Introduction to R - part 4

Tiago Nardi

University of Pavia

Statistical analysis - the basics

Statistical analysis is the science of collecting, exploring and presenting large amounts of data to discover underlying patterns and trends

Statistical analysis can be broken down into five discrete steps, as follows:

1. Describe the nature of the data to be analyzed
2. Explore the relation of the data to the underlying population
3. Create a model to summarize understanding of how the data relates to the underlying population
4. Prove (or disprove) the validity of the model
5. Employ predective analyses to run scenarios that will help guide future actions

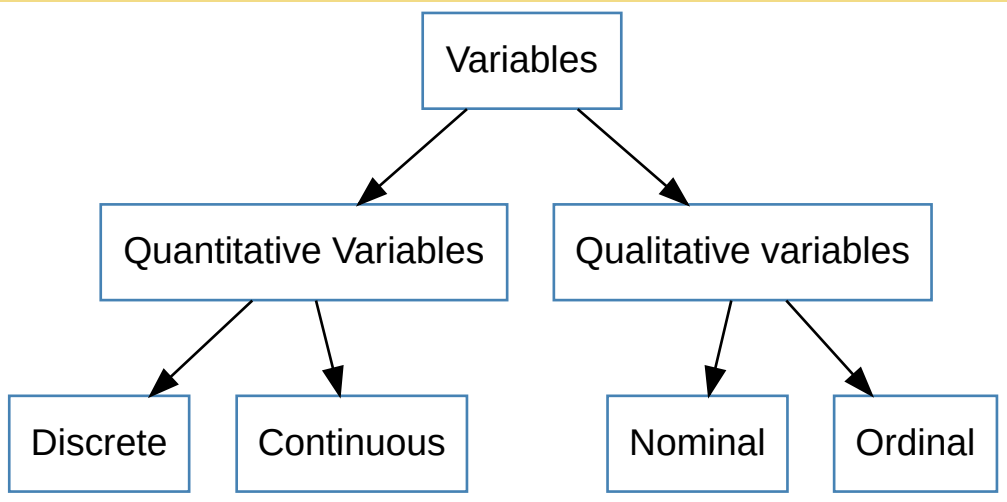
Statistical analysis - the basics

Variable: characteristic detected on each statistical unit, which can take different values in the different statistical units

