An introduction to R: **Basic Statistics with R**

Noémie Becker, Benedikt Holtmann & Dirk Metzler [1]

nbecker@bio.lmu.de - holtmann@bio.lmu.de

Winter semester 2016-17

[1] Special thanks to: Prof. Dr. Martin Hutzenthaler and Dr. Sonja Grath for course development

Contents

A simple example

- You want to show that a treatment is effective.

A simple example

- You want to show that a treatment is effective.
- You have data for 2 groups of patients with and without treatment.

A simple example

- You want to show that a treatment is effective.
- You have data for 2 groups of patients with and without treatment.
- 80% patients with treatment recovered whereas only 30% patients without recovered.

A simple example

- You want to show that a treatment is effective.
- You have data for 2 groups of patients with and without treatment.
- 80% patients with treatment recovered whereas only 30% patients without recovered.
- A pessimist would say that this just happened by chance.
- What do you do to convince the pessimist?

A simple example

- You want to show that a treatment is effective.
- You have data for 2 groups of patients with and without treatment.
- 80% patients with treatment recovered whereas only 30% patients without recovered.
- A pessimist would say that this just happened by chance.
- What do you do to convince the pessimist?
- You assume he is right and you show that under this hypothesis the data would be very unlikely.

In statistical words

- What you want to show is the alternative hypothesis H_1.
- The pessimist (by chance) is the null hypothesis H_0.

In statistical words

- What you want to show is the alternative hypothesis H_1.
- The pessimist (by chance) is the null hypothesis H_0.
- Show that the observation and everything more 'extreme' is sufficiently unlikely under this null hypothesis. Scientists have agreed that it suffices that this probability is at most 5%.
- This refutes the pessimist. Statistical language: We reject the null hypothesis on the significance level 5%.

In statistical words

- What you want to show is the alternative hypothesis H_1.
- The pessimist (by chance) is the null hypothesis H_0.
- Show that the observation and everything more 'extreme' is sufficiently unlikely under this null hypothesis. Scientists have agreed that it suffices that this probability is at most 5%.
- This refutes the pessimist. Statistical language: We reject the null hypothesis on the significance level 5%.
- $p = P$(observation and everything more 'extreme' $/\mathrm{H}_0$ is true)
- If the p value is over 5% you say you cannot reject the null hypothesis.

Statistical tests in R

There is a huge variety of statistical tests that you can perform in R.
We will cover on example of test in this lecture.

Contents

T test main idea

We have measured a variable in one or two samples and we want to test:

- One-sample: test if the mean is equal to a certain value (often 0)
- Two-samples: test if the two means are different

T test main idea

We have measured a variable in one or two samples and we want to test:

- One-sample: test if the mean is equal to a certain value (often 0)
- Two-samples: test if the two means are different

In case of two samples there are several possibilities:

- The same individuals are measured twice (ex: before and after a treatment): paired T test
- Independent samples: unpaired T test

T test main idea

We have measured a variable in one or two samples and we want to test:

- One-sample: test if the mean is equal to a certain value (often 0)
- Two-samples: test if the two means are different

In case of two samples there are several possibilities:

- The same individuals are measured twice (ex: before and after a treatment): paired T test
- Independent samples: unpaired T test

In case of unpaired there are several possibilities:

- Variance in the two samples is assumed equal
- Variance in the two samples is not assumed equal

More details about the One-sample T test

Given: Observations

$$X_1, X_2, \dots, X_n$$

More details about the One-sample T test

Given: Observations

$$X_1, X_2, \ldots, X_n$$

Null hypothesis H_0: $\mu_X = c$ (We test for a value c, usually $c = 0$)
Level of significance: α (usually $\alpha = 5\%$)

More details about the One-sample T test

Given: Observations

$$X_1, X_2, \ldots, X_n$$

Null hypothesis H_0: $\mu_X = c$ (We test for a value c, usually $c = 0$)
Level of significance: α (usually $\alpha = 5\%$)
Test: t-Test

