Сравнение двух групп

Основы биостатистики, осень 2022

Марина Варфоломеева

Сравнение двух групп

- Независимые выборки и парные наблюдения
- Парные наблюдения: два вида снотворного
    - Доверительный интервал к средней разнице значений
    - Парный t-тест
- Независимые выборки: гормоны и артериальная гипертензия
    - Доверительный интервал к разнице средних
    - Двухвыборочный t-тест
    - t-тест Стьюдента
    - t-тест Уэлча (разные SD)
- Множественное тестирование гипотез
- Интерпретация доверительных интервалов на графиках
- Сравнение дисперсий

Независимые выборки или парные наблюдения?

Независимые выборки или парные наблюдения

Два варианта снотворных. Какой из них лучше?

Различается ли изменение продолжительности сна (в часах) при применении этих двух снотворных?

Независимые выборки или парные наблюдения

Два варианта снотворных. Какой из них лучше?

Различается ли изменение продолжительности сна (в часах) при применении этих двух снотворных?

Ответить на вопрос могут два варианта эксперимента.

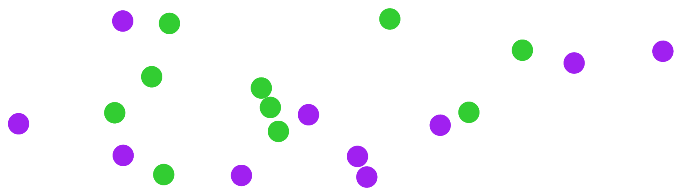
Независимые выборки или парные наблюдения

Два варианта снотворных. Какой из них лучше?

Различается ли изменение продолжительности сна (в часах) при применении этих двух снотворных?

Ответить на вопрос могут два варианта эксперимента.

Независимые выборки



- Наблюдения сделаны на независимых объектах.
- 20 человек случайным образом разделили на две группы (снотворное 1 или 2).
- **Разница средних значений** изменения продолжительности сна в двух группах — это разница влияния этих двух снотворных.

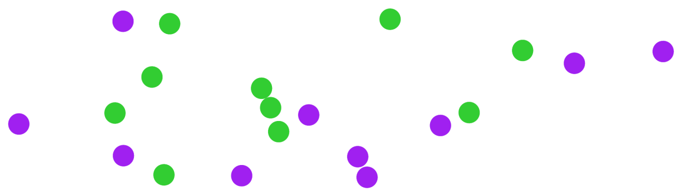
Независимые выборки или парные наблюдения

Два варианта снотворных. Какой из них лучше?

Различается ли изменение продолжительности сна (в часах) при применении этих двух снотворных?

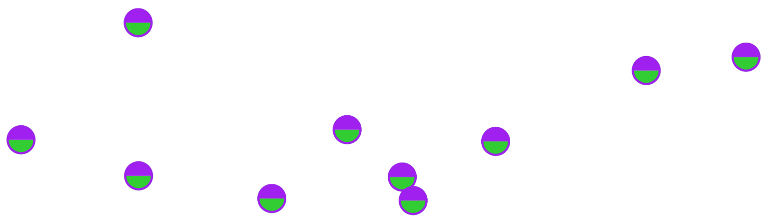
Ответить на вопрос могут два варианта эксперимента.

Независимые выборки



- Наблюдения сделаны на независимых объектах.
- 20 человек случайным образом разделили на две группы (снотворное 1 или 2).
- **Разница средних значений** изменения продолжительности сна в двух группах — это разница влияния этих двух снотворных.

Парные наблюдения



- Наблюдения взаимозависимые (парные).
- 10 человек, с каждым последовательно провели оба варианта эксперимента (снотворное 1 или 2).
- **Средняя разница** изменения продолжительности сна человека со снотворным 1 и со снотворным 2 — это разница влияния этих двух снотворных.

Парные наблюдения

Пример: Два вида снотворного

Пример: Два вида снотворного

В датасете `sleep` содержатся данные об увеличении продолжительности сна по сравнению с контролем после применения двух снотворных препаратов (Cushny, Peebles, 1905, Student, 1908).

Одинаково ли два снотворных влияют на увеличение продолжительности сна?

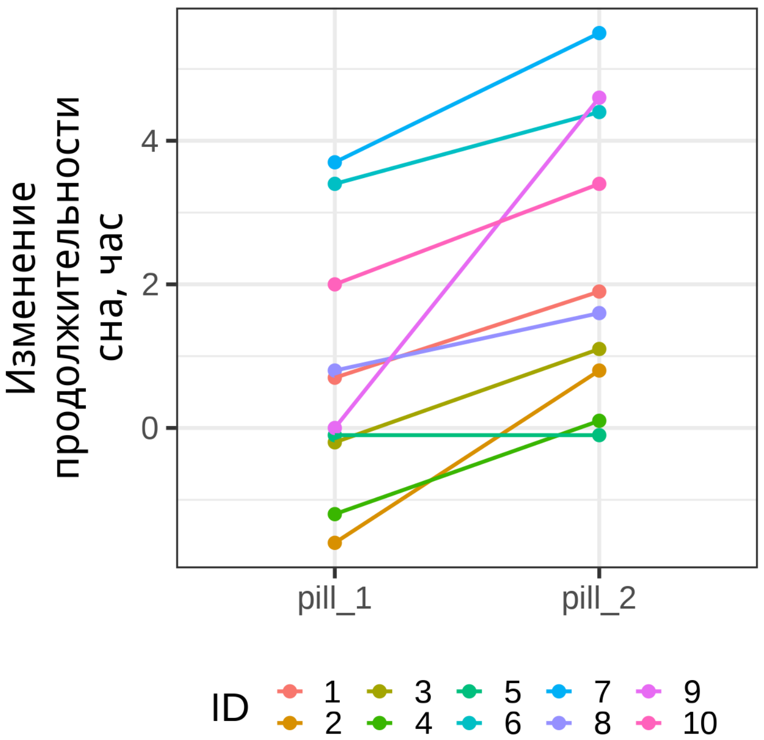
ID	pill_1	pill_2
1	0.7	1.9
2	-1.6	0.8
3	-0.2	1.1
4	-1.2	0.1
5	-0.1	-0.1
6	3.4	4.4
7	3.7	5.5
8	0.8	1.6
9	0.0	4.6
10	2.0	3.4

Пример: Два вида снотворного

В датасете `sleep` содержатся данные об увеличении продолжительности сна по сравнению с контролем после применения двух снотворных препаратов (Cushny, Peebles, 1905, Student, 1908).

Одинаково ли два снотворных влияют на увеличение продолжительности сна?

ID	pill_1	pill_2
1	0.7	1.9
2	-1.6	0.8
3	-0.2	1.1
4	-1.2	0.1
5	-0.1	-0.1
6	3.4	4.4
7	3.7	5.5
8	0.8	1.6
9	0.0	4.6
10	2.0	3.4



Средняя разница в парах значений

Если посчитать разницу между значениями для одного и того же объекта, то можно превратить парные данные в данные одной единственной выборки.

ID	pill_1	pill_2
1	0.7	1.9
2	-1.6	0.8
3	-0.2	1.1
4	-1.2	0.1
5	-0.1	-0.1
6	3.4	4.4
7	3.7	5.5
8	0.8	1.6
9	0.0	4.6
10	2.0	3.4

ID	d
1	1.2
2	2.4
3	1.3
4	1.3
5	0.0
6	1.0
7	1.8
8	0.8
9	4.6
10	1.4

Средняя разница в парах значений

Две возможных задачи:

1. Можно *оценить среднюю разницу* между изменением продолжительности сна после первого и второго снотворного (т.е. μ_d).
2. Можно проверить, *отличается ли эта средняя разница от нуля* (т.е. $H_0 : \mu_d = 0$).

ID	d
1	1.2
2	2.4
3	1.3
4	1.3
5	0.0
6	1.0
7	1.8
8	0.8
9	4.6
10	1.4

Доверительный интервал к средней разнице значений

Средняя разница значений

ID	pill_1	pill_2	d
1	0.7	1.9	1.2
2	-1.6	0.8	2.4
3	-0.2	1.1	1.3
4	-1.2	0.1	1.3
5	-0.1	-0.1	0.0
6	3.4	4.4	1.0
7	3.7	5.5	1.8
8	0.8	1.6	0.8
9	0.0	4.6	4.6
10	2.0	3.4	1.4

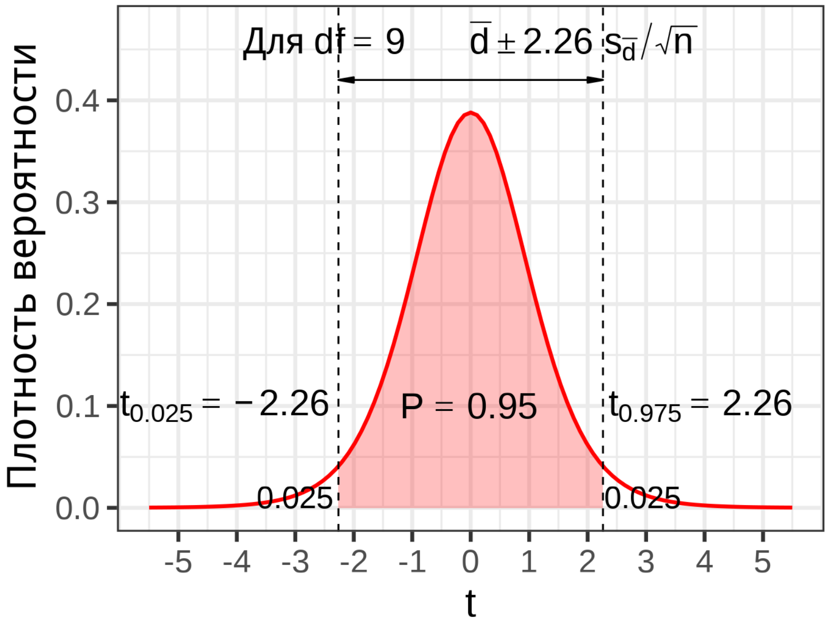
разница (между эффектом снотворных 2 и 1)
$d_i = pill_{2i} - pill_{1i}$

средняя разница
$\bar{d} = 1.58$

выборочное стандартное отклонение разницы
$s_d = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n-1}} = 1.23$

размер выборки
$n = 10$

Доверительный интервал к средней разнице

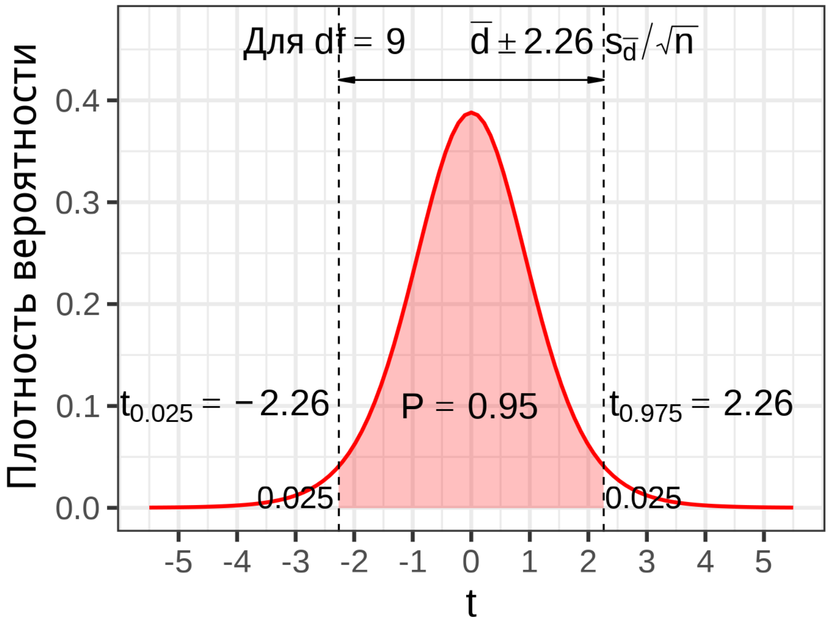


$\bar{d} \pm |t_{\alpha,df}| \cdot SE_{\bar{d}}$

$df = n - 1$

$SE_{\bar{d}} = \frac{s_d}{\sqrt{n}}$

Доверительный интервал к средней разнице



$\bar{d} \pm |t_{\alpha,df}| \cdot SE_{\bar{d}}$

$df = n - 1$

$SE_{\bar{d}} = \frac{s_d}{\sqrt{n}}$

В нашем примере

$SE_{\bar{d}} = \frac{1.23}{\sqrt{10}} = 0.389$

$df = 10 - 1 = 9$

Таким образом, 95% доверительный интервал для средней разницы эффектов двух снотворных

$\bar{d} \pm |t_{\alpha,df}| \cdot SE_{\bar{d}}$

$1.58 \pm 2.262 \cdot 0.389$

то есть

1.58 ± 0.88

Парный t-тест

Гипотезы в парном t-тесте

На человеческом языке:

H_0 : эффективность снотворных не различается.

