Hands-on Biostatistics 2

Analysis of non-parametric data in R

Eliana Ibrahimi

Department of Biology, University of Tirana

eliana.ibrahimi@fshn.edu.al

July 15-16, 2021

Outline

- Introduction
- Parametric vs. Nonparametric methods
- Two samples nonparametric tests
- Three or more samples nonparametric tests
- Spearman correlation

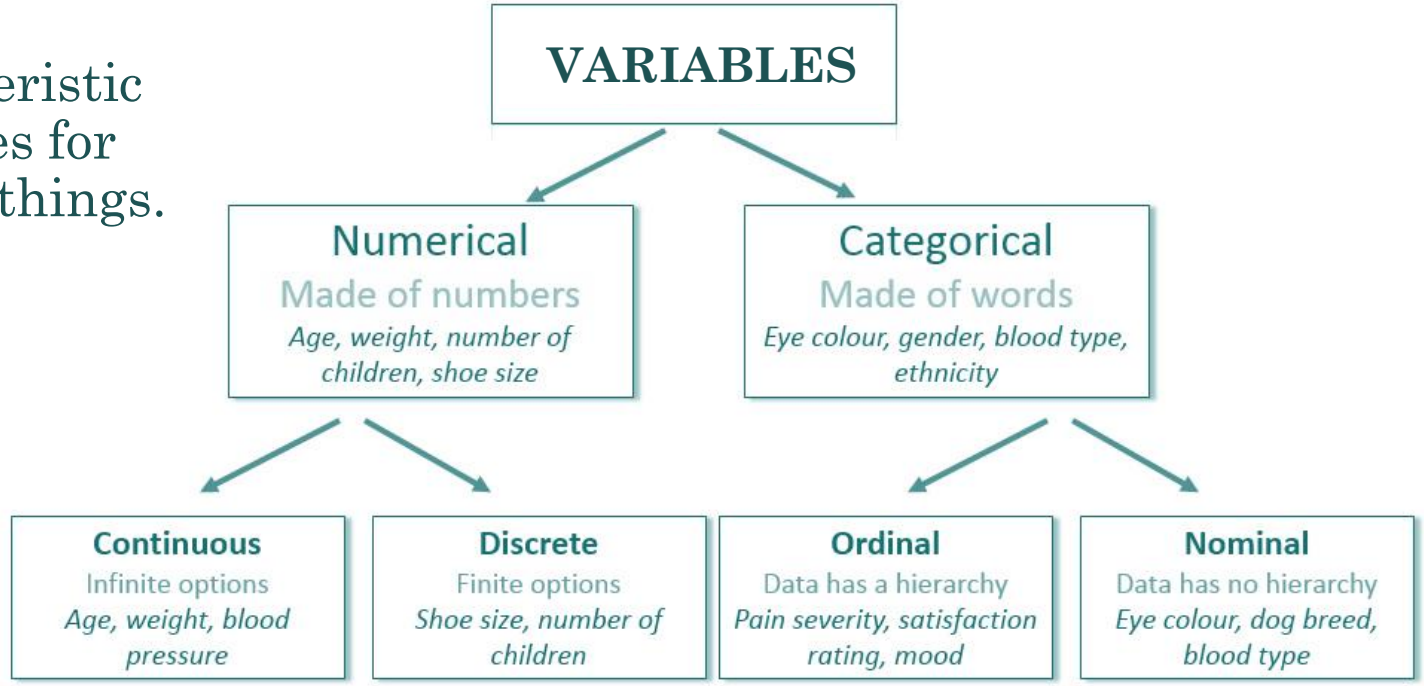
Introduction

Statistical thinking will one day be as necessary for citizenship as the ability to read and write.

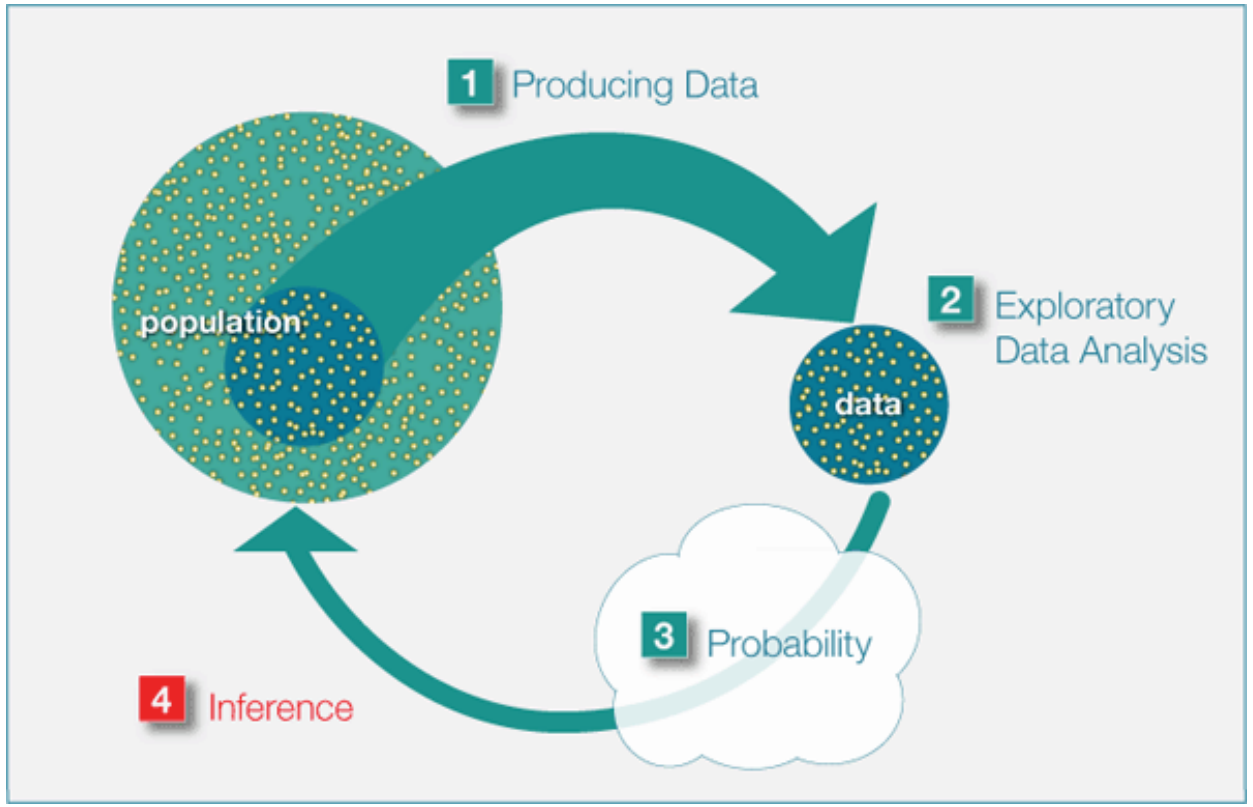
- Herbert George "H. G." Wells -
(1866 – 1946, English writer)

Variables in biostatistics

- A random variable is a characteristic that can take on different values for different individuals, places or things.



