Tutorial 5

Biological Data Analysis

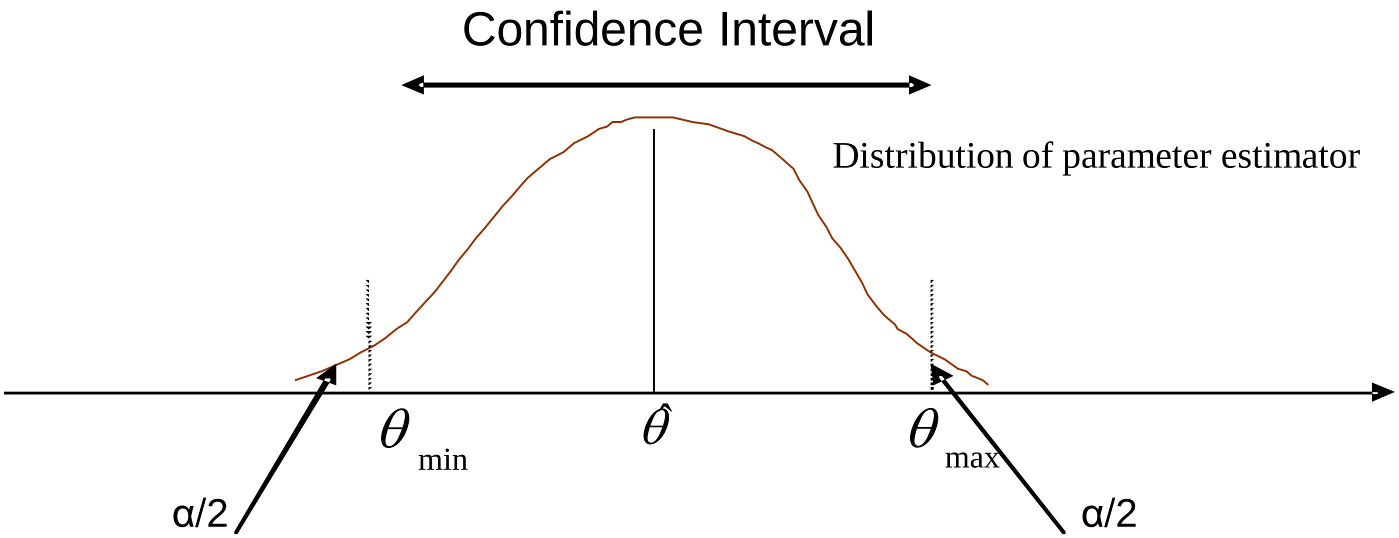
Spring 2023

Outline

- Parametric Confidence Interval
- t Distribution
- Two groups
- Permutations

Confidence Interval

- A range with 100(1-α)% of estimator distribution (α is the alpha error. Usually 0.05, but can also be 0.01 or 0.1, depending on a question)
- Probability that the real population parameter is outside the CI range is low



Hypothesis testing

P-value – a probability of getting a value as the Null Model (or more extreme) from the estimator distribution. Also called significance

If p-value is smaller than alpha error – null model is rejected. Otherwise, null model is not rejected

Effect size

Cohen's d For difference between two groups

$$\frac{\bar{x}_1 - \bar{x}_2}{s_{comb}}$$

$$s_{comb} = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}}$$

Small	0.2
Medium	0.5
Large	0.8

Parametric models

When we have a reasonable basis to assume that the population is distributed according to a particular distribution:

- Define a distribution
- Evaluate the parameters
- Look at Confidence Interval of the parameters
- Draw conclusion about the population behavior

Parametric models

Distribution	Parameters	Description	Mean	Variance
Normal(μ, σ)	μ , σ	Gaussian distribution with mean μ and standard deviation σ	μ	σ^2
Binomial(N, p)	N , p	Number of successful trials in N trials. Probability of success in a single trial – p	Np	$Np(1-p)$
Poisson(λ)	λ	Number of events in a time period with average number of events - λ	λ	λ
t(υ)	υ	Degree of freedom	0	$\upsilon/(\upsilon-2)$ for $\upsilon>2$

Exercise 1

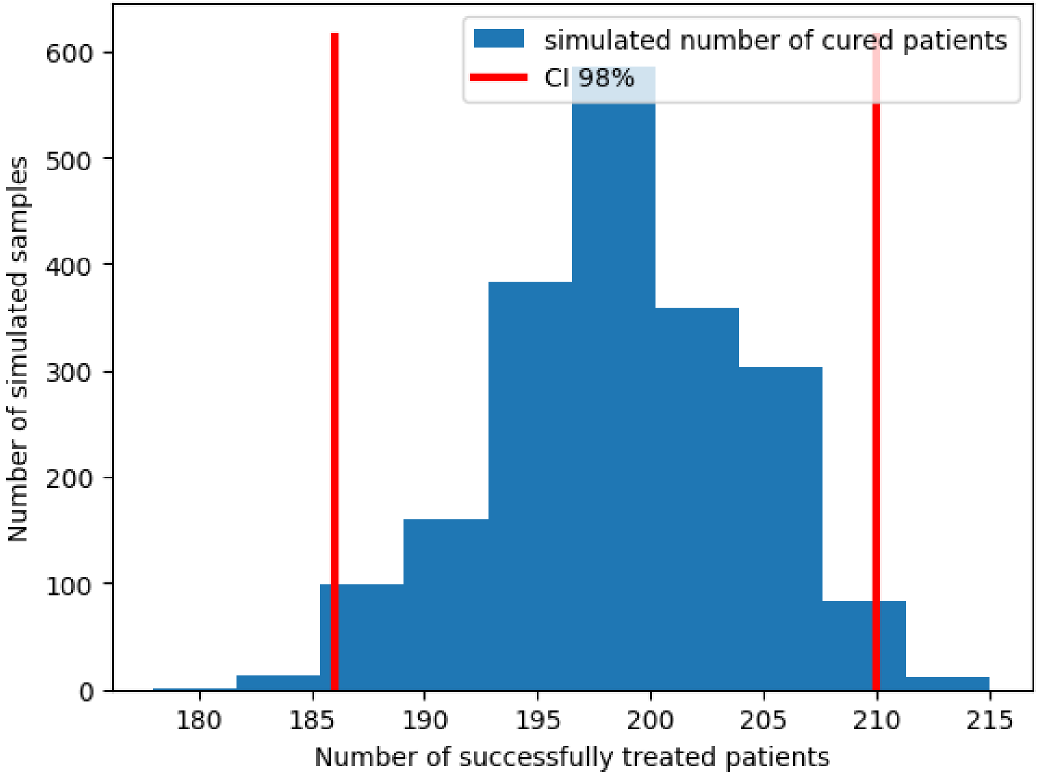
A newly developed drug has a 0.85 probability of being successful. 234 patients took the drug. How many of them will be treated successfully using this drug?

