Rcourse: **Basic statistics with R**

Sonja Grath, Noémie Becker & Dirk Metzler

Winter semester 2013-14

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%.
- $p = P(\text{observation and everything more 'extreme'} / \mathrm{H}_0 \text{ is true})$
- 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 the most basic ones in this lecture and you can find a non-exhaustive list in your lecture notes.

Contents

The Students T test: Underline

- **What is given?** Independent observations $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_m)$).

The Students T test: Underline

- **What is given?** Independent observations $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_m)$).
- **Null hypothesis:** x and y are samples from distributions having the same mean.

The Students T test: Underline

- **What is given?** Independent observations $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_m)$).
- **Null hypothesis:** x and y are samples from distributions having the same mean.
- **R command:** `t.test(x,y)`

The Students T test: Underline

- **What is given?** Independent observations $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_m)$).
- **Null hypothesis:** x and y are samples from distributions having the same mean.
- **R command:** `t.test(x,y)`
- **Idea of the test:** If the sample means are too far apart, then reject the null hypothesis.

The Students T test: Underline

- **What is given?** Independent observations $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_m)$).
- **Null hypothesis:** x and y are samples from distributions having the same mean.
- **R command:** `t.test(x,y)`
- **Idea of the test:** If the sample means are too far apart, then reject the null hypothesis.
- Approximative test but rather robust

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 abd 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 abd 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 abd 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.

Test for (un)equality of variances

In `t.test()` there is an option `var.equal=`.
This way we can control if the variances between the two samples are assumed to be equal or not. The default value is `FALSE`.

If you want to know before applying the T test you can apply a variance test with the command `var.test`.
Let's see an example on the R console.

Test for (un)equality of variances

In `t.test()` there is an option `var.equal=`.
This way we can control if the variances between the two samples are assumed to be equal or not. The default value is `FALSE`.

If you want to know before applying the T test you can apply a variance test with the command `var.test`.
Let's see an example on the R console.We cannot reject the null

hypothesis. We thus assume the variances are equal.

Contents

Testing for dependence

The test depends on the data type:

- **Nominal variables:** not ordered like eye colour or gender

Testing for dependence

The test depends on the data type:

- **Nominal variables:** not ordered like eye colour or gender
- **Ordinal variables:** ordered but not continuous like the result of a dice

Testing for dependence

The test depends on the data type:

- **Nominal variables:** not ordered like eye colour or gender
- **Ordinal variables:** ordered but not continuous like the result of a dice
- **Continuous variables:** like body height

Testing for dependence

The test depends on the data type:

- **Nominal variables:** not ordered like eye colour or gender
- **Ordinal variables:** ordered but not continuous like the result of a dice
- **Continuous variables:** like body height

Contents

Nominal variables: Underline

- **What is given?** Pairwise observations (x_1 , y_1) , (x_2 , y_2) ... (x_n , y_n)

Nominal variables: Underline

- **What is given?** Pairwise observations (x_1, y_1), (x_2, y_2) ... (x_n, y_n)
- **Null hypothesis:** x and y are independent

Nominal variables: Underline

- **What is given?** Pairwise observations $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$
- **Null hypothesis:** x and y are independent
- **Test:** χ^2

Nominal variables: Underline

- **What is given?** Pairwise observations (x_1, y_1), (x_2, y_2) ... (x_n, y_n)
- **Null hypothesis:** x and y are independent
- **Test:** χ^2
- **R command:** `chisq.test(x,y)` or `chisq.test(contingency table)`

Nominal variables: Underline

- **What is given?** Pairwise observations $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$
- **Null hypothesis:** x and y are independent
- **Test:** χ^2
- **R command:** `chisq.test(x,y)` or `chisq.test(contingency table)`
- **Idea of the test:** Calculate the expected abundances under the assumption of independence. If the observed abundances deviate too much from the expected abundances, then reject the null hypothesis.

Nominal variables: Underline

- **What is given?** Pairwise observations (x_1, y_1), (x_2, y_2) ... (x_n, y_n)
- **Null hypothesis:** x and y are independent
- **Test:** χ^2
- **R command:** `chisq.test(x,y)` or `chisq.test(contingency table)`
- **Idea of the test:** Calculate the expected abundances under the assumption of independence. If the observed abundances deviate too much from the expected abundances, then reject the null hypothesis.
- Approximative test, see the conditions on the lecture notes

Nominal variables: Example

```
contingency <- matrix( c(47,3,8,42,60,15,8,33,3),
nrow=3 )
chisq.test(contingency)$expected
```

See on the R console.

Nominal variables: Example

```
contingency <- matrix( c(47,3,8,42,60,15,8,33,3),
nrow=3 )
chisq.test(contingency)$expected
```

See on the R console.

All expected abundances are above 5, so we may apply the test.

```
chisq.test(contingency)
```

Nominal variables: Example

```
contingency <- matrix( c(47,3,8,42,60,15,8,33,3),
nrow=3 )
chisq.test(contingency)$expected
```

See on the R console.

All expected abundances are above 5, so we may apply the test.

```
chisq.test(contingency)
```

Reject the null hypothesis that the two variables are independent.

Nominal variables: Fishers exact test

In case of 2 by 2 contigency tables the chi square approximation is not needed and we can use the **Fisher's exact test**.

```
table <- matrix( c(14,10,21,3), nrow=2 )
fisher.test(table)
```

See on the R console.

Nominal variables: Fishers exact test

In case of 2 by 2 contigency tables the chi square approximation is not needed and we can use the **Fisher's exact test**.

```
table <- matrix( c(14,10,21,3), nrow=2 )
fisher.test(table)
```

See on the R console.