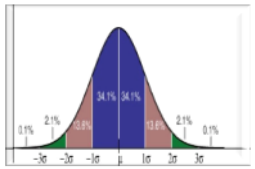
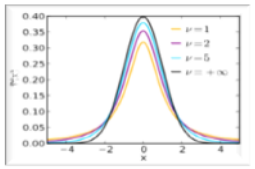
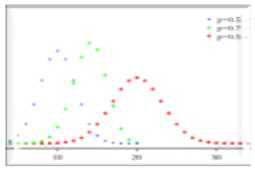
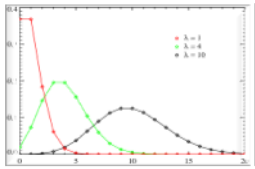
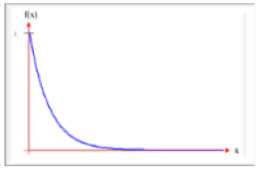
Statistical inference



Source: UF-Open learning

- Statistical inference: draw conclusions about a population based on the data obtained from a sample chosen from it.
  - Point Estimation
  - Interval estimation
  - Hypotheses testing

Probability distributions

	Normal Distribution	Student's T-Distribution	Binomial Distribution	Poisson Distribution	Exponential Distribution
What does it look like?					
Defining Characteristics	Distinctive Bell Shape	Shorter, fatter than the normal distribution.	Two outcomes: Success/Failure	Various shapes, but valid only for integers on the x-axis.	Models Time Between Events
Example of When to Use It	Modeling natural phenomena (height, weight, IQ, test scores etc.)	When you have small samples or don't know the population variance (σ^2).	Coin Toss Probability (Heads, Tails)	Gives probability of number of events in a fixed interval.	"How much time will go by before a major hurricane hits the Atlantic Seaboard?"
Example of DS Application	Least squares fitting or propagation of uncertainty.	Unknown σ^2 is common in real life data, you you'll have to use the T instead of the normal in that case.	Anywhere where binary (yes/no, black/white, vote/don't vote) data is used.	Anywhere there is a waiting time between events.	Building continuous-time Markov chains.

- To fully understand the theory behind statistical inference you will need some concepts related to probability distributions.
- We will not focus on this, but:
  - If you haven't taken any course in statistical theory during your studies, I recommend you check this great book by Rosner (2016).

Source: Data Science Central

Statistical hypothesis testing

- Assessing evidence provided by the data against the null hypothesis.

Step 1	Step 2	Step 3	Step 4
• ***Formulate the hypotheses:*** H_0 – null H_1- alternative	• Collect relevant data and summarize them.	• Test how likely it is to observe data we obtained, if null hypotheses is true. ***Compute test statistics***	• ***Compute p-value*** and make our decision.

Parametric vs. nonparametric methods

- Require assumptions about the distribution in the population

Parametric

- Distribution free, called **exact tests** due to the fact that their methods of calculating p-values require no mathematical approximation.

Nonparametric

- Note that when the assumptions are precisely satisfied, some "parametric" tests can also be considered "exact."

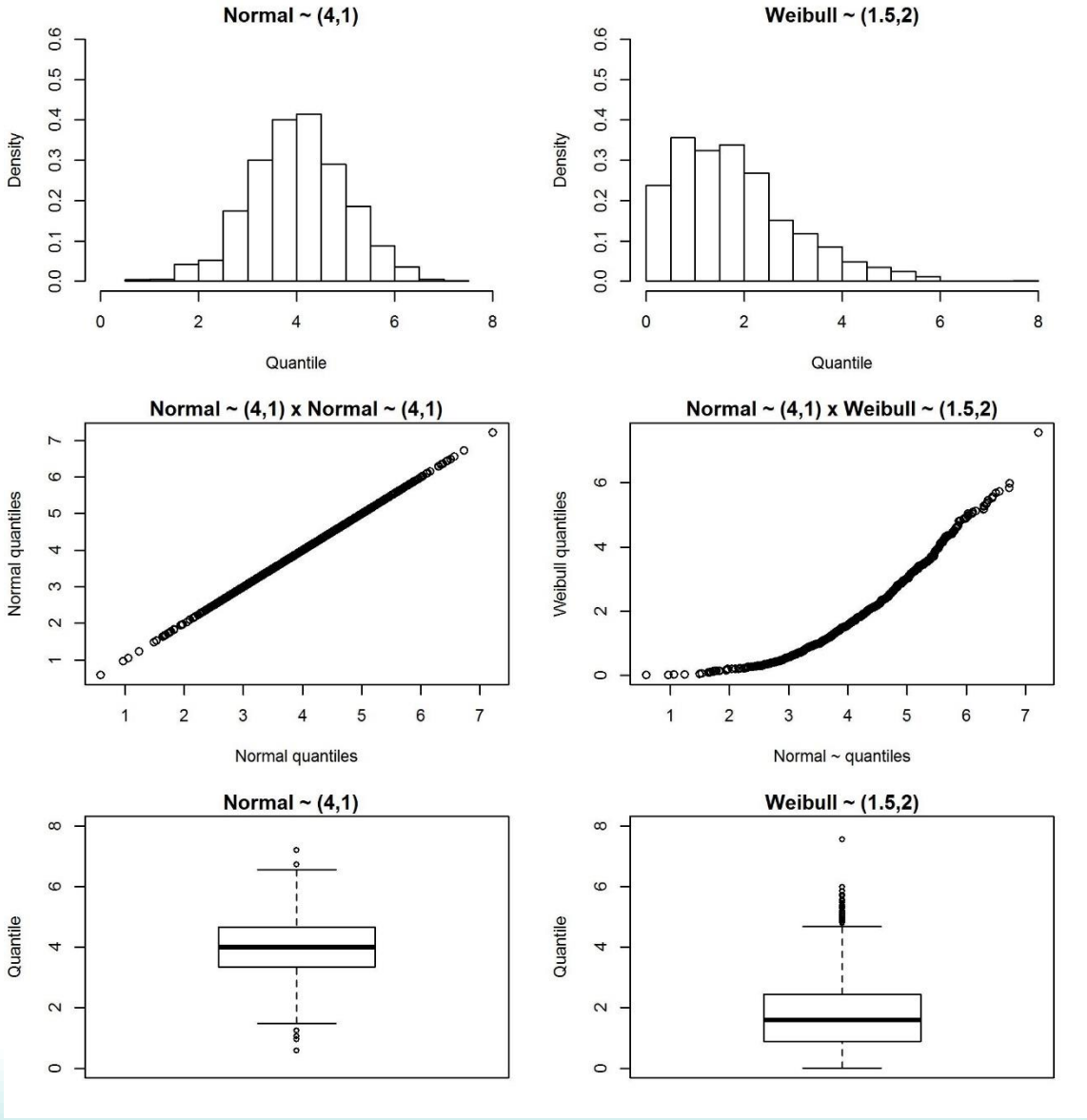
Exact tests

When and which nonparametric test to use?