Define the distribution and estimate the parameters

Binomial Distribution(N , p)
N = 234
P = 0.85
Significance level – 98%

Exercise 1

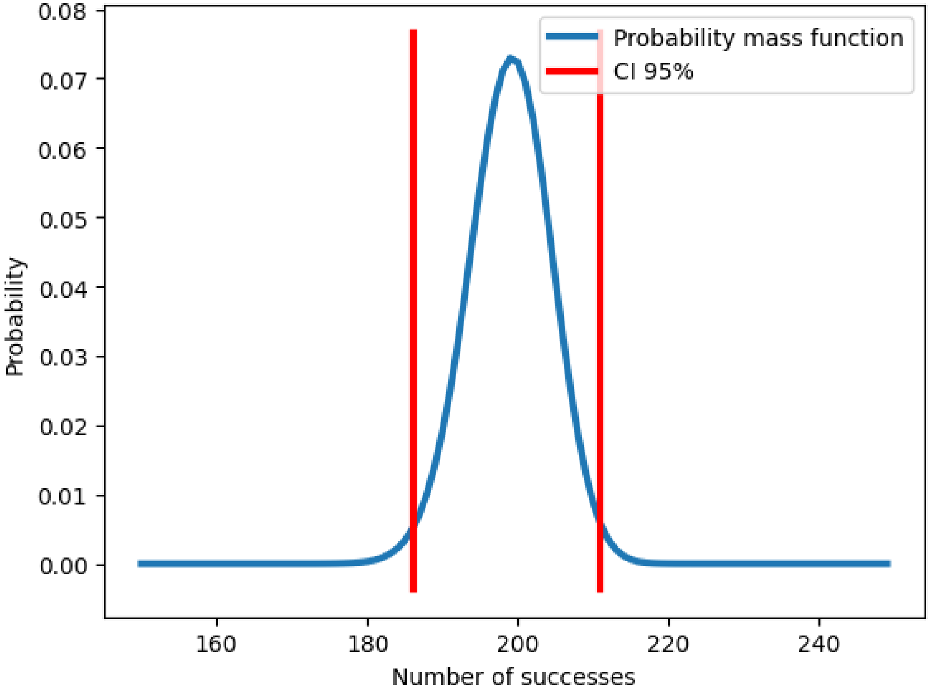
Method 1: parametric bootstrap – sampling from distribution



98% Confidence Interval = (186.0 , 210.0)

Exercise 1

Method 2: parametric CI: assume binomial distribution
Binomial (N=234 , p=0.85)



98% Confidence Interval = (186.0 , 211.0)

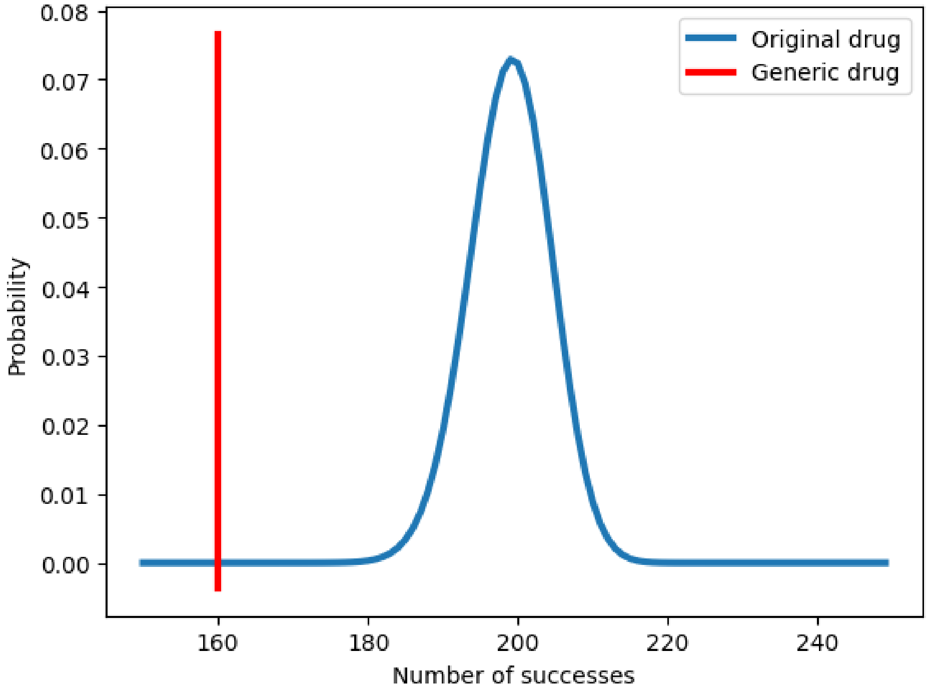
Exercise 1b

The company that produces generic drugs tested the drug that is supposed to be equivalent to the original. Out of 234 patients that took the generic 160 showed improvement.

Is the generic drug as effective as the original or the company should reevaluate its manufacturing process?

