Анализ мощности

Математические методы в зоологии с использованием R

Марина Варфоломеева

Экономим силы с помощью анализа мощности

- Тестирование гипотез (двухвыборочный t-критерий)
- Статистические ошибки при проверке гипотез
- Мощность статистического теста
- *A priori* анализ мощности, оценка величины эффекта
- Как влиять на мощность тестов

Вы сможете

- сравнивать средние значения при помощи t-критерия, интерпретировать и описывать результаты
- дать определение ошибок I и II рода, и графически изобразить их отношение к мощности теста
- оценивать величину эффекта и необходимый объем выборки по данным пилотного исследования
- загружать данные из .xlsx в R
- строить графики средних значений со стандартными отклонениями с помощью `ggplot2`

Тестирование гипотез

Тест Стьюдента (t-критерий)

Гипотезы: $H_0 : \bar{x}_1 = \bar{x}_2$, $H_A : \bar{x}_1 \neq \bar{x}_2$

Двухвыборочный тест Стьюдента (Student, 1908) используется для проверки значимости различий между средними значениями двух нормально распределенных величин.

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

Условия применимости:

- Выборки случайны и независимы друг от друга
- Величины нормально распределены
- **Дисперсии в группах одинаковы**

$$SE = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$df = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$$

t-тест Уэлча — это модификация теста Стьюдента для случая разных дисперсий

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

Условия применимости:

- Выборки случайны и независимы друг от друга
- Величины нормально распределены

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Приблизительное число степеней свободы рассчитывается по уравнению Уэлча-Саттеруэйта

$$df_{WelchSatterthwaite} \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2 \cdot df_1} + \frac{s_2^4}{n_2^2 \cdot df_1}}$$

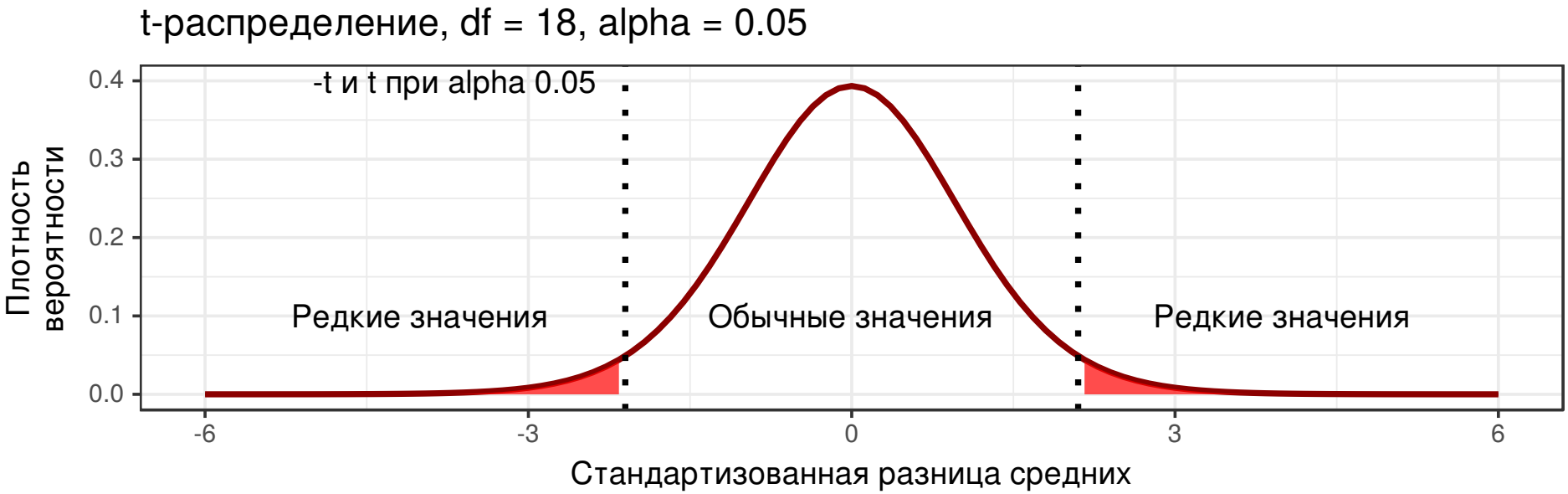
t-распределение — распределение разницы средних для выборок из одной совокупности

t-статистика подчиняется t-распределению.

Иными словами, если много раз взять выборки **из одной** совокупности (т.е. при условии, что H_0 верна) и посчитать между ними разницу, то она будет подчиняться t-распределению.

Форма t-распределения зависит только от одного параметра — числа степеней свободы *df*

В хвостах этого распределения находятся редкие значения



Обычно используется уровень значимости α 0.05 или 0.01.

Уровень значимости α — это вероятность ошибочно отвергнуть справедливую нулевую гипотезу. Т.е. это вероятность найти различия там, где их нет (**вероятность ошибки I рода**).

Для t-теста α — это вероятность ошибочно сделать вывод о том, что средние выборок различаются **при условии, что эти выборки получены из одной генеральной совокупности**.