	Data	
Objective	**Ordinal/Skewed**	**Continuous/ratio**
Compare two related samples	Wilcoxon Signed rank test	Paired t-test
Compare two independent samples	Mann-Whitney Test	Independent t-test
Compare more than two independent samples	Kruskal-Wallis Test	ANOVA
Discover association	Spearman rank correlation	Pearson correlation

How to test for normality?

- Graphically
  - Histogram
  - QQ plots
  - Boxplot
- Formal tests
  - Kolmogorov Smirnov
  - Shapiro-Wilk



Shapiro-Wilk test

1. Hypothesis

H_0: The population is normally distributed

H_1: the population is not normally distributed

2. Compute the test statistics (in R)

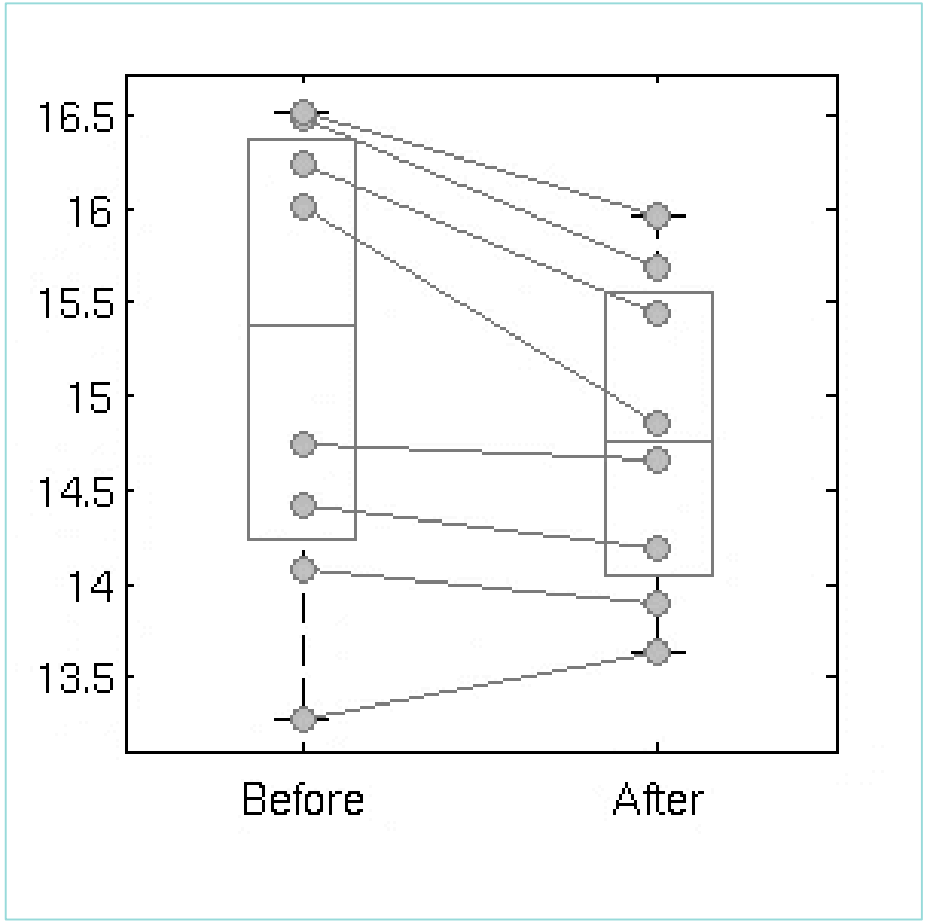
$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2},$$

See details on the formula here.

3. Decision based on p-value

If p<0.05 reject the null hypothesis (Ho)
The population is not normally distributed.

R code:
shapiro.test(data$variable)



Two related samples

Wilcoxon signed rank test

1. Hypothesis

   H_0: difference between the pairs follows a symmetric distribution around zero.

   H_1: difference between the pairs does not follow a symmetric distribution around zero.

2. Compute the test statistics (in R)

$$W = \sum_{i=1}^{N_r} [\operatorname{sgn}(x_{2,i} - x_{1,i}) \cdot R_i]$$

See details on the formula here

3. Decision based on p-value

If p<0.05 reject the null hypothesis (Ho)
If the difference is not symmetric around zero then there is difference between groups.

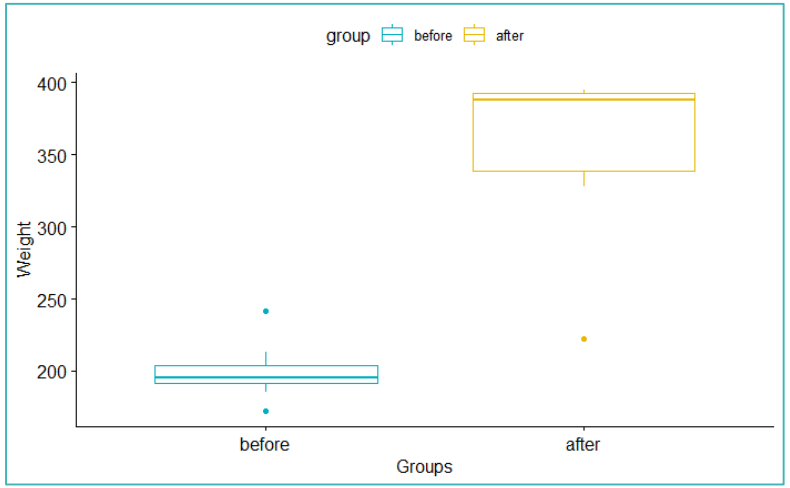
```
#In order to run the Wilcoxon signed rank test in R, use the code:

wilcoxon.test (variable1, variable2, Paired=TRUE, exact=FALSE)
```

