MML minor #10

Проверка гипотез

Рябенко Евгений
riabenko.e@gmail.com

[redacted]

Предсказание будущего

Как проверить?

Предсказание будущего

Эксперимент: записываются предсказания, генерируются события, проверяется правильность предсказаний.

$X^n = (X_1, \dots, X_n)$ — выборка результатов, например:

- $X = 1$, если предсказание сбылось, 0, если не сбылось
- X — точность предсказания (разность между фактом и прогнозом)

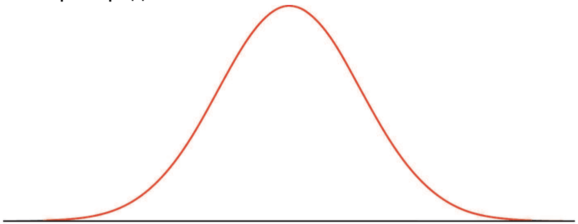
Предсказатель полезен, если он предсказывает лучше, чем генератор случайных чисел.
Гипотеза: предсказатель — и есть генератор случайных чисел.

Что говорят данные? Свидетельствуют ли они против такого предположения?

Проверка гипотез

$$
\begin{aligned}
\text{выборка:}\quad & X^n = (X_1, \dots, X_n),\ X \sim \mathbf{P};\\
\text{нулевая гипотеза:}\quad & H_0\colon\ \mathbf{P} \in \omega;\\
\text{альтернатива:}\quad & H_1\colon\ \mathbf{P} \notin \omega;\\
\text{статистика:}\quad & T(X^n),\quad T(X^n) \sim F(x)\ \text{при } H_0;\\
& \qquad\qquad\ \ T(X^n) \not\sim F(x)\ \text{при } H_1.
\end{aligned}
$$

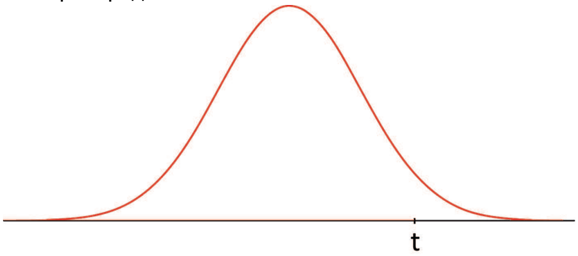
$F(x)$ — нулевое распределение статистики:



Вместе T и $F(x)$ — статистический критерий для проверки H_0 против H_1.

Нулевое распределение

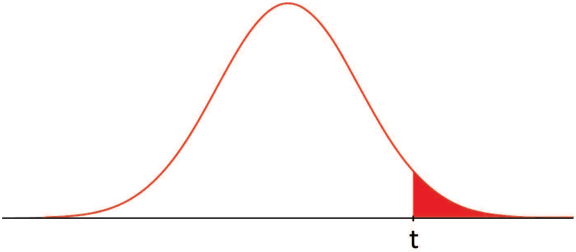
$F(x)$ — нулевое распределение статистики:



t — значение статистики на полученных данных.
Насколько оно вероятно при справедливости H_0?
Каким значениям статистики соответствует H_1?

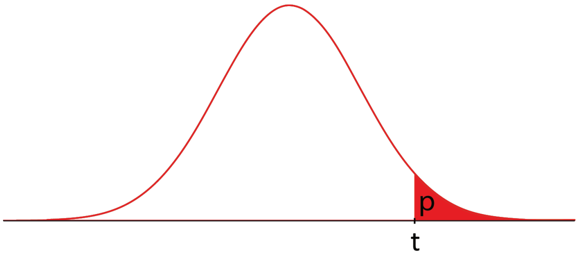
Нулевое распределение

Каким значениям статистики соответствует H_1?
Допустим, большим:



Какова вероятность при H_0 получить значение t или больше?

Достигаемый уровень значимости



Какова вероятность при H_0 получить значение t или больше?
Достигаемый уровень значимости (p-value):

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

p — вероятность при справедливости нулевой гипотезы получить значение статистики как в эксперименте или ещё более экстремальное.

p мало $\Rightarrow$ данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

α — уровень значимости; H_0 отвергается в пользу H_1 при $p \leqslant \alpha$.

Ошибки I и II рода

	H_0 верна	H_0 неверна
H_0 принимается	H_0 верно принята	Ошибка II рода
H_0 отвергается	Ошибка I рода	H_0 верно отвергнута

Ошибки I и II рода не равнозначны!

Задача несимметрична

Ошибка первого рода критичнее:

- $\mathbf{P}($отвергаем $H_0 \,|H_0\,)$ жёстко ограничивается:
  если H_0 отвергается при $p \leqslant \alpha$, то вероятность ошибки первого рода

$$\mathbf{P}(H_0 \text{ отвергнута} \mid H_0 \text{ верна}) = \mathbf{P}(p \leqslant \alpha \mid H_0) \leqslant \alpha.$$

- $\mathbf{P}($принимаем $H_0 \,|H_1\,)$ мягко минимизируется.
  Мощность критерия:

$$\text{pow} = \mathbf{P}(\text{отвергаем } H_0 | \, H_1) = 1 - \mathbf{P}(\text{принимаем } H_0 \,|H_1\,).$$

  Идеальный критерий имеет максимальную мощность.

H_0 и H_1 не равнозначны! Нельзя доказать, что H_0 верна:

- $p \leqslant \alpha \Rightarrow H_0$ отвергается в пользу H_1
- $p > \alpha \Rightarrow H_0$ не отвергается в пользу H_1

Отсутствие доказательств чего-то не является доказательством обратного!

Достигаемый уровень значимости

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

Вероятность получить значение статистики как в эксперименте или ещё более экстремальное при справедливости нулевой гипотезы.
Чем ниже p, тем сильнее данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

Неправильная интерпретация

$$
\begin{aligned}
p = \mathbf{P}(T \geqslant t \,|H_0\,) &\neq \mathbf{P}(H_0) \\
&\neq \mathbf{P}(H_0\,|T \geqslant t\,)
\end{aligned}
$$

Осьминог угадал результаты 11 из 13 матчей с участием сборной Германии на чемпионате мира по футболу 2010 г.

$p = 0.0112$ — не вероятность того, что осьминог выбирает кормушку наугад! Эта вероятность равна единице.

