbb{D}X_2$

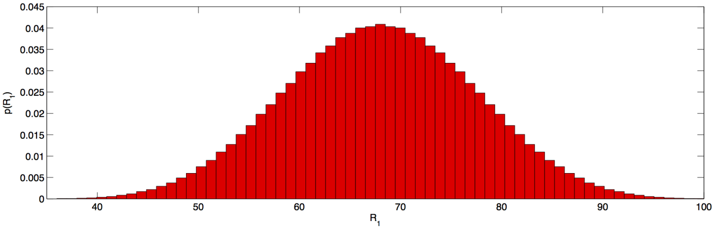
альтернатива: $H_1\colon \mathbb{D}X_1 <\neq> \mathbb{D}X_2$

статистика: $X_{(1)} \leq \ldots \leqslant X_{(N)}$ — вариационный ряд объединённой выборки $X^N = X_1^{n_1} \bigcup X_2^{n_2}, N = n_1 + n_2$

$$R_1\left(X_1^{n_1}, X_2^{n_2}\right) = \sum_{i=1}^{n_1} \widetilde{\mathrm{rank}}\left(X_{1i}\right)$$

нулевое распределение: табличное

в Python: `scipy.stats.ansari(x1, x2)`



(6) Критерий Ансари-Брэдли

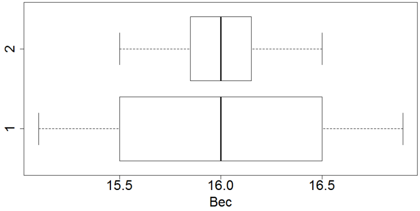
Ранги присваиваются от краёв к центру:

$$\begin{array}{ccccccccccccc} X_{(i)} & X_{(1)} & \leq & X_{(2)} & \leq & X_{(3)} & \leq & \ldots & \leq & X_{(N-2)} & \leq & X_{(N-1)} & \leq & X_{(N)} \\ \widetilde{\operatorname{rank}}\left(X_{(i)}\right) & 1 & & 2 & & 3 & & & & 3 & & 2 & & 1 \end{array}$$

(6) Критерий Ансари-Брэдли

Пример,Bonnini, табл. 2.1

Два поставщика шестнадцатикилограммовых свинцовых слитков выслали по выборке образцов. Средний вес образцов в обеих выборках соответствует норме; различаются ли дисперсий



$H_0:$ дисперсия веса слитков не отличается для двух поставщиков.
$H_1:$ дисперсия веса слитков для двух поставщиков отличается $\Rightarrow p = 0.014.$

Перестановочные критерии

Ранговые критерии:

1. выборки $\Rightarrow$ ранги
2. дополнительное предположение (о равенстве распределений / медиан и пр.)
3. перестановки $\Rightarrow$ нулевое распределение статистики

Что если пропустить пункт 1?

Пример (зеркала в клетках мышей):
$H_0:$ в клетке с зеркалом мыши проводят в среднем половину времени.
$H_1:$ в клетке с зеркалом мыши проводят в среднем не половину времени.

Проинтерпретируем задачу по-другому:
$H_0:$ *матожидание* времени в клетке с зералом равняется 0.5.
$H_1:$ *матожидание* времени в клетке с зералом не равняется 0.5

Предположение:
время, проведенное мышами в клетке с зеркалом симметрично относительно матожидания.
Тогда при верности H_0: $X - 0.5$ — симметрично относительно нуля.

Статистика:
$$T = \sum_{i=1}^{n} (X_i - 0.5) \, .$$

Как получить нулевое распределение:
будем переставлять элементы смещенной выборки $X - 0.5$ относительно нуля.

Пример:
H_0: в клетке с зеркалом мыши проводят в среднем половину времени.
H_1: в клетке с зеркалом мыши проводят в среднем не половину времени.

Статистика: $T = \sum\limits_{i=1}^{n} (X_i - 0.5)$; $t = -0.3784$.

$p = \frac{\#[|T| \geq |t|]}{2^n} = 0.2292.$
95% доверительный интервал для доли времени в клетке с зеркалом (BCa бутстреп) — $[0.447, 0.511]$.