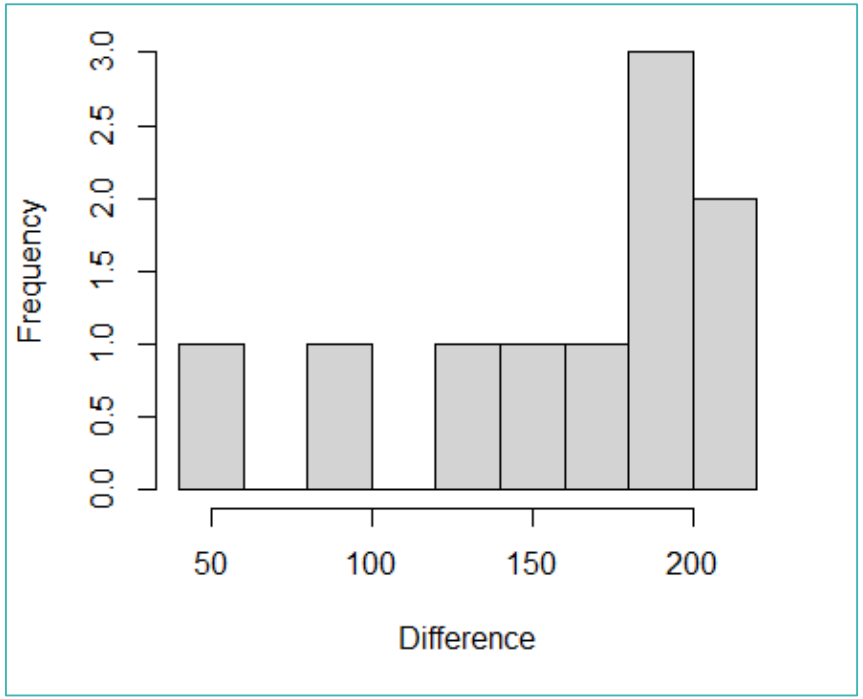
Dependent: Continuous or ordinal
Independent: Time/condition

Example 1

Data: We'll use an example data set, which contains the weight of 10 mice before and after a specific treatment.

Research question: Is there a difference between the mice mean weight before and after the treatment?





The differences:

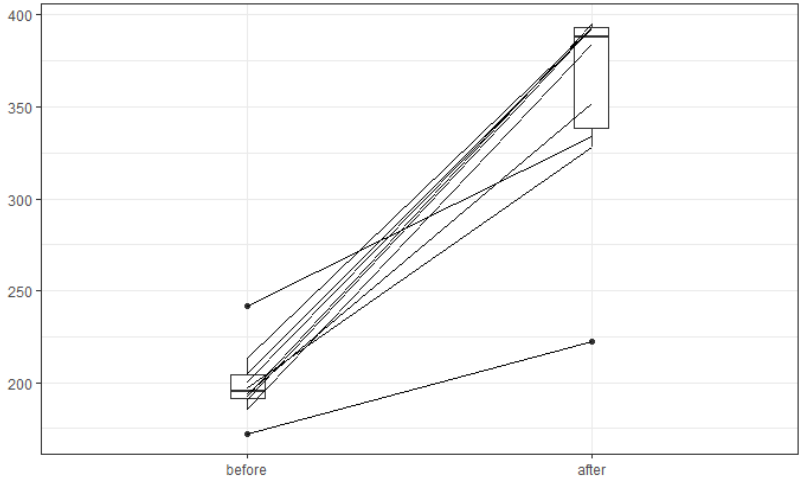
```
        Shapiro-Wilk normality test

data:  diff
W = 0.81975, p-value = 0.02516
```

➢ *See HOB_2 R Notebook for code and details*

Wilcoxon signed rank test: Example 1

#In order to run the Wilcoxon signed rank test in R, use the code:

wilcoxon.test (weight_after, weight_before, Paired=TRUE, exact=FALSE)

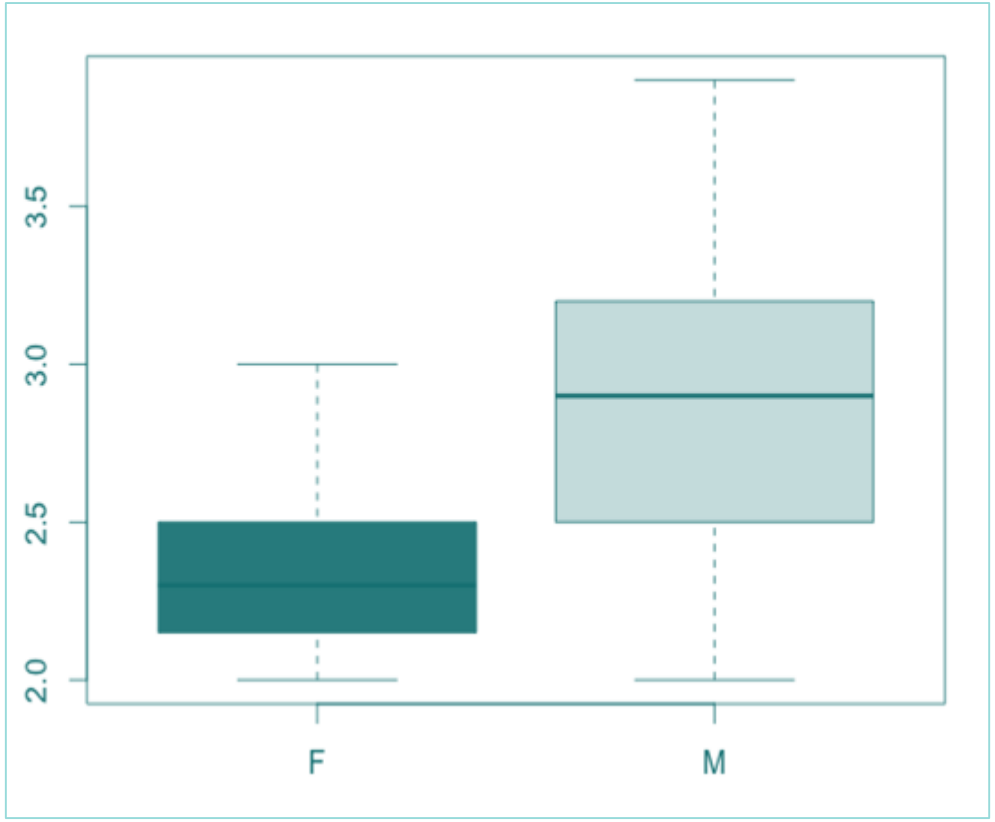


Output:

```
        Wilcoxon signed rank test with continuity correction

data:  weight_after and weight_before
V = 55, p-value = 0.005922
alternative hypothesis: true location shift is not equal to 0
```

Documentation

A Wilcoxon signed rank test showed that there was a significant difference (V=55, p=0.005) between weight before and after the treatment. The median weight after the treatment was 392.95 g compared to the baseline median weight of 195.3 g. Therefore, the scientist should start using the new treatment.