H_A : у снотворных разная эффективность.

Гипотезы в парном t-тесте

На человеческом языке:

H_0 : эффективность снотворных не различается.

H_A : у снотворных разная эффективность.

На языке статистики:

$H_0 : \mu_d = 0$ — средняя разница значений в парах равна нулю

$H_A : \mu_d \neq 0$ — средняя разница значений в парах не равна нулю

Парный t-тест

$H_0 : \mu_d = 0$ — средняя разница значений в парах равна нулю

$H_A : \mu_d \neq 0$ — средняя разница значений в парах не равна нулю

Парный t-тест

$H_0 : \mu_d = 0$ — средняя разница значений в парах равна нулю

$H_A : \mu_d \neq 0$ — средняя разница значений в парах не равна нулю

t-тест в общем виде:

Парный t-тест

$H_0 : \mu_d = 0$ — средняя разница значений в парах равна нулю

$H_A : \mu_d \neq 0$ — средняя разница значений в парах не равна нулю

t-тест в общем виде:

$$t = \frac{\text{Наблюдаемая величина - Ожидаемое значение при } H_0}{\text{Стандартная ошибка}}$$

Парный t-тест

$H_0 : \mu_d = 0$ — средняя разница значений в парах равна нулю

$H_A : \mu_d \neq 0$ — средняя разница значений в парах не равна нулю

t-тест в общем виде:

$$t = \frac{\text{Наблюдаемая величина - Ожидаемое значение при } H_0}{\text{Стандартная ошибка}}$$

$$t = \frac{\bar{d} - \mu_d}{SE_{\bar{d}}} \sim t(df)$$

$df = n - 1$

Парный t-тест

$H_0 : \mu_d = 0$ — средняя разница значений в парах равна нулю

$H_A : \mu_d \neq 0$ — средняя разница значений в парах не равна нулю

t-тест в общем виде:

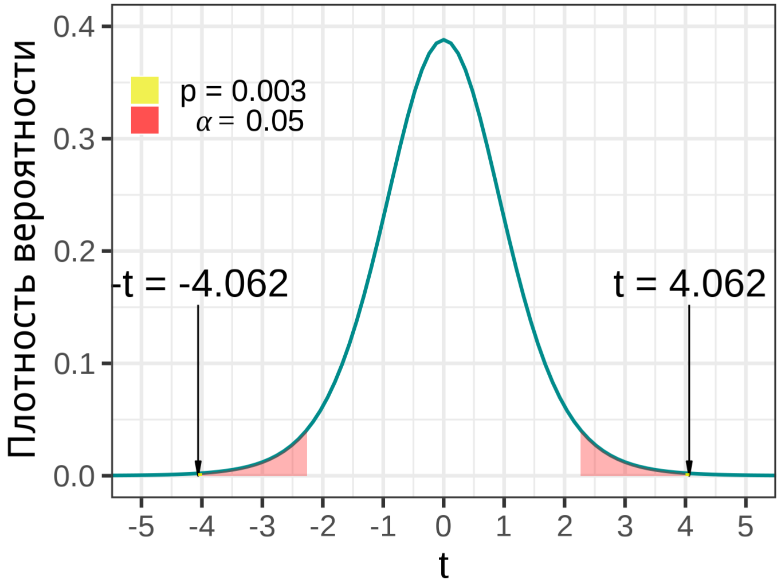
$$t = \frac{\text{Наблюдаемая величина - Ожидаемое значение при } H_0}{\text{Стандартная ошибка}}$$

$$t = \frac{\bar{d} - \mu_d}{SE_{\bar{d}}} \sim t(df)$$

$df = n - 1$

Парный t-тест — это всего лишь одновыборочный t-тест для средней разницы значений в парах.

Парный t-тест в нашем примере

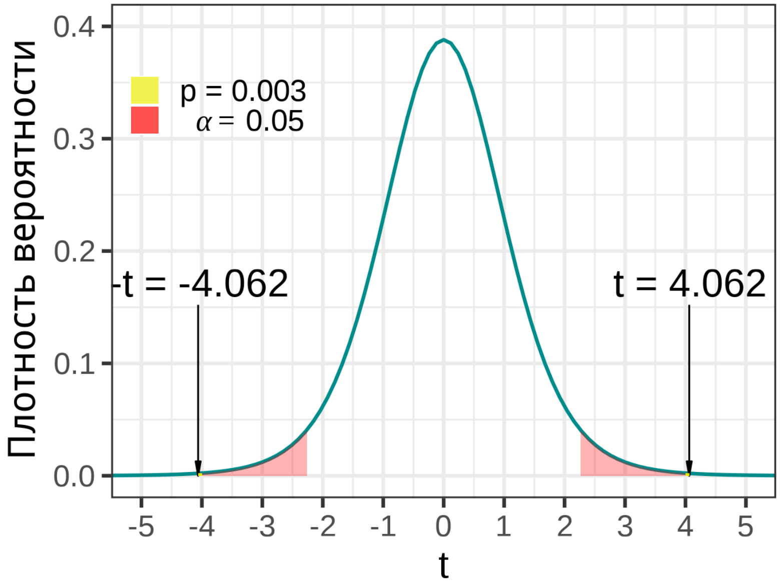


$$t = \frac{\bar{d} - \mu_d}{SE_{\bar{d}}}$$

$df = n - 1$

$SE_{\bar{d}} = \frac{s_d}{\sqrt{n}}$

Парный t-тест в нашем примере



$$t = \frac{\bar{d} - \mu_d}{SE_{\bar{d}}}$$

$df = n - 1$

$SE_{\bar{d}} = \frac{s_d}{\sqrt{n}}$

В нашем примере

$SE_{\bar{d}} = \frac{1.23}{\sqrt{10}} = 0.389$

$df = 10 - 1 = 9$

Таким образом, наблюдаемое значение t-статистики

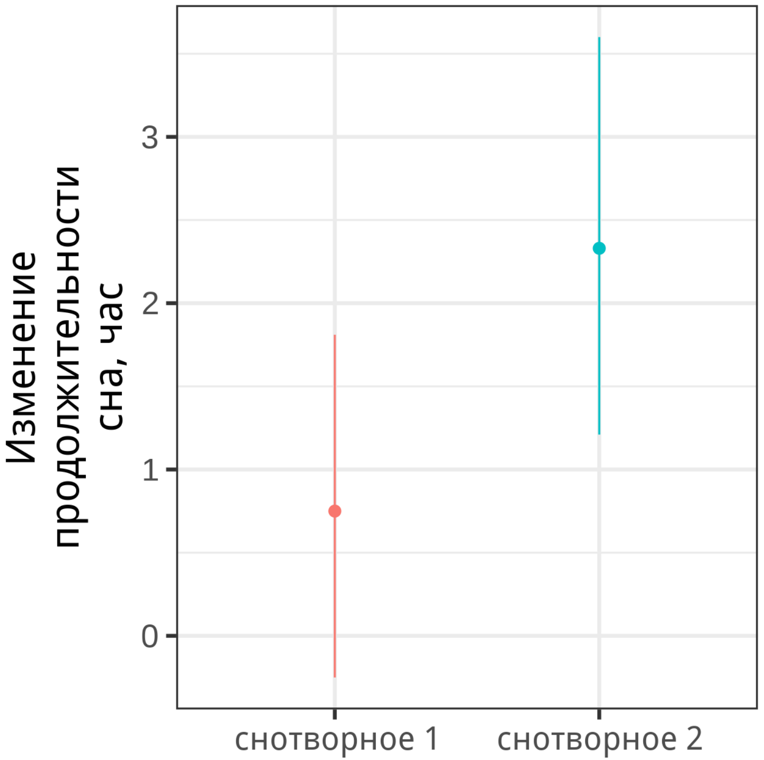
$t = \frac{\bar{d} - \mu_d}{SE_{\bar{d}}} = \frac{1.58 - 0}{0.389} = 4.062$

Ему соответствует $p = 0.003$

Отвергаем H_0

Опишем результаты

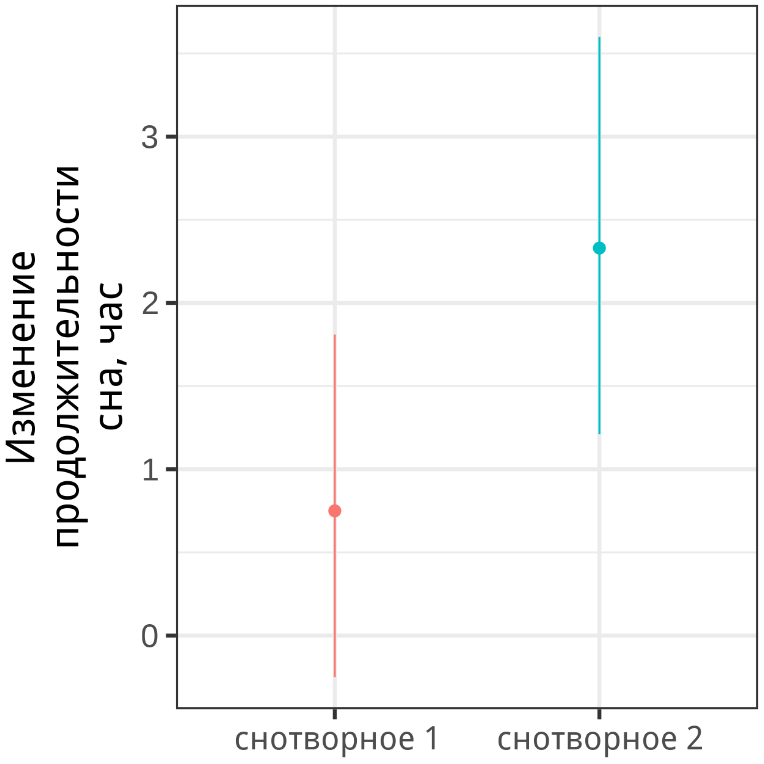
Различия изменения продолжительности сна при применении двух препаратов были статистически значимы ($t_9 = -4.06, p =< 0.01$)



Опишем результаты

Различия изменения продолжительности сна при применении двух препаратов были статистически значимы ($t_9 = -4.06, p =< 0.01$)

Осторожно! Если бы мы не учли зависимость между группами, то пришли бы к неверному выводу.

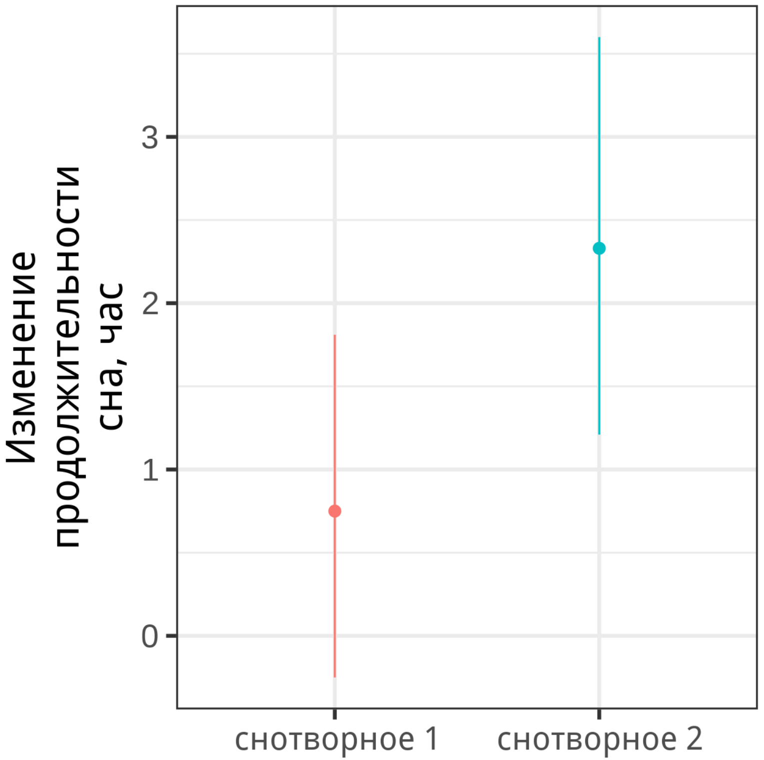


Опишем результаты

Различия изменения продолжительности сна при применении двух препаратов были статистически значимы ($t_9 = -4.06, p =< 0.01$)

Осторожно! Если бы мы не учли зависимость между группами, то пришли бы к неверному выводу.

