Univariate statistical analysis with R

Eliana Ibrahimi
Department of Biology
University of Tirana, Albania

MSB 2023 Training School
Tirana, Albania
October 18-20, 2023

Part 1: Introduction to Biostatistics

- What is Biostatistics and why do we need it?
- Variables in biostatistics
- Data analysis steps
- Inference statistics
  - Estimation

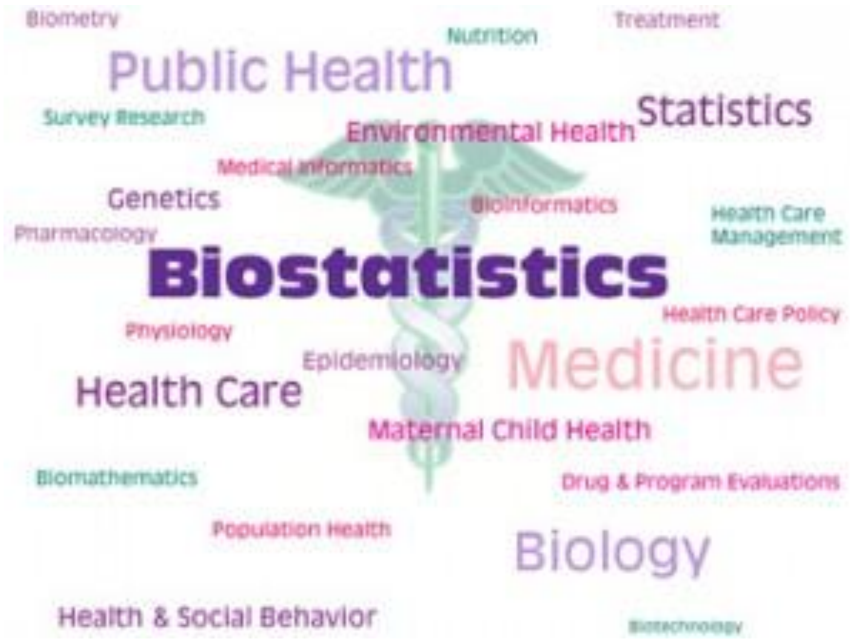
How knowledgeable are you in biostatistics and R?

- Go to wooclap.com and let me know by answering the questions.

What is Biostatistics?

Definition

- Biostatistics is a branch of applied statistics that applies statistical methods to collect, analyze, and interpret data related to biology, public health, and medicine.



--- Much more statistics than biology, however biostatisticians must learn the biology also.

Why do I need to learn biostatistics?

Three main reasons

1. To effectively conduct research
2. To be able to read scientific papers
3. To further develop critical and analytical thinking

Biostatistics in research

- A good way to learn about biostatistics and its role in the research process is to follow a research study from study design to its publication.
  - **The biostatistician should be present in each step**.

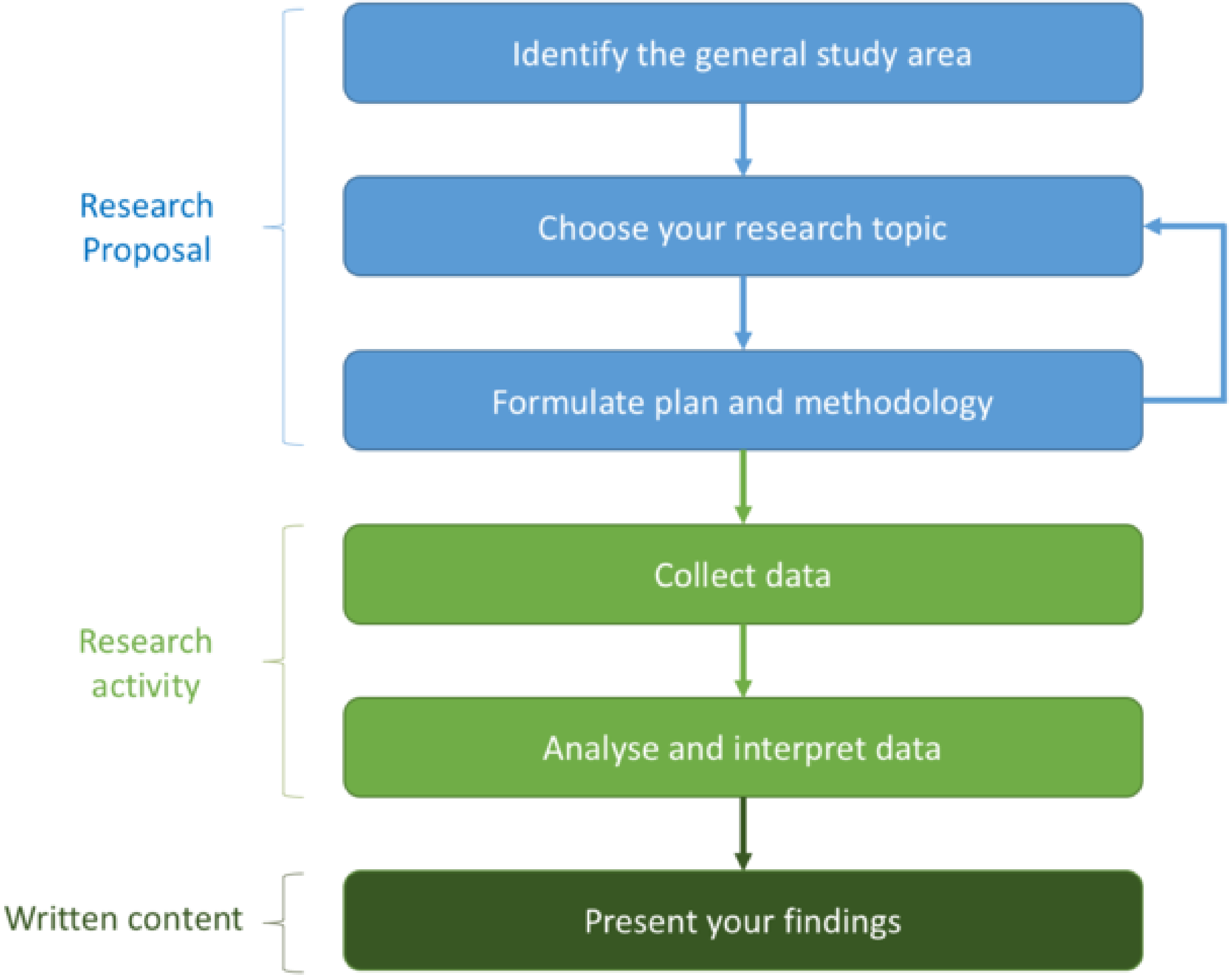
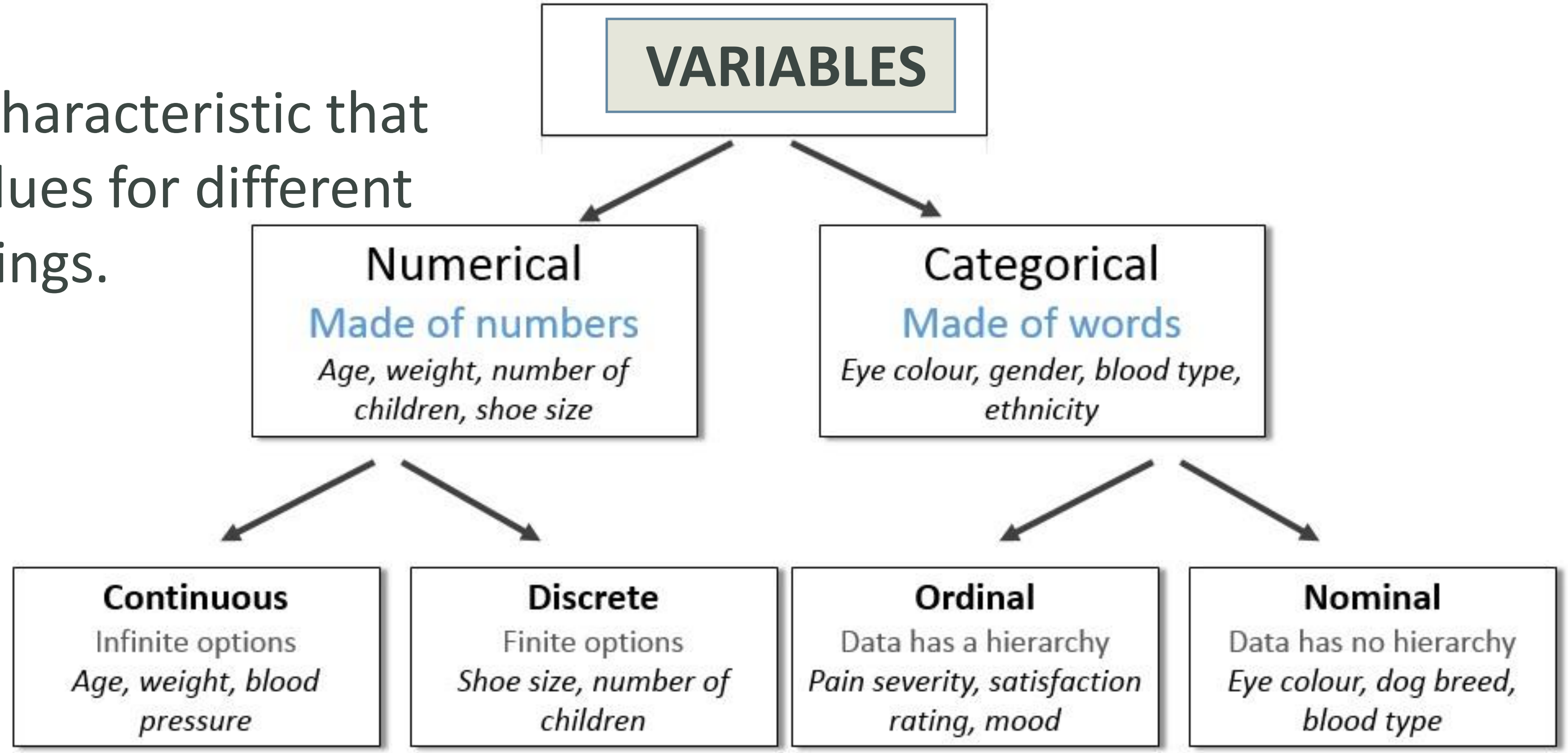