(8) Одновыборочный перестановочный критерий, гипотеза о среднем

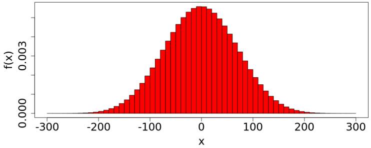
выборка:	$X_1^n = (X_1, \ldots, X_n)$
	$F(X)$ симметрично относительно матожидания
нулевая гипотеза:	$H_0\colon \mathbb{E}X = m_0$
альтернатива:	$H_1\colon \mathbb{E}X <\neq> m_0$
статистика:	$T(X^n) = \sum\limits_{i=1}^{n} (X_i - m_0)$
нулевое распределение:	порождается перебором 2^n знаков перед слагаемыми $X_i - m_0$

Достигаемый уровень значимости — доля перестановок знаков, на которых получилось такое же или ещё более экстремальное значение статистики.

(8) Одновыборочный перестановочный критерий, гипотеза о среднем

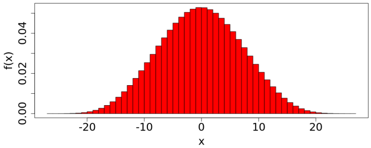
Пример (диаметры шайб):

Критерий знаковых рангов:



$p = 0.0673$

Перестановочный критерий:



$T = 14.6, p = 0.1026$

95% доверительный интервал для среднего диаметра (BCa бутстреп) — $[10.11, 11.20]$.

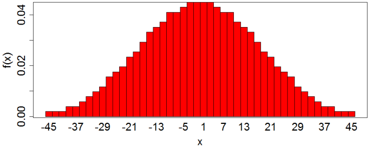
(9) Двухвыборочный перестановочный критерий, гипотеза о средних, связанные выборки

выборки:	$X_1^n = (X_{11}, \dots, X_{1n})$
	$X_2^n = (X_{21}, \dots, X_{2n})$
	выборки связанные
	распределение попарных разностей симметрично
нулевая гипотеза:	$H_0\colon \mathbb{E}\left(X_1 - X_2\right) = 0$
альтернатива:	$H_1\colon \mathbb{E}\left(X_1 - X_2\right) <\neq> 0$
статистика:	$D_i = X_{1i} - X_{2i}$
	$T\left(X_1^n, X_2^n\right) = \sum\limits_{i=1}^n D_i$
нулевое распределение:	порождается перебором 2^n знаков перед слагаемыми D_i

(9) Двухвыборочный перестановочный критерий, гипотеза о средних, связанные выборки

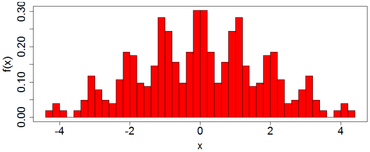
Пример (лечение депрессии):

Критерий знаковых рангов:



$p = 0.019$

Перестановочный критерий:



$T = 3.887, p = 0.0137$

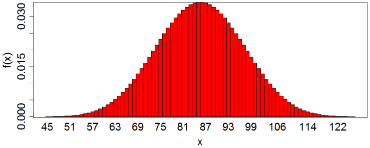
95% доверительный интервал для среднего уменьшения депрессивности (BCa бутстреп) — $[0.1658, 0.6834]$

(10) Двухвыборочный перестановочный критерий, гипотеза о средних, независимые выборки

выборки:	$X_1^{n_1} = (X_{11}, \dots, X_{1n_1})$
	$X_2^{n_2} = (X_{21}, \dots, X_{2n_2})$
нулевая гипотеза:	$H_0\colon F_{X_1}(x) = F_{X_2}(x)$
альтернатива:	$H_1\colon F_{X_1}(x) = F_{X_2}(x+\Delta), \Delta <\neq> 0$
статистика:	$T(X_1^{n_1}, X_2^{n_2}) = \frac{1}{n_1}\sum\limits_{i=1}^{n_1} X_{1i} - \frac{1}{n_2}\sum\limits_{i=1}^{n_2} X_{2i}$
нулевое распределение:	порождается перебором $C_{n_1+n_2}^{n_1}$ размещений объединённой выборки

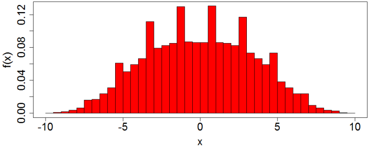
 Двухвыборочный перестановочный критерий, гипотеза о средних, независимые выборки

Пример (кофеин и респираторный обмен):
Критерий Манна-Уитни:



$p = 0.0521$

Перестановочный критерий:



$T = 6.33, p = 0.0578$
95% доверительный интервал для разности средних (BCa бутстреп) — $[1.556, 13.667]$

(11) Двухвыборочный перестановочный критерий, гипотеза о дисперсиях, статистика Али