Тестирование гипотезы о равенстве двух средних при помощи t-теста

1. Для конкретных данных считаем значение t-критерия
2. Сравниваем его с теоретическим распределением t (распределением при условии, что H_0 верна)
3. Принимаем решение, отвергнуть ли H_0

Пример: Снотворное

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

```
data(sleep)
# View(sleep)
```

Двухвыборочный t-критерий

Сравним увеличение продолжительности сна при помощи двухвыборочного t-критерия.

```
tt <- t.test(extra ~ group, sleep)
tt
```

```
#
#   Welch Two Sample t-test
#
# data:  extra by group
# t = -1.8608, df = 17.776, p-value = 0.07939
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#  -3.3654832  0.2054832
# sample estimates:
# mean in group 1 mean in group 2
#            0.75            2.33
```

Двухвыборочный t-критерий

Сравним увеличение продолжительности сна при помощи двухвыборочного t-критерия.

```
tt <- t.test(extra ~ group, sleep)
tt
```

```
#
#   Welch Two Sample t-test
#
# data:  extra by group
# t = -1.8608, df = 17.776, p-value = 0.07939
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#  -3.3654832  0.2054832
# sample estimates:
# mean in group 1 mean in group 2
#            0.75            2.33
```

Результаты можно описать, например, так:

- Различия изменения продолжительности сна при применении двух препаратов были недостоверны ($t_{17.78} = -1.86$, $p = 0.079394$)

Что спрятано в результатах?

Как называются отдельные элементы результатов можно узнать посмотрев их структуру при помощи функции `str()`

```
str(tt)
```

```
# List of 9
#  $ statistic  : Named num -1.86
#   ..- attr(*, "names")= chr "t"
#  $ parameter  : Named num 17.8
#   ..- attr(*, "names")= chr "df"
#  $ p.value    : num 0.0794
#  $ conf.int   : atomic [1:2] -3.365 0.205
#   ..- attr(*, "conf.level")= num 0.95
#  $ estimate   : Named num [1:2] 0.75 2.33
#   ..- attr(*, "names")= chr [1:2] "mean in group 1" "mean in group 2"
#  $ null.value : Named num 0
#   ..- attr(*, "names")= chr "difference in means"
#  $ alternative: chr "two.sided"
#  $ method     : chr "Welch Two Sample t-test"
#  $ data.name  : chr "extra by group"
#  - attr(*, "class")= chr "htest"
```

Можно получить элементы результатов в виде отдельных цифр

```
tt$parameter # степени свободы
```

```
#       df
# 17.77647
```

```
tt$p.value # доверительная вероятность
```

```
# [1] 0.07939414
```

```
tt$statistic # значение t-критерия
```

```
#         t
# -1.860813
```

Статистические ошибки при проверке гипотез

Типы ошибок при проверке гипотез

	$H0 == TRUE$	$H0 == FALSE$
Отклонить H_0	Ошибка I рода	Верно
Сохранить H_0	Верно	Ошибка II рода

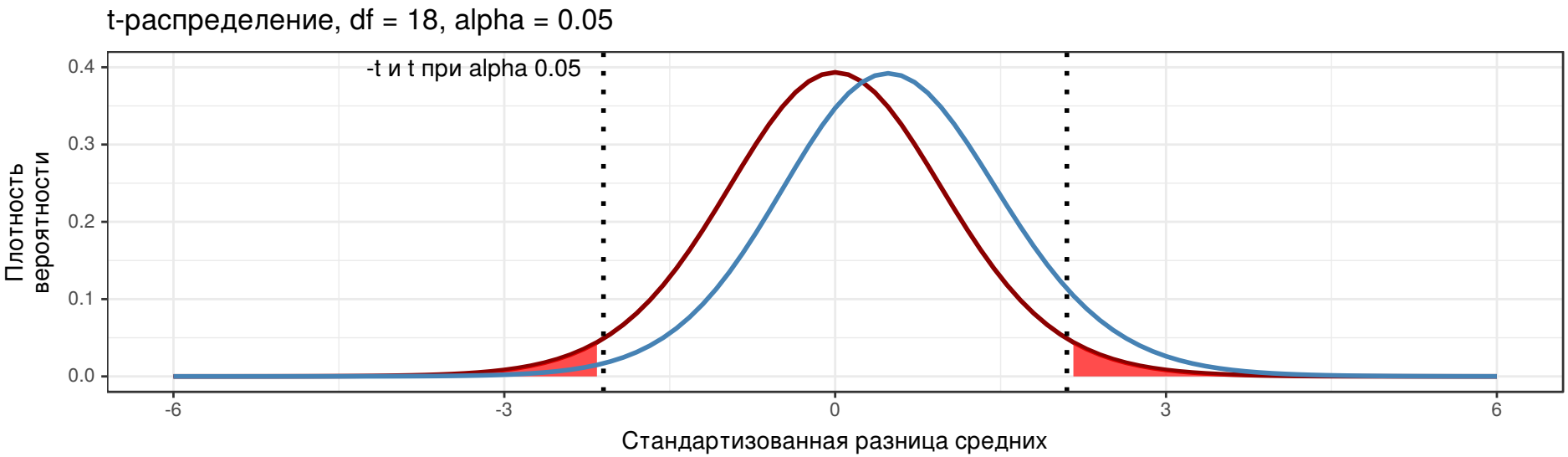
Ошибка I рода

	H0 == TRUE	*H0 == FALSE*
Отклонить H_0	Ошибка I рода	Верно
Сохранить H_0	Верно	Ошибка II рода

Ошибка I рода — вероятность отвергнуть H_0, когда верна H_0

Мы этого не знаем, но может быть верна H_A...