Image source: futurelearn.com

Variables in biostatistics

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



Data analysis

- According to Wikipedia, Data analysis is a process of inspecting, cleansing, transforming and modeling data to discover useful information and supporting decision-making.

Image source: vippng.com

✓ There are 5 steps in the process of data analysis

1. Data preparation (organize, transform, clean)
2. Data exploration
3. Data modeling
4. Draw conclusions
5. Communicating results

Data preparation

The collected data is generally not in the form to be analyzed.

- Data preparation includes editing, coding, data entry and is the process that ensures data accuracy and their transformation from raw to reduced and classified forms that are appropriate for analysis.
- Use excel to edit, code and tabulate your data if you are not expert in R or another programming software.
- Check this material for details.

Data preparation: Raw data

RAW DATA NP I		Number of colonies					
Treatment	**Concentration**	**dish 1**	**dish 2**	**dish 3**	**dish 4**	**dish 5**	**dish 6**
Positive Control	100µM	0	0	0	0	0	0
Control	0	122	132	120	134	123	154
Solvent Control	0.04%	152	139	132	118	148	142
I	1	145	134	144	149	138	129
I	5	137	133	143	155	141	135
I	10	129	124	135	138	146	143
I	12.5	146	113	131	138	130	145
I	15	72	75	75	82	96	101
I	20	55	28	17	77	41	10
I	25	0	0	0	0	0	0

Well organized data

- An example of well organized data

Month	Name	Gender	Diagnosis	Treatment
May	Jessica	F	Allergy	Eye Drops
May	Sam	M	Allergy	Eye Drops
May	Wes	M	Cataract	Cataract Surgery
May	Rachel	F	Ptergyium	Eye Drops
May	Lily	F	Allergy	Eye Drops
May	Hannah	F	Cataract	Cataract Surgery
May	Denise	F	Allergy	Eye Drops
May	Sharon	F	Allergy	Eye Drops
May	Robin	F	Allergy	Eye Drops
May	Lianna	F	Ptergyium	Eye Drops
May	Thomas	M	Presbyopia	Reading Glasses
May	Kimberly	F	Refractive Error	Distance Glasses
May	Michael	M	Refractive Error	Distance Glasses
May	Jacob	M	Conjunctivitis	Eye Drops
June	John	M	Presbyopia	Reading Glasses
June	Tim	M	Refractive Error	Distance Glasses
June	Allison	F	Cataract	Cataract Surgery
June	Laura	F	Ptergyium	Eye Drops
June	Scott	M	Cataract	Cataract Surgery
June	Sarah	F	Ptergyium	Eye Drops
June	Alex	M	Ptergyium	Eye Drops
June	Robert	M	Cataract	Cataract Surgery

Data exploration

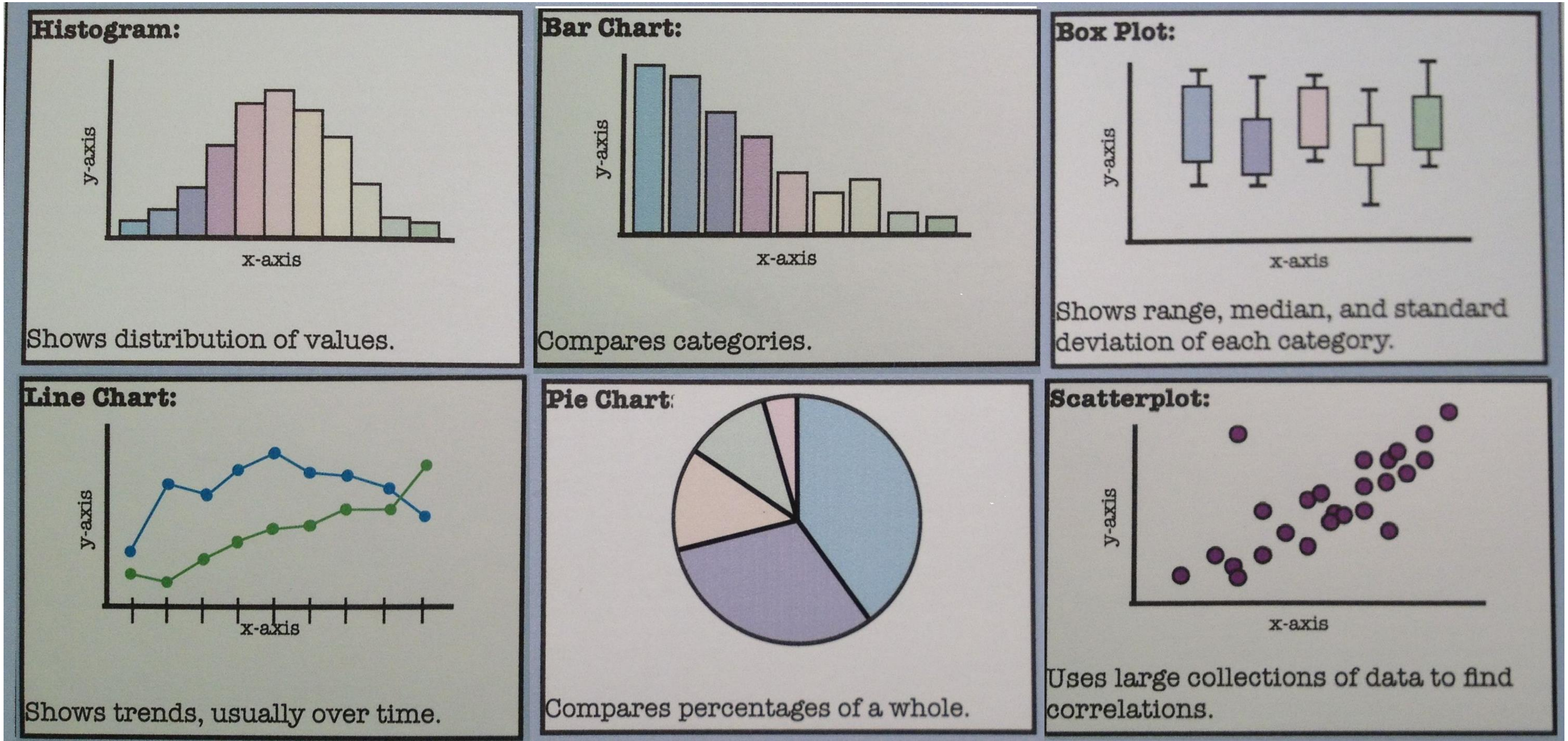
Nominal variables

- Frequency
  - Count (How frequent values occur)
  - Relative (The % of observations with a specific value)
- Graphs (simple and clustered)
  - Bar
  - Pie
  - Area
  - …

Numerical variables

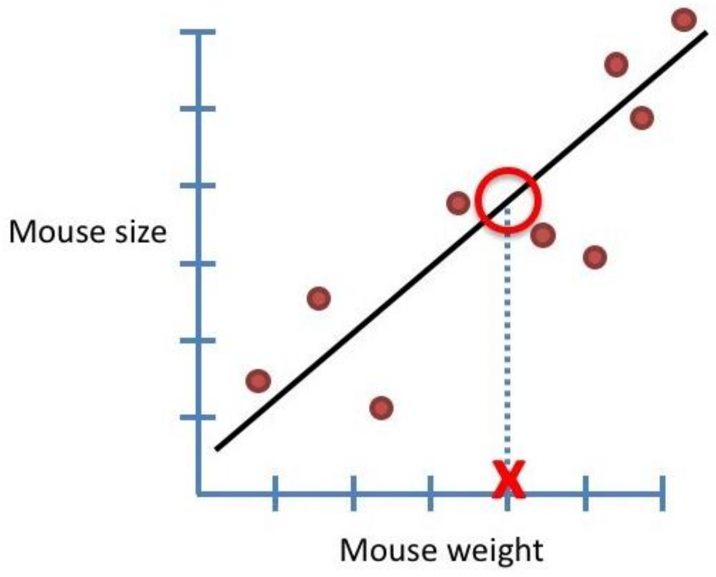
- Descriptive statistics
  - Mean, Mode, Median
  - Variance, Standard deviation, SE
  - Range, Percentiles
- Graphs (simple and clustered)
  - Histogram
  - Boxplot
  - Error bars
  - Scatterplot
  - …