выборки:	$X_1^n = (X_{11}, \ldots, X_{1n})$
	$X_2^n = (X_{21}, \ldots, X_{2n})$
	выборки независимые
нулевая гипотеза:	$H_0\colon \mathbb{D}X_1 = \mathbb{D}X_2$
альтернатива:	$H_1\colon \mathbb{D}X_1 <\neq> \mathbb{D}X_2$
статистика:	$\delta\left(D_1^{n-1}\right) = \sum\limits_{i=1}^{n-1} i(n-i)D_{1i},$
	$D_{1i} = X_{1(i+1)} - X_{1(i)}$
нулевое распределение:	порождается перебором 2^{n-1}
	попарных перестановок D_{1i} и D_{2i}

Особенности перестановочных критериев

- Статистику критерия можно выбрать разными способами. В некоторых случаях разные статистики приведут к одному и тому же достигаемому уровню значимости:

$$X^n, \quad H_0\colon \mathbb{E}X = 0, \; H_1\colon \mathbb{E}X \neq 0,$$

$$T_1\left(X^n\right) = \sum_{i=1}^{n} X_i \quad \sim \quad T_2\left(X^n\right) = \bar{X}.$$

  В других случаях достигаемый уровень значимости будет зависеть от выбора статистики:

$$T_2\left(X^n\right) = \bar{X} \quad \nsim \quad T_3\left(X^n\right) = \frac{\bar{X}}{S/\sqrt{n}}.$$

- Если множество перестановок G слишком велико, для оценки нулевого распределения T достаточно взять случайное подмножество $G' \in G$. При этом стандартное отклонение достигаемого уровня значимости будет равно примерно $\sqrt{\frac{p(1-p)}{|G'|}}$.

Перестановки и бутстреп

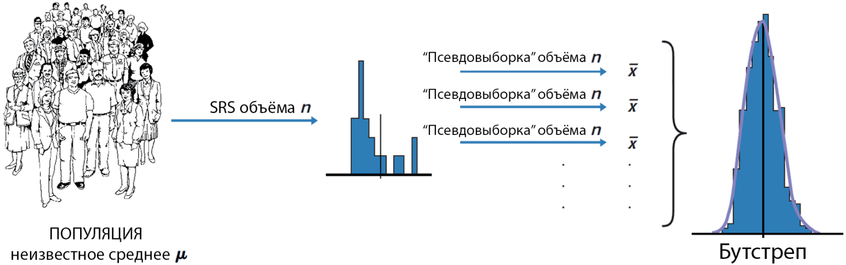
Перестановочные критерии:

1. выборки, статистика
2. дополнительное предположение
3. перестановки $\Rightarrow$ нулевое распределение статистики

Бутстреповые доверительные интервалы:

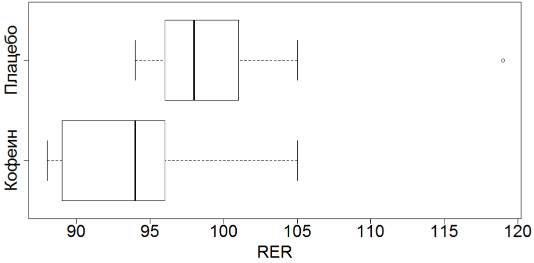
1. выборки, статистика, оценивающая параметр
2. бутстреп-псевдовыборки $\Rightarrow$ приближённое распределение статистики

- бутстреп:



Сгенерировать N «псевдовыборок» объёма n и оценить выборочное распределение $\hat{\theta}_n$ «псевдоэмпирическим».

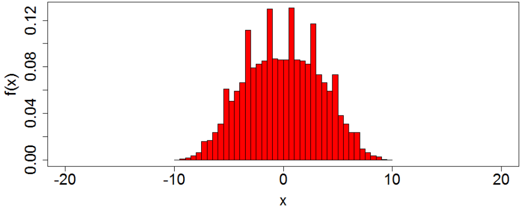
Кофеин и респираторный обмен



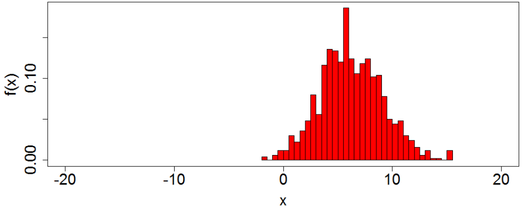
H_0: среднее значение показателя респираторного обмена не отличается в двух группах.
H_1: под воздействием кофеина среднее значение показателя респираторного обмена снижается.
$\bar{X}_{1n} - \bar{X}_{2n} = 6.33$