Кроме того, прежде чем описывать результаты, хорошо бы проверить выполнение условий применимости теста.



Условия применимости одновыборочного t-теста

Прежде чем описывать результаты, нужно проверить выполнение условий применимости теста.

Условия применимости одновыборочного t-теста

Прежде чем описывать результаты, нужно проверить выполнение условий применимости теста.

1.Объекты исследования независимы друг от друга?

Условия применимости одновыборочного *t*-теста

Прежде чем описывать результаты, нужно проверить выполнение условий применимости теста.

1.Объекты исследования независимы друг от друга?

- Да, независимы. Это случайная выборка людей.

Условия применимости одновыборочного t-теста

Прежде чем описывать результаты, нужно проверить выполнение условий применимости теста.

1.Объекты исследования независимы друг от друга?

- Да, независимы. Это случайная выборка людей.

2.Наблюдения зависимы друг от друга?

Условия применимости одновыборочного t-теста

Прежде чем описывать результаты, нужно проверить выполнение условий применимости теста.

1.Объекты исследования независимы друг от друга?

- Да, независимы. Это случайная выборка людей.

2.Наблюдения зависимы друг от друга?

- Да, одни и те же люди участвуют в обоих вариантах исследования (10 человек, каждый пил оба снотворных). Эти данные подходят для парного t-теста.

Условия применимости одновыборочного t-теста

Прежде чем описывать результаты, нужно проверить выполнение условий применимости теста.

1.Объекты исследования независимы друг от друга?

- Да, независимы. Это случайная выборка людей.

2.Наблюдения зависимы друг от друга?

- Да, одни и те же люди участвуют в обоих вариантах исследования (10 человек, каждый пил оба снотворных). Эти данные подходят для парного t-теста.

3.Объем выборки достаточно велик или разницы значений d нормально распределены?

Условия применимости одновыборочного t-теста

Прежде чем описывать результаты, нужно проверить выполнение условий применимости теста.

1.Объекты исследования независимы друг от друга?

- Да, независимы. Это случайная выборка людей.

2.Наблюдения зависимы друг от друга?

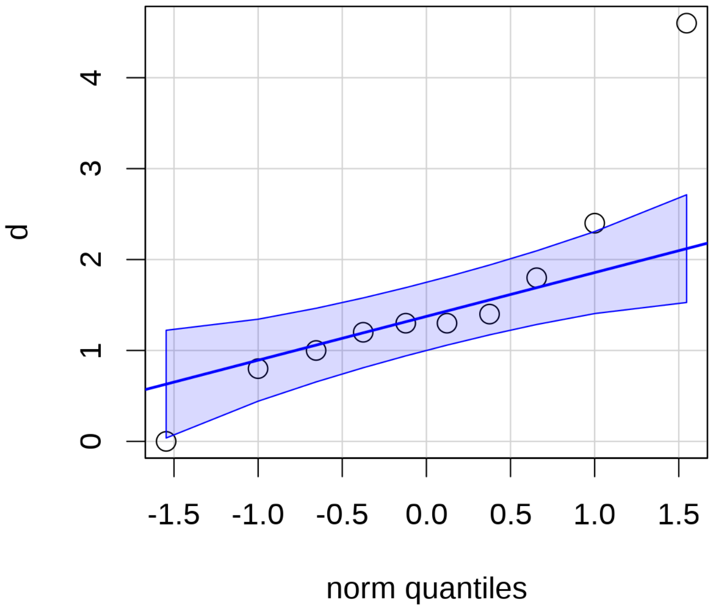
- Да, одни и те же люди участвуют в обоих вариантах исследования (10 человек, каждый пил оба снотворных). Эти данные подходят для парного t-теста.

3.Объем выборки достаточно велик или разницы значений d нормально распределены?

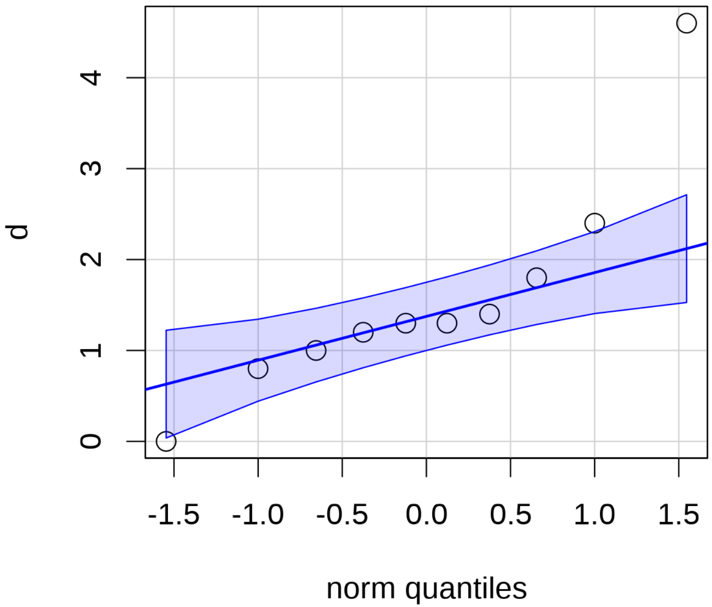
- Объем выборки мал, $n = 10$

Нужно проверить форму распределения разницы значений.

Нормально ли распределена разница значений?



Нормально ли распределена разница значений?



Могло бы быть и лучше. Одно наблюдение сильно отклоняется от ожиданий для нормального распределения.

Для этих данных хорошо бы использовать другой, непараметрический тест, у которого нет таких жестких требований к данным. Позже мы с такими тестами познакомимся.

Независимые выборки

Пример: Гормоны и артериальная гипертензия

Пример: Гормоны и артериальная гипертензия

Синдром Кушинга — это нарушения уровня артериального давления, вызванные гиперсекрецией кортизола надпочечниками.

В датасете `Cushings` (пакет `MASS`) записаны данные о секреции его метаболита тетрагидрокортизона с мочой (мг/сут.) при разных типах синдрома Кушинга (данные из кн. Aitchison, Dunsmore, 1975):

- `a` — аденома
- `b` — двусторонняя гиперплазия
- `c` — карцинома
- `u` — другие патологии (без уточнения)

Пусть нас интересует только две из этих групп: аденома и двусторонняя гиперплазия надпочечников.

Пример: Гормоны и артериальная гипертензия

Синдром Кушинга — это нарушения уровня артериального давления, вызванные гиперсекрецией кортизола надпочечниками.

В датасете `Cushings` (пакет `MASS`) записаны данные о секреции его метаболита тетрагидрокортизона с мочой (мг/сут.) при разных типах синдрома Кушинга (данные из кн. Aitchison, Dunsmore, 1975):

- `a` — аденома
- `b` — двусторонняя гиперплазия
- `c` — карцинома
- `u` — другие патологии (без уточнения)

Пусть нас интересует только две из этих групп: аденома и двусторонняя гиперплазия надпочечников.

Две возможных задачи:

1. Можно оценить разницу среднего уровня секреции гидрокортизона при этих двух патологиях (т.е. $\mu_1 - \mu_2$).
2. Можно проверить, отличается ли эта разница от нуля (т.е. $H_0: \mu_1 - \mu_2 = 0$).

Доверительный интервал к разнице средних

Разница выборочных средних

Чему равна $\mu_1 - \mu_2$ разница среднего уровня секреции гидрокортизона при аденоме и двусторонней гиперплазии надпочечников в генеральной совокупности?

Это можно оценить по разнице средних в выборках

$\bar{x}_1 - \bar{x}_2 = 2.967 - 8.18 = -5.213$

Чтобы построить к ней доверительный интервал, нужно понять, как она распределена.

Распределение разницы выборочных средних

Если $x_1 \sim N(\mu_1, \sigma_1)$ и $x_2 \sim N(\mu_2, \sigma_2)$ или размер выборки достаточно велик,

то **согласно ЦПТ** мы можем сказать, что их средние значения...

Распределение разницы выборочных средних

Если $x_1 \sim N(\mu_1, \sigma_1)$ и $x_2 \sim N(\mu_2, \sigma_2)$ или размер выборки достаточно велик,

то **согласно ЦПТ** мы можем сказать, что их средние значения...

$\bar{x}_1 \sim N(\mu_1, \sigma_1/\sqrt{n_1})$ и $\bar{x}_2 \sim N(\mu_2, \sigma_2/\sqrt{n_2})$

Распределение разницы выборочных средних

Если $x_1 \sim N(\mu_1, \sigma_1)$ и $x_2 \sim N(\mu_2, \sigma_2)$ или размер выборки достаточно велик,

то **согласно ЦПТ** мы можем сказать, что их средние значения...

$\bar{x}_1 \sim N(\mu_1, \sigma_1/\sqrt{n_1})$ и $\bar{x}_2 \sim N(\mu_2, \sigma_2/\sqrt{n_2})$

Тогда разница выборочных средних (двух нормально распределенных величин) будет нормально распределена

$$\bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2, SE_{\bar{x}_1 - \bar{x}_2})$$

со стандартной ошибкой

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Стандартную ошибку придется оценить по данным выборок.

Оценка стандартной ошибки разницы средних

Стандартная ошибка разницы средних:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Оценка стандартной ошибки разницы средних

Стандартная ошибка разницы средних:

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Если предположить, что дисперсии в группах равны
(вернее, в генеральных совокупностях, откуда были сделаны выборки), то

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})}$$

Оценка стандартной ошибки разницы средних

Стандартная ошибка разницы средних:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Если предположить, что дисперсии в группах равны
(вернее, в генеральных совокупностях, откуда были сделаны выборки), то

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma^2 (\frac{1}{n_1} + \frac{1}{n_2})}$$

Тогда σ^2 можно оценить как $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$

s_p^2 — это **обобщенная дисперсия** (pooled variance), т.е. усредненная дисперсия в двух выборках, взвешенная по их числу степеней свободы.

Оценка стандартной ошибки разницы средних

Стандартная ошибка разницы средних:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Если предположить, что дисперсии в группах равны
(вернее, в генеральных совокупностях, откуда были сделаны выборки), то

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma^2 (\frac{1}{n_1} + \frac{1}{n_2})}$$

Тогда σ^2 можно оценить как $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$

s_p^2 — это **обобщенная дисперсия** (pooled variance), т.е. усредненная дисперсия в двух выборках, взвешенная по их числу степеней свободы.

Т.е.

$$SE_{\bar{x}_1 - \bar{x}_2} \approx \sqrt{s_p^2 (\frac{1}{n_1} + \frac{1}{n_2})}$$

Распределение выборочных средних для построения доверительного интервала разницы средних

Распределение разницы выборочных средних

$$\bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2, SE_{\bar{x}_1 - \bar{x}_2})$$

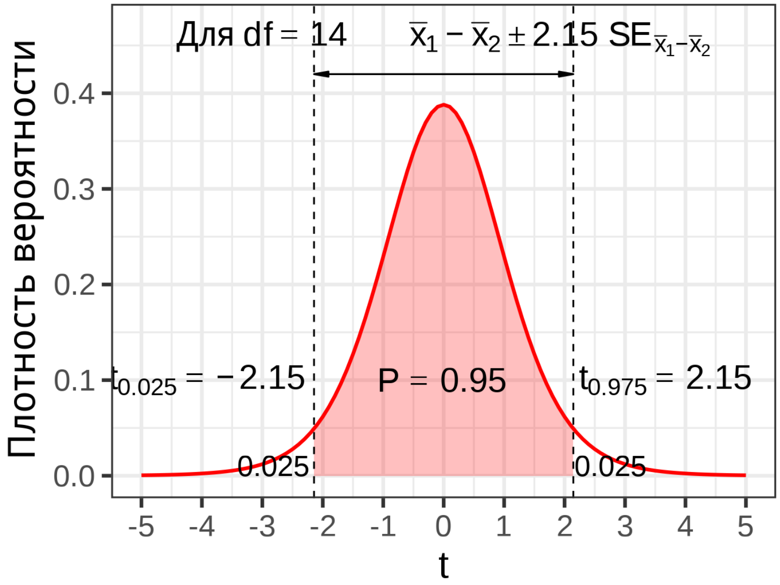
После стандартизации превращается в

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{\bar{x}_1 - \bar{x}_2}} \sim t(df)$$

Т.е. стандартизованная разница средних подчиняется t-распределению с числом степеней свободы $df = n_1 + n_2 - 2$. (Это справедливо с учетом сделанных ранее предположений, т.е. если выполняется ЦПТ и дисперсии в выборках равны).