Two independent samples

Mann Whitney U test

1. Hypothesis

H_0: the distributions of both populations are equal

H_1: the distributions are not equal

2. Compute the test statistics (in R)

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i, Y_j),$$

See details on the formula here

3. Decision based on p-value

If p<0.05 reject the null hypothesis (Ho)

#In order to run the Mann Whitney test in R, use the code:

wilcox.test(dependent~independent)

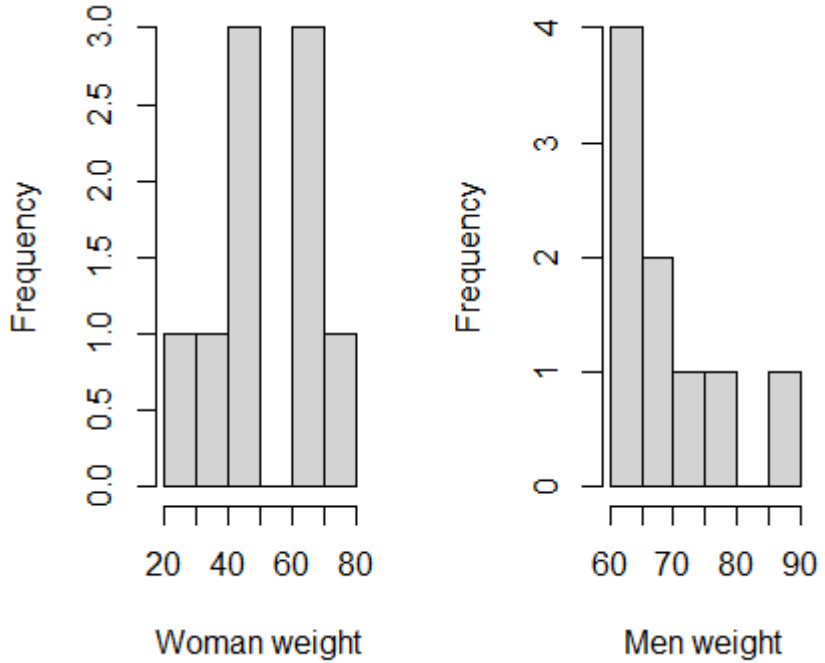
Dependent variable:
Numerical/continuous (skewed) or ordinal
Independent: Nominal (binary)

Example 2

Data: We'll use an example data set, which contains the weight of 18 individuals (9 women and 9 men).

Research question: Is there a difference between the mean weight for the woman and men?

Dependent variable: Weight
Independent: group/gender




```
        Shapiro-Wilk normality test

data:  data$weight[group == "Woman"]
W = 0.94266, p-value = 0.6101
```

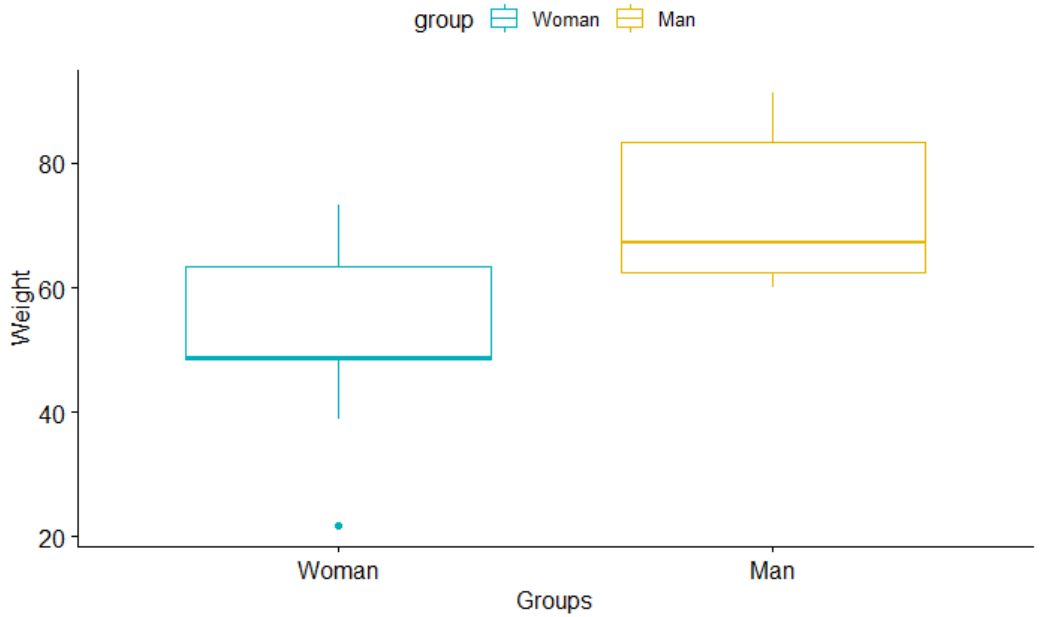
```
        Shapiro-Wilk normality test

data:  data$weight[group == "Man"]
W = 0.81403, p-value = 0.0295
```

Mann Whitney test: Example 2

#In order to run the Mann Whitney test in R, use the code:

wilcox.test(weight~group)

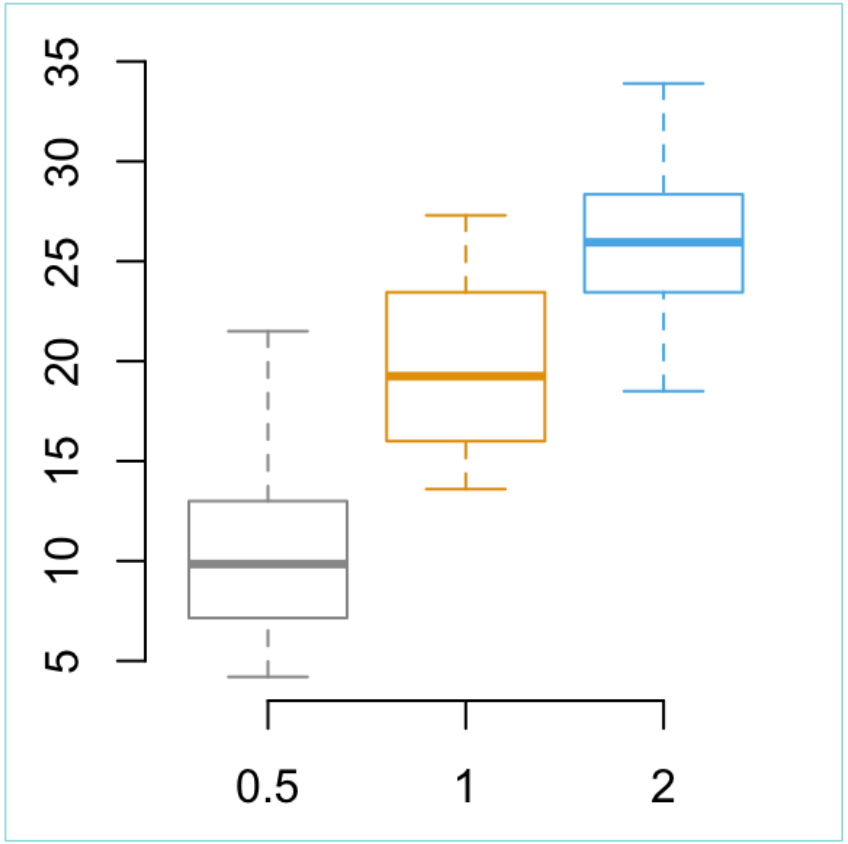
```
        Wilcoxon rank sum test with continuity correction

data:  weight by group
W = 65.5, p-value = 0.03042
alternative hypothesis: true location shift is not equal to 0
```