Observation: value assumed by a variable in a given statistical unit

Statistical_unit	Variable
Individual1	a
Individual2	b
Individual3	c

Variables classification



Variables classification

Quantitative variables

expressed by numbers

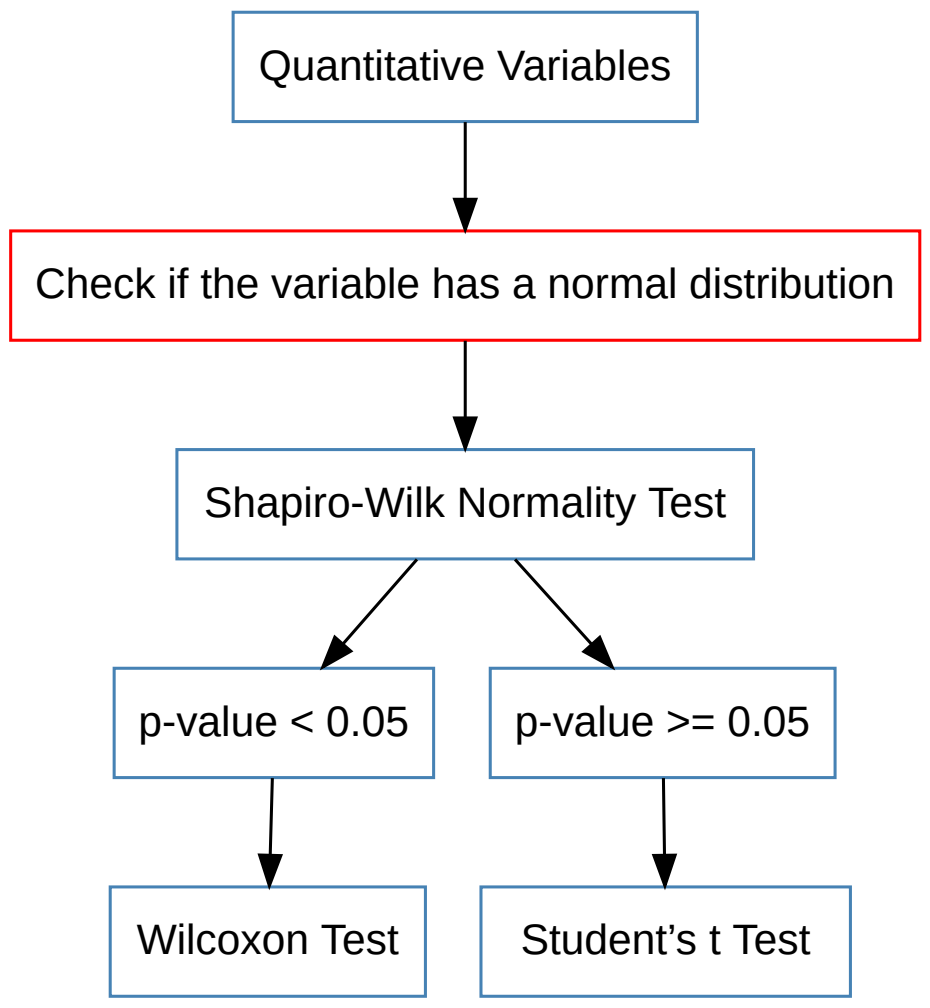
- Discrete: can only be particular values within a certain interval (often derived from counts). They are natural numbers
  - Number of colonies per plate
  - Number of chromosomes of a species
- Continuous: can be any value within a given range (they come from measurements). They are real numbers
  - Weight
  - Height

Variables classification

Qualitative variables

- **Nominal are characterized by different modes that can not be ordered:**
  - Sex: male / female
  - Survival: alive / dead
  - Blood groups: A, B, AB, 0
- **Ordinal are characterized by different modes that can be ordered:**
  - Obesity levels: overweight, obesity I, obesity II, obesity III
  - Intensity of reaction to an antigen: zero, medium, high
  - Educational qualification: compulsory school, bachelor, master, PhD

Analysis flowchart



Statistical analysis - the basics

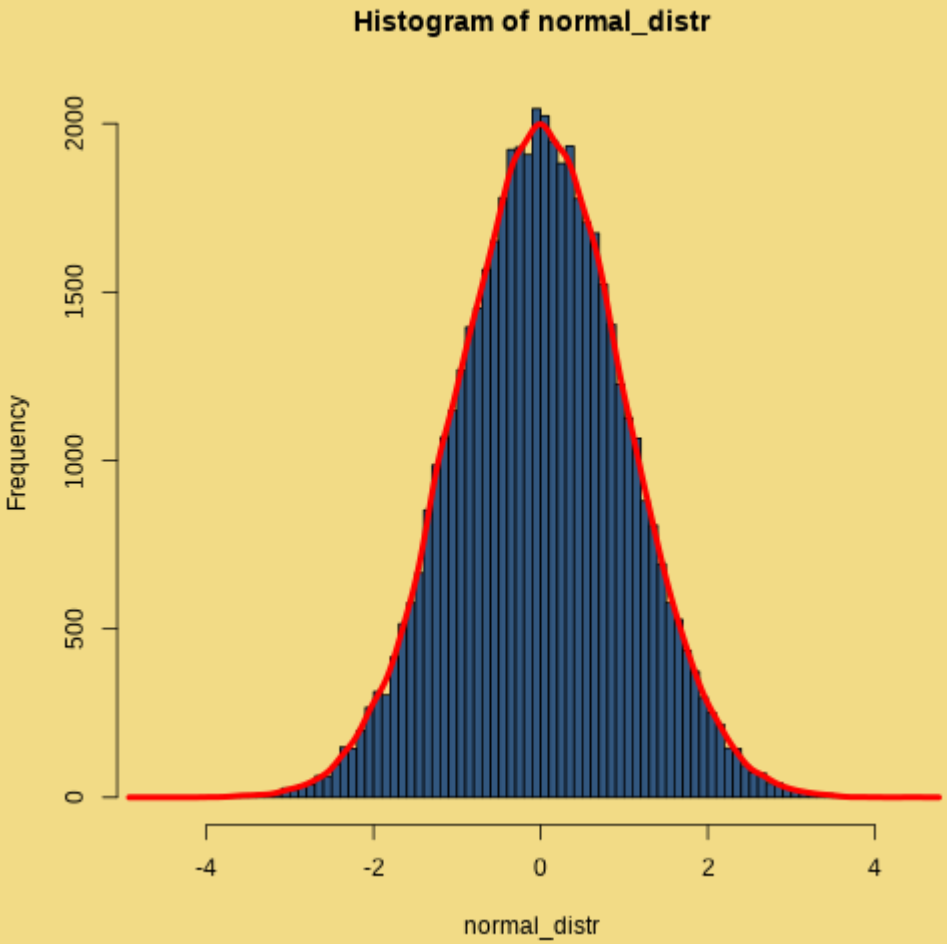
What is a normal distribution?