Exercise 1b

What is the probability of getting 160 cured patient (or less) out of 234 if the number of cured patients distributes Binomial(234,0.85)



Probability =

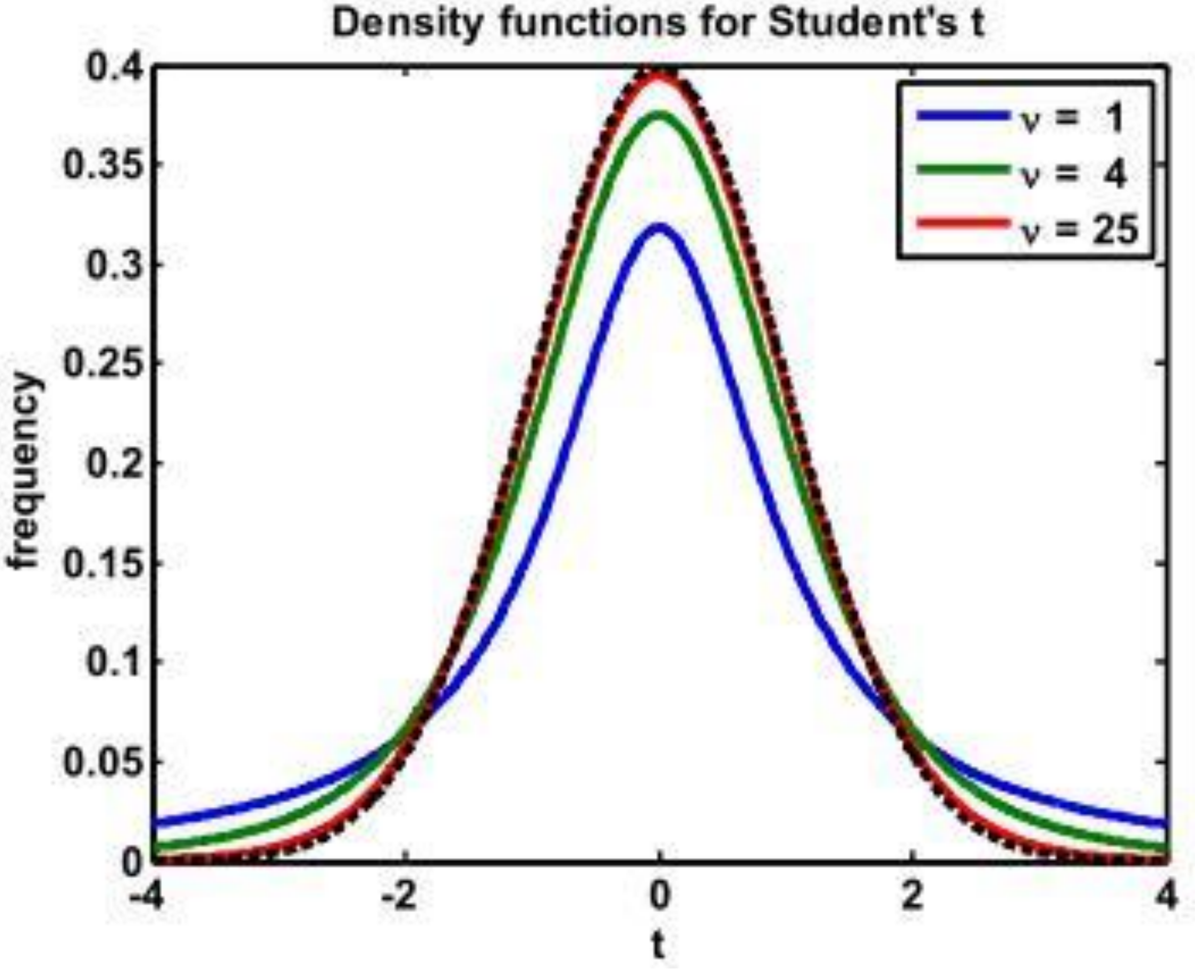
1.27e-10

Useful function

Distribution	Probability density/mass function	Cumulative distribution function	Inverse Cumulative distribution function	Generate random variables
Normal(μ, σ)	scipy.stats.norm.pdf	~.cdf	~.ppf	~.rvs
Binomial(N, p)	scipy.stats.binom.pmf	~.cdf	~.ppf	~.rvs
Poisson(λ)	scipy.stats.poisson.pmf	~.cdf	~.ppf	~.rvs
t(nu - degrees of freedom)	scipy.stats.t.pdf	~.cdf	~.ppf	~.rvs

You can also generate random numbers from distributions using numpy.random

Student’s t distribution



Student’s t distribution

- A variable from Normal distribution with expected value 0 and divided by standard error has t distribution
- t distribution is defined by degrees of freedom
- For the mean estimator

$$\frac{\bar{x}-\mu}{s\hat{e}} \sim t_{\nu=N-1} \quad ; \quad s\hat{e} = {std(x)}\Big/{\sqrt{N}}$$

- Confidence Interval for the mean estimator

$$\bar{x} - s\hat{e}\cdot t_{n-1}^{(\alpha/2)} \leq \mu \leq \bar{x} + s\hat{e}\cdot t_{n-1}^{(\alpha/2)}$$

Exercise 2

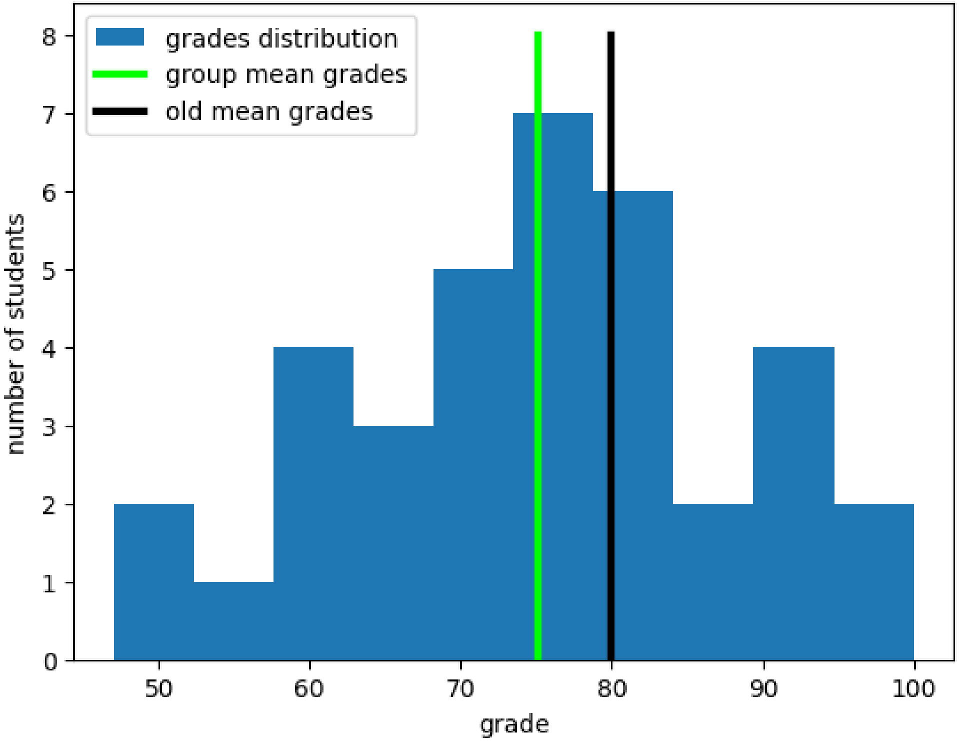
One country evaluates the education of primary school children by giving one national test in mathematics in all schools at the end of the sixth grade. In 1994, after composing a test, the committee wanted to check if the test has the same difficulty as in previous years. They chose 36 random student and gave them the test to solve.