Кофеин и респираторный обмен

Нулевое распределение перестановочного критерия со статистикой $\bar{X}_{1n} - \bar{X}_{2n}$:

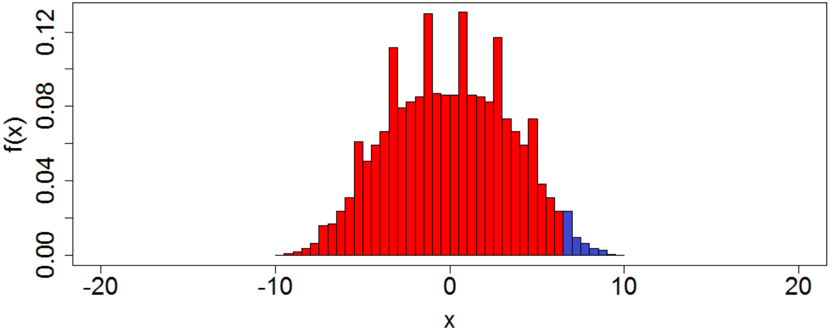


Бутстреп-распределение статистики $\bar{X}_{1n} - \bar{X}_{2n}$:



Кофеин и респираторный обмен

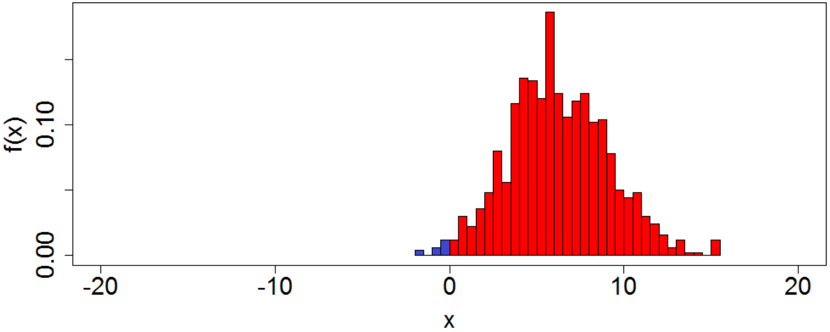
Нулевое распределение перестановочного критерия со статистикой $\bar{X}_{1n} - \bar{X}_{2n}$:



Доля перестановок, на которых среднее больше либо равно 6.33 — 0.0289.
Это точный достигаемый уровень значимости перестановочного критерия.

Кофеин и респираторный обмен

Бутстреп-распределение статистики $\bar{X}_{1n} - \bar{X}_{2n}$:

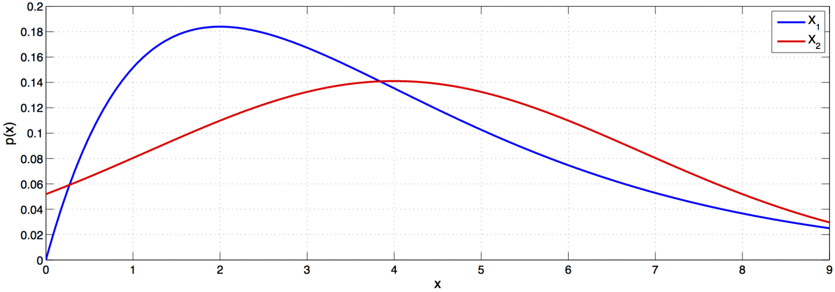


Доля псевдовыборок, на которых среднее меньше либо равно нулю — 0.011.
Это приближённый достигаемый уровень значимости бутстреп-критерия.

Перестановки vs. бутстреп

Перестановочный критерий	**Бутстреп-критерий**
Центр в нуле	Центр в точечной оценке
Точный	Приближенный
$H_0\colon F_{X_1}(x) = F_{X_2}(x)$ $H_1\colon F_{X_1}(x) = F_{X_2}(x+\Delta), \Delta > 0$	$H_0\colon \mathbb{E}X_1 = \mathbb{E}X_2$ $H_1\colon \mathbb{E}X_1 > \mathbb{E}X_2$

Различия между моментами высокого порядка



$$X_1 \sim \chi^2_4, \quad X_2 \sim N\left(4, \sqrt{8}\right);$$
$$\mathbb{E}X_1 = \mathbb{E}X_2, \quad \mathbb{D}X_1 = \mathbb{D}X_2.$$

Двухвыборочные критерии согласия

выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1})$
$X_2^{n_2} = (X_{21}, \dots, X_{2n_2})$
выборки независимые