The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side

The area under the normal distribution curve represents probability and the total area under the curve sums to one.

Most of the continuous data values in a normal distribution tend to cluster around the mean, and the further a value is from the mean, the less likely it is to occur

Normal distribution

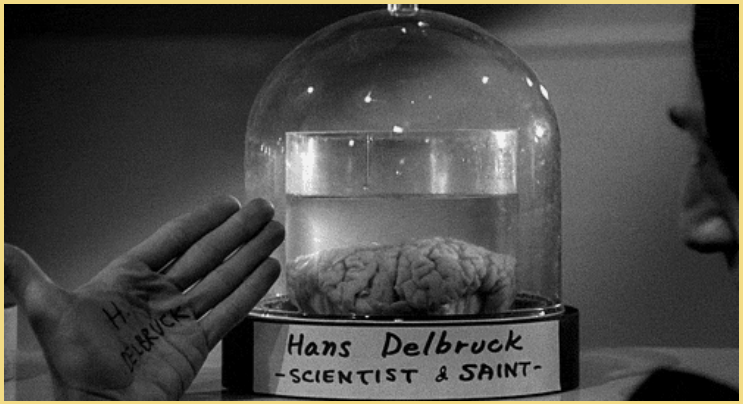


Testing the normality

To the test the normality we use the Shapiro-Wilk test

```
shapiro.test(numeric_vector)
```

- The null hypothesis for this test is that the data are normally distributed
- The alternative hypothesis for this test is that the data are not normally distributed

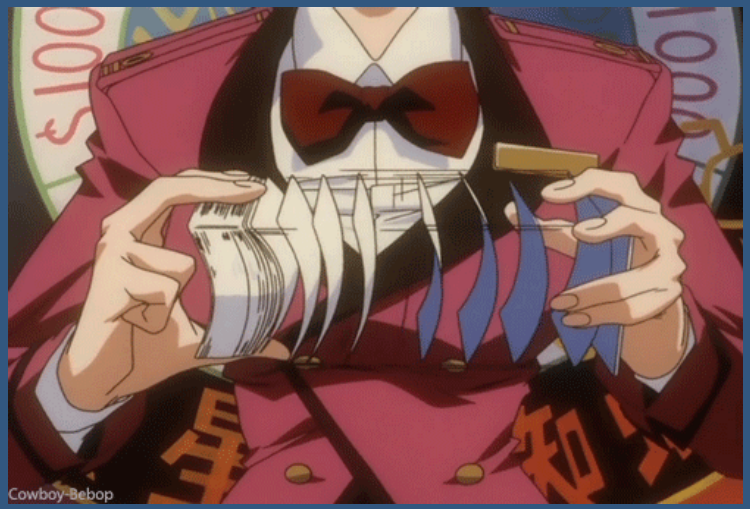


Testing the normality

- If the **p-value** is less than **0.05** (5%), we rejected the null hypothesis and we assume the variable does not have a **normal distribution**
- If the **p-value** is more than **0.05** (5%), we assume the null hypothesis is true and we assume that the variable has a **normal distribution**