The average grade for the test in previous years was 80.

Is the test more difficult than in previous years?

Exercise 2

`grades =`
[73 74 82 78 70 66 99 80 92 79 60 62 91 74 56 75 79 83 82 68 90 61 71
100 52 69 77 58 90 89 77 47 68 69 87 77]



Exercise 2

- Define population, sample, estimator

- Find 95% CI range for students grades

- Given that the average grade in previous tests was 80, can with say the test more difficult than in previous years?

Exercise 2 – parametric CI

$$\overline{x} - s\hat{e} \cdot t_{n-1}^{(\alpha/2)} \le \mu \le \overline{x} + s\hat{e} \cdot t_{n-1}^{(\alpha/2)}$$

Group grades mean
75.14

grades standard error
2.11

N = 36

t range [-2.03 2.03]

Confidence Interval 95% [70.85 79.43]

80 is outside the CI

The test is harder than in previous years

Exercise 2
group comparison to value significance

Find t-statistics $$\frac{\bar{x}-\mu}{s\hat{e}} \sim t_{\nu=N-1} \quad ; \quad s\hat{e} = {std(x)} \Big/ {\sqrt{N}}$$

Significance: p-value = 0.0138

Null model is rejected

Exercise 2
group comparison to value
effect size

Effect Size for comparison of two groups $$\frac{\bar{x}_1 - \bar{x}_2}{s_{comb}}$$

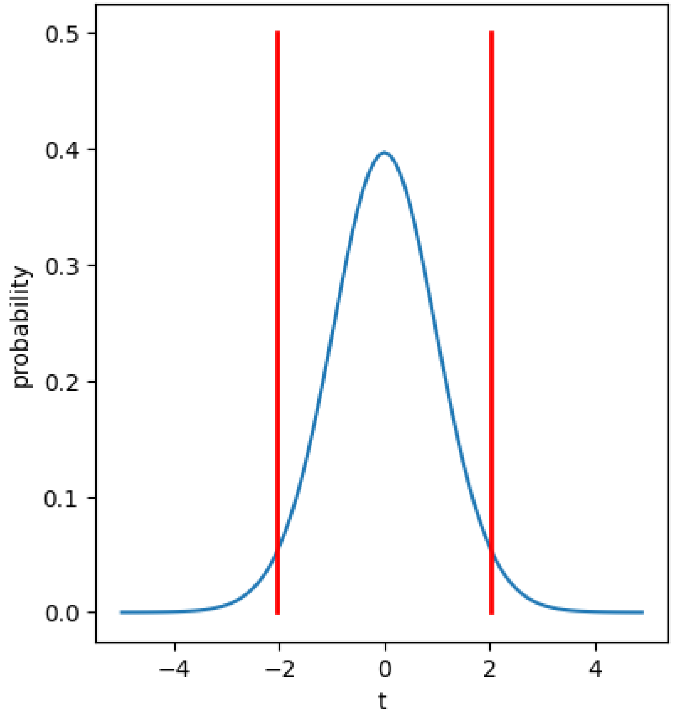
Effect Size for comparison of one group to number $$\frac{\bar{x} - c}{s_x}$$

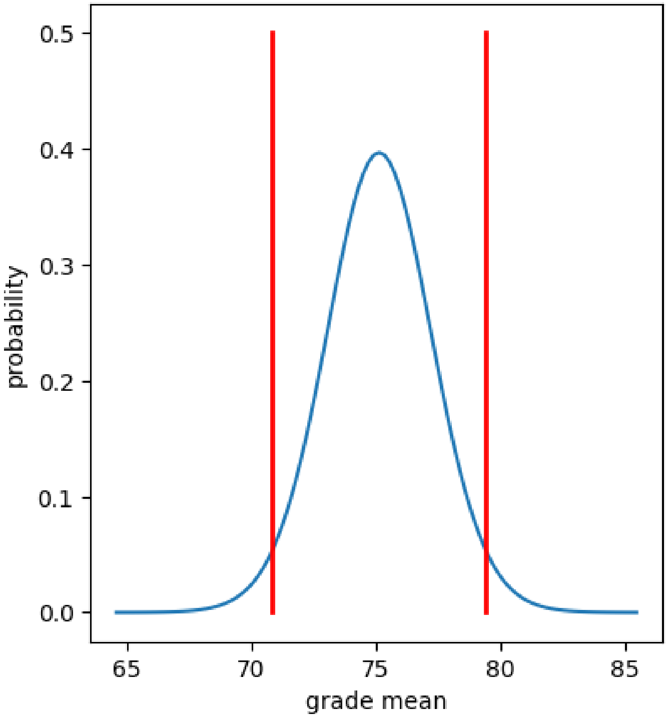
Effect Size for difference in group grades to 80 = -0.3834

Small effect

Exercise 2 – parametric CI demonstration

$$\frac{\bar{x}-\mu}{s\hat{e}} \sim t_{\nu=N-1} \quad ; \quad s\hat{e} = {}^{std(x)}\!\big/_{\sqrt{N}}$$





Compare two groups

Difference between means of two groups divided by combined standard error has t distribution if the expected value of the difference is 0

$$\bar{x}_{12} = \bar{x}_1 - \bar{x}_2$$

$$s\hat{e}_{12}^2 = \mathrm{var}\left\{\bar{x}_1 - \bar{x}_2\right\}$$

$$= \mathrm{var}\,\bar{x}_1 + \mathrm{var}\,\bar{x}_2 = s\hat{e}_1{}^2 + s\hat{e}_2{}^2 = \frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}$$

$$\frac{\bar{x}_{12} - \mu_{12}}{s\hat{e}_{12}} \to t(\upsilon = N_1 + N_2 - 2)$$