Compute test statistic

$$t := \frac{\overline{X} - c}{s(X)/\sqrt{n}}$$

More details about the One-sample T test

Given: Observations

$$X_1, X_2, \ldots, X_n$$

Null hypothesis H_0: $\mu_X = c$ (We test for a value c, usually $c = 0$)
Level of significance: α (usually $\alpha = 5\%$)
Test: t-Test

Compute test statistic

$$t := \frac{\overline{X} - c}{s(X)/\sqrt{n}}$$

p-value $= \Pr(|T_{n-1}| \geq |t|)$ ($n - 1$ degrees of freedom)
Reject null hypothesis, if p-value $\leq \alpha$

More details about the paired T test

Given: paired observations

$$(Y_1, Z_1), (Y_2, Z_2), \ldots, (Y_n, Z_n)$$

More details about the paired T test

Given: paired observations

$$(Y_1, Z_1), (Y_2, Z_2), \ldots, (Y_n, Z_n)$$

Null hypothesis H_0: $\mu_Y = \mu_Z$
Level of significance: α (usually $\alpha = 5\%$)

More details about the paired T test

Given: paired observations

$$(Y_1, Z_1), (Y_2, Z_2), \ldots, (Y_n, Z_n)$$

Null hypothesis H_0: $\mu_Y = \mu_Z$
Level of significance: α (usually $\alpha = 5\%$)
Test: paired t-Test (more precisely: two-sided paired t-Test)

Compute the difference $X := Y - Z$
Compute test statistic

$$t := \frac{\overline{X}}{s(X)/\sqrt{n}}$$

More details about the paired T test

Given: paired observations

$$(Y_1, Z_1), (Y_2, Z_2), \ldots, (Y_n, Z_n)$$

Null hypothesis H_0: $\mu_Y = \mu_Z$
Level of significance: α (usually $\alpha = 5\%$)
Test: paired t-Test (more precisely: two-sided paired t-Test)

Compute the difference $X := Y - Z$
Compute test statistic

$$t := \frac{\overline{X}}{s(X)/\sqrt{n}}$$

p-value $= \Pr(|T_{n-1}| \geq |t|)$ ($n - 1$ degrees of freedom)
Reject null hypothesis, if p-value $\leq \alpha$

More details about the unpaired T test with equal variance

Given: unpaired observations

$$X_1, X_2, \ldots, X_n and Y_1, Y_2, \ldots, Y_m$$

More details about the unpaired T test with equal variance

Given: unpaired observations

$$X_1, X_2, \ldots, X_n and Y_1, Y_2, \ldots, Y_m$$

Null hypothesis H_0: $\mu_X = \mu_Y$
Level of significance: α (usually $\alpha = 5\%$)

More details about the unpaired T test with equal variance

Given: unpaired observations

$$X_1, X_2, \ldots, X_n and Y_1, Y_2, \ldots, Y_m$$

Null hypothesis H_0: $\mu_X = \mu_Y$
Level of significance: α (usually $\alpha = 5\%$)
Test: unpaired t-Test (more precisely: two-sided unpaired t-Test)

Compute common variance of the sample

$$s_p^2 = \frac{(n-1)\cdot s_X^2 + (m-1)\cdot s_Y^2}{m+n-2}$$

Compute test statistic

$$t = \frac{\overline{X} - \overline{Y}}{s_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

More details about the unpaired T test with equal variance

Given: unpaired observations

$$X_1, X_2, \dots, X_n and Y_1, Y_2, \dots, Y_m$$

Null hypothesis H_0: $\mu_X = \mu_Y$
Level of significance: α (usually $\alpha = 5\%$)
Test: unpaired t-Test (more precisely: two-sided unpaired t-Test)

Compute common variance of the sample

$$s_p^2 = \frac{(n-1) \cdot s_X^2 + (m-1) \cdot s_Y^2}{m+n-2}$$

Compute test statistic

$$t = \frac{\overline{X} - \overline{Y}}{s_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

p-value $= \Pr(|T_{n+m-2}| \geq |t|)$ ($n+m-2$ degrees of freedom)
Reject null hypothesis, if p-value $\leq \alpha$

Contents

Function `t.test()`

The function used in R is called `t.test()`.

```
?t.test
 t.test(x, y = NULL,
alternative = c("two.sided", "less", "greater"),
mu = 0, paired = FALSE, var.equal = FALSE,
conf.level = 0.95, ...)
```

Martian example

Dataset containing height of martian of different colours.
See the code on the R console.