Important Reminder

The p-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct



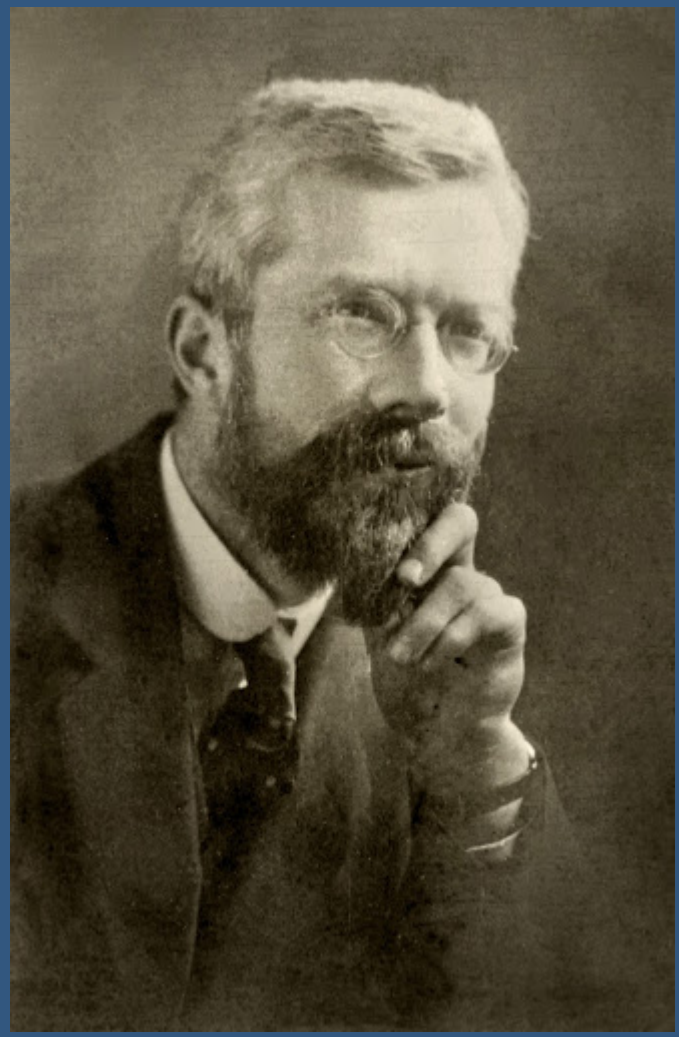
Other Reminder

The 0.05 (1/20) is a commonly used threshold, often considered adeguate. It's used a standard value, but there is no inherent reasons to prefer this specific value

In RA Fisher own words

"If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty or one in a hundred. Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fails to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance"

Fisher, R. A. 1926. The arrangement of field experiments. Journal of the Ministry of Agriculture. 33, pp. 503-515



Have a try

Load the usual tableand see if the *Contigs* variable has a normal distribution

```
db <- read.csv("patric_redux.csv")
shapiro.test(vector_name)
```

Shapiro test

```
db <- read.csv("patric_redux.csv")
shapiro.test(db$Contigs)
```

```
##
##     Shapiro-Wilk normality test
##
## data:  db$Contigs
## W = 0.91364, p-value = 0.0005409
```

p-value lesser than 0.05

We assume that the distribution is not normal

This information is used to decide which test to use for hypothesis testing

Hypothesis testing

We test if the **mean** of two (or more) values are **significantly different**

- The null hypothesis for this test is that the two mean values are not statistically different
- The alternative hypothesis for this test is that the two mean values are statistically different