Exercise 3

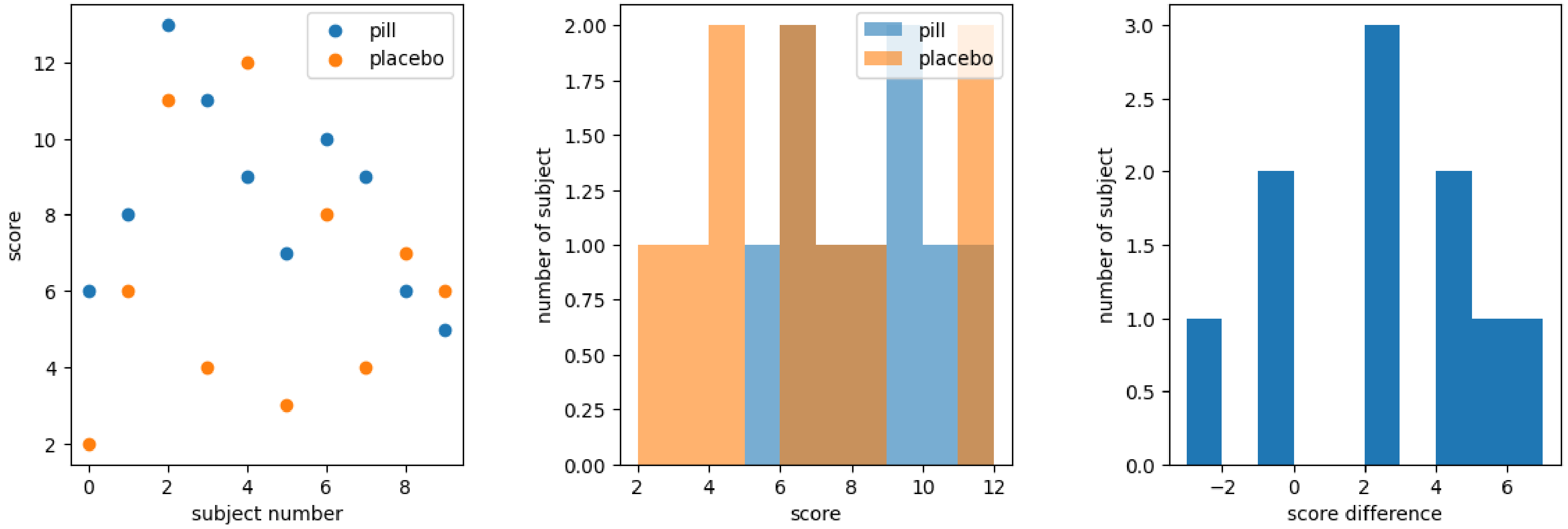
A pill for memory improvement was tested on a group of students. They were given a set of formulas and after a period of memorization had to write down all formulas they remembered.

10 students were tested twice on different days – with a pill and with placebo

Assuming the number of formulas has normal distribution, is the pill effective?

Exercise 3
present the data

pill = [6,8,13,11,9,7,10,9,6,5]
placebo = [2,6,11,4,12,3,8,4,7,6]



Exercise 3 paired case confidence interval 95% significance

Parametric CI – t-distribution

$$\bar{x} - s\hat{e} \cdot t_{n-1}^{(\alpha/2)} \le \mu \le \bar{x} + s\hat{e} \cdot t_{n-1}^{(\alpha/2)}$$

$$\frac{\bar{x} - \mu}{s\hat{e}} \sim t_{\nu = N-1} \quad ; \quad s\hat{e} = {std(x)} \big/ {\sqrt{N}}$$

In paired test we calculate standard error of one group of difference

CI 95% = [-0.0971 4.2971]

Null model inside

Significance: p-value = 0.0294

Null model not rejected

Exercise 3 unpaired case confidence interval 95% significance

20 students were tested:
10 – after taking a pill, 10 – after taking placebo

$$\bar{x} - s\hat{e} \cdot t_{n-1}^{(\alpha/2)} \leq \mu \leq \bar{x} + s\hat{e} \cdot t_{n-1}^{(\alpha/2)}$$

$$\frac{\bar{x}_{12} - \mu_{12}}{s\hat{e}_{12}} \rightarrow t(\upsilon = N_1 + N_2 - 2)$$

In unpaired test we calculate combined standard error of two groups

CI 95% = [-0.6527 4.8527]

Null model inside

Significance: p-value = 0.0717

Null model not rejected

Exercise 3
Effect size

Effect Size for comparison of two groups paired

$$\frac{\overline{x_1 - x_2}}{s_{(x_1 - x_2)}}$$

Effect Size = 0.6837

Effect Size for comparison of two groups unpaired

$$\frac{\bar{x}_1 - \bar{x}_2}{s_{comb}}$$

Effect Size = 0.7168

Permutations

Distribution of Null model

Given two groups,

Null Model – there is no difference between the groups, meaning they come from one population

Create new samples from data under assumption that there is no difference between the groups

Hypothesis testing using permutations

1. Define Null Model
2. Choose the statistic
3. Calculate the statistic of the data
4. Permute the data
5. Calculate the statistic of the permuted data
6. Repeat 4-5 till you have the distribution of the statistic of the permuted data
7. Look at the statistic of the original data compared to the permuted distribution. P-value is the probability of getting the statistic of the original data or more extreme (removed from the null model) under distribution of the null model
8. Reject/not reject the null model

Exercise 4

Serotonin is a chemical that influences mood balance.