Documentation:

A Mann-Whitney U test showed that there was a significant difference (W= 65.5, p = 0.03) between the weights for the man compared to woman. The median weight was 67.3 for man compared to 48.8 for woman.

➢ *See HOB_2 R Notebook for details*



Two or more independent samples

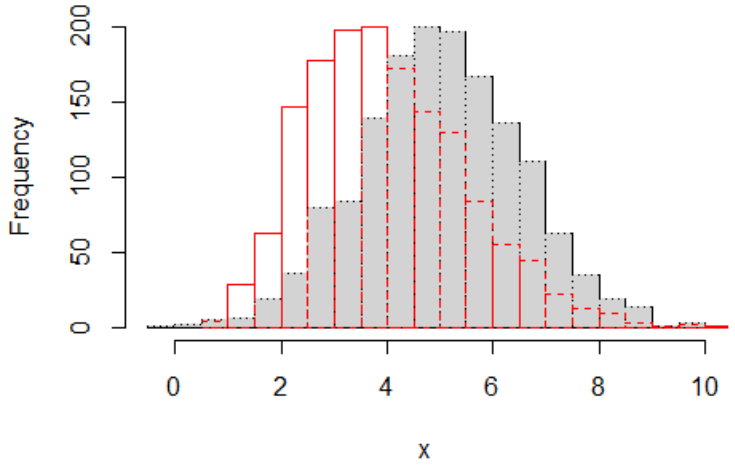
Kruskal-Wallis test

1. Hypothesis

   Under the assumption of an identically shaped and scaled distribution for all groups,

   Ho: the medians of all groups are equal, and

   H_1: at least one population median of one group is different from the population median of at least one other group.



2. Compute the test statistics (in R)

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2};$$

See details for the test statistics here

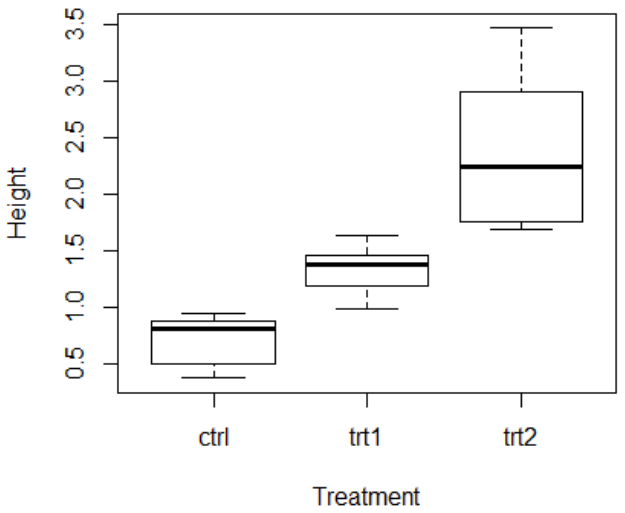
3. Decision based on p-value

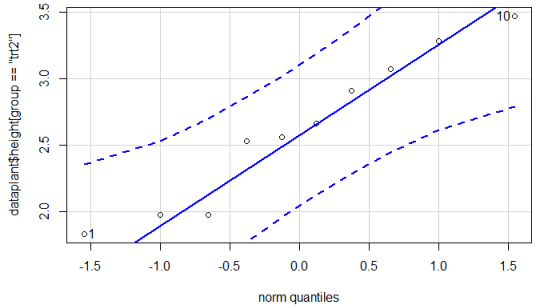
If p<0.05 reject the null hypothesis (Ho)

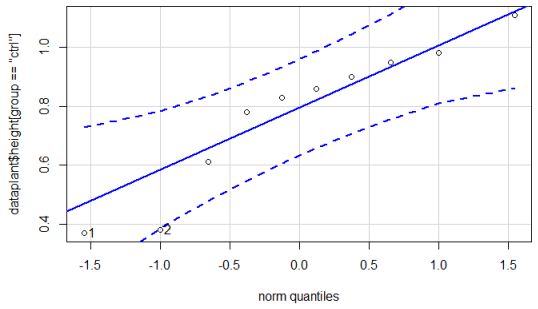
Example 3

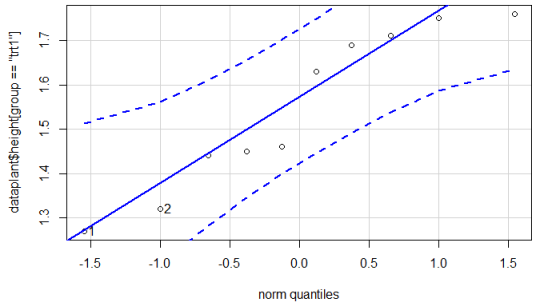
Data: Here, we'll use the data set named *PlantGrowth*. It contains the height of plants (cm) obtained under a control and two different treatment conditions.

Research question: Is there a difference between the mean height for the treatments? Or, which is the best treatment?