Data exploration: Graphs



Statistical modeling

- “A statistical model is usually specified as a mathematical relationship between one or more random variables and other non-random variables. As such, a statistical model is "a formal representation of a theory” and represents, often in considerably idealized form, the data-generating process” (Wikipedia).
- Includes
  - T-tests
  - ANOVA
  - Linear Regression
  - General linear regression
  - And many more…



https://www.youtube.com/watch?v=yQhTtdq_y9M

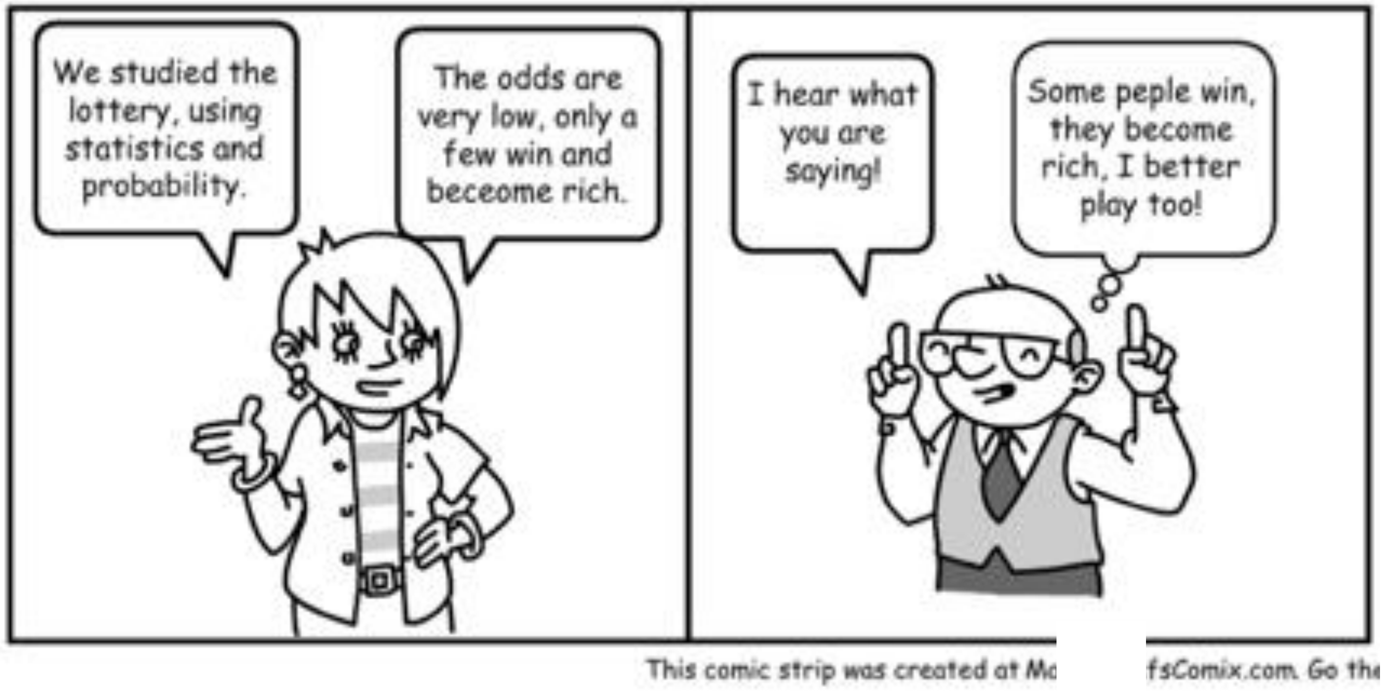
Draw conclusions

While interpreting the results ask yourself:

- Did the analysis answer my research question?
- Was there any limitation in my analysis which would affect my conclusions?
- Was the analysis sufficient enough to help decision making?

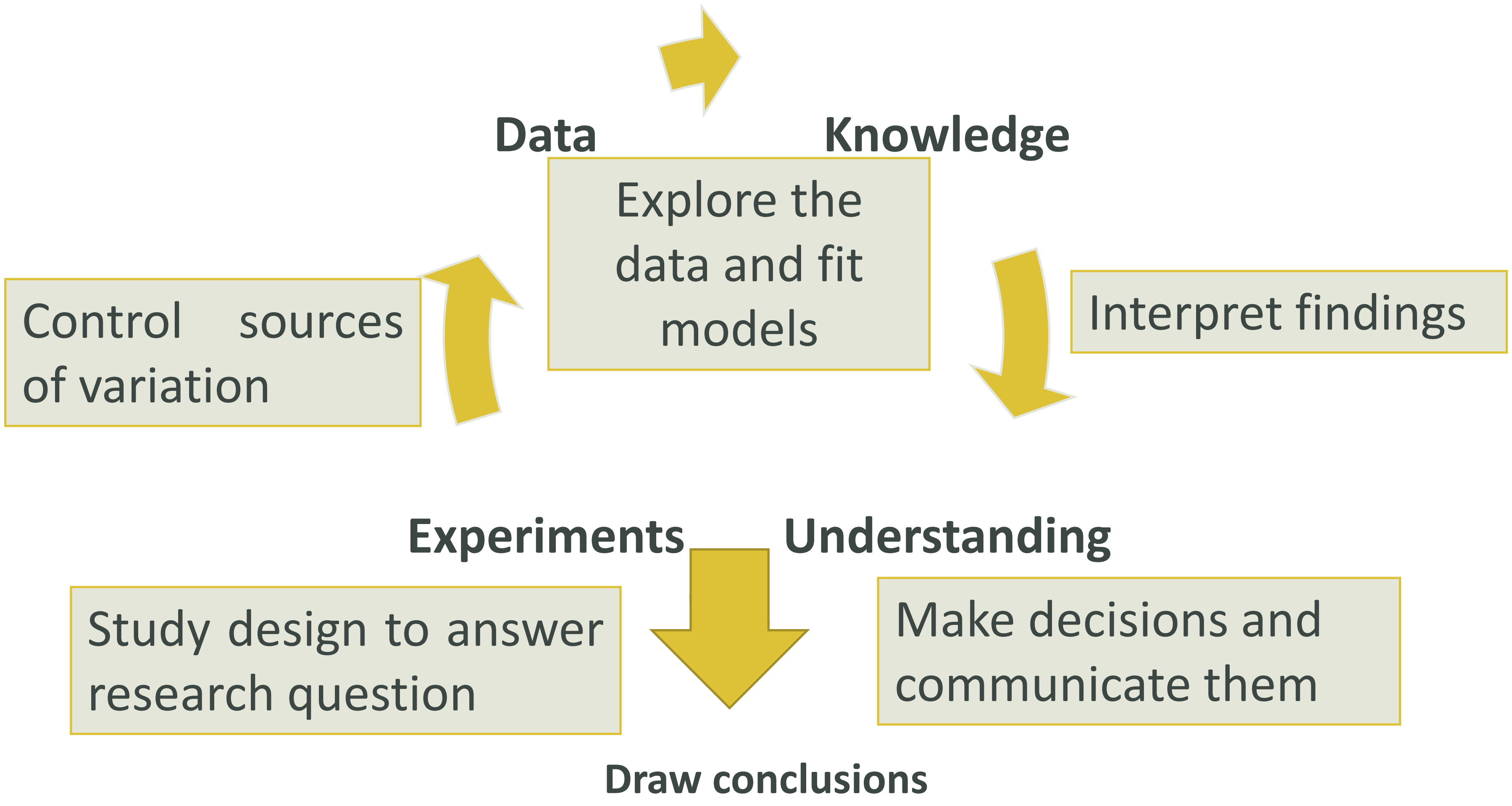
Communicate the results

- Now it's time to communicate your findings
  - Documentation of statistics
  - Making presentations
  - Writing reports, or blogs
  - Writing manuscripts