И это распределение как раз можно использовать для построения доверительного интервала.

Доверительный интервал к разнице средних



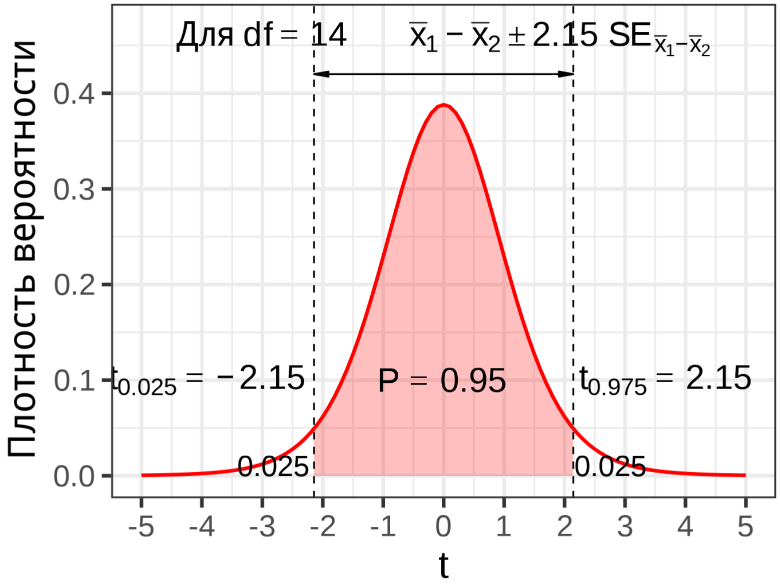
$$\bar{x}_1 - \bar{x}_2 \pm t(\alpha, df) \cdot SE_{\bar{x}_1 - \bar{x}_2}$$

$$df = n_1 + n_2 - 2$$

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Доверительный интервал к разнице средних



$$\bar{x}_1 - \bar{x}_2 \pm t(\alpha, df) \cdot SE_{\bar{x}_1-\bar{x}_2}$$

$$df = n_1 + n_2 - 2$$

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

В нашем примере

$x_1 = 2.967$, $x_2 = 8.18$
$n_1 = 6$, $n_2 = 10$
$sd_1 = 0.924$, $sd_2 = 3.789$

$$s_p^2 = \frac{(6-1)0.924^2 + (10-1)3.789^2}{(6-1)+(10-1)} =$$

$$\frac{133.489}{14} = 9.535$$

$$SE_{\bar{d}} = \sqrt{9.535\left(\frac{1}{6} + \frac{1}{10}\right)} = 1.595$$

$$df = 6 + 10 - 2 = 14$$

95% доверительный интервал для разницы среднего уровня секреции тетрагидрокортизона (мг/сут.) при аденоме и двусторонней гиперплазии надпочечников
$2.967 - 8.18 \pm 2.145 \cdot 1.595$

то есть

$$-5.213 \pm 3.42$$

Двухвыборочный t-тест

Гипотезы в двухвыборочном t-тесте

На человеческом языке:

H_0 : средний уровень секреции тетрагидрокортизона одинаков при аденоме и двусторонней гиперплазии надпочечников.

H_A : при этих двух патологиях средний уровень секреции тетрагидрокортизона разный.

Гипотезы в двухвыборочном t-тесте

На человеческом языке:

H_0 : средний уровень секреции тетрагидрокортизона одинаков при аденоме и двусторонней гиперплазии надпочечников.

H_A : при этих двух патологиях средний уровень секреции тетрагидрокортизона разный.

На языке статистики:

$H_0 : \mu_1 - \mu_2 = 0$ — разность между средними в двух группах равна нулю.

$H_A : \mu_1 - \mu_2 \neq 0$ — между средними в группах есть ненулевая разница.

Двухвыборочный t-тест

$H_0 : \mu_1 - \mu_2 = 0$ — разность между средними в двух группах равна нулю.

$H_A : \mu_1 - \mu_2 \neq 0$ — между средними в группах есть ненулевая разница.

t-тест в общем виде:

$$t = \frac{\text{Наблюдаемая величина - Ожидаемое значение}}{\text{Стандартная ошибка}}$$

Двухвыборочный *t*-тест

$H_0 : \mu_1 - \mu_2 = 0$ — разность между средними в двух группах равна нулю.

$H_A : \mu_1 - \mu_2 \neq 0$ — между средними в группах есть ненулевая разница.

t-тест в общем виде:

$$t = \frac{\text{Наблюдаемая величина - Ожидаемое значение}}{\text{Стандартная ошибка}}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{SE_{\bar{x}_1 - \bar{x}_2}} \sim t(df)$$

Двухвыборочный t-тест

$H_0 : \mu_1 - \mu_2 = 0$ — разность между средними в двух группах равна нулю.

$H_A : \mu_1 - \mu_2 \neq 0$ — между средними в группах есть ненулевая разница.

t-тест в общем виде:

$$t = \frac{\text{Наблюдаемая величина - Ожидаемое значение}}{\text{Стандартная ошибка}}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{SE_{\bar{x}_1 - \bar{x}_2}} \sim t(df)$$

$SE_{\bar{x}_1 - \bar{x}_2}$ — стандартная ошибка разности двух средних, может рассчитываться по-разному

- t-тест Стьюдента — если считать, что дисперсии в группах равны (как мы это только что делали для расчета доверительного интервала)
- t-тест Уэлча — если считать, что дисперсии могут быть разными

Стандартная ошибка разности средних в t-тесте Стьюдента

Student 1908

Если группы независимы и дисперсии в них равны, то по центральной предельной теореме

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} \approx \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

где $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$ — это **обобщенная дисперсия** по двум выборкам.

Результирующая t-статистика подчиняется t-распределению с $df = n_1 + n_2 - 2$.

Стандартная ошибка разности средних в t-тесте Стьюдента

Student 1908

Если группы независимы и дисперсии в них равны, то по центральной предельной теореме

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} \approx \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

где $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$ — это **обобщенная дисперсия** по двум выборкам.

Результирующая t-статистика подчиняется t-распределению с $df = n_1 + n_2 - 2$.

Осторожно! Равенство дисперсий в группах — это часто нереалистичное предположение!

Стандартная ошибка разности средних в t-тесте Уэлча

Если группы независимы и дисперсии в них неизвестны, то получается

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Стандартная ошибка разности средних в t-тесте Уэлча

Если группы независимы и дисперсии в них неизвестны, то получается

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Проблема в том, что эта величина **лишь приблизительно следует t-распределению**, если считать его степени свободы как обычно для двух групп $df = n_1 + n_2 - 2$.

Это из-за того, что мы оцениваем **две** дисперсии при помощи их стандартных отклонений.

Стандартная ошибка разности средних в t-тесте Уэлча

Если группы независимы и дисперсии в них неизвестны, то получается

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Проблема в том, что эта величина **лишь приблизительно следует t-распределению**, если считать его степени свободы как обычно для двух групп $df = n_1 + n_2 - 2$.

Это из-за того, что мы оцениваем **две** дисперсии при помощи их стандартных отклонений.

Приблизительное число степеней свободы можно рассчитать по уравнению Уэлча-Саттеруэйта (Welch-Satterthwaite). Это решит проблему.

$$df_{WS} \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Стандартная ошибка разности средних в t-тесте Уэлча

Если группы независимы и дисперсии в них неизвестны, то получается

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Проблема в том, что эта величина **лишь приблизительно следует t-распределению**, если считать его степени свободы как обычно для двух групп $df = n_1 + n_2 - 2$.

Это из-за того, что мы оцениваем **две** дисперсии при помощи их стандартных отклонений.

Приблизительное число степеней свободы можно рассчитать по уравнению Уэлча-Саттеруэйта (Welch-Satterthwaite). Это решит проблему.

$$df_{WS} \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

t-тестом Уэлча можно пользоваться, даже если дисперсии равны.

Он немного консервативнее, чем тест Стьюдента.

Условия применимости двухвыборочного t-теста

Почти такие же, как условия справедливости ЦПТ

Условия применимости двухвыборочного t-теста

Почти такие же, как условия справедливости ЦПТ

- Наблюдения независимы друг от друга.
- Выборки независимы друг от друга (новое условие).
- Объем выборки достаточно велик или величины нормально распределены.
- Для теста Стьюдента требуется равенство дисперсий в выборках.
  Для теста Уэлча это не требуется.

Проверяем условия применимости...

1. Наблюдения независимы друг от друга?

Проверяем условия применимости...

1.Наблюдения независимы друг от друга?

- Да, независимы. Это случайная выборка.

Проверяем условия применимости...

1.Наблюдения независимы друг от друга?

- Да, независимы. Это случайная выборка.

2.Выборки независимы друг от друга?

Проверяем условия применимости...

1.Наблюдения независимы друг от друга?

- Да, независимы. Это случайная выборка.

2.Выборки независимы друг от друга?

- Да, независимы. В группах разные люди (естественно, т.к. тип синдрома у человека может быть только какой-то один).

Проверяем условия применимости…

1. Наблюдения независимы друг от друга?

- Да, независимы. Это случайная выборка.

2. Выборки независимы друг от друга?

- Да, независимы. В группах разные люди (естественно, т.к. тип синдрома у человека может быть только какой-то один).

3. Объем выборки достаточно велик или величины нормально распределены?

Проверяем условия применимости...

1.Наблюдения независимы друг от друга?

- Да, независимы. Это случайная выборка.

2.Выборки независимы друг от друга?

- Да, независимы. В группах разные люди (естественно, т.к. тип синдрома у человека может быть только какой-то один).

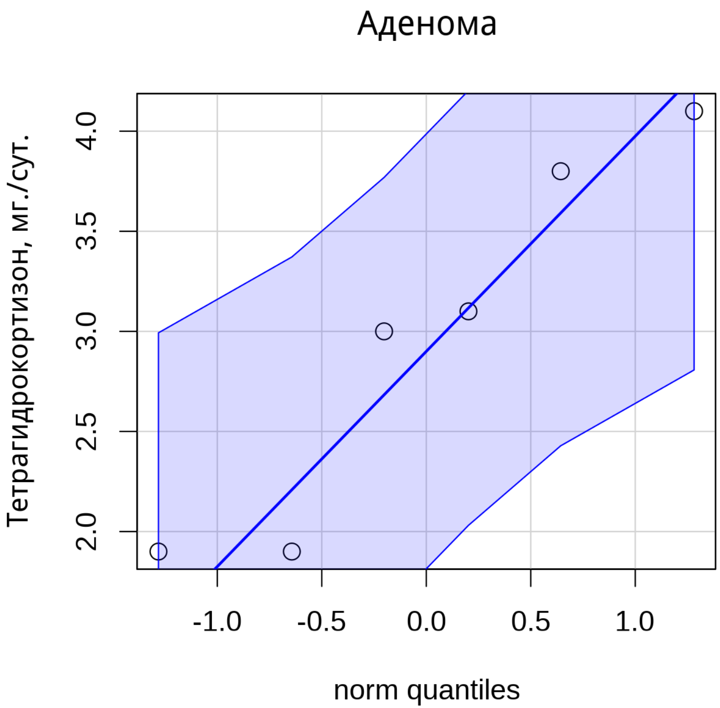
3.Объем выборки достаточно велик или величины нормально распределены?