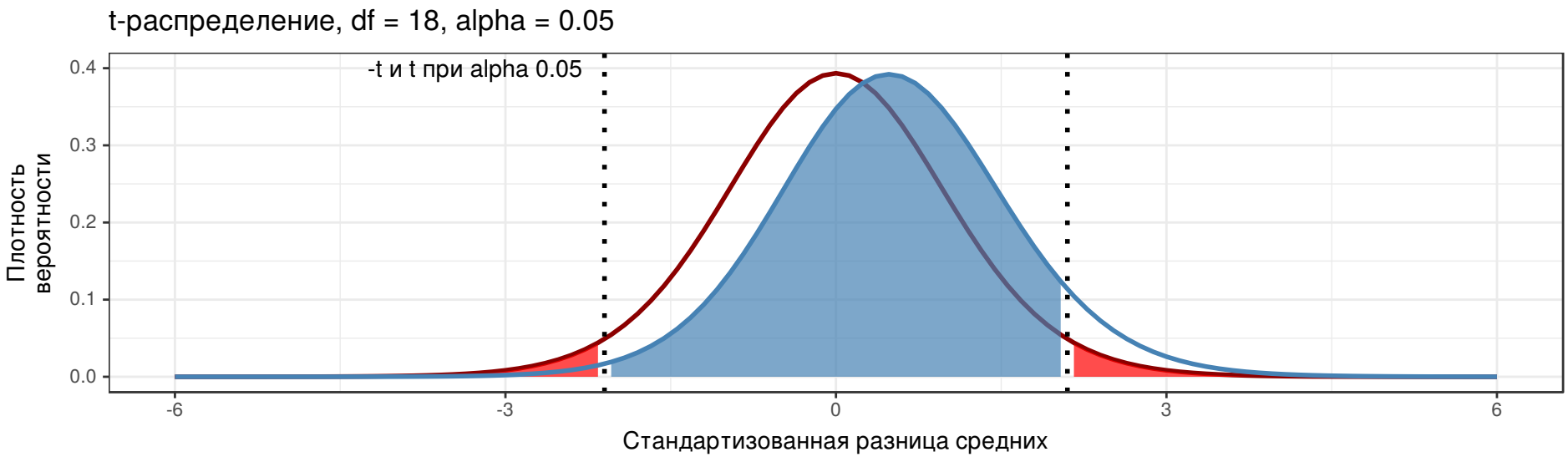
	H0 == TRUE	*H0 == FALSE*
Отклонить H_0	Ошибка I рода	Верно
Сохранить H_0	Верно	Ошибка II рода



Можно построить еще одно распределение статистики — распределение, при условии того, что верна H_A

Ошибка II рода

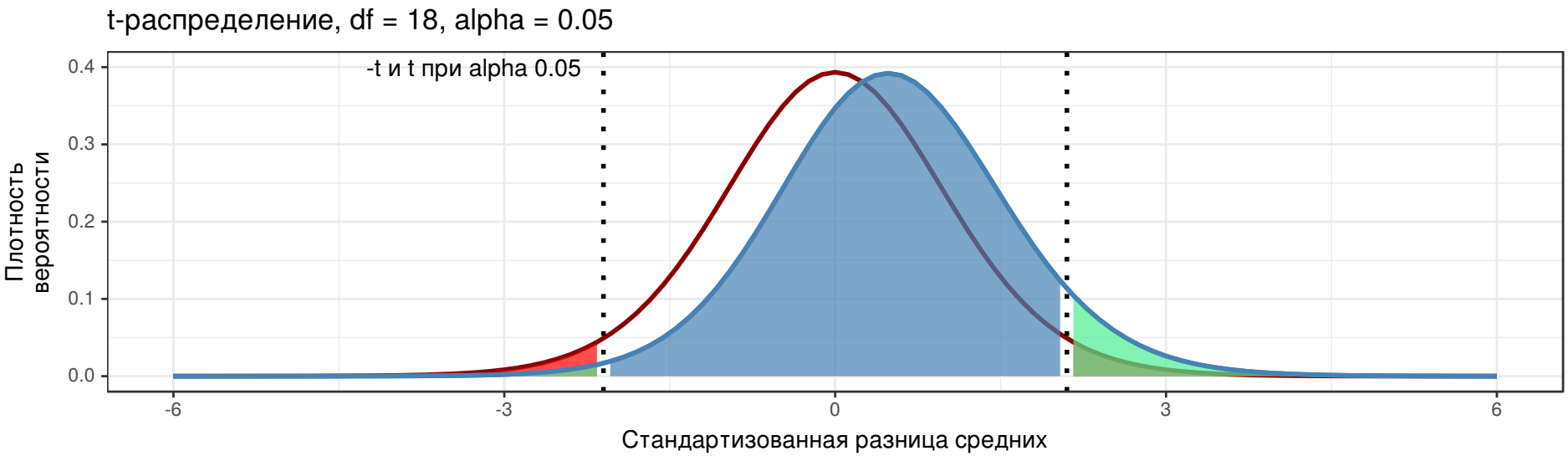
	H0 == TRUE	*H0 == FALSE*
Отклонить H_0	Ошибка I рода	Верно
Сохранить H_0	Верно	Ошибка II рода