Martian example

Dataset containing height of martian of different colours.
See the code on the R console.

We cannot reject the null hypothesis.
It was an unpaired test because the two samples are independent.

Shoe example

Dataset containing wear of shoes of 2 materials A and B. The same persons have weared the two types of shoes and we have a measure of use of the shoes.

Shoe example

Dataset containing wear of shoes of 2 materials A and B. The same persons have weared the two types of shoes and we have a measure of use of the shoes.
Paired test because some persons will cause more damage to the shoe than others.
See the code on the R console.

Shoe example

Dataset containing wear of shoes of 2 materials A and B. The same persons have weared the two types of shoes and we have a measure of use of the shoes.
Paired test because some persons will cause more damage to the shoe than others.
See the code on the R console.

We can reject the null hypothesis.

Contents

Definition

There are two types of error for a statistical test:

- Type I error (or first kind or alpha error or false positive): rejecting H_0 when it is true.

Definition

There are two types of error for a statistical test:

- Type I error (or first kind or alpha error or false positive): rejecting H_0 when it is true.
- Type II error (or second kind or beta error or false negative): failing to reject H_0 when it is not true.

Definition

There are two types of error for a statistical test:

- Type I error (or first kind or alpha error or false positive): rejecting H_0 when it is true.
- Type II error (or second kind or beta error or false negative): failing to reject H_0 when it is not true.

Power of a test = 1 - β

If power=0: you will never reject H_0.

Definition

There are two types of error for a statistical test:

- Type I error (or first kind or alpha error or false positive): rejecting H_0 when it is true.
- Type II error (or second kind or beta error or false negative): failing to reject H_0 when it is not true.

Power of a test = 1 - β
If power=0: you will never reject H_0.
The choice of H_1 is important because it will influence the power.
In general the power increases with sample size.

Power in R

Use the functions `power.t.test()` to calculate the minimal sample size needed to show a certain difference.

Power in R

Use the functions `power.t.test()` to calculate the minimal sample size needed to show a certain difference.
We will try this with the following example:

- We know there is a third colour of martians (yellow) and we want to test whether their size is different from that of the green ones (64 cm).
- We assume they have the same standard deviation in size (8 cm).
- We would like to be able to find significant a difference of 5 cm with power 90%.

What is the planned test?

Power in R

Use the functions `power.t.test()` to calculate the minimal sample size needed to show a certain difference.
We will try this with the following example:

- We know there is a third colour of martians (yellow) and we want to test whether their size is different from that of the green ones (64 cm).
- We assume they have the same standard deviation in size (8 cm).
- We would like to be able to find significant a difference of 5 cm with power 90%.

What is the planned test?

One-sample t test.

Power in R

One-sample t test.

```
power.t.test(n=NULL, delta=5, sd=8,
sig.level=0.005,power=0.9, type="one.sample",
alternative="two.sided")
```

Power in R

One-sample t test.

```
power.t.test(n=NULL, delta=5, sd=8,
sig.level=0.005,power=0.9, type="one.sample",
alternative="two.sided")
One-sample t test power calculation
n = 46.77443
delta = 5
sd = 8
sig.level = 0.005
power = 0.9
alternative = two.sided
```

Contents

Questions for the exam

- The exam will be on paper (no computer).
- You still need to know the precise commands and script structure
- You are allowed to bring a two-sided A4 formula sheet (on your own handwriting only).
- The exam takes place on Friday at 10am in B00.019
- Be on time!

Do you have questions?