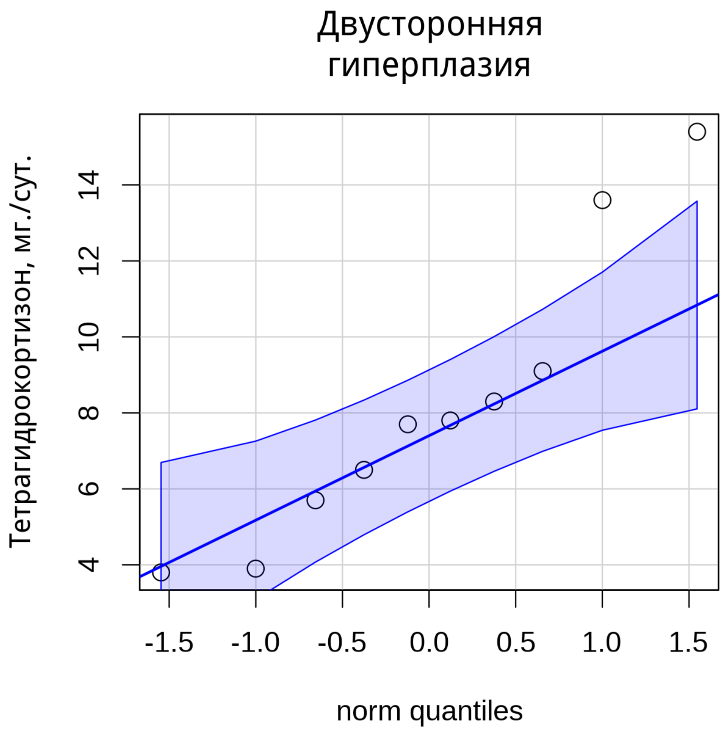
- Объем выборки мал

Пациентов с аденомой всего $n_1 = 6$,
а с двусторонней гиперплазией надпочечников — $n_2 = 10$.

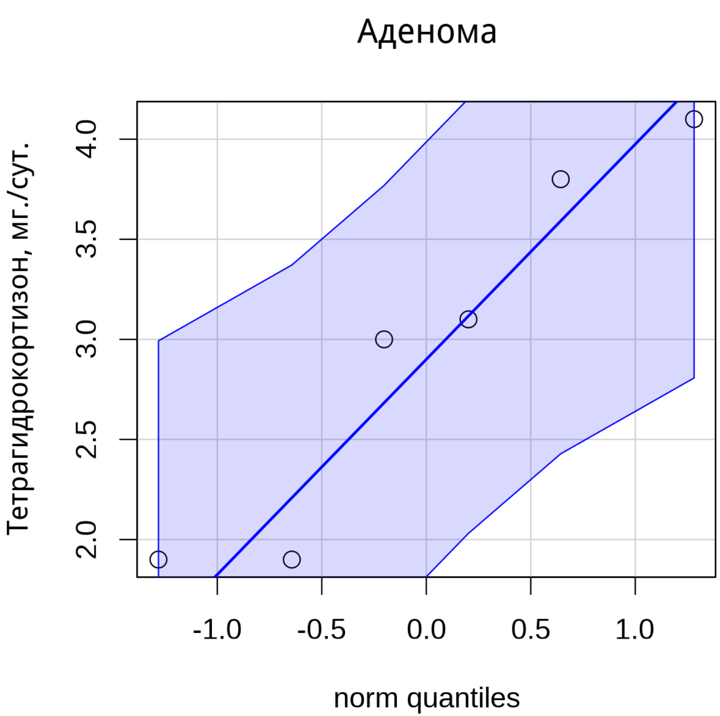
Нужно проверить форму распределения в обеих группах.

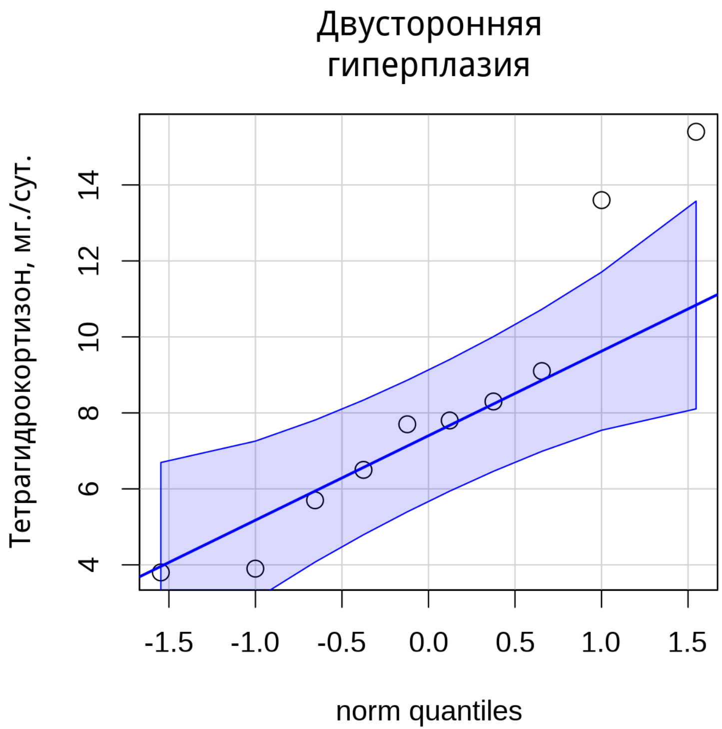
Нормально ли распределены концентрации тетрагидрокортизона в группах?





Нормально ли распределены концентрации тетрагидрокортизона в группах?

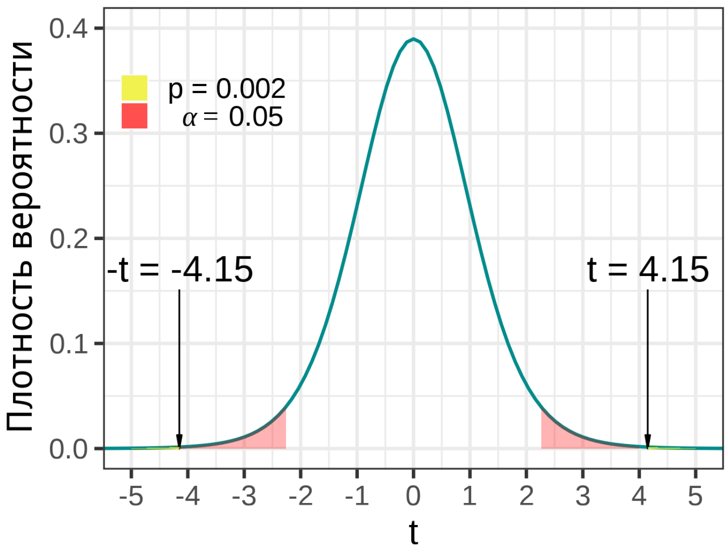




В одной из групп несколько значений выбиваются. Но при таких ничтожно малых объемах выборки сложно ожидать лучшего.

В этом учебном примере будем считать, что можно аппроксимировать концентрацию тетрагидрокортизона нормальным распределением.

Двухвыборочный *t*-тест в нашем примере

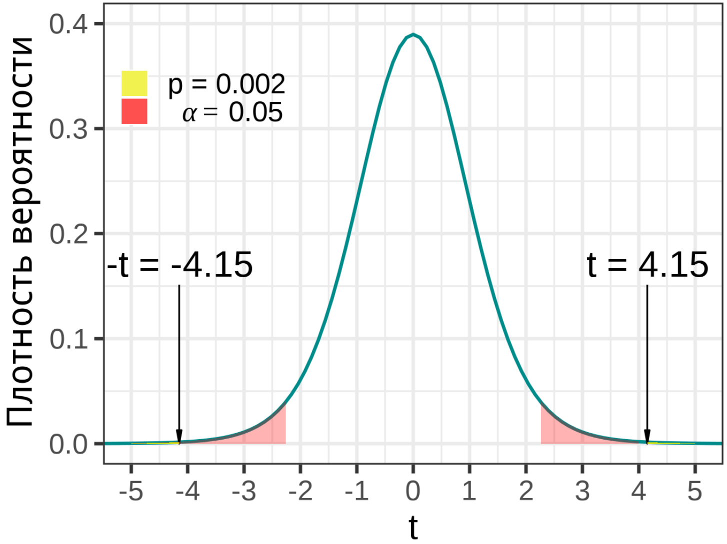


$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE_{\bar{x}_1 - \bar{x}_2}}$$

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$df_{WS} \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Двухвыборочный *t*-тест в нашем примере



$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE_{\bar{x}_1 - \bar{x}_2}}$$

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$df_{WS} \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$$

В нашем примере

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{0.924^2}{6} + \frac{3.789^2}{10}} = 1.256$$

$$df_{WS} \approx \frac{\left(\frac{0.924^2}{6} + \frac{3.789^2}{10}\right)^2}{\frac{1}{6-1}\left(\frac{0.924^2}{6}\right)^2 + \frac{1}{10-1}\left(\frac{3.789^2}{10}\right)^2} =$$

$$= 10.685$$

Таким образом, значение t-статистики

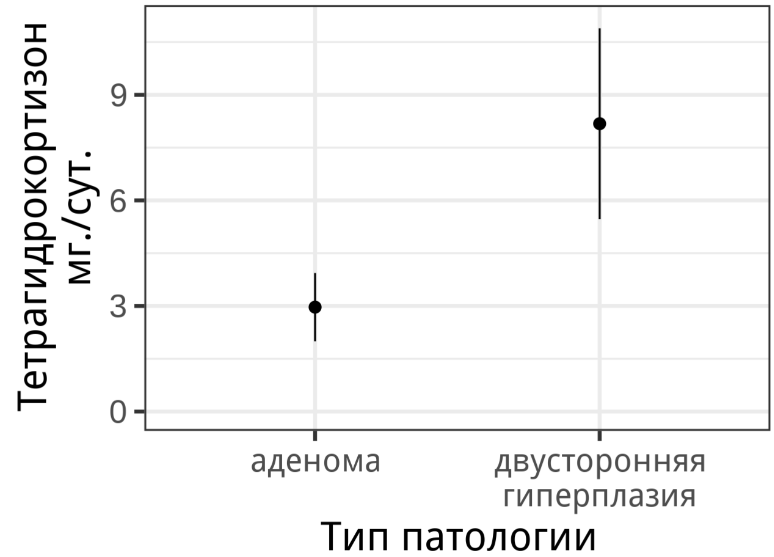
$$t = \frac{2.967 - 8.18}{1.256} = -4.15$$

Ему соответствует $p = 0.002$

Отвергаем H_0

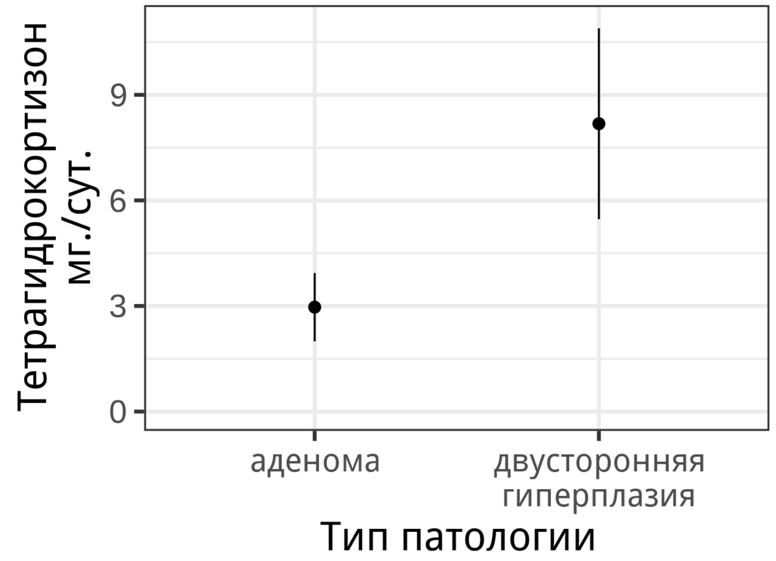
Опишем результаты

Уровень секреции тетрагидрокортизона статистически значимо отличается у пациентов с аденомой и двусторонней гиперплазией надпочечников ($t_{10.69} = -4.15, p =< 0.05$).



Опишем результаты

Уровень секреции тетрагидрокортизона статистически значимо отличается у пациентов с аденомой и двусторонней гиперплазией надпочечников ($t_{10.69} = -4.15, p =< 0.05$).



Можно указать в скобках не сравнение с α, а само значение p:
($t_{10.69} = -4.15, p = 0.0017$).

Только не надо безумствовать и указывать слишком много знаков...

Множественное тестирование гипотез

Есть ли взаимосвязь между поведением бабочек в Швейцарии и дождем в Париже?

Inaudi et al. 1995. Chaos: evidence for the butterfly effect. Annals of Improbable research 1 (6).

Есть ли взаимосвязь между поведением бабочек в Швейцарии и дождем в Париже?

Inaudi et al. 1995. Chaos: evidence for the butterfly effect. Annals of Improbable research 1 (6).

Если проверить много-много взаимосвязей, то сколько-то из них окажутся статистически значимыми.

Чем больше тестов, тем больше вероятность ошибочно обнаружить различия.