- *See HOB_2 R Notebook for details*

Kruskal-Wallis test: Example 3

#In order to run the Mann Whitney test in R, use the code:

kruskal.test(height ~ group, data = dataplant)

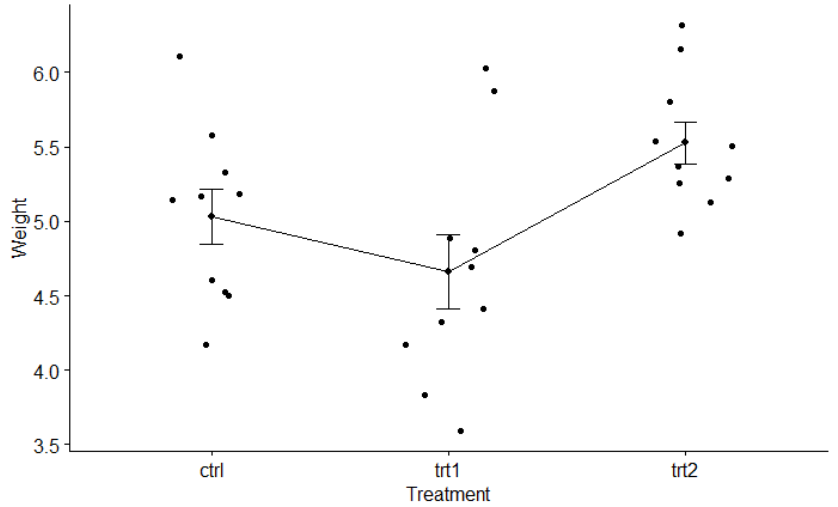
As the p-value < 0.001, there is very strong evidence to suggest a difference between at least one pair of groups but which pairs? To find out produce pairwise Wilcoxon signed rank comparisons for each pair of groups.

```
        Kruskal-Wallis rank sum test

data:  height by group
Kruskal-Wallis chi-squared = 24.986, df = 2, p-value = 3.752e-06
```

- *See HOB_2 R Notebook for details*

Pairwise Wilcoxon signed rank test: Example 3

```
#In order to run the Mann Whitney test in R, use the code:

pairwise.wilcox.test(dataplant$height,dataplant$group,p.adj='bonferroni',exact=F)
```



Output:

```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  dataplant$height and dataplant$group

     ctrl    trt1
trt1 0.00282 -
trt2 0.00045 0.00045

P value adjustment method: bonferroni
```

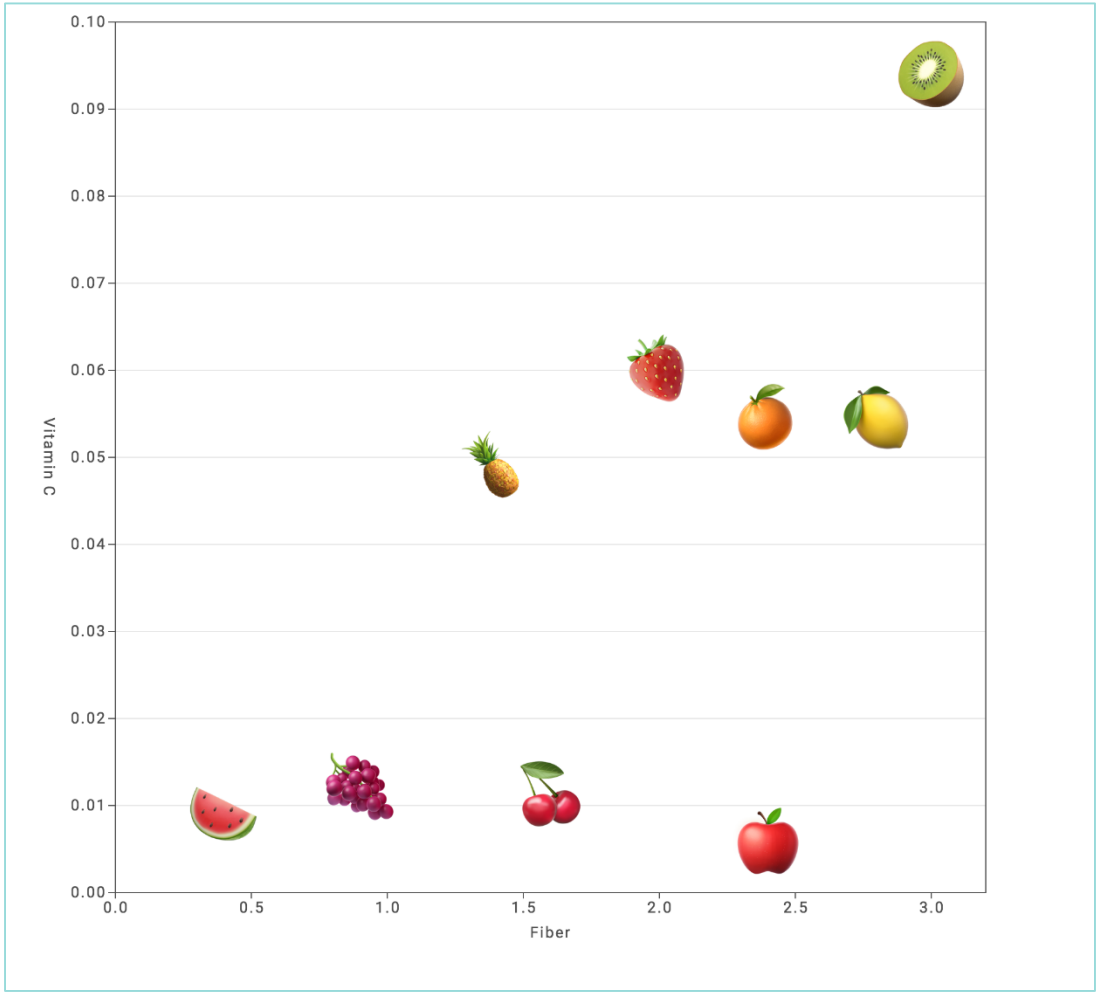
➢ *See HOB_2 R Notebook for details*

Pairwise Wilcoxon signed rank test: Example 3

Reporting the results:

A Kruskal-Wallis test was carried out to compare plant height after two treatments and control (no treatment). There was very strong evidence of a difference (p-value < 0.001) between the mean ranks of at least one pair of groups. Wilcoxon signed rank pairwise tests were carried out for the three pairs of groups. There was strong evidence (p-value < 0.05, adjusted using the Bonferroni correction) of all the differences between the groups.

Treatment 2 was the most efficient treatment for plant development.



Discover association

Spearman correlation

1. Let ρ be the Spearman's population correlation coefficient, then we can express this test as:

   H_0: $\rho = 0$

   H_1: $\rho \neq 0$

2. Compute the Spearman correlation coefficient and p-value in R.

$$rho = \frac{\sum(x' - m_{x'})(y'_i - m_{y'})}{\sqrt{\sum(x' - m_{x'})^2 \sum(y' - m_{y'})^2}}$$

Where $x' = rank(x)$ and $y' = rank(y)$.

3. Decision based on p-value

If p<0.05 reject the null hypothesis (Ho) There is an association between variables.