Ошибка II рода — вероятность принять H_0, когда верна H_A

Мощность теста — способность выявлять различия

	H0 == TRUE	*H0 == FALSE*
Отклонить H_0	Ошибка I рода	Верно
Сохранить H_0	Верно	Ошибка II рода



Мощность теста - вероятность отвергнуть H_0, когда верна H_A

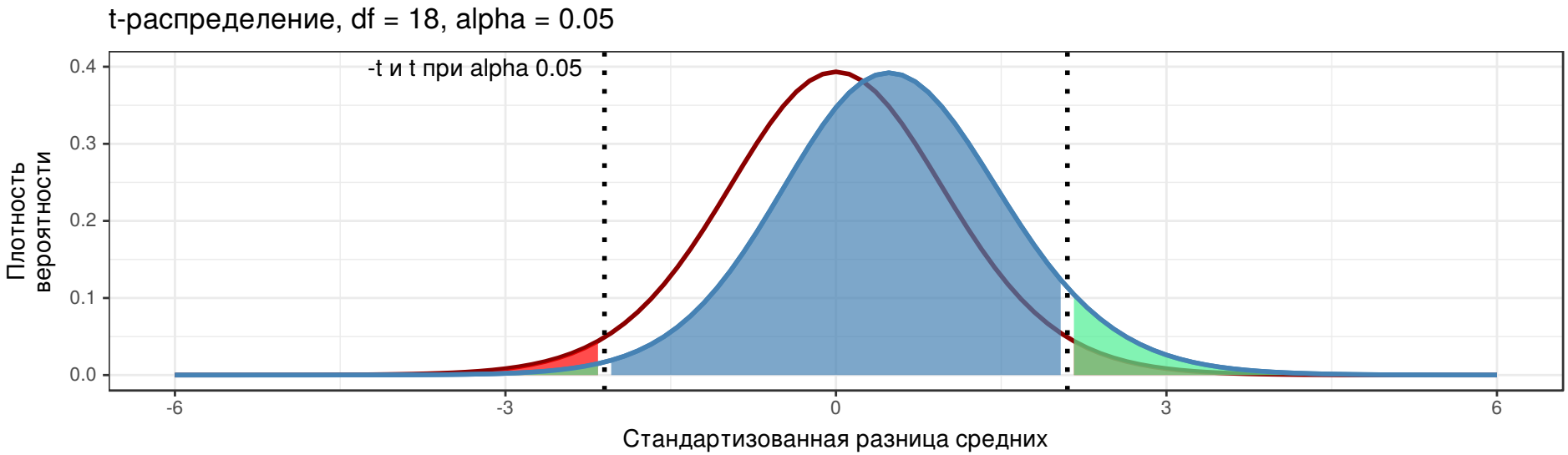
$$Power = 1 - \beta$$

Мощность теста

$$Power = 1 - \beta$$

Обычно считается, что хорошо, когда мощность не меньше 0.8

Т.е. что в 80% случаев мы можем найди различия заданной величины, если они есть.



Анализ мощности

A priori

- какой нужен объем выборки, чтобы найти различия с разумной долей уверенности?
- различия какой величины мы можем найти, если известен объем выборки?

Post hoc

- смогли бы мы найти различия при помощи нашего эксперимента (α, n), если бы величина эффекта была X?

A priory анализ мощности

A priori анализ мощности

Что нужно

- тест
- уровень значимости
- желаемая мощность теста
- ожидаемая величина эффекта

A priori анализ мощности

Что нужно

- тест
- уровень значимости
- желаемая мощность теста
- ожидаемая величина эффекта

Что есть

- t-критерий
- $\alpha = 0.05$
- $Power = 0.8$
- ?

Величина эффекта

d Коэна (Cohen's d)

$$d = \frac{\bar{x}_1 - \bar{x}_2}{SD_{pooled}}$$

где SD_{pooled} — обобщенное стандартное отклонение

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

Величина эффекта

Яков Коэн предложил делить эффекты на сильные, умеренные и слабые (Cohen, 1982)

```
library(pwr)
cohen.ES(test = "t", size = "large")
```

```
#
#      Conventional effect size from Cohen (1982)
#
#            test = t
#            size = large
#     effect.size = 0.8
```

Рассчет объема выборки для обнаружения эффекта известной величины

Функции для анализа мощности t-критерия:

- при одинаковых объемах групп `pwr.t.test()`
- при разных объемах групп `pwr.t2n.test()`

Какая нужна выборка, чтобы обнаружить *сильный эффект* с вероятностью 0.8 при уровне значимости 0.05?

```
pwr.t.test(n = NULL, d = 0.8, power = 0.8, sig.level = 0.01,
           type = "two.sample", alternative = "two.sided")
```

```
#
#      Two-sample t test power calculation
#
#               n = 38.18831
#               d = 0.8
#       sig.level = 0.01
#           power = 0.8
#     alternative = two.sided
#
# NOTE: n is number in *each* group
```

Задание

Какая нужна выборка, чтобы обнаружить *слабый эффект* с вероятностью 0.8 при уровне значимости 0.05?