Размер эффекта

Интерес представляет не p, а размер эффекта — степень отклонения данных от нулевой гипотезы.

- вероятность верного предсказания
- вероятность выздоровления пациента, принимавшего лекарство, минус вероятность выздоровления пациента, принимавшего плацебо
- увеличение среднего чека интернет-магазина при подключении программы лояльности

Оценка размера эффекта по выборке — случайная величина;
p показывает, с какой вероятностью такую оценку можно было получить случайно.
При этом p зависит не только от размера эффекта, но и от размера выборки: по мере увеличения n H_0 может сначала приниматься, но потом выявятся более тонкие несоответствия выборки гипотезе H_0, и она будет отвергнута.

Статистическая и практическая значимость

- (Lee et al, 2010): за три года женщины, упражнявшиеся не меньше часа в день, набрали значимо меньше веса, чем женщины, упражнявшиеся меньше 20 минут в день ($p < 0.001$). Разница в набранном весе составила 150 г. Практическая значимость такого эффекта сомнительна.
- (Ellis, 2010, гл. 2): в 2002 году клинические испытания гормонального препарата Премарин, облегчающего симптомы менопаузы, были досрочно прерваны. Было 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%. Формально эффект крайне мал, но с учётом численности населения он превращается в тысячи дополнительных смертей.
- (Kirk, 1996): 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.

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.

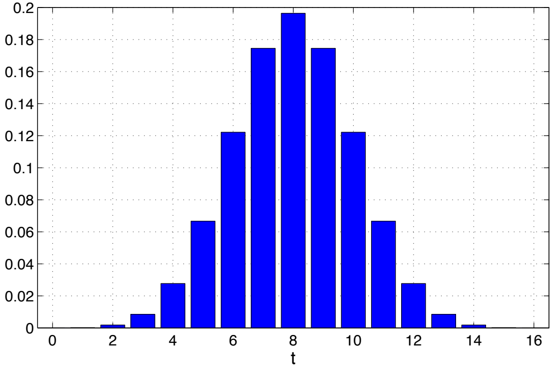
Нулевая гипотеза: Джеймс Бонд не различает два вида мартини, т. е., выбирает наугад.

Статистика T — число единиц в выборке.

Нулевое распределение

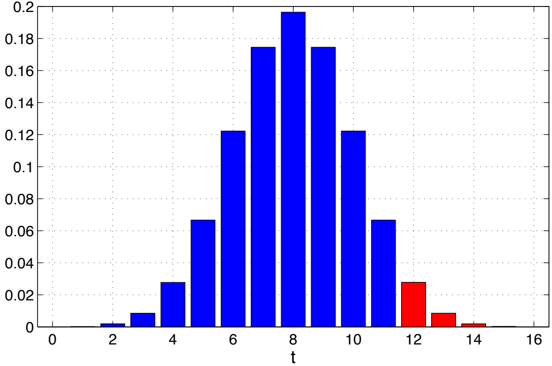
Если нулевая гипотеза справедлива и Джеймс Бонд не различает два вида мартини, то равновероятны все выборки длины n из нулей и единиц.

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



Односторонняя альтернатива

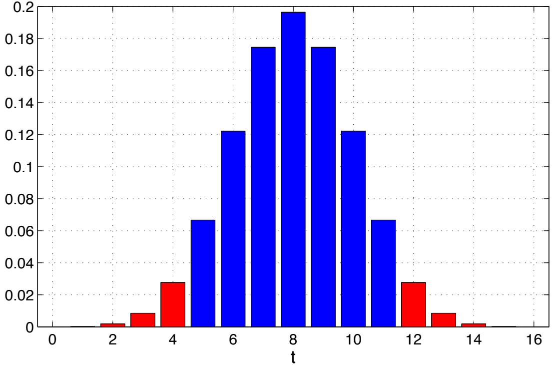
H_1: Джеймс Бонд предпочитает взболтанный мартини.
При справедливости такой альтернативы более вероятны большие значения T (т.е., большие T свидетельствуют против H_0 в пользу H_1).
Вероятность того, что Джеймс Бонд предпочтёт взболтанный мартини в 12 или более случаях из 16 при справедливости H_0, равна $\frac{2517}{65536} \approx 0.0384$.



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

Двусторонняя альтернатива

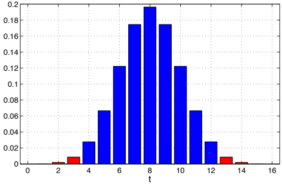
H_1: Джеймс Бонд предпочитает какой-то определённый вид мартини.
При справедливости такой альтернативы и большие, и маленькие значения T свидетельствуют против H_0 в пользу H_1).
Вероятность того, что Джеймс Бонд предпочтёт взболтанный мартини в $\geqslant 12$ случаях из 16 при справедливости H_0, равна $\frac{5034}{65536} \approx 0.0768$.



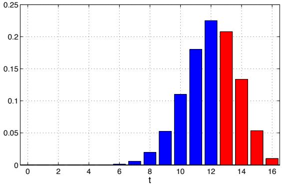
0.0768 — достигаемый уровень значимости при реализации $t = 12$.

Мощность

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



Пусть Джеймс Бонд выбирает взболтанный мартини в 75% случаев.



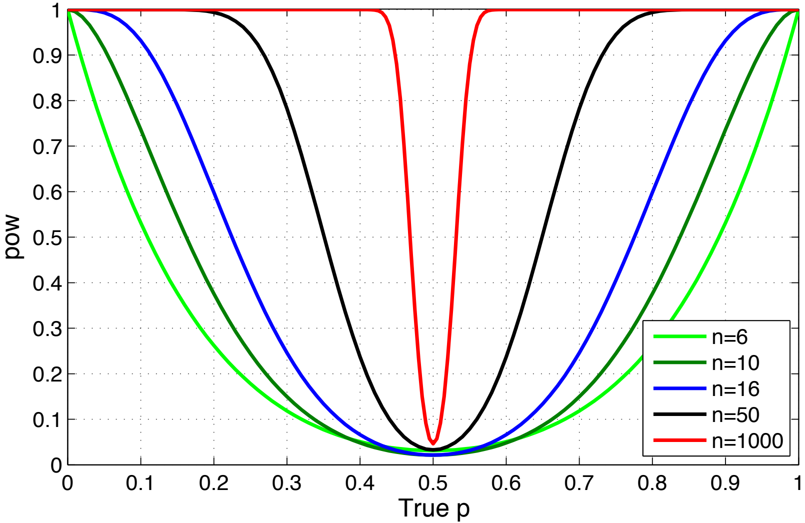
$\text{pow} \approx 0.6202$, т. е., при многократном повторении эксперимента гипотеза будет отклонена только в 62% случаев.