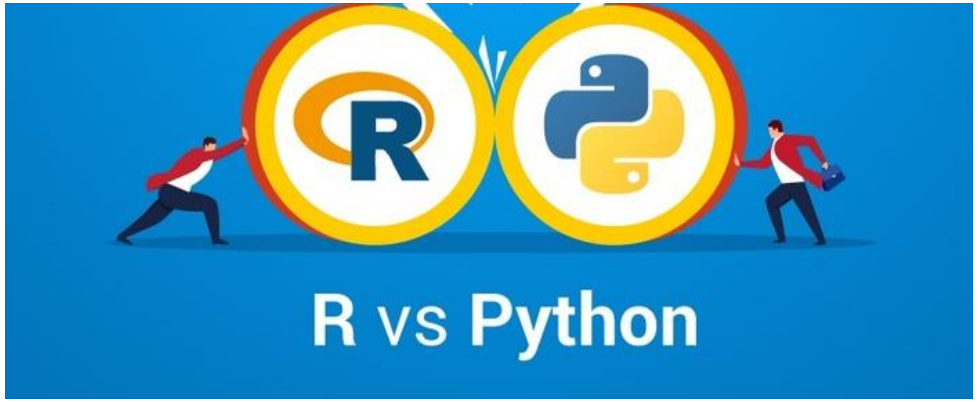
- Extra skills needed at this stage
  - Writing
  - Presenting
  - Communication

Let’s summarize

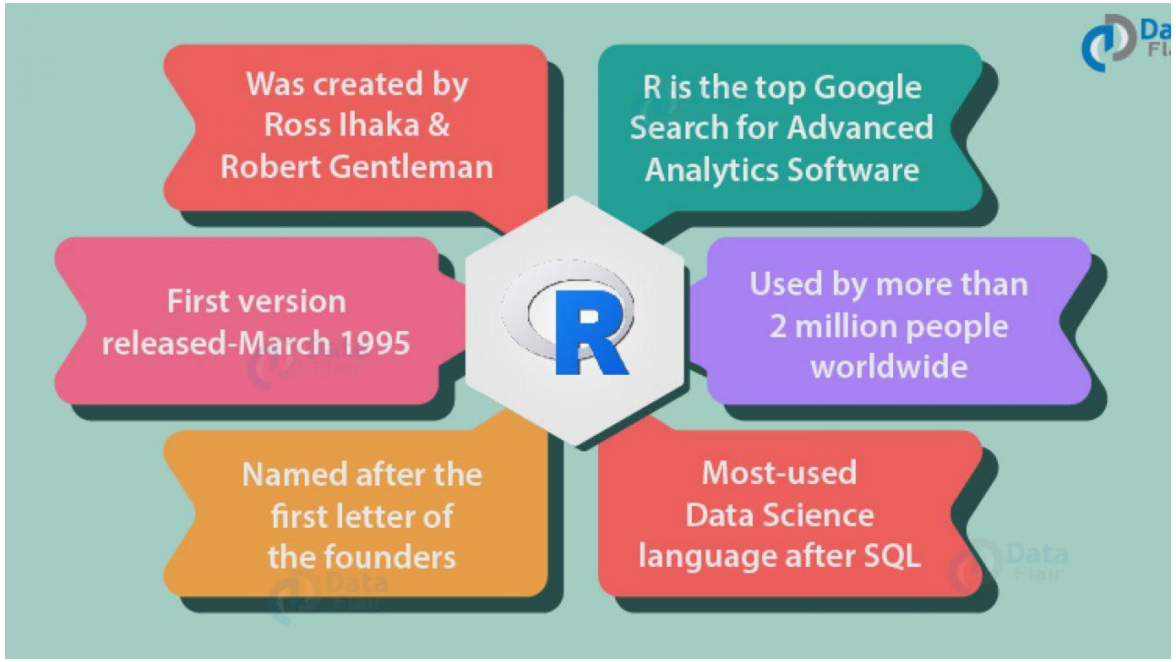


Software

The battle is …

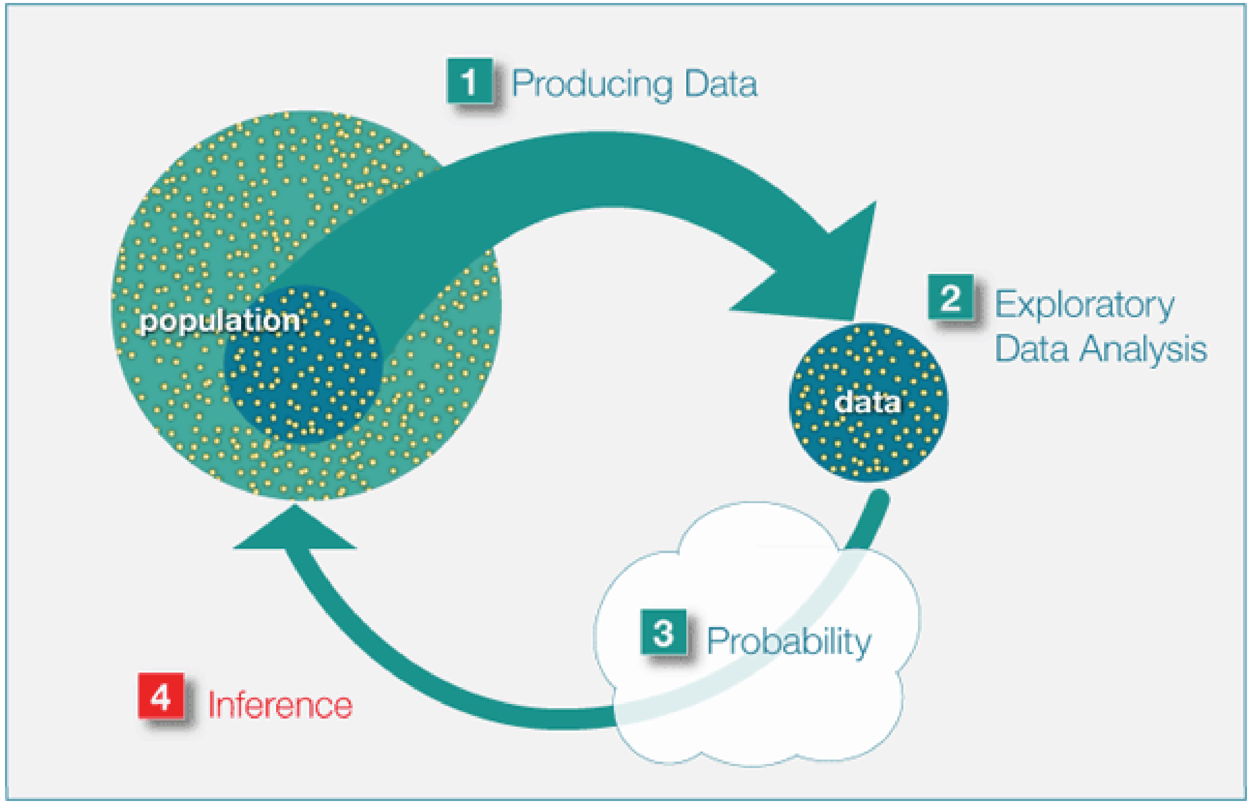


Why R?



https://data-flair.training/blogs/why-learn-r/

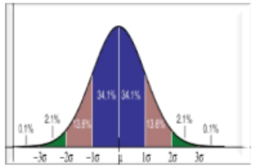
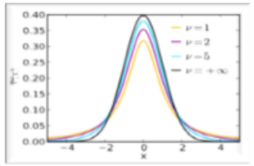
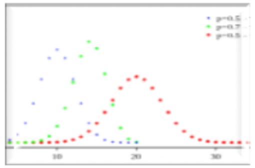
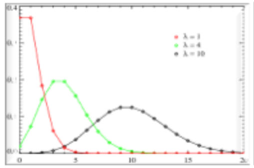
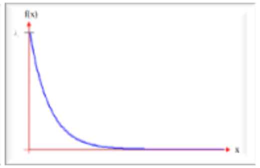
Statistical inference



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

Source: [UF-Open learning]

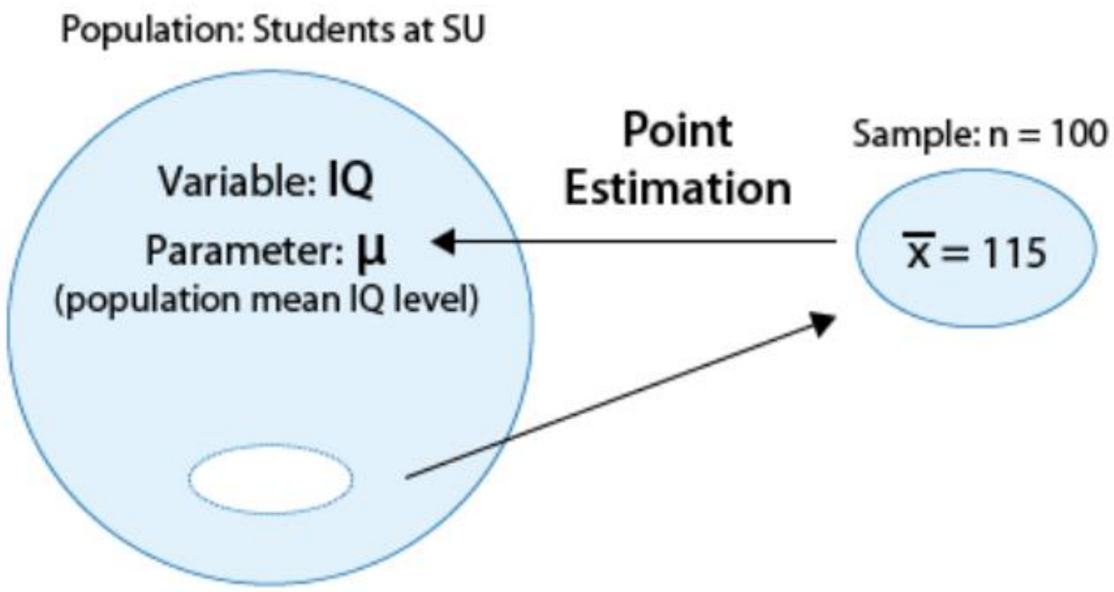
Statistical inference is based on probability distributions

	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.