Вам понадобятся функции `cohen.ES()` и `pwr.t.test()`

Решение

```
cohen.ES(test = "t", size = "small")
```

```
#
#      Conventional effect size from Cohen (1982)
#
#            test = t
#            size = small
#     effect.size = 0.2
```

```
pwr.t.test(n = NULL, d = 0.2, power = 0.8, sig.level = 0.05,
           type = "two.sample", alternative = "two.sided")
```

```
#
#      Two-sample t test power calculation
#
#               n = 393.4057
#               d = 0.2
#       sig.level = 0.05
#           power = 0.8
#     alternative = two.sided
#
# NOTE: n is number in *each* group
```

A priory анализ мощности по данным пилотного исследования

Пример: Морфометрия жуков-листоедов

Измерения 43 самцов жуков-листоедов двух видов жуков из подсемейства козявок (Galerucinae) в семействе листоедов (Chrysomelidae): *Chaetocnema concinna*, *Ch. heptapotamica*.

Переменные

- fjft — ширина первого членика первой лапки в микронах (сумма измерений для обеих лапок)
- species — вид блох (1=Ch. concinna, 2= Ch. heptapotamica)

Есть ли морфологические различия между видами?

```
library(readxl)
flea <- read_excel(path = "data/fleabeetles-subset.xlsx", sheet = "dat")
```

Фрагмент данных из работы Lubischew, A.A., 1962. On the use of discriminant functions in taxonomy. Biometrics, pp.455-477.

Все ли правильно открылось?

```
str(flea)  # Структура данных
```

```
# Classes 'tbl_df', 'tbl' and 'data.frame': 43 obs. of  2 variables:
#  $ fjft   : num  191 185 200 173 171 160 188 186 174 163 ...
#  $ species: num  1 1 1 1 1 1 1 1 1 1 ...
```

```
head(flea) # Первые несколько строк файла
```

```
# # A tibble: 6 × 2
#    fjft species
#   <dbl>   <dbl>
# 1   191       1
# 2   185       1
# 3   200       1
# 4   173       1
# 5   171       1
# 6   160       1
```

Делаем фактором переменную, где записан вид

```
flea$species <- factor(flea$species, levels = c(1, 2),
                       labels = c("cocin", "hept"))
```

Знакомимся с данными

Есть ли пропущенные значения?

```
colSums(is.na(flea))
```

```
#    fjft species
#       0       0
```

Каковы объемы выборок? Поскольку нет пропущенных значений, можно посчитать так

```
table(flea$species)
```

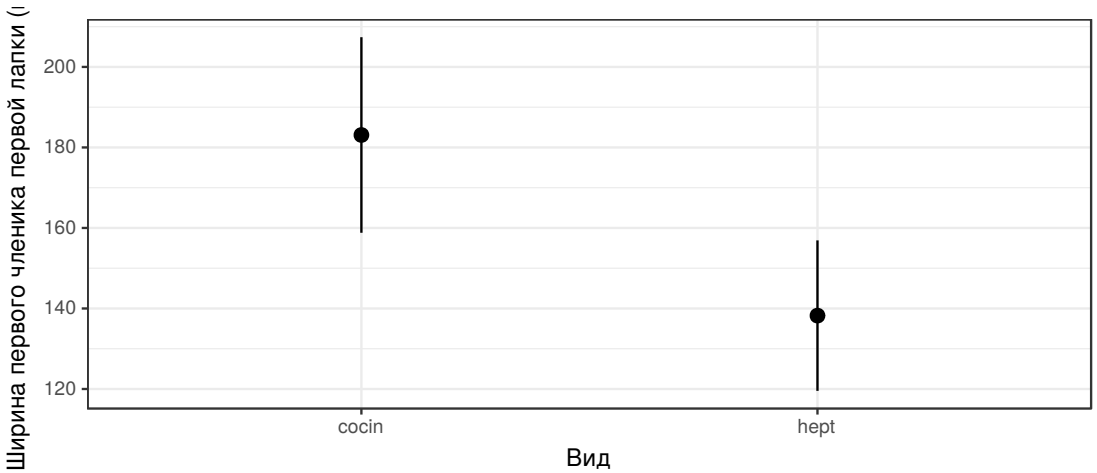
```
#
# cocin  hept
#    21    22
```

Представим, что эти данные — это данные пилотного исследования.