Мощность

Мощность критерия зависит от следующих факторов:

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

Мощность



1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
pow
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
True p
n=6
n=10
n=16
n=50
n=1000

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.

Вес детей при рождении

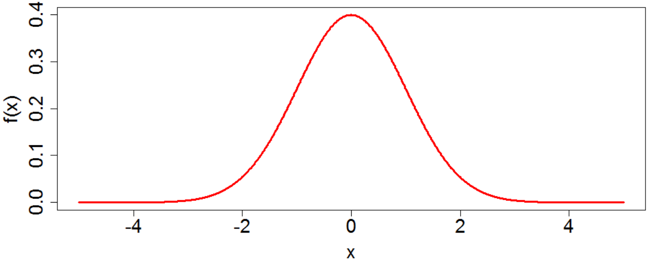
Средний вес детей при рождении — 3.3 кг, у женщин, живущих за чертой бедности — 2.8 кг.

25 женщин, живущих за чертой бедности, участвовали в экспериментальной программе ведения беременности. Средний вес их детей при рождении составил 3075 г, стандартное отклонение 500 г.

Эффективна ли программа?

Z-критерий

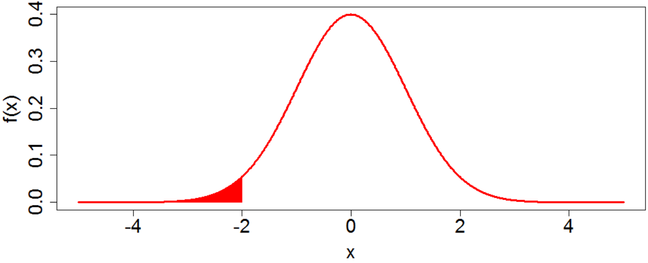
выборка:	$X^n = (X_1, \dots, X_n)$,
	$X \sim N(\mu, \sigma^2)$, σ известна;
нулевая гипотеза:	$H_0\colon \mu = \mu_0$;
альтернатива:	$H_1\colon \mu <\neq> \mu_0$;
статистика:	$Z(X^n) = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$;
нулевое распределение:	$Z(X^n) \sim N(0, 1)$.



Z-критерий

Достигаемый уровень значимости:

- при $H_1 : \mu < \mu_0$

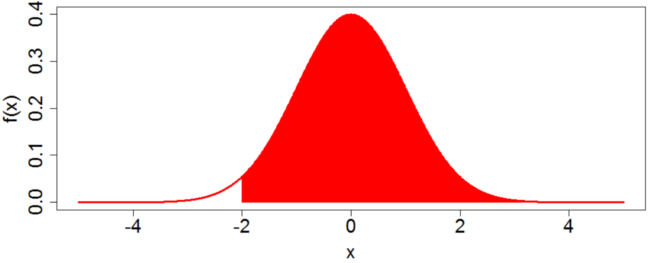


$$p = F_{N(0,1)}\left(z\right).$$

Z-критерий

Достигаемый уровень значимости:

- при $H_1 : \mu > \mu_0$

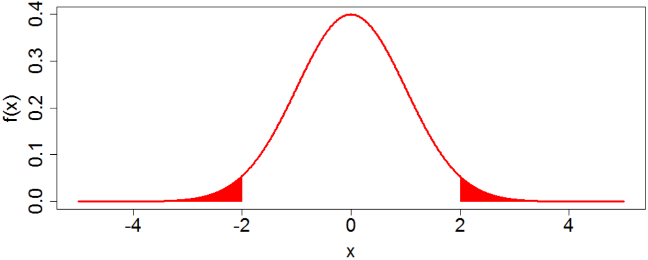


$$p = 1 - F_{N(0,1)}\left(z\right).$$

Z-критерий

Достигаемый уровень значимости:

- при $H_1 : \mu \neq \mu_0$



$$p = 2\left(1 - F_{N(0,1)}\left(|z|\right)\right).$$

t-критерий

выборка: $X^n = (X_1, \dots, X_n)$,

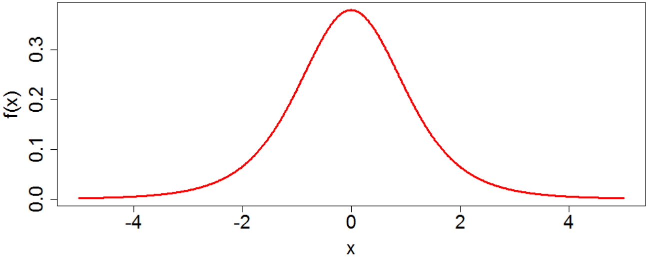
$X \sim N(\mu, \sigma^2)$, σ неизвестна;

нулевая гипотеза: $H_0\colon \mu = \mu_0$;

альтернатива: $H_1\colon \mu <\neq> \mu_0$;

статистика: $T(X^n) = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$;

нулевое распределение: $T(X^n) \sim St(n-1)$.



t-критерий

Достигаемый уровень значимости:

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

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

Вес детей при рождении

$H_0:$ программа неэффективна, $\mu = 2800$.
$H_1:$ программа как-то влияет на вес детей, $\mu \neq 2800$.

t-критерий: $p = 0.0111$, средний вес детей увеличивается на 275 г (95% доверительный интервал — $[233.7, 316.3]$ г).

Вес детей при рождении

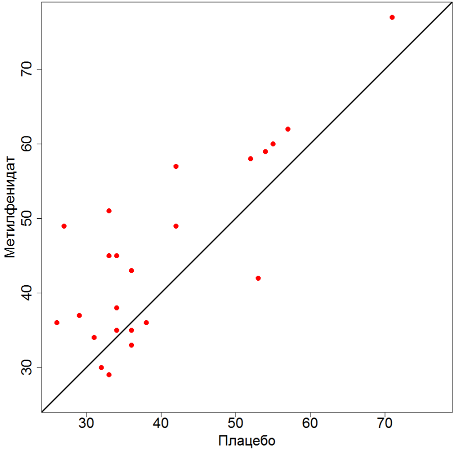
$H_0:$ программа неэффективна, $\mu = 2800$.
$H_1:$ программа эффективна, $\mu > 2800$.

t-критерий: $p = 0.0056$, средний вес детей увеличивается на 275 г (нижний 95% доверительный предел — 240.7 г.).