Множественное тестирование гипотез

В датасете `Cushings` (пакет `MASS`) есть информация о секреции тетрагидрокортизона с мочой (мг/сут.) при синдроме Кушинга (данные из кн. Aitchison, Dunsmore, 1975). Всего 4 типа патологии:

- `a` — аденома
- `b` — двусторонняя гиперплазия
- `c` — карцинома
- `u` — другие патологии (без уточнения)

Мы сравнили уровень секреции только при двух из них.

А что, если мы хотим сравнить все друг с другом?

Сколько всего нужно будет сделать сравнений?

Множественное тестирование гипотез

В датасете `Cushings` (пакет `MASS`) есть информация о секреции тетрагидрокортизона с мочой (мг/сут.) при синдроме Кушинга (данные из кн. Aitchison, Dunsmore, 1975). Всего 4 типа патологии:

- `a` — аденома
- `b` — двусторонняя гиперплазия
- `c` — карцинома
- `u` — другие патологии (без уточнения)

Мы сравнили уровень секреции только при двух из них.

А что, если мы хотим сравнить все друг с другом?

Сколько всего нужно будет сделать сравнений?

m — количество попарных сравнений k групп между собой:

$m = \frac{k^2 - k}{2}$, т.е. в нашем случае $m = \frac{4^2 - 4}{2} = 6$

Тестируем множество гипотез

Сравним уровень секреции тетрагидрокортизона при разных типах патологий при помощи двухвыборочного t-теста.

Группа 1	Группа 2	p в t-тесте
b	a	0.304
c	a	0.008
c	b	0.039
u	a	0.058
u	b	0.252
u	c	0.337

Представим, что мы проводим каждый из тестов при $\alpha = 0.05$, желая зафиксировать вероятность ошибок I рода на уровне 5%.

В этом случае количество тестов, где различия оказались статистически значимыми: 2

Тестируем множество гипотез

Сравним уровень секреции тетрагидрокортизона при разных типах патологий при помощи двухвыборочного t-теста.

Группа 1	Группа 2	p в t-тесте
b	a	0.304
c	a	0.008
c	b	0.039
u	a	0.058
u	b	0.252
u	c	0.337

Представим, что мы проводим каждый из тестов при $\alpha = 0.05$, желая зафиксировать вероятность ошибок I рода на уровне 5%.

В этом случае количество тестов, где различия оказались статистически значимыми: 2

Но вот вероятность ошибок I рода в итоге вовсе не 5%.

При множественных тестах увеличивается вероятность ошибочно найти различия

α — это вероятность совершить хотя бы одну ошибку I рода (найти различия там, где их нет).

Если сделать n тестов, то вероятность совершить хотя бы одну ошибку I рода во всем семействе тестов (family-wise error rate, FWER) значительно возрастает.

$$1 - (1 - \alpha)^m$$

При множественных тестах увеличивается вероятность ошибочно найти различия

α — это вероятность совершить хотя бы одну ошибку I рода (найти различия там, где их нет).

Если сделать n тестов, то вероятность совершить хотя бы одну ошибку I рода во всем семействе тестов (family-wise error rate, FWER) значительно возрастает.

$$1 - (1 - \alpha)^m$$

Например, если мы проведем 6 тестов на уровне значимости 0.05, то с вероятностью 26.5 мы совершим хотя бы одну ошибку I рода, т.е. практически в 1/4 всех тестов.

Поправка Бонферрони

Простой, но очень жесткий способ коррекции. Зная общее число тестов, можно вычислить скорректированный уровень значимости и использовать его.

$$\alpha^* = \frac{\alpha}{m}$$

Поправка Бонферрони

Простой, но очень жесткий способ коррекции. Зная общее число тестов, можно вычислить скорректированный уровень значимости и использовать его.

$$\alpha^* = \frac{\alpha}{m}$$

Например, чтобы сохранить в группе из 6 тестов вероятность ошибки I рода 0.05, нужно проводить каждый тест при $\alpha^* = 0.00833$.

Поправка Бонферрони

Простой, но очень жесткий способ коррекции. Зная общее число тестов, можно вычислить скорректированный уровень значимости и использовать его.

$$\alpha^* = \frac{\alpha}{m}$$

Например, чтобы сохранить в группе из 6 тестов вероятность ошибки I рода 0.05, нужно проводить каждый тест при $\alpha^* = 0.00833$.

А если бы нужно было сделать 10 тестов, то при $\alpha^* = 0.005$, и т.д.

Поправка Бонферрони

Простой, но очень жесткий способ коррекции. Зная общее число тестов, можно вычислить скорректированный уровень значимости и использовать его.

$$\alpha^* = \frac{\alpha}{m}$$

Например, чтобы сохранить в группе из 6 тестов вероятность ошибки I рода 0.05, нужно проводить каждый тест при $\alpha^* = 0.00833$.

А если бы нужно было сделать 10 тестов, то при $\alpha^* = 0.005$, и т.д.

При проведении множества тестов (не обязательно попарных)
резко возрастает вероятность ошибки II рода,
т.е. не найти различий там, где они есть.

Применим поправку Бонферрони

Bonferroni, 1936

Каждый из 6 тестов проводим при $\alpha^* = 0.00833$.

Группа 1	Группа 2	p в t-тесте	С поправкой Бонферрони
b	a	0.304	Сохраняем \`$H_0$\`
c	a	0.008	Сохраняем \`$H_0$\`
c	b	0.039	Сохраняем \`$H_0$\`
u	a	0.058	Сохраняем \`$H_0$\`
u	b	0.252	Сохраняем \`$H_0$\`
u	c	0.337	Сохраняем \`$H_0$\`

После поправки Бонферрони мы сохранили для всего семейства сравнений вероятность ошибки I рода на уровне 5%. Но при этом не осталось значимых различий — такая это жесткая поправка.

Метод Хольма-Бонферрони

Holm, 1979

Метод Хольма-Бонферрони

Holm, 1979

Чтобы зафиксировать $FWER \leq \alpha$:

1.Сортируем в порядке возрастания m значений p, полученные в тестах, и присваиваем им ранги j от 1 до m
$p_1 \leq p_2 \leq \cdots \leq p_{m-1} \leq p_m$

Метод Хольма-Бонферрони

Holm, 1979

Чтобы зафиксировать $FWER \le \alpha$:

1.Сортируем в порядке возрастания m значений p, полученные в тестах, и присваиваем им ранги j от 1 до m
$p_1 \le p_2 \le \cdots \le p_{m-1} \le p_m$

2.Вводим поправку для значений p
$p_j^* = min\{(m - j + 1) \cdot p_j,\ 1\}$

Метод Хольма-Бонферрони

Holm, 1979

Чтобы зафиксировать $FWER \leq \alpha$:

1.Сортируем в порядке возрастания m значений p, полученные в тестах, и присваиваем им ранги j от 1 до m
$p_1 \leq p_2 \leq \dots \leq p_{m-1} \leq p_m$

2.Вводим поправку для значений p
$p_j^* = min\{(m - j + 1) \cdot p_j,\ 1\}$

3.Если на каком-то шаге получилось p_{j+1}^* меньше, чем предыдущее p_j^*, то записываем большее из них.

Метод Хольма-Бонферрони

Holm, 1979

Чтобы зафиксировать $FWER \leq \alpha$:

1.Сортируем в порядке возрастания m значений p, полученные в тестах, и присваиваем им ранги j от 1 до m
$p_1 \leq p_2 \leq \cdots \leq p_{m-1} \leq p_m$

2.Вводим поправку для значений p
$p_j^* = min\{(m - j + 1) \cdot p_j,\ 1\}$

3.Если на каком-то шаге получилось p_{j+1}^* меньше, чем предыдущее p_j^*, то записываем большее из них.

4.Сравниваем с α

Метод Хольма-Бонферрони

Holm, 1979

Чтобы зафиксировать $FWER \leq \alpha$:

1.Сортируем в порядке возрастания m значений p, полученные в тестах, и присваиваем им ранги j от 1 до m
$p_1 \leq p_2 \leq \cdots \leq p_{m-1} \leq p_m$

2.Вводим поправку для значений p
$p_j^* = min\{(m - j + 1) \cdot p_j,\ 1\}$

3.Если на каком-то шаге получилось p_{j+1}^* меньше, чем предыдущее p_j^*, то записываем большее из них.

4.Сравниваем с α

Группа 1	Группа 2	p в t-тесте	Ранг \`$j$\`	\`$p^*$\`	С поправкой Хольма-Бонферрони
c	a	0.008	1	0.051	Сохраняем \`$H_0$\`
c	b	0.039	2	0.194	Сохраняем \`$H_0$\`
u	a	0.058	3	0.234	Сохраняем \`$H_0$\`
u	b	0.252	4	0.755	Сохраняем \`$H_0$\`
b	a	0.304	5	0.755	Сохраняем \`$H_0$\`
u	c	0.337	6	0.755	Сохраняем \`$H_0$\`

Бонферрони, Хольм или что-то другое?

Поправку Борферрони имеет смысл использовать, если тестов не больше 5. Иначе она становится слишком жесткой.

Поправку Хольма-Бонферрони можно делать, если тестов несколько. М.б. пара десятков.

Бонферрони, Хольм или что-то другое?

Поправку Борферрони имеет смысл использовать, если тестов не больше 5. Иначе она становится слишком жесткой.

Поправку Хольма-Бонферрони можно делать, если тестов несколько. М.б. пара десятков.

Если тестов еще больше — десятки или сотни, как это бывает для молекулярных данных, то придется использовать принципиально иной способ коррекции.
Не Family-Wise Error Rate (FWER), а False Discovery Rate (FDR).

Поправки False Discovery Rate

False Discovery Rate corrections — поправки, которые фиксируют долю ложно-положительных срабатываний теста не из всех тестов, а только из всех срабатываний.

Эти поправки мягче, чем традиционные FWER.

Поправки False Discovery Rate

False Discovery Rate corrections — поправки, которые фиксируют долю ложно-положительных срабатываний теста не из всех тестов, а только из всех срабатываний.

Эти поправки мягче, чем традиционные FWER.

Поправки FDR:

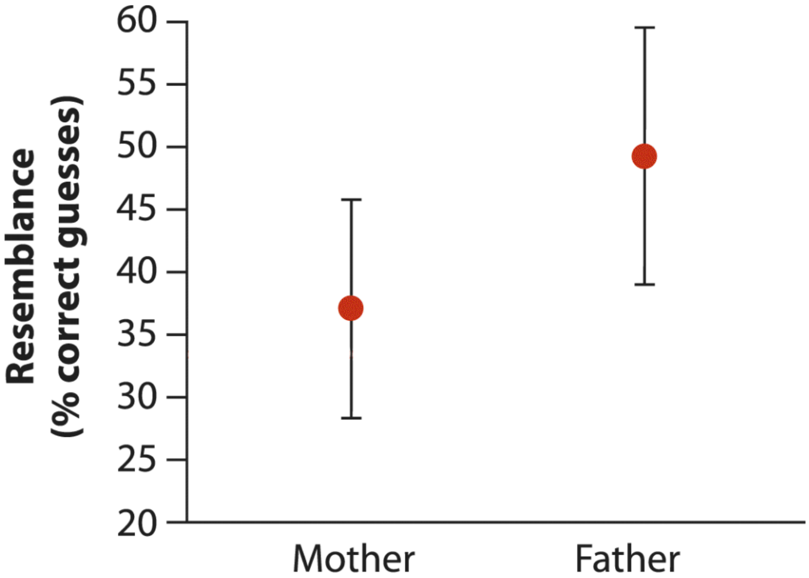
- Метод Беньямини-Хохберга (Benjamini, Hochberg, 1995)
- Метод Беньямини-Йекутили (Benjamini, Yekutieli, 2001)

Интерпретация доверительных интервалов на графиках

Опасности косвенных сравнений

Участникам эксперимента (Christenfeld, Hill, 1995) показывали фото младенцев и предлагали выбрать наиболее похожее из 3 фото их матерей и 3 фото отцов.

На кого больше похожи дети: на отцов или на матерей?