Мы хотим выяснить, сколько нужно жуков, чтобы показать, что ширина первого членика первой лапки различается у этих двух видов

График средних и стандартных отклонений

```
library(ggplot2)
theme_set(theme_bw())
ggplot(data = flea, aes(x = species, y = fjft)) +
  stat_summary(geom = "pointrange", fun.data = mean_sdl) +
  labs(y = "Ширина первого членика первой лапки (мкм)",
       x = "Вид")
```



Величина эффекта по исходным данным

```
library(effsize)
eff_flea <- cohen.d(flea$fjft, flea$species)
eff_flea
```

```
#
# Cohen’s d
#
# d estimate: 4.153819 (large)
# 95 percent confidence interval:
#      inf      sup
# 3.059340 5.248298
```

Вычислим модуль, поскольку для `pwr.t.test()` эффект должен быть положительным

```
effect_size_flea <- abs(eff_flea$estimate)
```

Задание

Рассчитайте объем выборки, чтобы показать различия размеров с вероятностью 0.8 на уровне значимости 0.05

Используйте функцию `pwr.t.test()`

Решение

```
pwr_flea <- pwr.t.test(n = NULL, d = effect_size_flea,
                       power = 0.8, sig.level = 0.05,
                       type = ”two.sample”,
                       alternative = ”two.sided”)
pwr_flea
```

```
#
#      Two-sample t test power calculation
#
#               n = 2.354027
#               d = 4.153819
#       sig.level = 0.05
#           power = 0.8
#     alternative = two.sided
#
# NOTE: n is number in *each* group
```

Решение

```
pwr_flea <- pwr.t.test(n = NULL, d = effect_size_flea,
                       power = 0.8, sig.level = 0.05,
                       type = "two.sample",
                       alternative = "two.sided")
pwr_flea
```

```
#
#      Two-sample t test power calculation
#
#               n = 2.354027
#               d = 4.153819
#       sig.level = 0.05
#           power = 0.8
#     alternative = two.sided
#
# NOTE: n is number in *each* group
```

- Нужна выборка из **3 жуков каждого вида**, чтобы с вероятностью 0.8 обнаружить различия размеров между видами.

Задание

Представьте, что в датасете `sleep` содержатся данные пилотного исследования.

Оцените, какой объем выборки нужно взять, чтобы показать, что число часов дополнительного сна после применения двух препаратов различается?

Решение

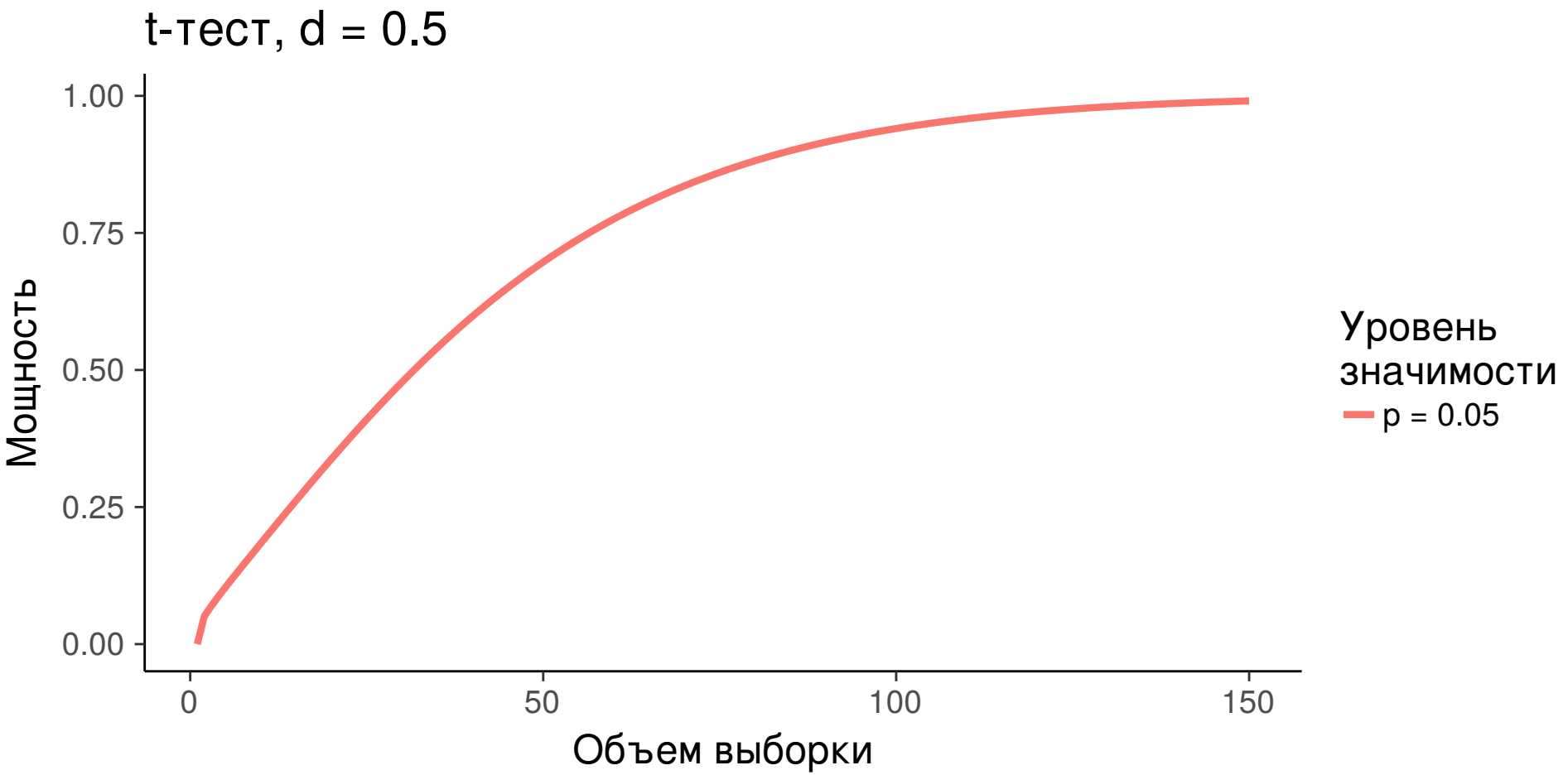
```
eff_sleep <- cohen.d(sleep$extra, sleep$group)
effect_sleep <- abs(eff_sleep$estimate)
pwr_sleep <- pwr.t.test(n = NULL, d = effect_sleep,
                        power = 0.8, sig.level = 0.05,
                        type = "two.sample",
                        alternative = "two.sided")
pwr_sleep
```

```
#
#      Two-sample t test power calculation
#
#               n = 23.6672
#               d = 0.8321811
#       sig.level = 0.05
#           power = 0.8
#     alternative = two.sided
#
# NOTE: n is number in *each* group
```