Одностороннюю альтернативу можно использовать, если знак изменения среднего известен заранее.
Альтернатива должна выбираться до получения данных!

Лечение СДВГ

24 ребёнка прошли тест на способность к подавлению импульсивных поведенческих реакций после недели приёма метилфенидата и после недели приёма плацебо.



Каков эффект препарата?

t-критерий для связанных выборок

выборки: $X_1^n = (X_{11}, \dots, X_{1n}), X_1 \sim N(\mu_1, \sigma_1^2)$,

$X_2^n = (X_{21}, \dots, X_{2n}), X_2 \sim N(\mu_2, \sigma_2^2)$,

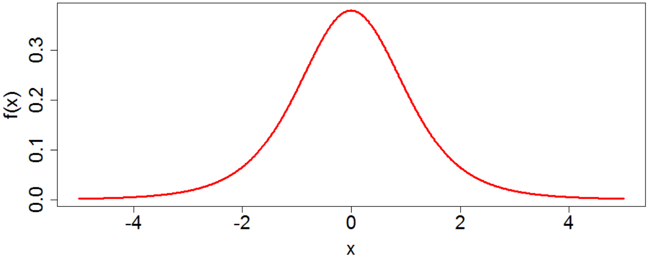
нулевая гипотеза: $H_0\colon \mu_1 = \mu_2$;

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

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

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

нулевое распределение: $T(X_1^n, X_2^n) \sim St(n-1)$.



$\Leftrightarrow$ Переходим от пары связанных выборок к выборке их попарных разностей и применяем одновыборочный t-критерий.

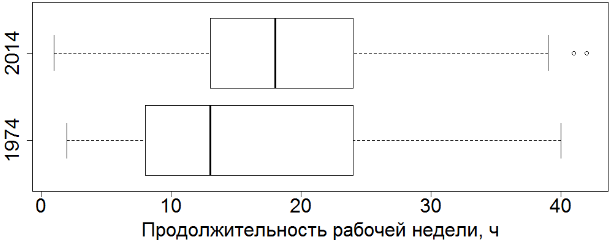
Лечение СДВГ

H_0: способность к подавлению импульсивных поведенческих реакций не изменилась, $\mu_1 = \mu_2$.
H_1: способность к подавлению импульсивных поведенческих реакций изменилась, $\mu_1 \neq \mu_2$.

t-критерий: $p = 0.00377$, средняя способность к подавлению импульсивных поведенческих реакций увеличилась на 4.95 пунктов (95% доверительный интервал — $[1.78, 8.14]$ пунктов).

Продолжительность рабочей недели

В 1974 году 108 респондентов GSS работали неполный день, в 2014 — 196. Для каждого из них известно количество рабочих часов за неделю, предшествующую опросу.



Изменилось ли среднее время работы у работающих неполный день?

t-критерий

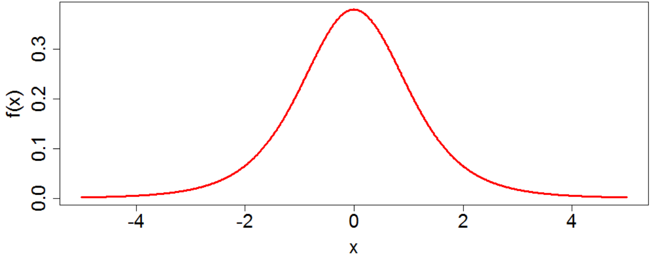
выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1})$,
$X_2^{n_2} = (X_{21}, \dots, X_{2n_2})$,
$X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$,
σ_1, σ_2 неизвестны;

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

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

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

нулевое распределение: $T(X_1^{n_1}, X_2^{n_2}) \approx\sim St(\nu)$.



t-критерий

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

Нулевое распределение приближённое, а не точное.
Точного решения не существует! (проблема Беренца-Фишера)

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

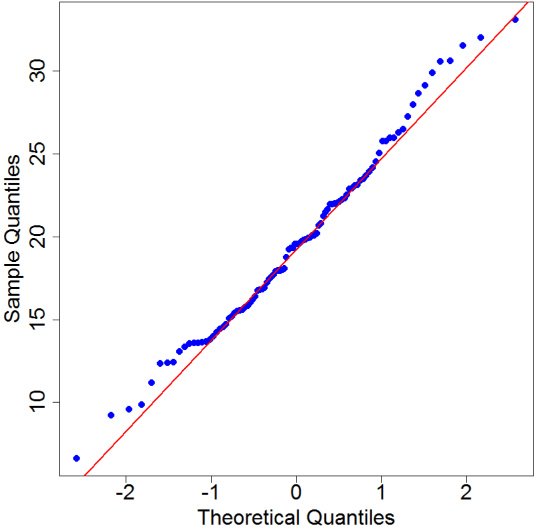
Продолжительность рабочей недели

H_0: среднее время работы не изменилось, $\mu_1 = \mu_2$.
H_1: среднее время работы изменилось, $\mu_1 \neq \mu_2$.

t-критерий: $p = 0.02707$, средняя продолжительность рабочей недели увеличилась на 2.57 часов (95% доверительный интервал — $[0.29, 4.85]$ ч).

Q-Q plot

Визуальный метод проверки согласия выборки и распределения — ку-ку график:



Критерий Шапиро-Уилка

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

a_i основаны на матожиданиях порядковых статистик нормального распределения и также табулированы.

Критерий проверяет, сильно ли точки на ку-ку графике отклоняются от прямой.

Другие критерии для проверки нормальности

Хи-квадрат, Харке-Бера, Колмогорова (Лиллиефорса), Крамера-фон Мизеса, Андерсона-Дарлинга, ...

???

Зачем проверять нормальность?

- на маленьких выборках нормальность, скорее всего, не отвергается
- на больших выборках нормальность, скорее всего, отвергается
- многие методы нечувствительны к отклонениям от нормальности (например, критерии Стьюдента)

«Все модели неверны, но некоторые полезны» (Джордж Бокс)

Как проверять нормальность?