- If the **p-value** is less than **0.05** (5%), we rejected the null hypothesis and we assume the two mean values are **significantly different**

- If the **p-value** is more than **0.05** (5%), we assume the null hypothesis is true and we assume that the two mean values are not **significantly different**

Hypothesis testing

When the sample is normally distributed we use the T test

When the sample is not normally distributed we use the Wilcoxon test

Wilcoxon test measures medians, not means

```
wilcox.test(first_numeric_vector, second_numeric_vector)

t.test(first_numeric_vector, second_numeric_vector)
```

Exercise

We want to check if the samples with Source from *Hobbit* have a different mean of *Contigs* compared to sample with a Source from *Human*

- Use the *patric_redux.csv* dataset and use subset to have the *Contigs* for the two categories
- Check if the *Contigs* variable has a normal distribution and test the hypothesis with the correct test

```
db <- read.csv("patric_redux.csv")
shapiro.test(numeric_vector)
wilcox.test(first_numeric_vector, second_numeric_vector)
t.test(first_numeric_vector, second_numeric_vector)
```

Results

```
db <- read.csv("patric_redux.csv")
hobbit <- subset(db,db$Source=="Hobbit")
human <- subset(db,db$Source=="Human")
shapiro.test(hobbit$Contigs)
```

```
##
##     Shapiro-Wilk normality test
##
## data:  hobbit$Contigs
## W = 0.93522, p-value = 0.2946
```

```
shapiro.test(human$Contigs)
```

```
##
##     Shapiro-Wilk normality test
##
## data:  human$Contigs
## W = 0.83538, p-value = 0.0007502
```

Results

As one of them is not normal we use the Wilcoxon

```
wilcox.test(hobbit$Contigs,human$Contigs)
```

```
## Warning in wilcox.test.default(hobbit$Contigs, human$Contigs): cannot compu
## exact p-value with ties

##
##     Wilcoxon rank sum test with continuity correction
##
## data:  hobbit$Contigs and human$Contigs
## W = 340.5, p-value = 0.0006272
## alternative hypothesis: true location shift is not equal to 0
```

Chi squared test

Test whether there is an association between categorical variables

Chi-square test

Chi-square test requires a contingency table

Contingency table (two-way table): data is classified with two categorical variables. In the rows we have the categories for one variable, and in the columns the categories for the other variable.

Each variable must have two or more categories.

Each cell reflects the total count of cases for a specific pair of categories

```
db_red <- subset(db,(db$Isolation_location=="The Shire" |
                       db$Isolation_location=="Gondor"))
cont_db <- table(db_red$Isolation_location, db_red$Source)
cont_db
```

```
##
##              Hobbit Human
##   Gondor          1    10
##   The Shire      12     1
```

Chi-squared test

- The null hypothesis for this test is that the categorical variables are independent (no correlation)
- The alternative hypothesis for this test is that the categorical variables are correlated

```
chisq.test(cont_db)
```

```
##
##     Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cont_db
## X-squared = 13.437, df = 1, p-value = 0.0002467
```

Warning message as the dataset is limited

Exercise

Using the subsetted database (only from *The Shire* and *Gondor*), we want to check if there is an association between the *Species* of the bacteria and the *Source* of the isolate

- Prepare a contingency table
- Check the hypothesis Chi-squared

```
subsetted_table <- subset(table,logical_condition)
contigency_table <- table(var1,var2)
chisq.test(contigency_table)
```

Results

```
db_red <- subset(db,(db$Isolation_location=="The Shire" |
                       db$Isolation_location=="Gondor"))
cont_db <- table(db_red$Species, db_red$Source)
cont_db
```

```
##
##                            Hobbit Human
##   Yersinia enterocolitica       3     9
##   Yersinia pestis              10     2
```

```
chisq.test(cont_db)
```

```
##
##     Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cont_db
## X-squared = 6.042, df = 1, p-value = 0.01397
```

So there is a significant association between these two categories