Source: Data Science Central

- 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 (2010).

Point Estimation



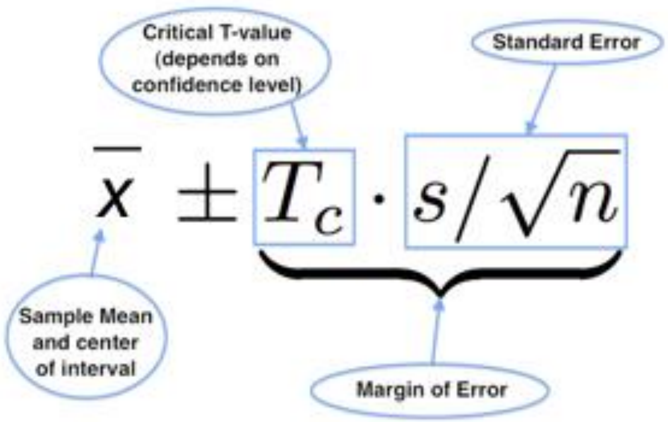
Reference: UF-Open learning

- In **point estimation**, we estimate an unknown parameter using a single number that is calculated from the sample data.

- Point estimates are totally unbiased estimates for the population parameter only if the sample is random and the study design is not flawed.

Interval estimation

- In **interval estimation**, we estimate an unknown parameter using an interval of values that is likely to contain the true value of that. Parameter.
- For example: We are 95% confident that μ for IQ in the previous sample is covered by the interval (112, 118).



Statistical hypothesis testing

Statistical hypothesis testing is defined as:

Assessing evidence provided by the data against the null hypothesis.

Step 1	Step 2	Step 3	Step 4
• ***Formulate the hypotheses***: • H_0 – null and 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.

Hypotheses testing: Type I and II errors

- The probability of a **type I error** is the probability of rejecting the null hypothesis when H0 is true. Is denoted by **α** and is commonly referred to as the **significance level of a test**.

- The probability of a **type II error** is the probability of accepting the null hypothesis when *H*1 is true. Is denoted by β and is highly affected by the sample size.

- The power of a test is defined as 1 - β

or 1 - probability of a type II error = P(rejecting H0|H1 true)

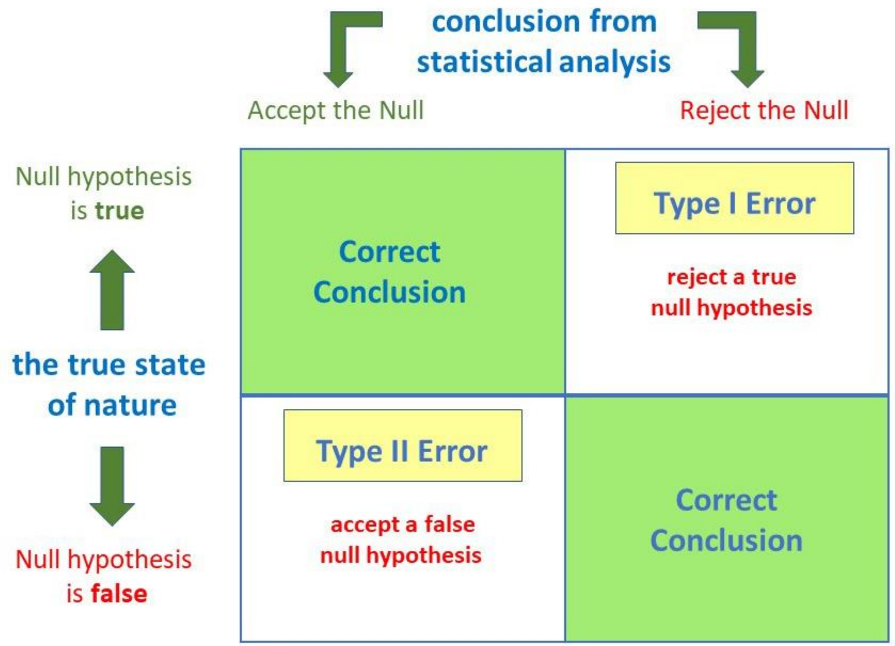
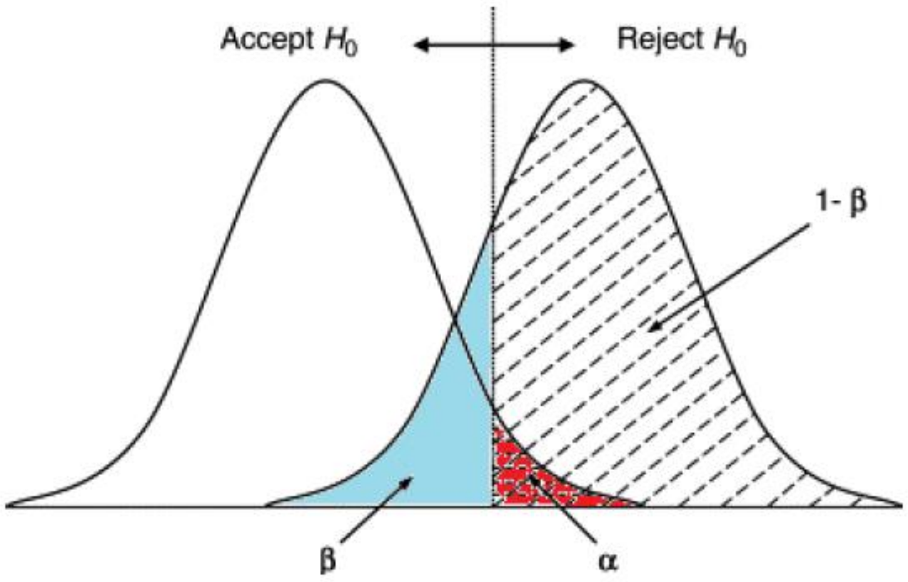


Image source: simplypsychology.org

Power and sample size calculation

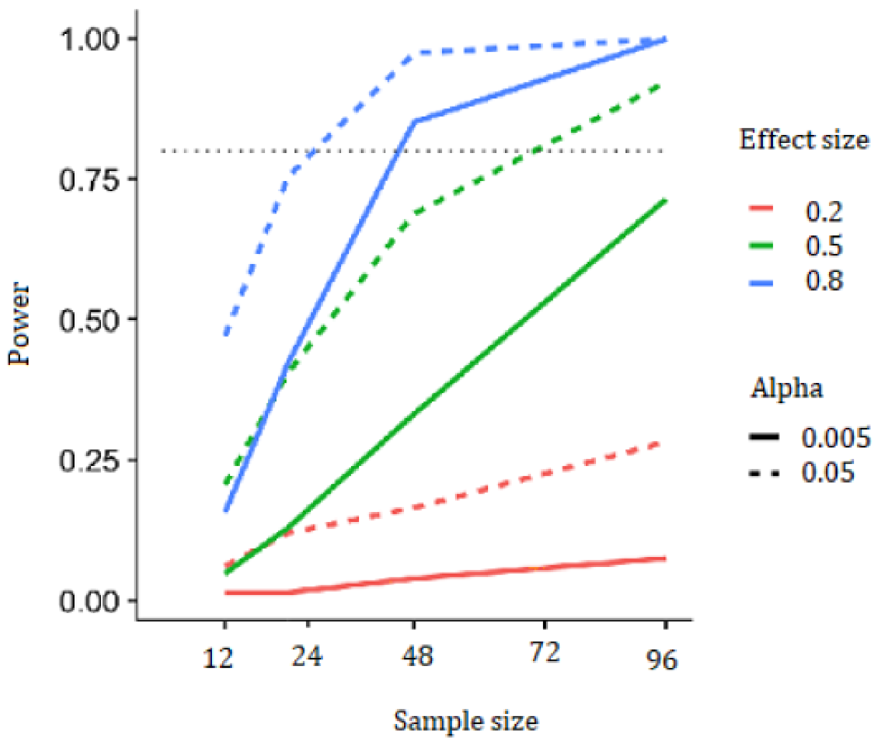
How many subjects do we need?