- 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, критерии для дисперсии), проверять её рекомендуется критерием Шапиро-Уилка
- если нормальность отвергается, чувствительные методы, предполагающие нормальность, использовать нельзя!

Непараметрические критерии

$X^n = (X_1, \dots, X_n)\,, X \sim F(x)$
Равно ли среднее X нулю?

Статистика T; нулевое распределение — ?

Проблемы:

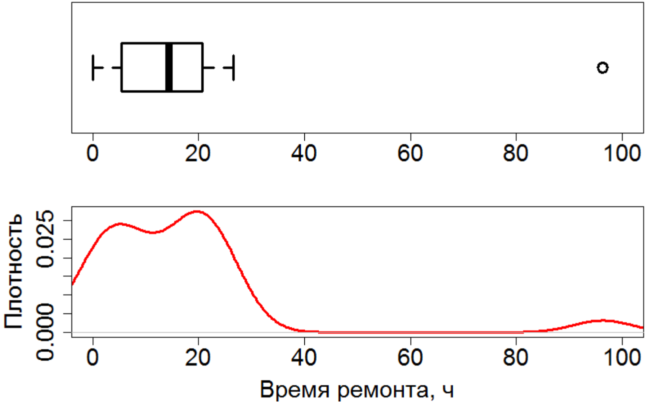
- распределение $F(x)$ может быть нестандартным
- ЦПТ работает не всегда

Решения:

- превратить выборку во что-то более понятное
- сделать какие-то предположения о $F(x)$

Время ремонта

Время ремонта оборудования местных клиентов провайдера Verizon $(n = 23)$:



Можно ли утверждать, что среднее время больше восьми часов?

Одновыборочный критерий знаков

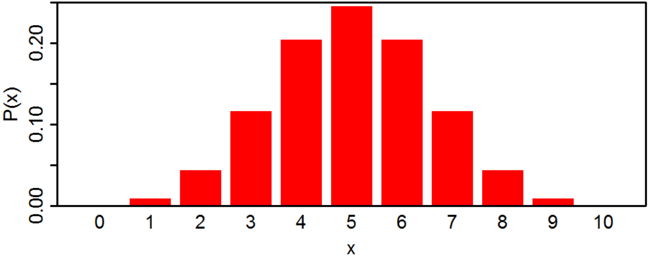
выборка: $X^n = (X_1, \ldots, X_n)\,, X_i \neq m_0;$

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

альтернатива: $H_1\colon\ \operatorname{med} X <\neq> m_0;$

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

нулевое распределение: $T(X^n) \sim Bin(n, \frac{1}{2}).$



Время ремонта

$H_0\colon$ среднее время ремонта — 8 часов, $\operatorname{med} X = 8$.
$H_1\colon$ ремонт в среднем длится дольше 8 часов, $\operatorname{med} X > 8$.

Ремонт занял больше 8 часов в 15 случаях из 23.

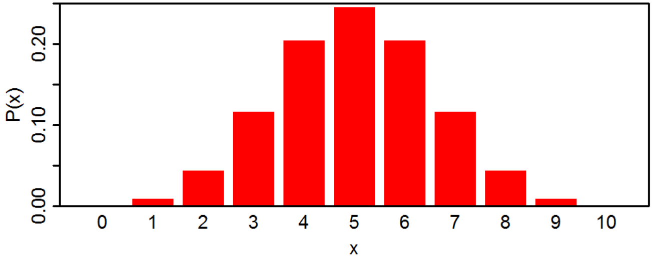
Критерий знаков: $p = 0.105$, нельзя утверждать, что ремонт в среднем длится дольше 8 часов.

Качество классификаторов

	$\mathrm{AUC}_{C4.5}$	$\mathrm{AUC}_{C4.5+m}$
adult (sample)	0.763	**0.768**
breast cancer	**0.599**	0.591
breast cancer wisconsin	0.954	**0.971**
cmc	0.628	**0.661**
ionosphere	0.882	**0.888**
iris	**0.936**	0.931
liver disorders	0.661	**0.668**
lung cancer	0.583	0.583
lymphography	0.775	**0.838**
mushroom	1.000	1.000
primary tumor	0.940	**0.962**
rheum	0.619	**0.666**
voting	0.972	**0.981**
wine	0.957	**0.978**

Двухвыборочный критерий знаков

выборки:	$X_1^n = (X_{11}, \dots, X_{1n})$, $X_2^n = (X_{21}, \dots, X_{2n})$, $X_{1i} \neq X_{2i}$, 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:	$Bin(n, \frac{1}{2})$.



Качество классификаторов

$H_0:$ у классификаторов одинаковое среднее качество,
$\mathbf{P}(\mathrm{AUC}_{C4.5+m} > \mathrm{AUC}_{C4.5}) = \frac{1}{2}$.
$H_1:$ среднее качество модифицированного классификатора выше,
$\mathbf{P}(\mathrm{AUC}_{C4.5+m} > \mathrm{AUC}_{C4.5}) > \frac{1}{2}$.

Модифицированный алгоритм выигрывает на 10 датасетах из 14, ещё на 2 ничья.

Критерий знаков: $p = 0.019$, модифицированный алгоритм лучше на 83% датасетов (95% нижний доверительный предел — 56.2%).

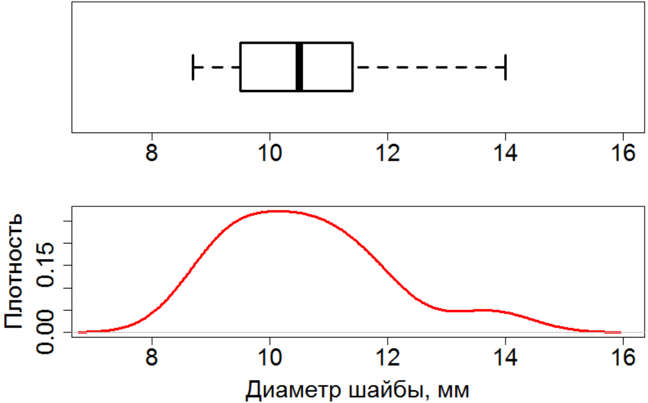
Вариационный ряд

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

Ранг наблюдения X_i:
если X_i не в связке, то $\operatorname{rank}(X_i) = r\colon X_i = X_{(r)}$,
если X_i в связке $X_{(k_1)}, \dots, X_{(k_2)}$, то $\operatorname{rank}(X_i) = \frac{k_1+k_2}{2}$.