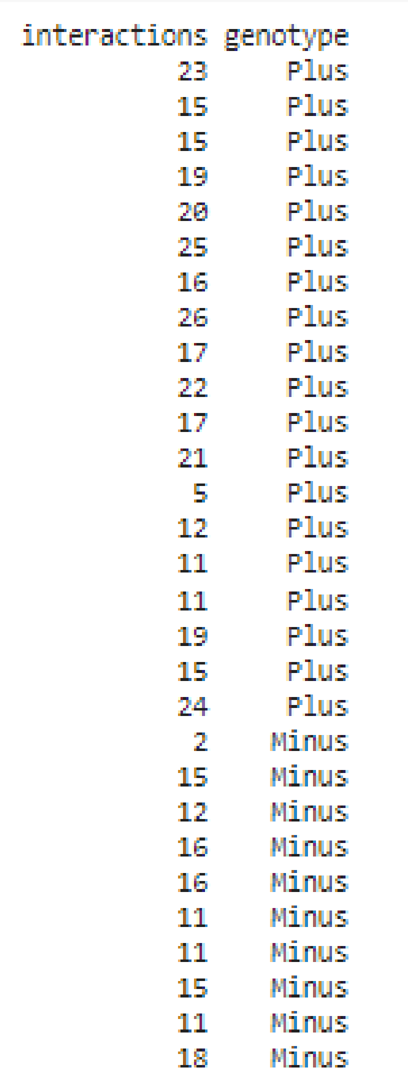
How does it affect mice behavior?

Scientists genetically altered mice by "knocking out" the expression of a gene, tryptophan hydroxylase 2 (Tph2), that regulates serotonin production.

With careful breeding, the scientists produced two types of mice that we label as “Minus” for Tph2-/-, “Plus” for Tph2+/+.

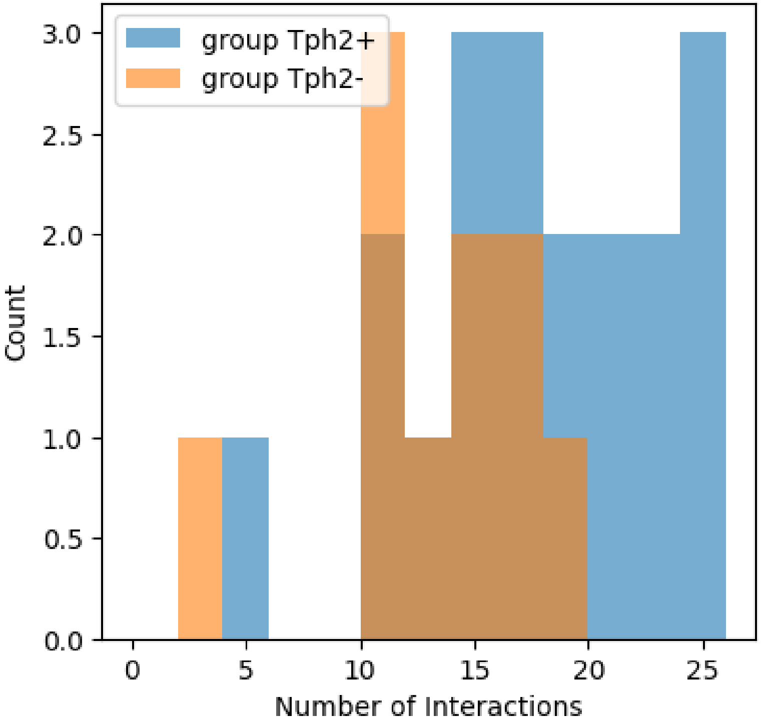
The variable ‘genotype’ records Minus/Plus.

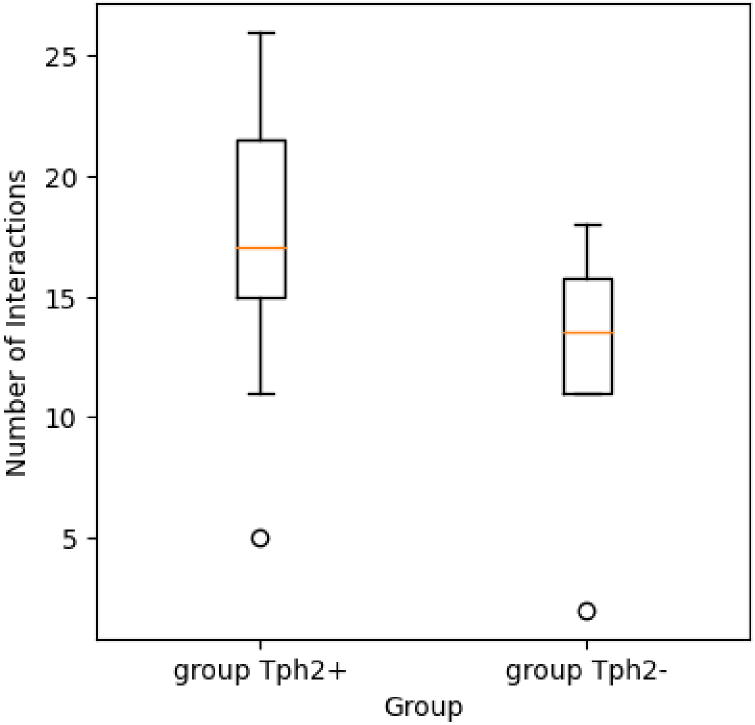
The variable ‘interactions’ is the number of social contacts that a mouse had with other mice during an experiment.



```
interactions genotype
          23     Plus
          15     Plus
          15     Plus
          19     Plus
          20     Plus
          25     Plus
          16     Plus
          26     Plus
          17     Plus
          22     Plus
          17     Plus
          21     Plus
           5     Plus
          12     Plus
          11     Plus
          11     Plus
          19     Plus
          15     Plus
          24     Plus
           2    Minus
          15    Minus
          12    Minus
          16    Minus
          16    Minus
          11    Minus
          11    Minus
          15    Minus
          11    Minus
          18    Minus
```

Exercise 4


group Tph2+
group Tph2-
Count
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
5
10
15
20
25
Number of Interactions


Number of Interactions
5
10
15
20
25
group Tph2+
group Tph2-
Group

Exercise 4

Null Model: Serotonin has no influence on behavior

Effects Model: Serotonin has influence on behavior. (Mice with normal serotonin production interact with other mice more often than the mice with impaired serotonin production)