Reject the null hypothesis that the two variables are independent.

Contents

Continuous variables: Underline

- **What is given?** Pairwise observations (x_1 , y_1) , (x_2 , y_2) ... (x_n , y_n)

Continuous variables: Underline

- **What is given?** Pairwise observations (x_1, y_1) , (x_2, y_2) ... (x_n, y_n)
- **Null hypothesis:** x and y are independent

Continuous variables: Underline

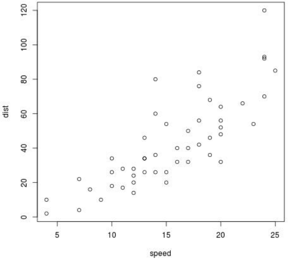
- **What is given?** Pairwise observations $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$
- **Null hypothesis:** x and y are independent
- **Test:** Pearsons correlation test for independence
- **Assumption:** x and y are samples from a normal distribution.

Continuous variables: Underline

- **What is given?** Pairwise observations $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$
- **Null hypothesis:** x and y are independent
- **Test:** Pearsons correlation test for independence
- **Assumption:** x and y are samples from a normal distribution.
- **R command:** `cor.test(x,y)`

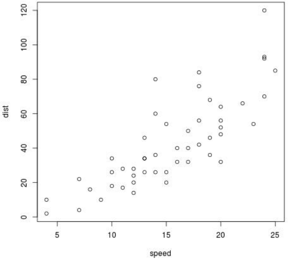
Continuous variables: Example

Distance needed to stop from a certain speed for cars. This dataset is pre-installed in R and can be loaded with the command `data(cars)`



Continuous variables: Example

Distance needed to stop from a certain speed for cars. This dataset is pre-installed in R and can be loaded with the command `data(cars)`



Reject the null hypothesis that the correlation is equal to 0.

Testing for neutrality

The Pearsons correlation assumes normal distribition of the variables.
When this is not true you can modify the option `method = "pearson"` to use another type of correlation test (Kendall or Spearman).

If you want to test for deviation from the normality you can apply a Shapiro test with the command `shapiro.test`.
Let's see an example on the R console.

Testing for neutrality

The Pearsons correlation assumes normal distribution of the variables.
When this is not true you can modify the option `method = "pearson"` to use another type of correlation test (Kendall or Spearman).

If you want to test for deviation from the normality you can apply a Shapiro test with the command `shapiro.test`.
Let's see an example on the R console.The measure of speed

does not deviate significantly from normality, but the distance variable does deviate.

Contents

Ordinal variables: Underline

- **What is given?** Pairwise observations (x_1, y_1), (x_2, y_2) ... (x_n, y_n), values can be ordered.

Ordinal variables: Underline

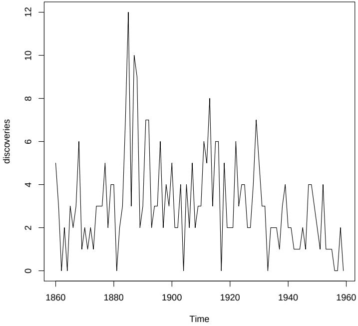
- **What is given?** Pairwise observations $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$, values can be ordered.
- **Null hypothesis:** x and y are uncorrelated

Ordinal variables: Underline

- **What is given?** Pairwise observations $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$, values can be ordered.
- **Null hypothesis:** x and y are uncorrelated
- **Test:** spearmans rank correlation rho
- **R command:** `cor.test(x,y, method="spearman")`

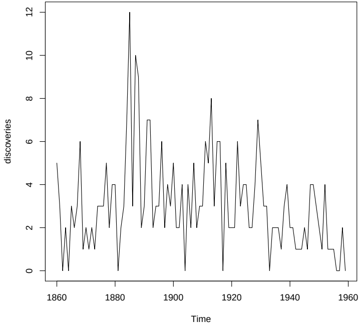
Ordinal variables: Example

Number of important scientific discoveries or inventions per year. This dataset is pre-installed in R and can be loaded with the command `data(discoveries)`



Ordinal variables: Example

Number of important scientific discoveries or inventions per year. This dataset is pre-installed in R and can be loaded with the command `data(discoveries)`



Reject the null hypothesis that the correlation is equal to 0. There is a significant negative correlation.

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 H1 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()` or `power.fisher.test()` (in package statmod) to calculate the minimal sample size needed to show a certain difference.

We will try this during the exercise session.

Contents

Concept

You may have noticed that we see a value named `df` in our test results.

Concept

You may have noticed that we see a value named `df` in our test results.
Do you know what degrees of freedom are?

Concept

You may have noticed that we see a value named `df` in our test results.
Do you know what degrees of freedom are?
Lets try with an example:
Degrees of freedom of a vector x(x1,x2,x3,x4,x5)?

Concept

You may have noticed that we see a value named `df` in our test results.
Do you know what degrees of freedom are?
Lets try with an example:
Degrees of freedom of a vector x(x1,x2,x3,x4,x5)?5
Degrees of freedom of the vector x - mean(x)?

Concept

You may have noticed that we see a value named `df` in our test results.
Do you know what degrees of freedom are?
Lets try with an example:
Degrees of freedom of a vector x(x1,x2,x3,x4,x5)?5
Degrees of freedom of the vector x - mean(x)?4

Concept

You may have noticed that we see a value named `df` in our test results.
Do you know what degrees of freedom are?
Lets try with an example:
Degrees of freedom of a vector x(x1,x2,x3,x4,x5)?5
Degrees of freedom of the vector x - mean(x)?4

Definition: degrees of freedom of a sample = the sample size minus the number of parameters estimated from the sample.