Диаметр шайбы

Диаметры шайб на производстве ($n = 24$):



Соответствуют ли шайбы стандартному размеру 10 мм?

Критерий знаковых рангов

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

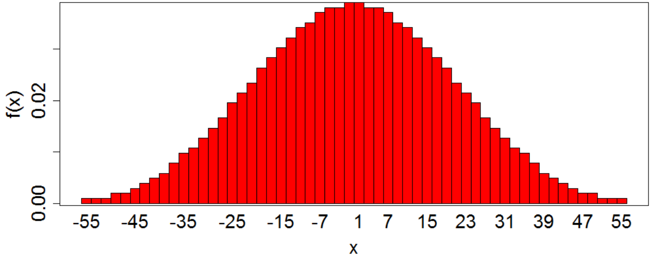
F_X симметрично относительно медианы;

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

альтернатива: $H_1\colon\ \operatorname{med} X <\neq> m_0;$

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

нулевое распределение: табличное.



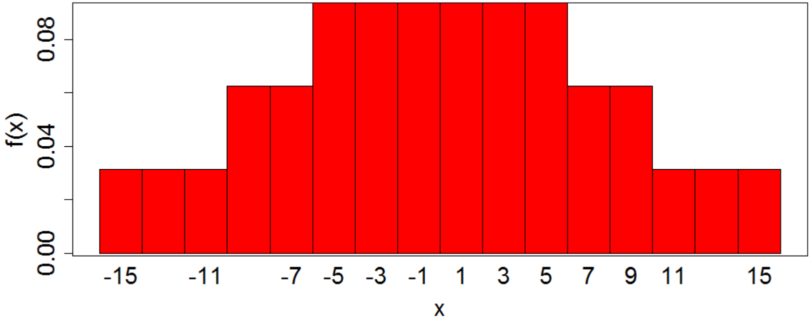
Нулевое распределение

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

Всего 2^n вариантов.

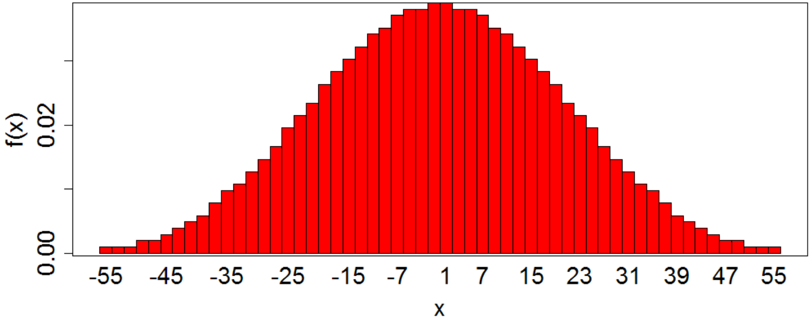
Нулевое распределение

$n = 5$:



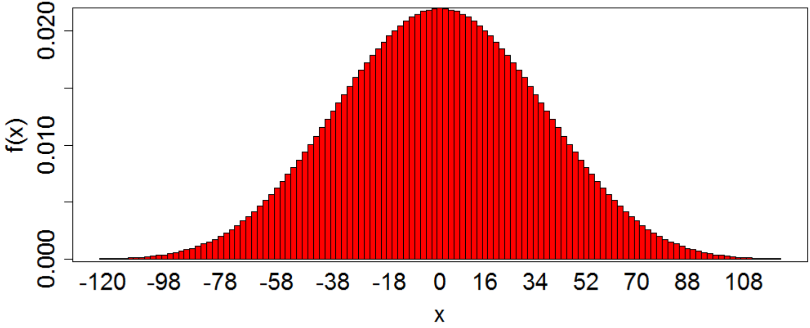
Нулевое распределение

$n = 10$:



Нулевое распределение

$n = 15$:



Аппроксимация для $n > 20$:

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

Диаметр шайбы

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

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

Критерий знаковых рангов

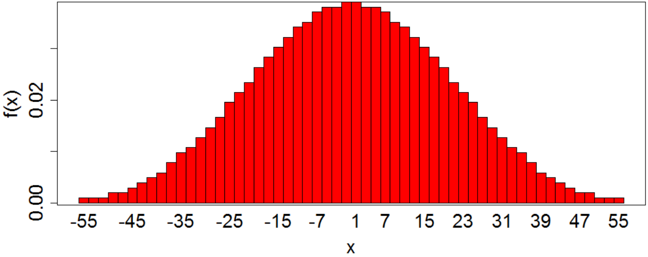
выборки: $X_1^n = (X_{11}, \dots, X_{1n})$,
$X_2^n = (X_{21}, \dots, X_{2n})$,
$X_{1i} \neq X_{2i}$, выборки связанные;

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

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

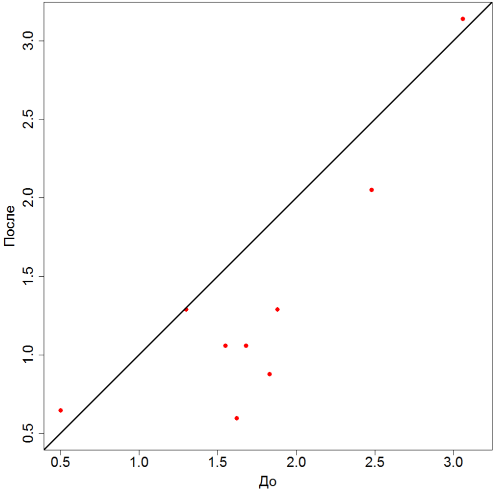
статистика: $W(X_1^n, X_2^n) = \sum\limits_{i=1}^{n} \operatorname{rank}(|X_{1i} - X_{2i}|) \cdot \operatorname{sign}(X_{1i} - X_{2i})$;

нулевое распределение: табличное.