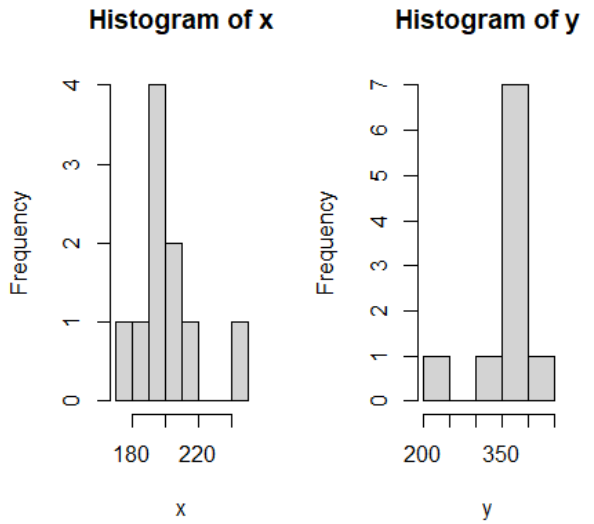
0-0.19 "very weak"
0.20-0.39 "weak"
0.40-0.59 "moderate"
0.60-0.79 "strong"
0.80-1.0 "very strong"

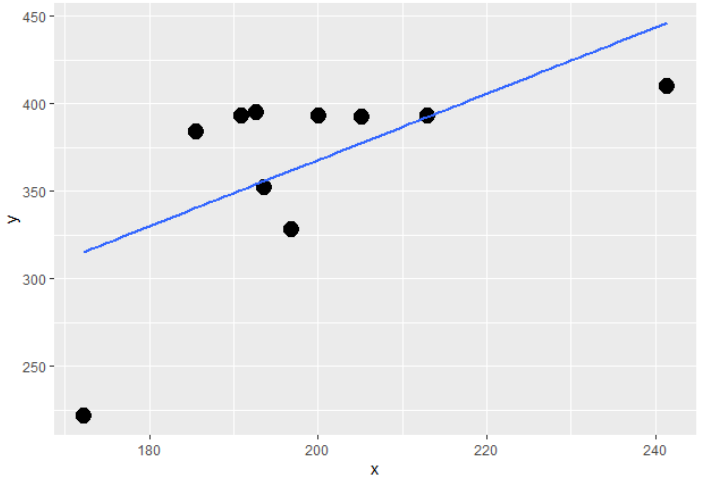
https://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf

Example 4

Data: We'll use an example data set, which contains the weight of 10 mice before and after a specific treatment.

Research question: Is there a correlation between the mice mean weight before and after the treatment?





Note: Spearman's correlation coefficient is a statistical measure of the strength of a monotonic relationship between paired data. Read more on monotonic relationships here.

Spearman correlation: Example 4

#In order to calculate the Spearman correlation in R, use the code:

cor(x, y, method = c("spearman"))
cor.test(x, y, method=c("spearman"))

Output:

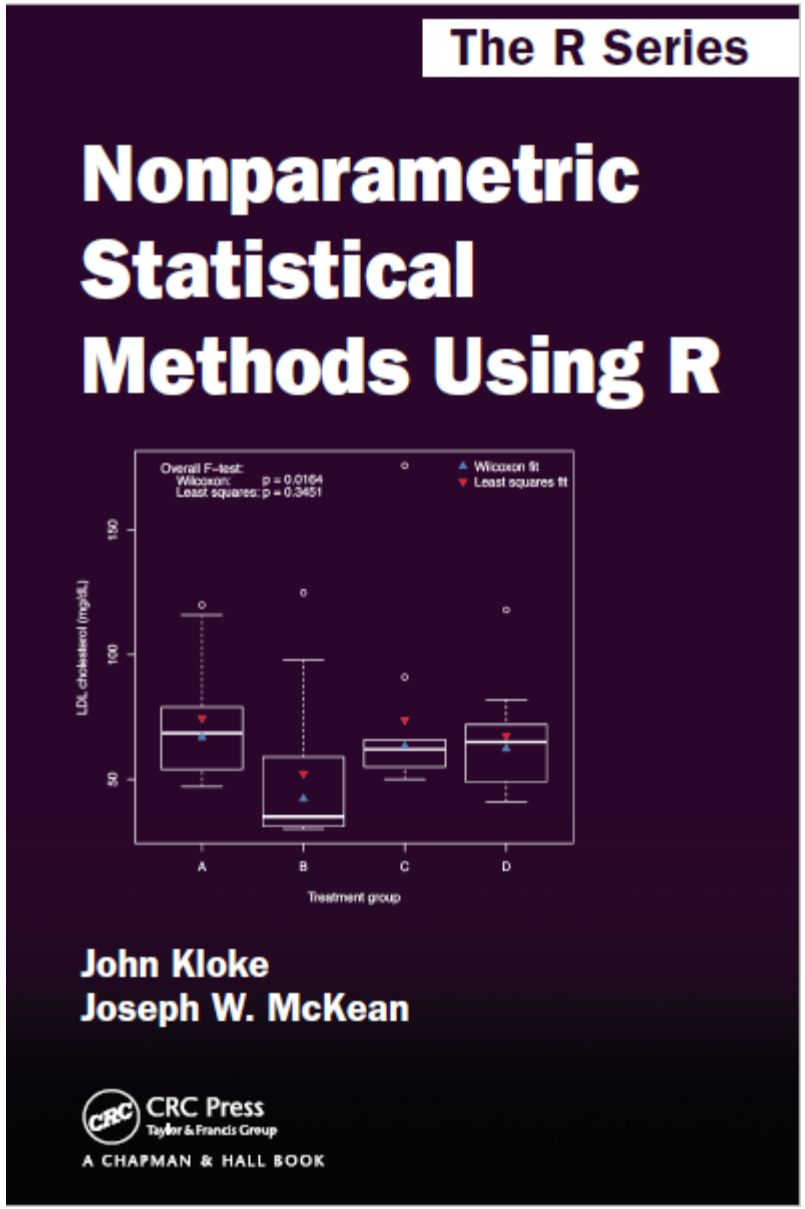
```
[1] 0.4666667

        Spearman's rank correlation rho

data:  x and y
S = 88, p-value = 0.1782
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4666667
```

This could be formally reported as follows: "A Spearman's correlation was run to determine the relationship between values of weight before and after treatment. There was no significant monotonic correlation between weight values (r_s= 0.467, n = 10, p=0.178)."

Other nonparametric methods

- Friedman test
- Adaptive Rank Scores Tests
- Aligned Rank Tests
- Nonparametric Regression
- Time to Event Analysis

Recommend book for more



Available statistical software?

The battle is …

References/Useful links

1. Rosner, Bernard. Fundamentals of Biostatistics. Cengage Learning, 2016.
2. Kloke, J., & McKean, J.W. (2014). Nonparametric Statistical Methods Using R (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b17501
3. Pezzullo, John. Biostatistics For Dummies. Wiley, 2013.
4. https://bolt.mph.ufl.edu/6050-6052
5. http://www.biostathandbook.com/HandbookBioStatThird.pdf
6. https://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf
7. http://www.statstutor.ac.uk//
8. https://data-flair.training/blogs/why-learn-r/


Any questions?