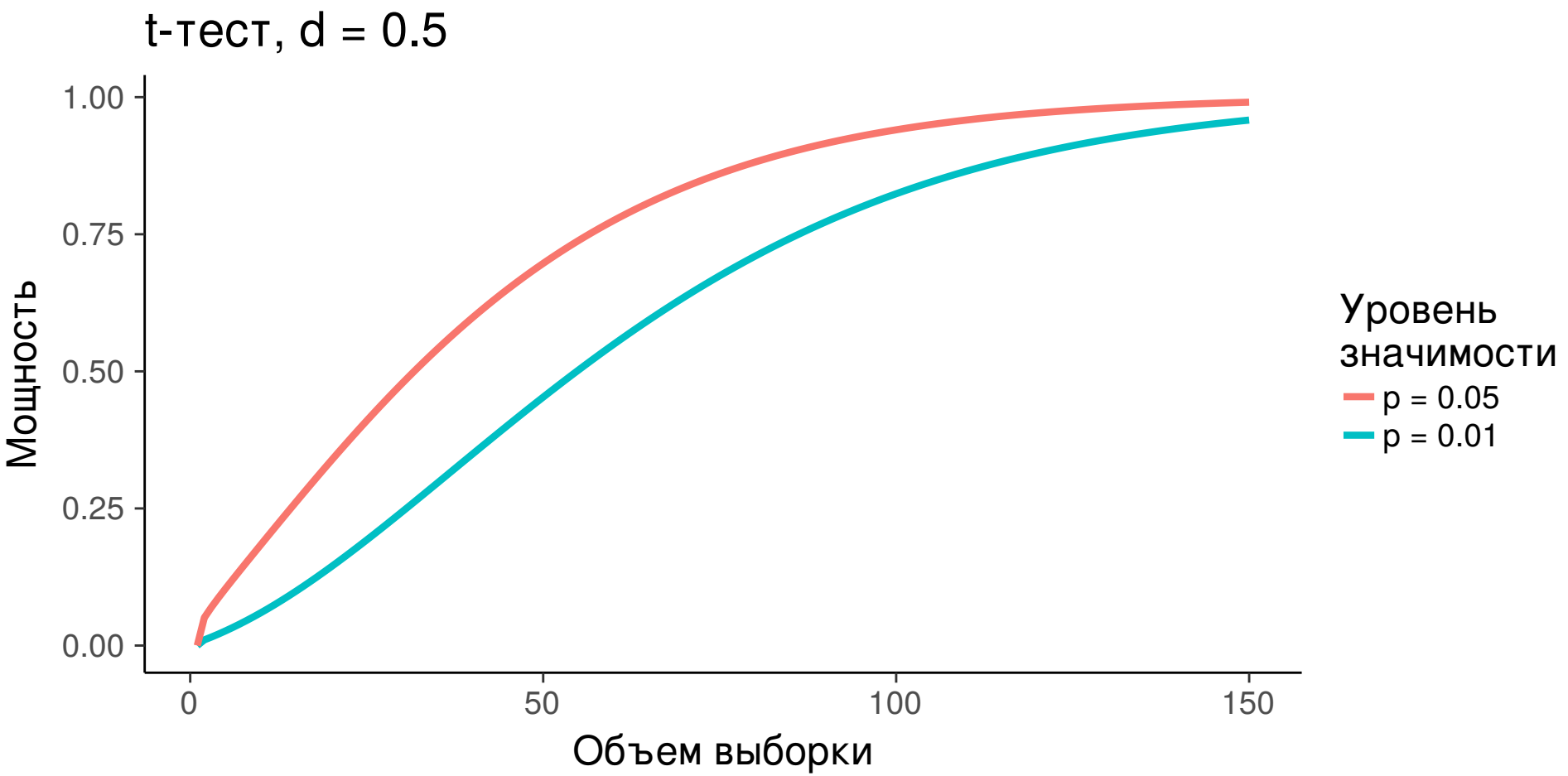
Нужна выборка **24 человека в каждой из групп**, чтобы с вероятностью 0.8 обнаружить различия числа часов дополнительного сна после применения двух препаратов.