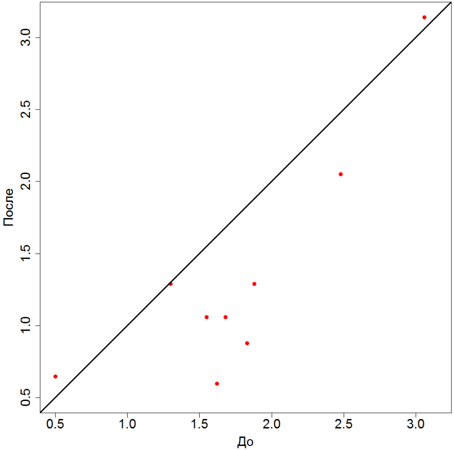


Лечение депрессии

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



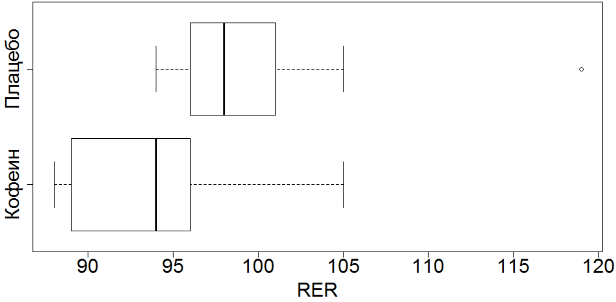
Лечение депрессии



$H_0:$ депрессивность не изменилась, $\operatorname{med}(X_2 - X_1) = 0$.
$H_1:$ депрессивность снизилась, $\operatorname{med}(X_2 - X_1) < 0$.
Критерий знаковых рангов: $p = 0.019$, медиана снижения — 0.49 пт (95% нижний доверительный предел — 0.175 пт).

Кофеин и респираторный обмен

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?

Критерий Манна-Уитни

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

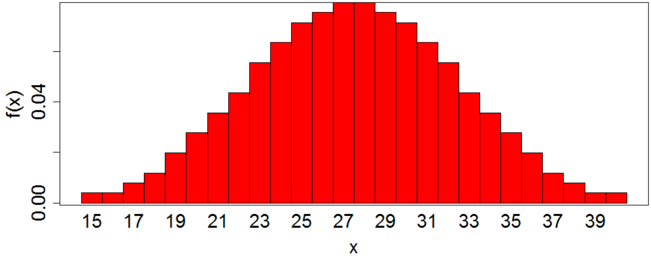
нулевая гипотеза: $H_0\colon F_{X_1}(x) = F_{X_2}(x)$;

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

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

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

нулевое распределение: табличное.



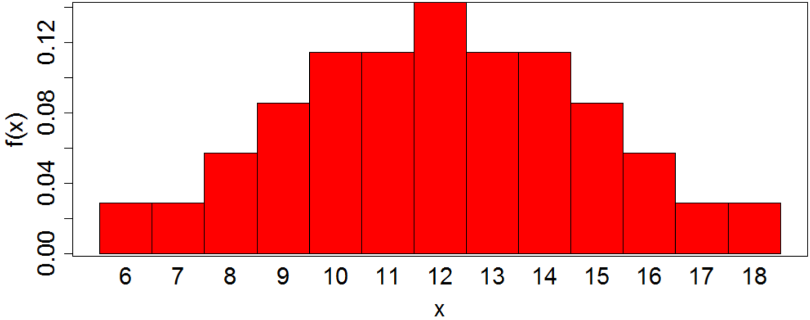
Нулевое распределение

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.

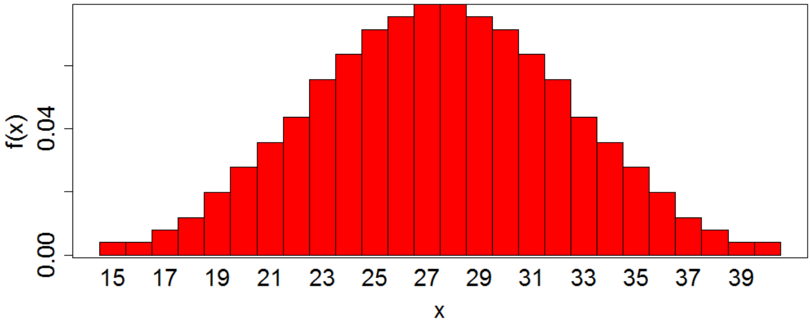
Нулевое распределение

$n_1 = 3, n_2 = 4$:



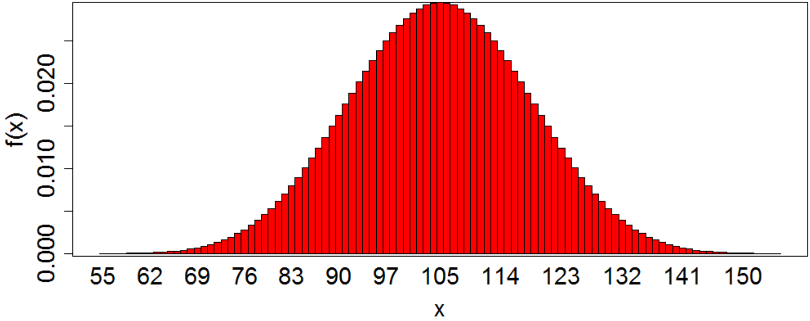
Нулевое распределение

$n_1 = n_2 = 5$:



Нулевое распределение

$n_1 = n_2 = 10$:



Аппроксимация для $n_1, n_2 > 10$:

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

Кофеин и респираторный обмен

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

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

Перестановочные критерии

Ранговые критерии:

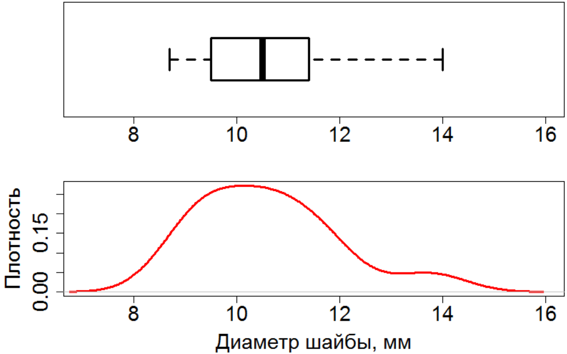
1. выборки $\Rightarrow$ ранги
2. дополнительное предположение
3. перестановки $\Rightarrow$ нулевое распределение статистики

Что если пропустить пункт 1?

Одновыборочный критерий

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 $X_i - m_0$.