A study with a small or large sample can be a waste of resources, and the truth will be hard to show.



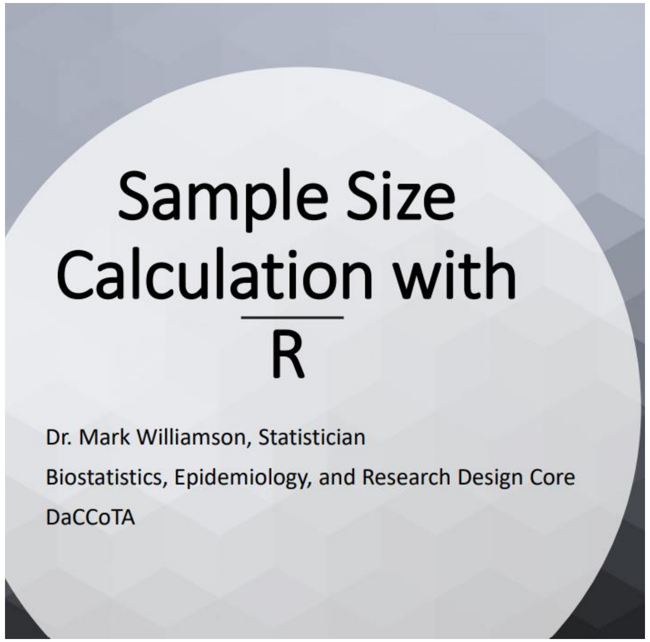
Power of the test $= 1\text{-}\beta$

Power and sample size calculation



Ref: stanford.edu

- Power, effect size, sample size and alpha are inter-related.



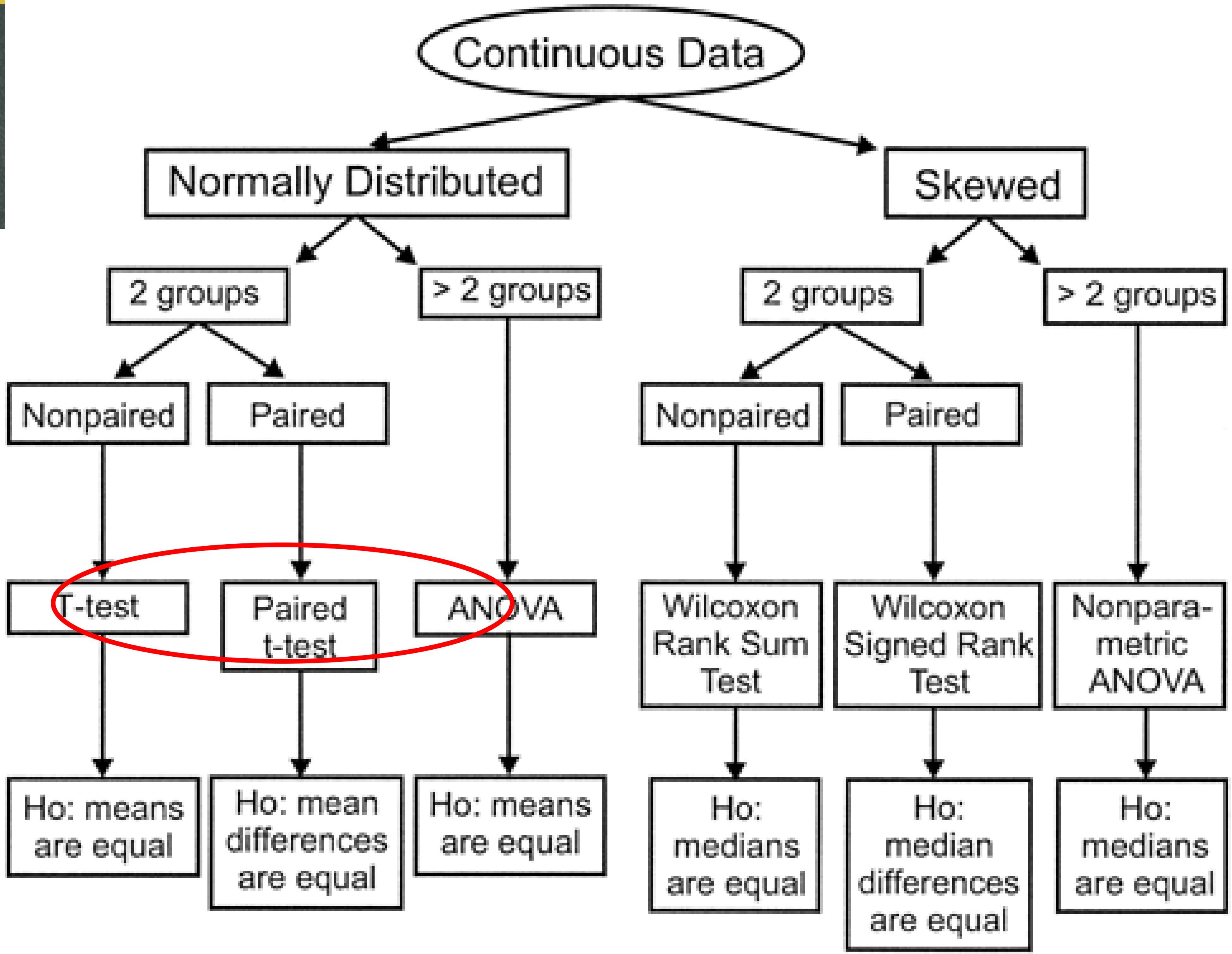
Download handout [here]


Any questions?

Part 2: Parametric methods

- Which test to use?
- Student t-test
  - One-sample t-test
  - Paired samples t-test
  - Independent samples t-test
- One-way ANOVA

Choosing the test …



Simple hypotheses testing: One-sample t-test

1. **Hypotheses (two-sided)**

H_0: $\mu = \mu_0$ vs. H_1: $\mu \neq \mu_0$

2. **Test statistic**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

3. Compute the p-value

$$p = \begin{cases} 2 \times Pr(t_{n-1} \leq t), \text{ if } t \leq 0 \\ 2 \times [1 - Pr(t_{n-1} \leq t)], \text{ if } t > 0 \end{cases}$$

4. Decision

$p<0.05$ Reject H_0

$P>0.05$ Accept H0

Guidelines for Judging the Significance of p-value

- **Guidelines for Judging the Significance of a p-Value**
  - If $0.01 < p < 0.05$, then the results are significant
  - If $p < 0.01$, then the results are highly significant
  - If $p > 0.05$, then the results are not significant

Paired Samples t-test

- The paired t-test is used to determine whether the mean of a dependent variable is the same in two related groups of the independent variable:

- Paired/related groups mean that same individuals are measured at two different "time points" or under two different "conditions"

- **Assumptions**

1. Your dependent variable should be continuous.
2. Your independent variable should consist of two categorical, related groups.
3. The differences between pairs should be normally distributed

Paired Samples t-test

1. Hypotheses

$H_0: \Delta = 0$ vs. $H_1: \Delta \neq 0$

2. Test statistics

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

- *Has a $t_{(n-1)}$ distribution*

3. Decision

p-value > 0.05 accept H_0

- The difference is not significantly different from zero.

P-value < 0.05 reject H_0

- The difference is significantly different from zero.

Test for Normality

- Shapiro Wilk or Kolmogorov Smirnov test
- Perform one of them in R and decide based on p-value.
  - If p-value > 0.05 the data is normally distributed
  - If p-value < 0.05 the data is skewed or not normally distributed.

Independent Samples t-test

- The independent-samples t-test compares the means between two unrelated groups on the same continuous, dependent variable.

- **Assumptions**

1. Your dependent variable should be measured on a continuous scale.
2. Your independent variable should consist of two categorical, independent groups (i.e., gender).
3. There should be no significant outliers.
4. Your dependent variable should be approximately normally distributed for each group of the independent variable.
5. Test the homogeneity of variances.

Test homogeneity of variances

- Levene's test
- Hypotheses: H_0: $\sigma_1^2 = \sigma_2^2$ vs. H_1 : $\sigma_1^2 \neq \sigma_2^2$
- *Test statistics:* $F = s_1^2 / s_2^2 \sim F_{(n1-1;n2-1)}$ distribution
- Decision

p-value > 0.05 accept Ho – equal variances

P-value < 0.05 reject H0 – unequal variances

Independent Samples t-test: Equal variances assumed

1. Hypotheses

H_0: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 \neq \mu_2$.

2. Test statistics

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \sim t_{(n1+n2-2)} \text{ distribution}$$

where $s = \sqrt{\left[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right] / (n_1 + n_2 - 2)}$

3. Decision

p-value > 0.05 accept H_0

- The populations means are not significantly different from each other.

P-value < 0.05 reject H_1

- The populations means are significantly different from each other.

Independent Samples t-test: Unequal variances assumed

1. Hypotheses

H_0: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 \neq \mu_2$.