нулевая гипотеза: $H_0\colon F_{X_1}(x) = F_{X_2}(x)$

альтернатива: $H_1\colon H_0$ неверна

Критерий Смирнова

статистика: $D\left(X_1^{n_1}, X_2^{n_2}\right) = \sup\limits_{-\infty<x<\infty} \left|F_{n_1 X_1}(x) - F_{n_2 X_2}(x)\right|$

Критерий Андерсона (модификация критерия Смирнова-Крамера-фон Мизеса)

статистика:
$$T\left(X_1^{n_1}, X_2^{n_2}\right) = \frac{1}{n_1 n_2 (n_1+n_2)} \left(n_1 \sum_{i=1}^{n_1} \left(\operatorname{rank}\left(X_{1i}\right) - i\right)^2 + \right.$$
$$\left. + n_2 \sum_{j=1}^{n_1} \left(\operatorname{rank}\left(X_{2j}\right) - j\right)^2 \right) - \frac{4 n_1 n_2 - 1}{6(n_1+n_2)}$$

Где что искать? I

1. 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
2. Параметрические критерии 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
3. Критерии, основанные на правдоподобии: Bilder, раздел B.5
4. Критерии для распределения Бернулли:
   - всё про одновыборочную задачу — Agresti, 1.3, 1.4
   - Z-критерии (Z-tests) — Kanji, №№ 4, 5

Где что искать? II

- точный критерий (exact binomial test) — McDonald, `http://www.biostathandbook.com/exactgof.html`
- доверительные интервалы Уилсона (score confidence intervals) — Newcombe, 1998a, 1998b, 1998c

5. непараметрические критерии:
   - критерии знаков (sign tests) — Kanji, №№ 45, 46;
   - критерии знаковых рангов (signed-rank tests) — Kanji, №№ 47, 48;
   - критерий Манна-Уитни-Уилкоксона (Mann-Whitney-Wilcoxon test) — Kanji, № 52;
   - перестановочные критерии (permutation tests) — Good, 3.2.1, 3.6.4, 3.7.2 (с ошибкой, исправлено в Ramsey);
   - двухвыборочные критерии согласия (two-sample goodness-of-fit tests) — Кобзарь, 3.1.2.8.

Литература I

- О сравнении ассимптотических критериев
- Кобзарь А.И. *Прикладная математическая статистика*, 2006.
- Королёв В.Ю. *Теория вероятностей и математическая статистика*, 2008.

- Dinse G.E. (1982). *Nonparametric estimation for partially-complete time and type of failure data*. Biometrics, 38, 417–431.
- Ramsey P.H., Ramsey P.P. (2008). *Brief investigation of tests of variability in the two-sample case*. Journal of Statistical Computation and Simulation, 78(12), 1125–1131.
- Bonnini S., Corain L., Marozzi M., Salmaso S. *Nonparametric Hypothesis Testing - Rank and Permutation Methods with Applications in R*, 2014.
- Kanji G.K. *100 statistical tests*, 2006.
- Agresti A. *Categorical Data Analysis*, 2013.

Литература II

- Bilder C.R., Loughin T.M. *Analysis of Categorical Data with R*, 2013.
- McDonald J.H. *Handbook of Biological Statistics*, 2008.
- 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.
- Newcombe R.G. (1998). *Two-sided confidence intervals for the single proportion: comparison of seven methods*. Statistics in Medicine, 17, 857–872.
- NIST/SEMATECH. *e-Handbook of Statistical Methods*. `http://www.itl.nist.gov/div898/handbook/`
- Efron B., Tibshirani R. *An Introduction to the Bootstrap*, 1993.
- Good P. *Permutation, Parametric and Bootstrap Tests of Hypotheses: A Practical Guide to Resampling Methods for Testing Hypotheses*, 2005.

Литература III

- Hollander M., Wolfe D.A. *Nonparametric statistical methods*, 1973.
- Kanji G.K. *100 statistical tests*, 2006.
- Laureysens I., Blust R., De Temmerman L., Lemmens C., Ceulemans R. (2004). *Clonal variation in heavy metal accumulation and biomass production in a poplar coppice culture. I. Seasonal variation in leaf, wood and bark concentrations.* Environmental Pollution, 131, 485-494.
- Shekin D. *Handbook of Parametric and Nonparametric Statistical Procedures*, 2007.
- Shervin C.M. (2004) *Mirrors as potential environmental enrichment for individually housed laboratory mice*. Applied Animal Behaviour Science, 87(1-2), 95–103.