Диаметр шайбы

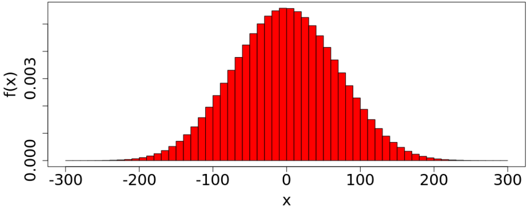


H_0: средний диаметр шайбы — 10 мм, $\mathbb{E}X = 10$.

H_1: средний диаметр шайбы не соответствует стандарту, $\mathbb{E}X \neq 10$.

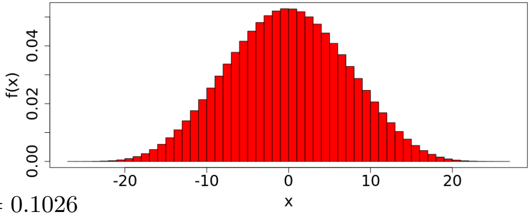
Диаметр шайбы

Критерий знаковых рангов:



$p = 0.0673$

Перестановочный критерий:



$T = 14.6, p = 0.1026$

Для связанных выборок

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

нулевая гипотеза: $H_0 \colon \mathbb{E}(X_1 - X_2) = 0$;

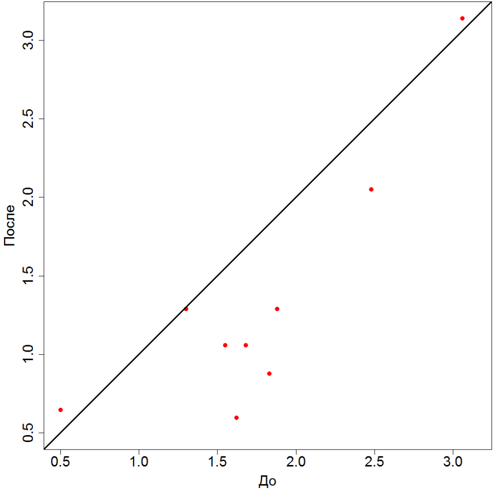
альтернатива: $H_1 \colon \mathbb{E}(X_1 - X_2) <\neq> 0$;

статистика: $D^n = (X_{1i} - X_{2i})$,

$$T(X_1^n, X_2^n) = T(D^n) = \sum_{i=1}^{n} D_i,$$

нулевое распределение: порождается перебором 2^n знаков перед слагаемыми D_i.

Лечение депрессии

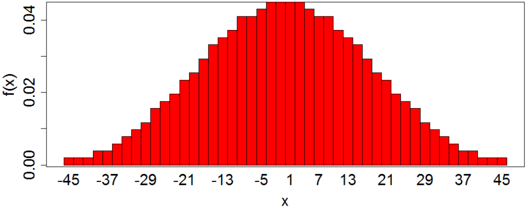


$H_0:$ депрессивность не изменилась, $\mathbb{E}\left(X_1 - X_2\right) = 0.$

$H_1:$ депрессивность снизилась, $\mathbb{E}\left(X_1 - X_2\right) > 0.$

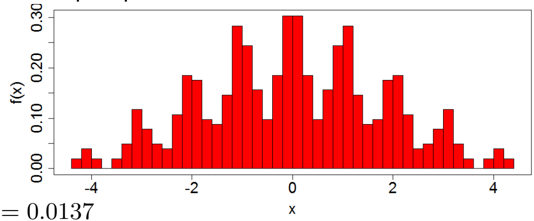
Лечение депрессии

Критерий знаковых рангов:



$p = 0.019$

Перестановочный критерий:



$T = 3.887, p = 0.0137$

Для независимых выборок

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

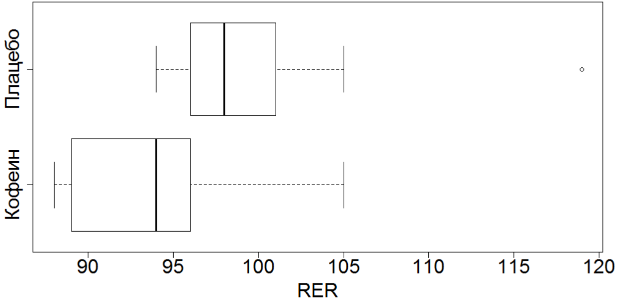
нулевая гипотеза: $H_0\colon F_{X_1}(x) = F_{X_2}(x)$;

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.

Кофеин и респираторный обмен

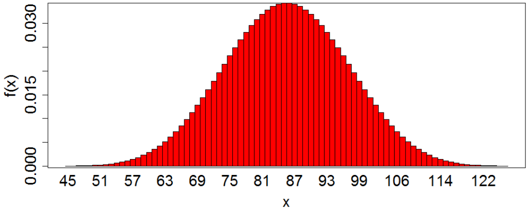


$H_0:$ среднее значение показателя респираторного обмена не отличается в двух группах.

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

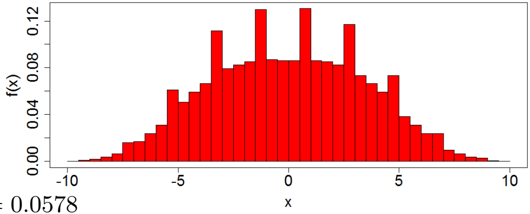
Кофеин и респираторный обмен

Критерий Манна-Уитни:



$p = 0.0521$

Перестановочный критерий:



$T = 6.33, p = 0.0578$

Особенности перестановочных критериев

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

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

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

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

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

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

Литература

Критерии:

- нормальные — Kanji, 1-3, 7-9
- проверка нормальности — Кобзарь, 3.2.2.1
- знаковые — Kanji, 45, 46
- ранговые — Kanji, 47, 48, 52
- перестановочные — Good, 3.2.1, 3.6.4

Кобзарь А.И. *Прикладная математическая статистика*, 2006.

Kanji G.K. *100 statistical tests*, 2006.

Ellis P.D. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*, 2010.

Good P. *Permutation, Parametric and Bootstrap Tests of Hypotheses: A Practical Guide to Resampling Methods for Testing Hypotheses*, 2005.

Kirk R.E. (1996). *Practical Significance: A Concept Whose Time Has Come*. Educational and Psychological Measurement, 56(5), 746–759.

Lee I.-M., Djoussè L., Sesso H.D., Wang L., Buring J.E. (2010). *Physical Activity and Weight Gain Prevention*. JAMA: the Journal of the American Medical Association, 303(12), 1173–1179.