Как влиять на мощность теста?

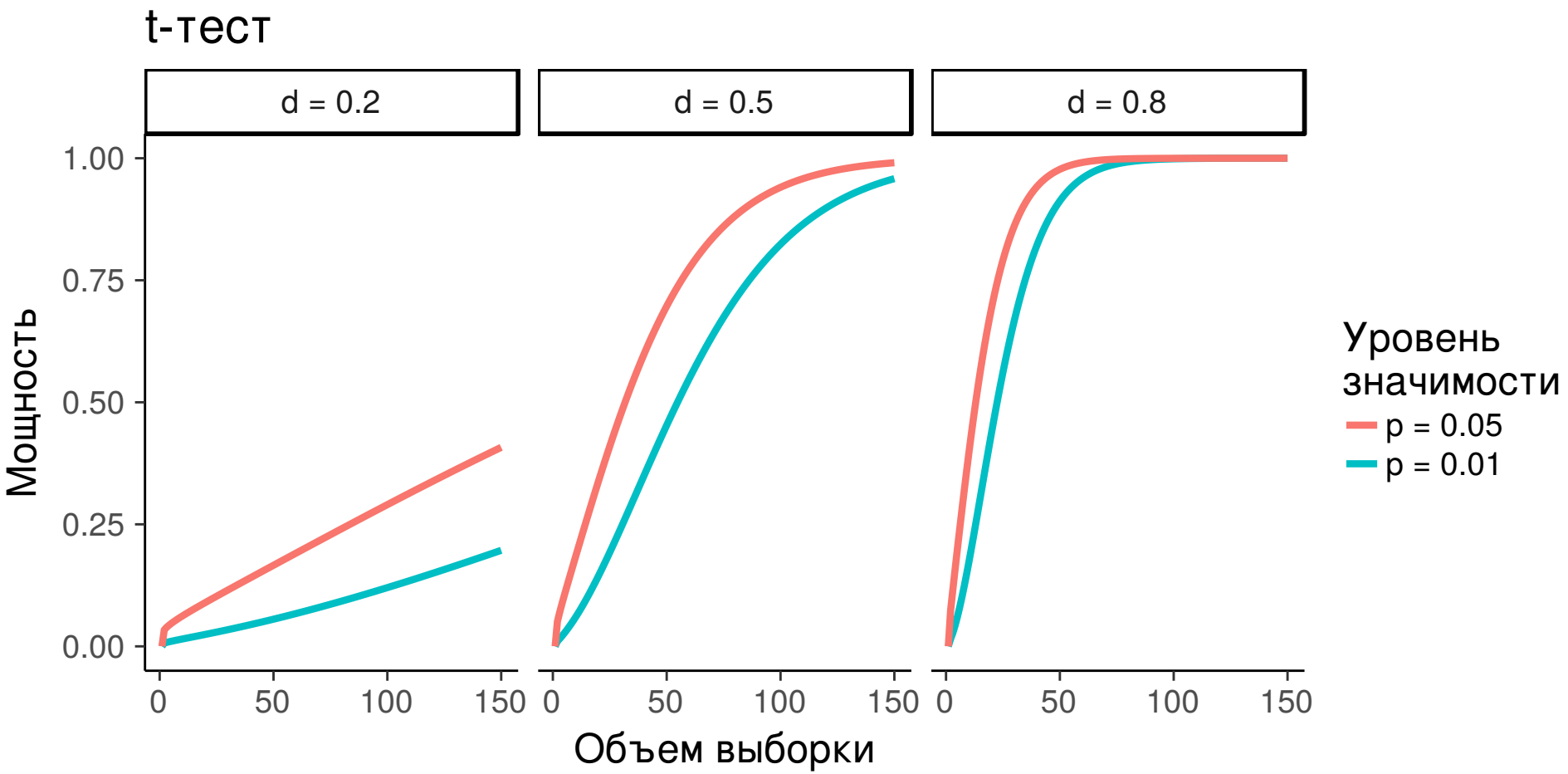
Чем больше объем выборки — тем больше мощность



Чем больше уровень значимости — тем больше мощность



Чем больше величина различий — тем больше мощность



Каким образом можно повлиять на мощность теста?

- Мощность теста можно регулировать, если
  - изменить число повторностей
  - выбрать другой уровень значимости (α)
  - определиться, какие эффекты действительно важны (*ES*)

Take-home messages

- Чтобы не находить несуществующих эффектов, фиксируем уровень значимости
- Чтобы не пропустить значимое, рассчитываем величину эффекта, объем выборки и мощность теста
- Способность выявлять различия зависит
  - от объема выборки,
  - от уровня значимости
  - от величины эффекта

Дополнительные ресурсы

- Quinn, Keough, 2002, pp. 164-170
- OpenIntro: Statistics
- Sokal, Rohlf, 1995, pp. 167-169.
- Zar, 1999, p. 83.
- R Data Analysis Examples - Power Analysis for Two-group Independent sample t-test. UCLA: Statistical Consulting Group.
- R Data Analysis Examples - Power Analysis for One-sample t-test. UCLA: Statistical Consulting Group.
- FAQ - How is effect size used in power analysis? UCLA: Statistical Consulting Group.