2. Test statistics

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \sim t_{(d')} \text{ distribution}$$

Compute the approximate degrees of freedom d', where

$$d' = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left(s_1^2/n_1\right)^2 \big/ (n_1 - 1) + \left(s_2^2/n_2\right)^2 \big/ (n_2 - 1)}$$

3. Decision

p-value > 0.05 accept H_0

- The populations means are not significantly different from each other.

P-value < 0.05 reject H_1

- The populations means are significantly different from each other.

One-way ANOVA

- The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of two or more independent (unrelated) groups.

- **Assumptions**

1. Your dependent variable should be measured on a continuous scale.
2. Your independent variable should consist of two or more categorical, independent groups.
3. There should be no significant outliers.
4. Your dependent variable should be approximately normally distributed for each group of the independent variable.
5. There is need to test homogeneity of variances.

One-way ANOVA

1. **Hypotheses**

H_0: $\mu_1 = \mu_2 = \ldots = \mu_a$

H_1: At least two means are different

2. **Test statistics**

$F = s_b^2 / s_w^2$

$\sim F_{(a-1;n-a)}$ distribution

3. Decision

- p-value > 0.05 accept H0

Means are statistically equal.

- p-value < 0.05

We can reject H_0, that all the means are equal, and can conclude that at least two of the means are significantly different. These results are displayed in an ANOVA table (we discuss it in R).

One-way ANOVA: Post-hoc analysis

- If H_0 is rejected, we should perform a post-hoc analysis to see which groups are different.

- We will discuss and interpret this in R


Any questions?

Part 3: Nonparametric analysis

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

Parametric vs. nonparametric methods

- Require assumptions about the distribution in the population

Parametric

- Distribution free, called **exact tests.**

Nonparametric

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

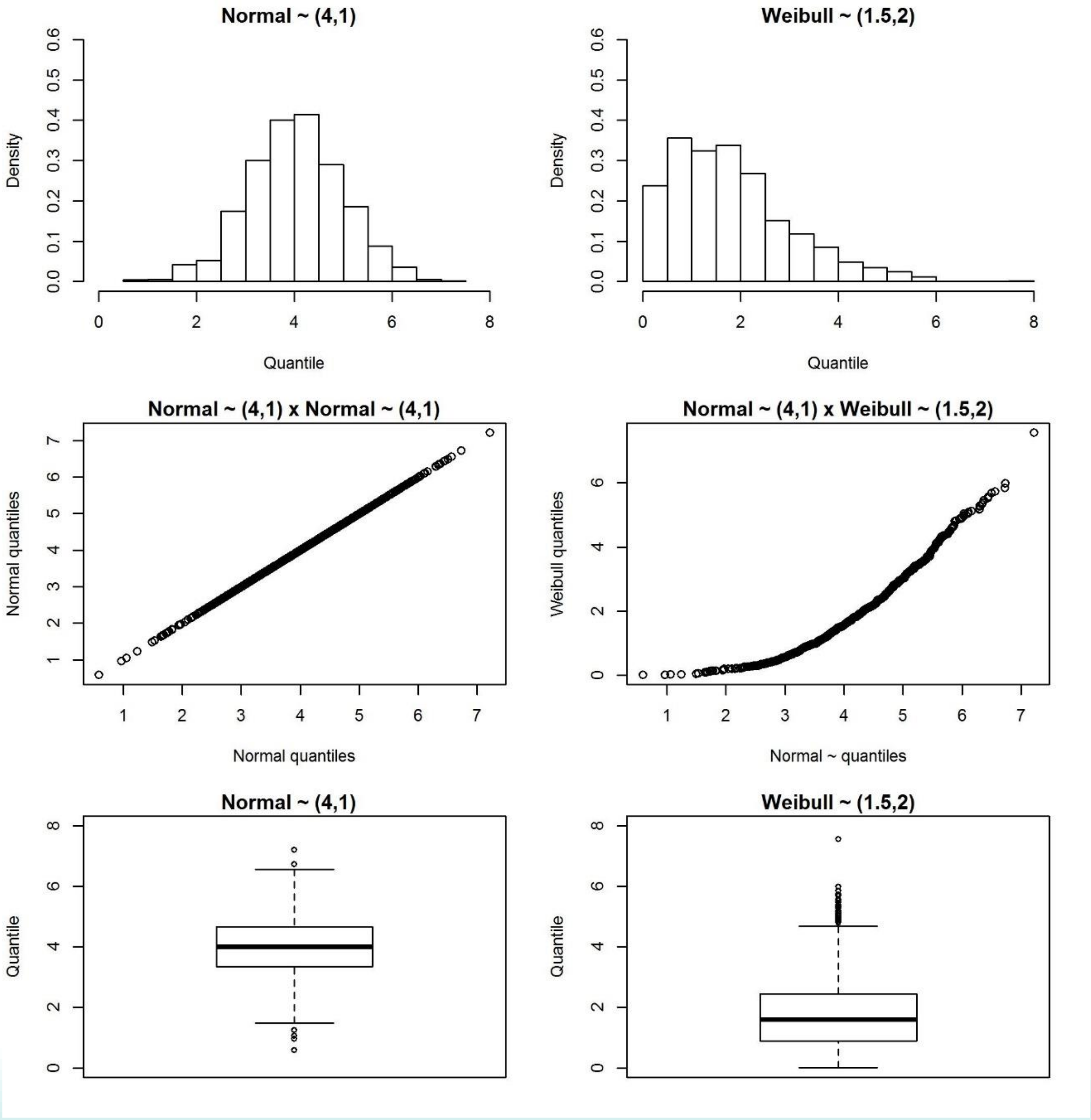
Exact tests

When and which nonparametric test to use?

Objective	Data	
	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)

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

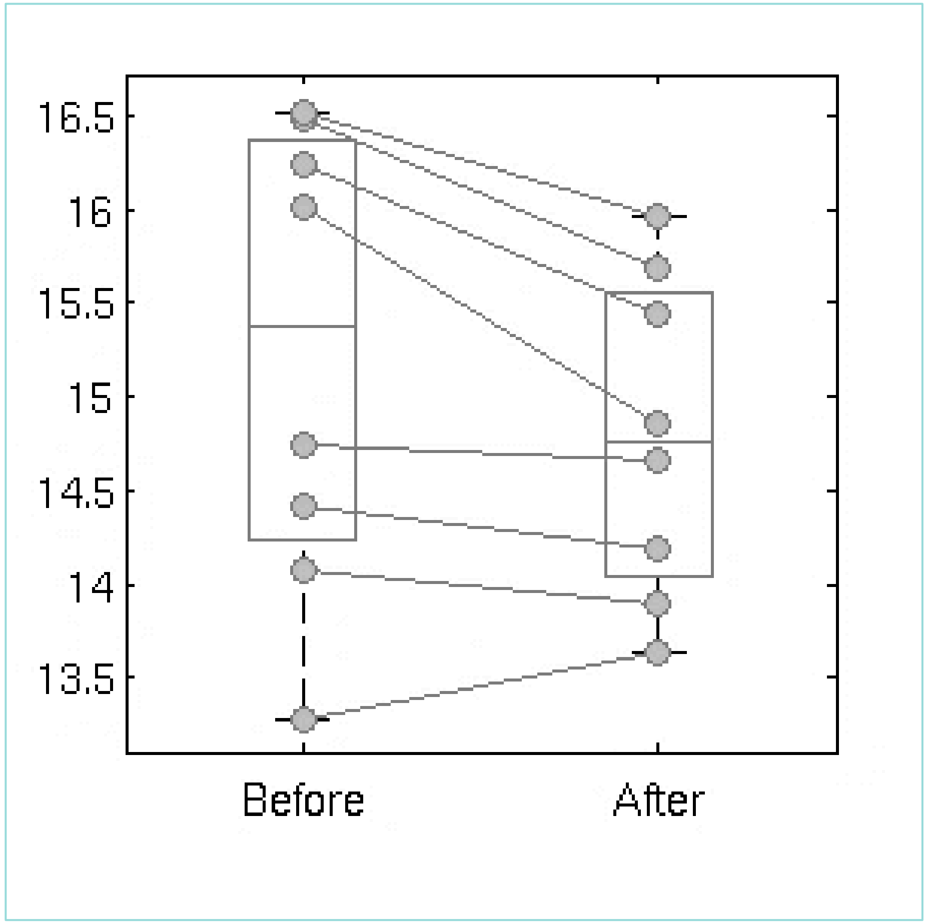
See details on the formula here.