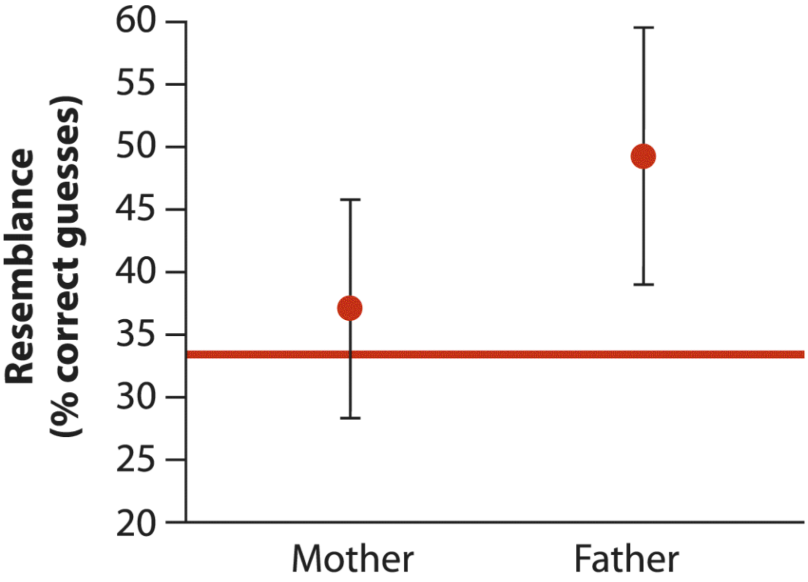
Whitlock, Schluter, 2015, fig.12.5-1

Какой процент правильных догадок можно было бы ожидать при случайном выборе фото?

Опасности косвенных сравнений

Участникам эксперимента (Christenfeld, Hill, 1995) показывали фото младенцев и предлагали выбрать наиболее похожее из 3 фото их матерей и 3 фото отцов.

На кого больше похожи дети: на отцов или на матерей?

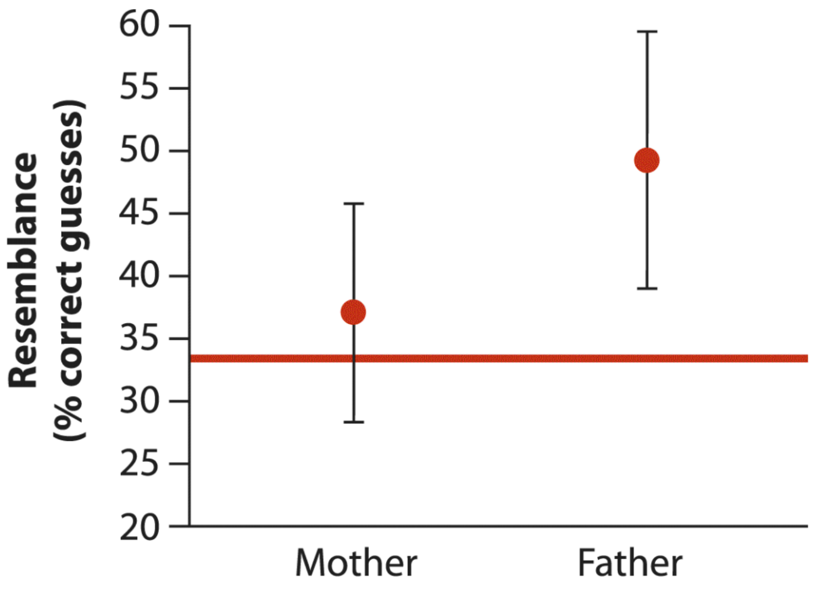


Whitlock, Schluter, 2015, fig.12.5-1

Опасности косвенных сравнений

Участникам эксперимента (Christenfeld, Hill, 1995) показывали фото младенцев и предлагали выбрать наиболее похожее из 3 фото их матерей и 3 фото отцов.

На кого больше похожи дети: на отцов или на матерей?



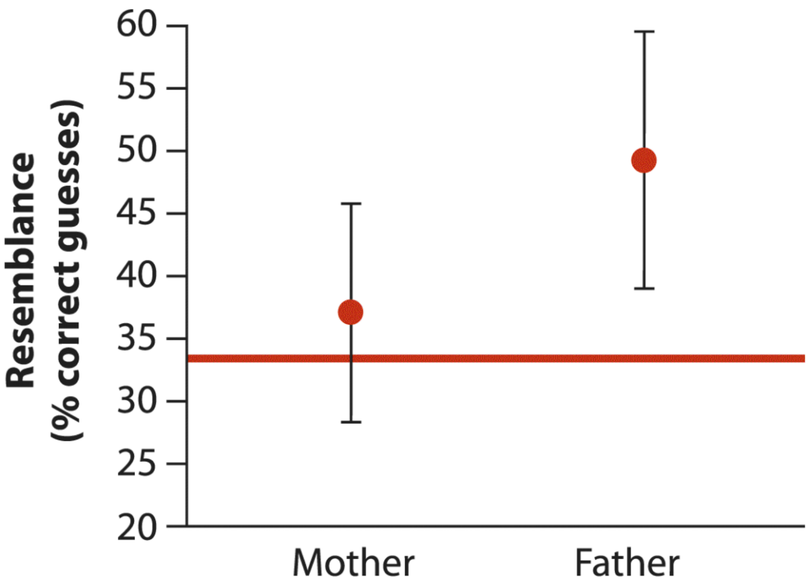
Процент правильных догадок о матерях не значимо отличался от случайного, а правильные фото отцов выбирали значимо чаще, чем при случайном выборе.

Whitlock, Schluter, 2015, fig.12.5-1

Опасности косвенных сравнений

Участникам эксперимента (Christenfeld, Hill, 1995) показывали фото младенцев и предлагали выбрать наиболее похожее из 3 фото их матерей и 3 фото отцов.

На кого больше похожи дети: на отцов или на матерей?



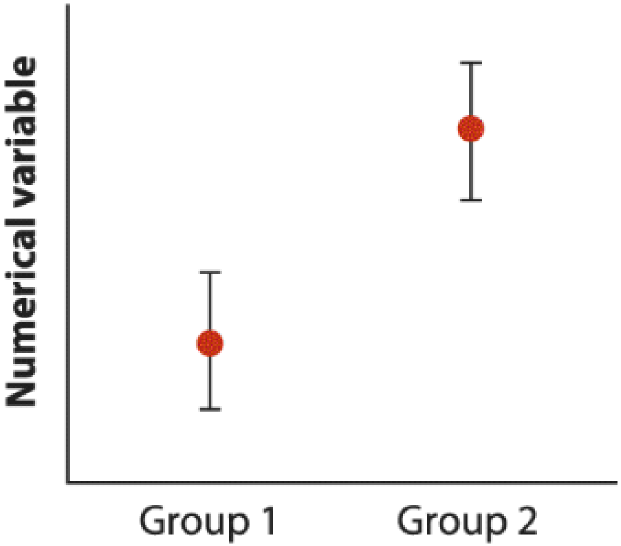
Процент правильных догадок о матерях не значимо отличался от случайного, а правильные фото отцов выбирали значимо чаще, чем при случайном выборе.

Whitlock, Schluter, 2015, fig.12.5-1

Однако нельзя заключить, что отцов правильные фото отцов выбирали чаще, чем матерей, т.к. частота правильных угадываний не различается.

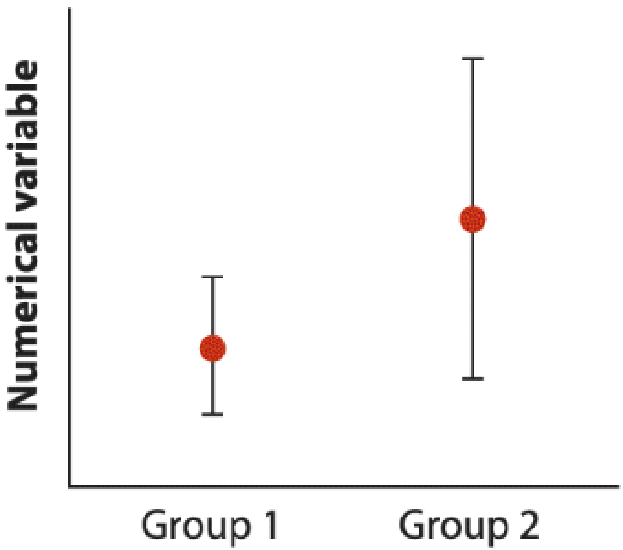
Из того, что A не значимо отличается от C,
а B значимо отличается от C,
не следует, что A значимо отличается от B.

Интерпретация перекрывания доверительных интервалов



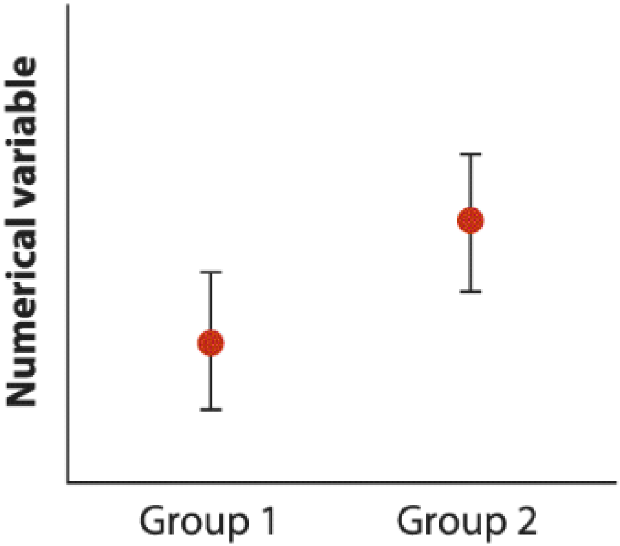
Whitlock, Schluter, 2015, fig.12.6-1

Интерпретация перекрывания доверительных интервалов



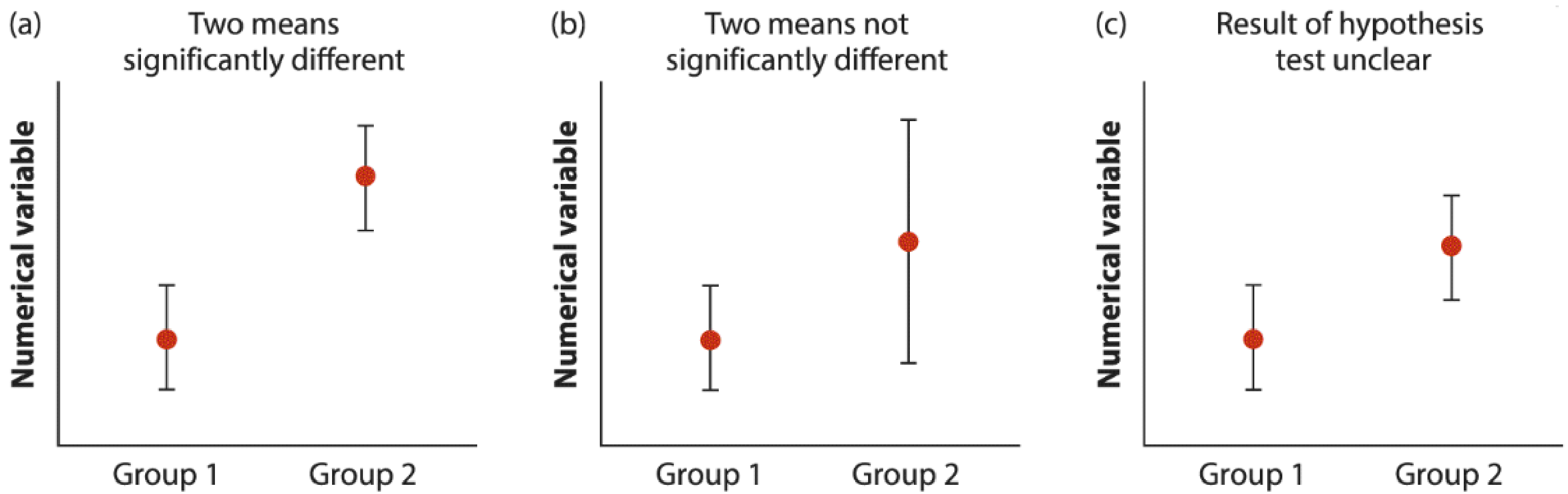
Whitlock, Schluter, 2015, fig.12.6-1

Интерпретация перекрывания доверительных интервалов



Whitlock, Schluter, 2015, fig.12.6-1

Интерпретация перекрывания доверительных интервалов



Whitlock, Schluter, 2015, fig.12.6-1

Сравнение дисперсий

Тесты для сравнения дисперсий

Тесты для сравнения дисперсий

F-тест равенства дисперсий

(F-test of equal variances)

Предложен Р.Фишером в 20е годы (назван в честь Фишера в работе Snedecor, Cochran, 1989)

- Сравнение двух дисперсий друг с другом.