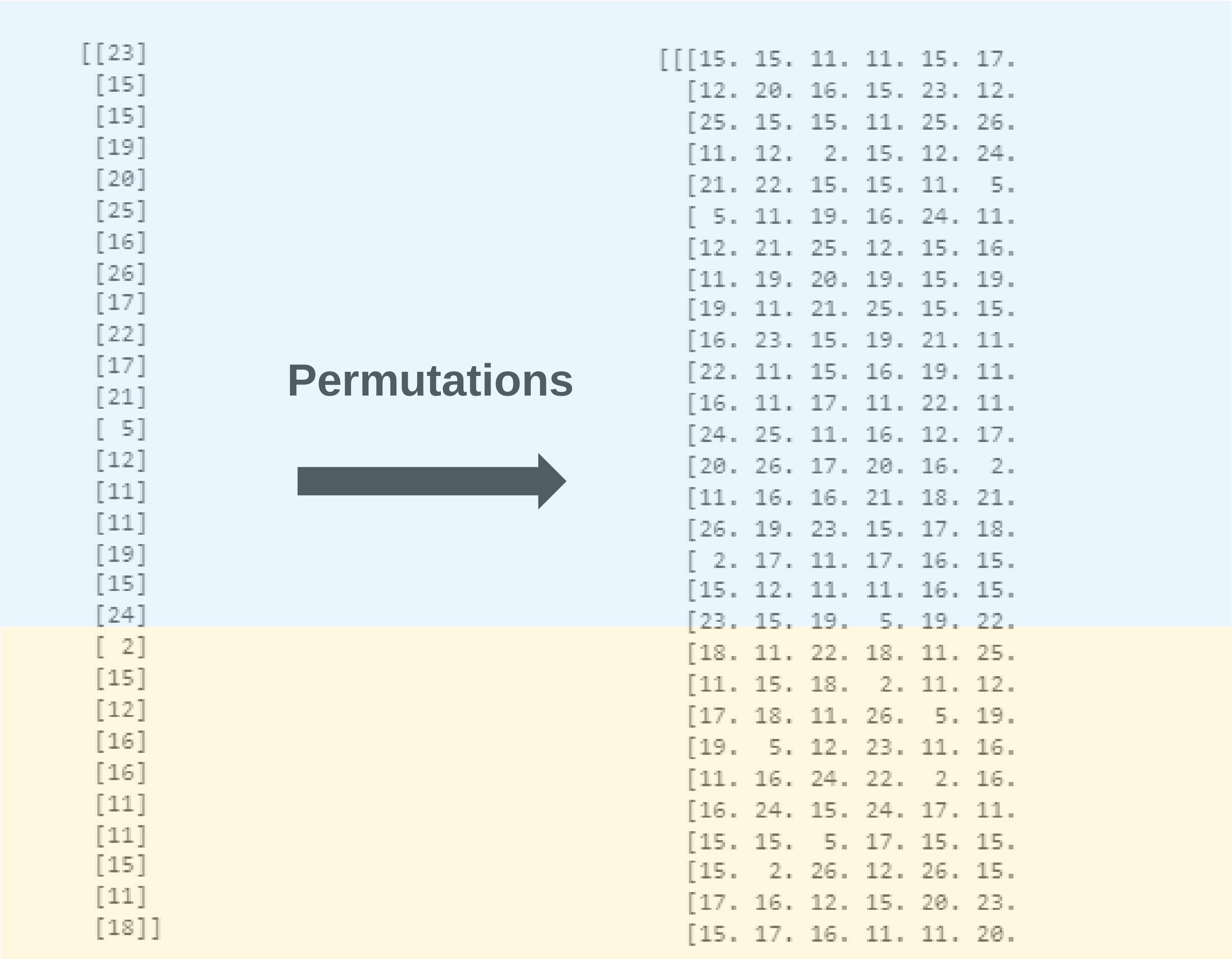
Statistic: difference of means between two groups – number of interactions with other mice

Significance level: 0.05

Exercise 4: permutations

Group Tph+

Group Tph+

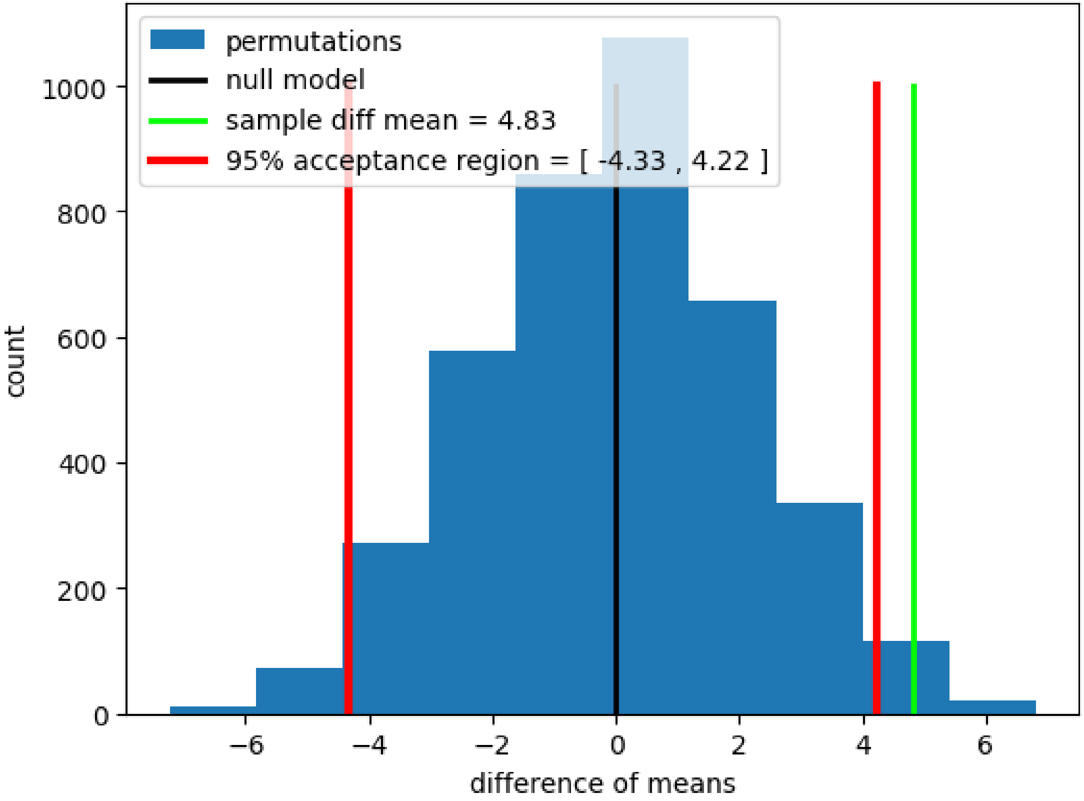


```
[[23]
 [15]
 [15]
 [19]
 [20]
 [25]
 [16]
 [26]
 [17]
 [22]
 [17]
 [21]
 [ 5]
 [12]
 [11]
 [11]
 [19]
 [15]
 [24]
 [ 2]
 [15]
 [12]
 [16]
 [16]
 [11]
 [11]
 [15]
 [11]
 [18]]
```