Тесты для сравнения дисперсий

F-тест равенства дисперсий

(F-test of equal variances)

Предложен Р.Фишером в 20е годы (назван в честь Фишера в работе Snedecor, Cochran, 1989)

- Сравнение двух дисперсий друг с другом.

Тест Левина на гомогенность дисперсий

(Levene's test for homogeneity of variances)

Levene, Howard, 1960

- Сравнение сразу нескольких дисперсий друг с другом.

Гипотезы в F-тесте равенства дисперсий

На человеческом языке:

H_0 : — дисперсии в двух сравниваемых группах равны

H_A : — дисперсии в двух сравниваемых группах не равны

Гипотезы в F-тесте равенства дисперсий

На человеческом языке:

H_0 : — дисперсии в двух сравниваемых группах равны

H_A : — дисперсии в двух сравниваемых группах не равны

На языке статистики:

$H_0 : \sigma_1^2 = \sigma_2^2$ — дисперсии в двух сравниваемых группах равны

$H_A : \sigma_1^2 \neq \sigma_2^2$ — дисперсии в двух сравниваемых группах не равны

F-тест равенства дисперсий

$H_0 : \sigma_1^2 = \sigma_2^2$ — дисперсии в двух сравниваемых группах равны

$H_A : \sigma_1^2 \neq \sigma_2^2$ — дисперсии в двух сравниваемых группах не равны

$$F = \frac{s_1^2}{s_2^2}$$

При H_0 F-статистика подчиняется F-распределению с двумя параметрами: $df_1 = n_1 - 1$ и $df_2 = n_2 - 1$.

F-тест равенства дисперсий

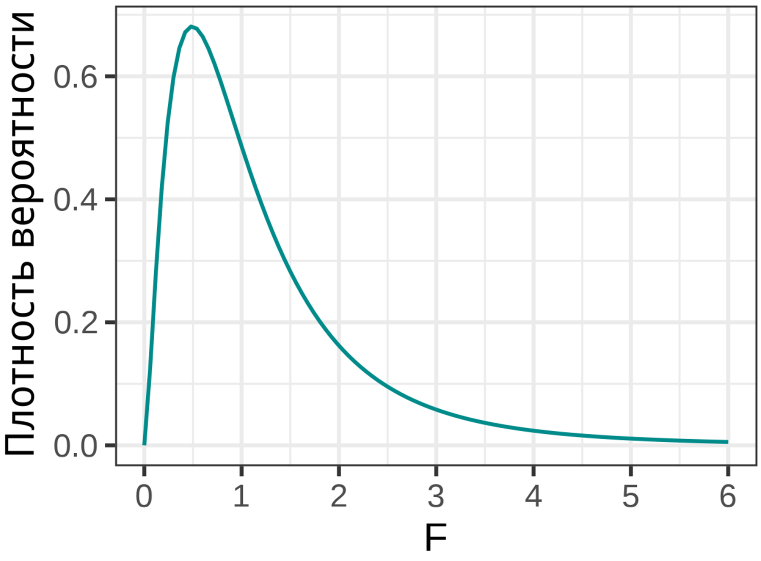
$H_0 : \sigma_1^2 = \sigma_2^2$ — дисперсии в двух сравниваемых группах равны

$H_A : \sigma_1^2 \neq \sigma_2^2$ — дисперсии в двух сравниваемых группах не равны

$$F = \frac{s_1^2}{s_2^2}$$

При H_0 F-статистика подчиняется F-распределению с двумя параметрами: $df_1 = n_1 - 1$ и $df_2 = n_2 - 1$.

F-распределение при $df_1 = 5$ и $df_2 = 9$



F-тест равенства дисперсий

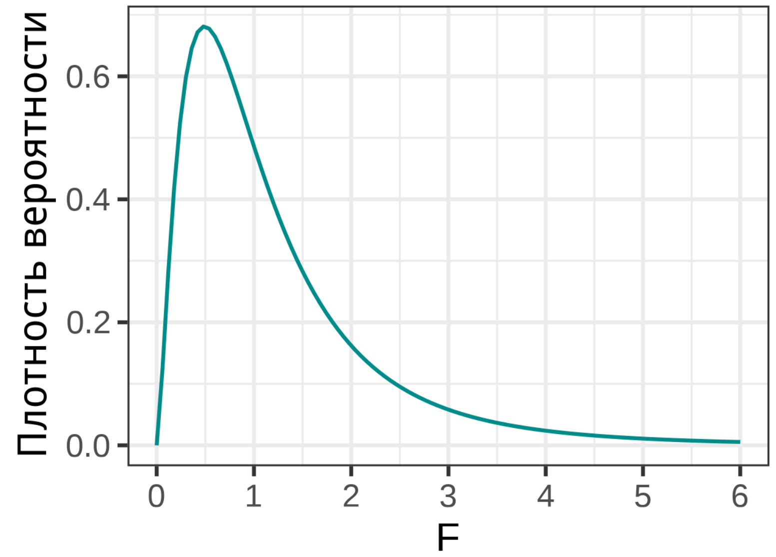
$H_0 : \sigma_1^2 = \sigma_2^2$ — дисперсии в двух сравниваемых группах равны

$H_A : \sigma_1^2 \neq \sigma_2^2$ — дисперсии в двух сравниваемых группах не равны

$$F = \frac{s_1^2}{s_2^2}$$

При H_0 F-статистика подчиняется F-распределению с двумя параметрами: $df_1 = n_1 - 1$ и $df_2 = n_2 - 1$.

F-распределение при $df_1 = 5$ и $df_2 = 9$



Условие применимости теста: нормальное распределение в обеих группах.

F-тест равенства дисперсий

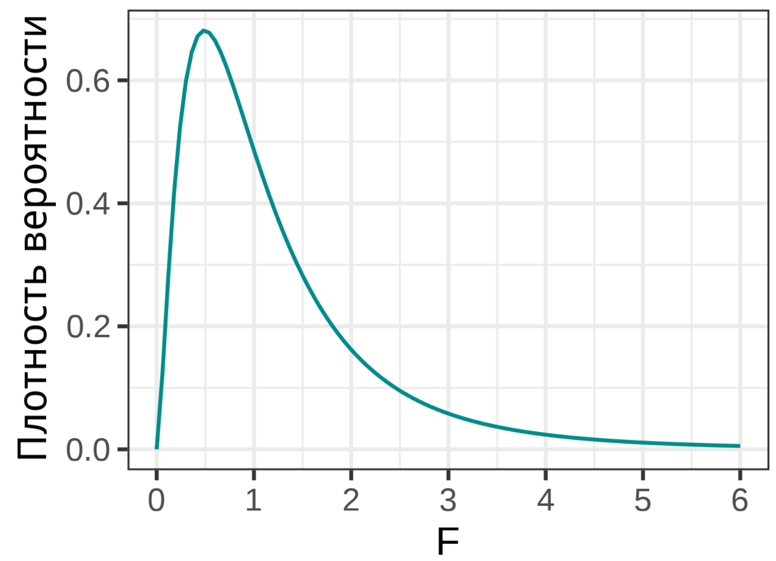
$H_0: \sigma_1^2 = \sigma_2^2$ — дисперсии в двух сравниваемых группах равны

$H_A: \sigma_1^2 \neq \sigma_2^2$ — дисперсии в двух сравниваемых группах не равны

$$F = \frac{s_1^2}{s_2^2}$$

При H_0 F-статистика подчиняется F-распределению с двумя параметрами: $df_1 = n_1 - 1$ и $df_2 = n_2 - 1$.

F-распределение при $df_1 = 5$ и $df_2 = 9$



Условие применимости теста: нормальное распределение в обеих группах.

F-тест равенства дисперсий крайне чувствителен к отклонениям от нормального распределения. Если в одной из групп переменная не подчиняется нормальному распределению, резко возрастает вероятность ошибки I рода.

Гипотезы в тесте Левина на гомогенность дисперсий

На человеческом языке:

$H_0:$ — дисперсии во всех сравниваемых группах равны

$H_A:$ — **хотя бы одна дисперсия отличается**

Гипотезы в тесте Левина на гомогенность дисперсий

На человеческом языке:

H_0 : — дисперсии во всех сравниваемых группах равны

H_A : — **хотя бы одна дисперсия отличается**

На языке статистики:

$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$ — дисперсии во всех группах равны.

$H_A : \exists\, i, j : \sigma_i^2 \neq \sigma_j^2$ — среди всех выбранных групп существуют такие две группы i и j, дисперсии в которых не равны.

Тест Левина на гомогенность дисперсий

Levene, Howard, 1960

$H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

$H_A: \exists\, i, j: \sigma_i^2 \neq \sigma_j^2$ — **хотя бы одна** дисперсия отличается

$$W = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} N_i \left(\bar{Z}_i - \bar{Z}\right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} \left(Z_{ij} - \bar{Z}_i\right)^2}$$

- k — число групп
- $N = n_1 + n_2$ — общий объем выборки
- $Z_{ij} = |x_{ij} - \bar{x}_i|$ — модуль отклонения значения x_{ij} от среднего в своей i-той группе $\bar{x}_i$
- $\bar{Z}_i$ — среднее Z_{ij} в i-той группе
- $\bar{Z}$ — общее среднее Z_{ij} (без учета группы)

При H_0 W-статистика подчиняется F-распределению с двумя параметрами: $df_1 = k - 1$ и $df_2 = N - k$.

Тест Левина на гомогенность дисперсий

Levene, Howard, 1960

$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

$H_A : \exists\, i, j : \sigma_i^2 \neq \sigma_j^2$ — **хотя бы одна** дисперсия отличается

$$W = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} N_i \left(\bar{Z}_i - \bar{Z}\right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} \left(Z_{ij} - \bar{Z}_i\right)^2}$$

- k — число групп
- $N = n_1 + n_2$ — общий объем выборки
- $Z_{ij} = |x_{ij} - \bar{x}_i|$ — модуль отклонения значения x_{ij} от среднего в своей i-той группе $\bar{x}_i$
- $\bar{Z}_i$ — среднее Z_{ij} в i-той группе
- $\bar{Z}$ — общее среднее Z_{ij} (без учета группы)

При H_0 W-статистика подчиняется F-распределению с двумя параметрами: $df_1 = k - 1$ и $df_2 = N - k$.

Условие применимости: распределение значений во всех группах приблизительно симметрично.

Тест Левина более устойчив к отклонениям от условий применимости.

Summary

Summary:

- При сравнении значений статистик в группах важно обращать внимание, зависимы ли наблюдения.
- Для сравнения средних в двух зависимых группах используется парный t-тест. Он оценивает статистическую значимость средней разницы значений измерения у одного и того же объекта.
- Для сравнения средних в двух независимых группах используется двухвыборочный t-тест. Он оценивает статистическую значимость разницы средних в каждой из двух групп.

Summary:

- Двухвыборочный t-тест бывает двух разновидностей:
  - классический t-тест Стьюдента подразумевает, что дисперсии в группах одинаковы, что часто нереалистично.
  - t-тест Велча позволяет разные дисперсии в группах. Приближенное число степеней свободы для получившейся t-статистики приходится расчитывать по сложной формуле. Но зато этот тест можно использовать даже, когда дисперсии в группах равны.

Summary:

Если приходится делать множество тестов, то увеличивается вероятность допустить хотябы одну ошибку I рода. (Например, если 6 сравнений, то ошибка I рода будет в 26.5% случаев).

Для снижения вероятности ошибки I рода в группе сравнений p-значения считают с поправкой.

Поправки на множественное тестирование бывают двух типов:

- family-wise error rate, FWER
- false discovery rate, FDR

Summary:

Доверительные интервалы на графиках нужно интерпретировать с осторожностью:

- Избегайте косвенных сравнений
- Перекрывание доверительных интервалов однозначно говорит о значимости различий средних только, если интервалы сравниваемых групп захватывают средние из другой группы.

Summary:

Можно сравнивать не только средние значения в группах, но и дисперсии.

F-тест равенства дисперсий сравнивает дисперсии в двух группах.

Тест Левина на гомогенность дисперсий позволяет сравнить сразу несколько дисперсий и выяснить, что хотя бы одна из них отличается от других.

Статистики в обоих этих тестах подчиняются F-распределению с двумя параметрами.

Что почитать

- **Quinn, G.G.P., Keough, M.J., 2002. Experimental design and data analysis for biologists. Cambridge University Press.**
- Sokal, R.R., Rohlf, F.J., 1995. Biometry (3rd edn). WH Freeman and company: New York.
- Zar, J.H., 2010. Biostatistical Analysis. Prentice Hall: New York.