Permutations

```
[[[15. 15. 11. 11. 15. 17.
  [12. 20. 16. 15. 23. 12.
  [25. 15. 15. 11. 25. 26.
  [11. 12.  2. 15. 12. 24.
  [21. 22. 15. 15. 11.  5.
  [ 5. 11. 19. 16. 24. 11.
  [12. 21. 25. 12. 15. 16.
  [11. 19. 20. 19. 15. 19.
  [19. 11. 21. 25. 15. 15.
  [16. 23. 15. 19. 21. 11.
  [22. 11. 15. 16. 19. 11.
  [16. 11. 17. 11. 22. 11.
  [24. 25. 11. 16. 12. 17.
  [20. 26. 17. 20. 16.  2.
  [11. 16. 16. 21. 18. 21.
  [26. 19. 23. 15. 17. 18.
  [ 2. 17. 11. 17. 16. 15.
  [15. 12. 11. 11. 16. 15.
  [23. 15. 19.  5. 19. 22.
  [18. 11. 22. 18. 11. 25.
  [11. 15. 18.  2. 11. 12.
  [17. 18. 11. 26.  5. 19.
  [19.  5. 12. 23. 11. 16.
  [11. 16. 24. 22.  2. 16.
  [16. 24. 15. 24. 17. 11.
  [15. 15.  5. 17. 15. 15.
  [15.  2. 26. 12. 26. 15.
  [17. 16. 12. 15. 20. 23.
  [15. 17. 16. 11. 11. 20.
```

Exercise 4: permutations Acceptance Region



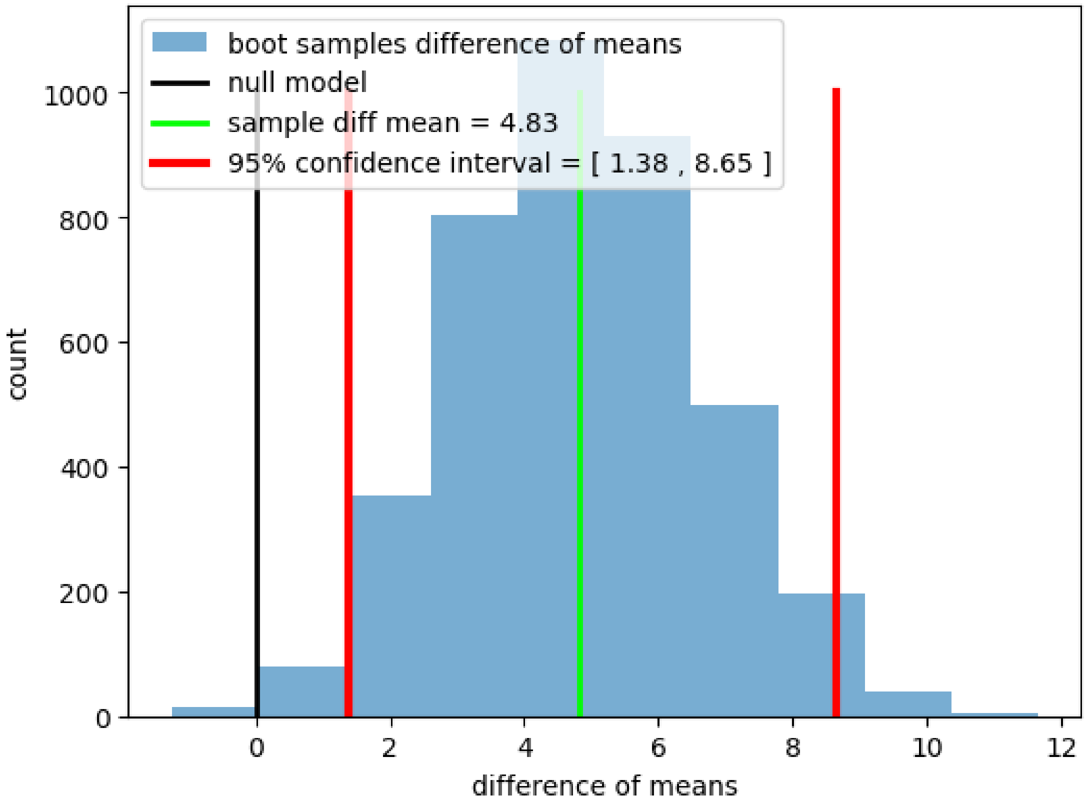
acceptance region 95%
[-4.33, 4.22]

Sample value outside

p value = 0.0135
Smaller than 0.05

Null model rejected

Exercise 4: bootstrap Confidence Interval



Confidence interval 95%
[1.38 , 8.65]

Null model outside

p value = 0.0035
Smaller than 0.05

Null model rejected

Exercise 4: t-distribution Parametric Confidence Interval

$$\bar{x} - s\hat{e} \cdot t_{n-1}^{(\alpha/2)} \leq \mu \leq \bar{x} + s\hat{e} \cdot t_{n-1}^{(\alpha/2)}$$

95% confidence interval = [0.93 , 8.73]

Null model outside

$$\frac{\bar{x}_{12} - \mu_{12}}{s\hat{e}_{12}} \rightarrow t(\upsilon = N_1 + N_2 - 2)$$

P value = 0.0086

Smaller than wanted significance level

Null model rejected

Exercise 4: Effect Size

$$\frac{\bar{x}_1 - \bar{x}_2}{s_{comb}}$$

$$s_{comb} = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}}$$

Effect Size = 0.9341

Large effect

The effect is statistically significant and large.

Serotonin can influence the behavior of mice.
Mice with defected serotonin production have significantly lower number of contacts with other mice.