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:

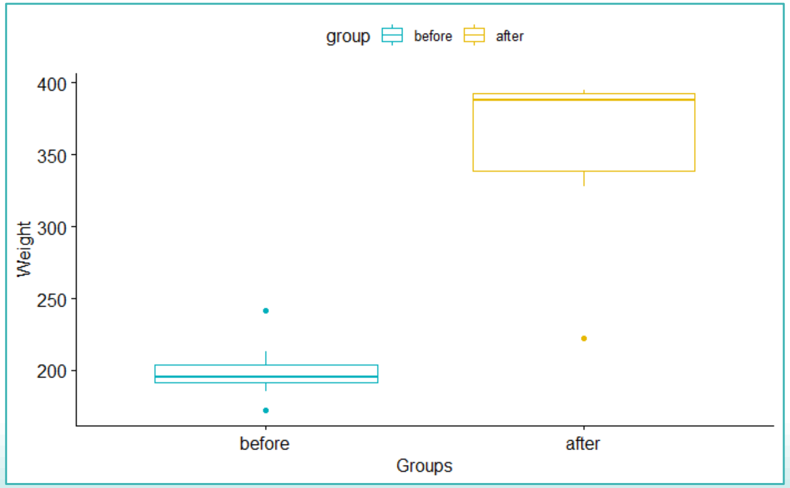
```
wilcoxon.test (variable1, variable2, Paired=TRUE, exact=FALSE)
```

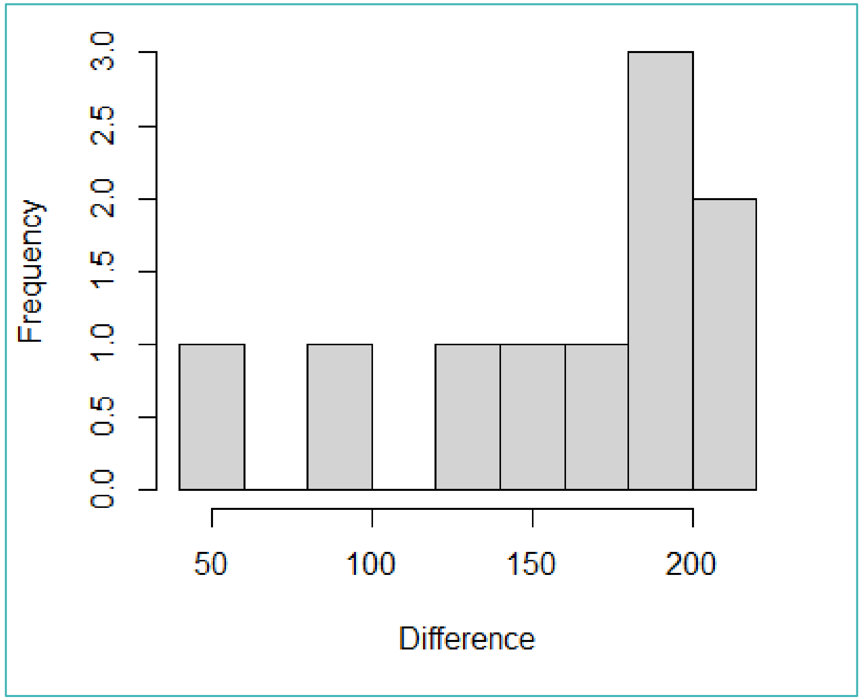
Dependent: Continuous or ordinal
Independent: Time/condition

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 R Notebook for code and details*

Wilcoxon signed rank test: Example 1

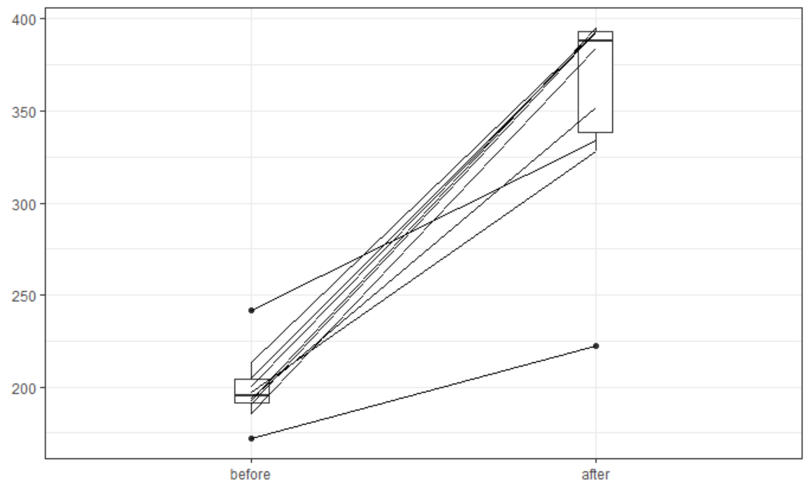
#In order to run the Wilcoxon signed rank test in R, use the code:

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

Output:

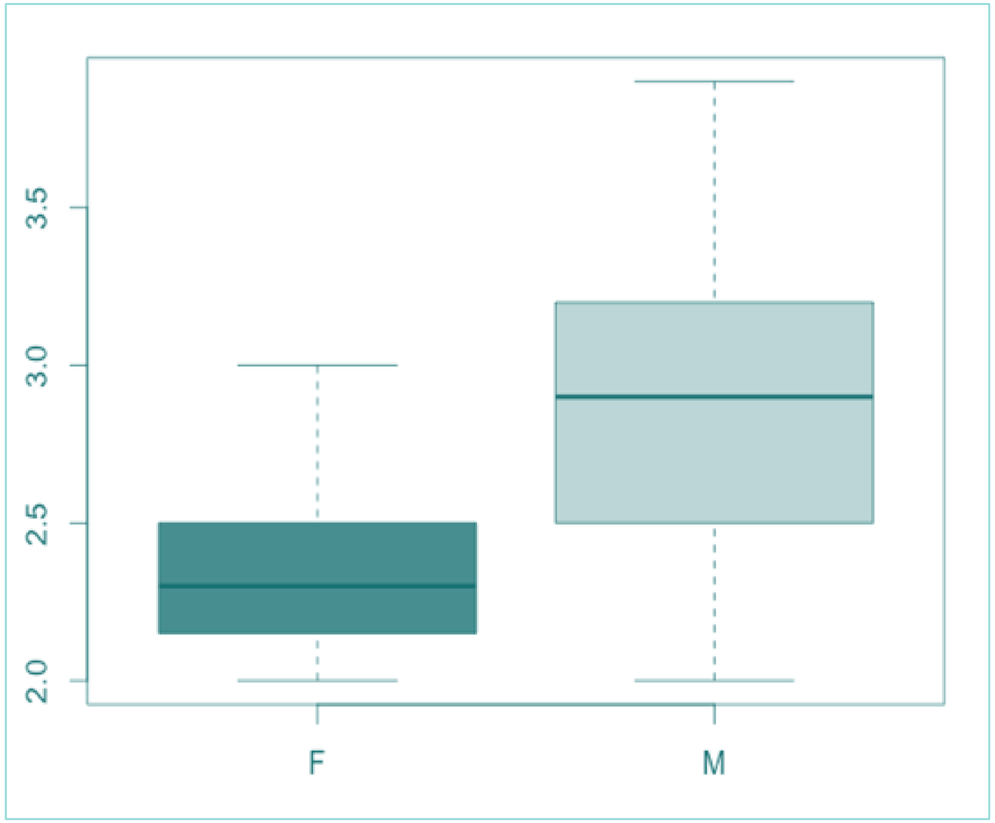
```
        Wilcoxon signed rank test with continuity correction

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:

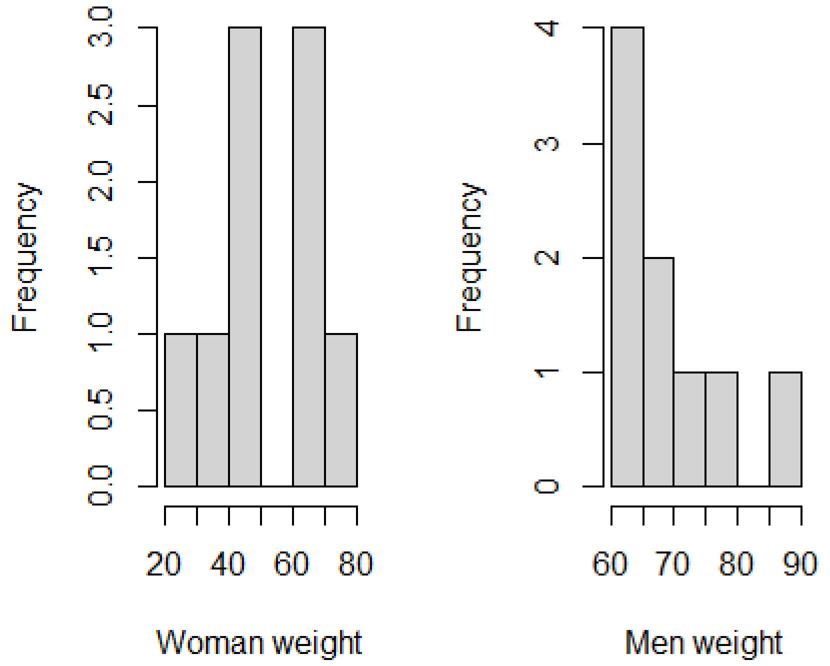
```
wilcox.test(dependent~independent)
```

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
```

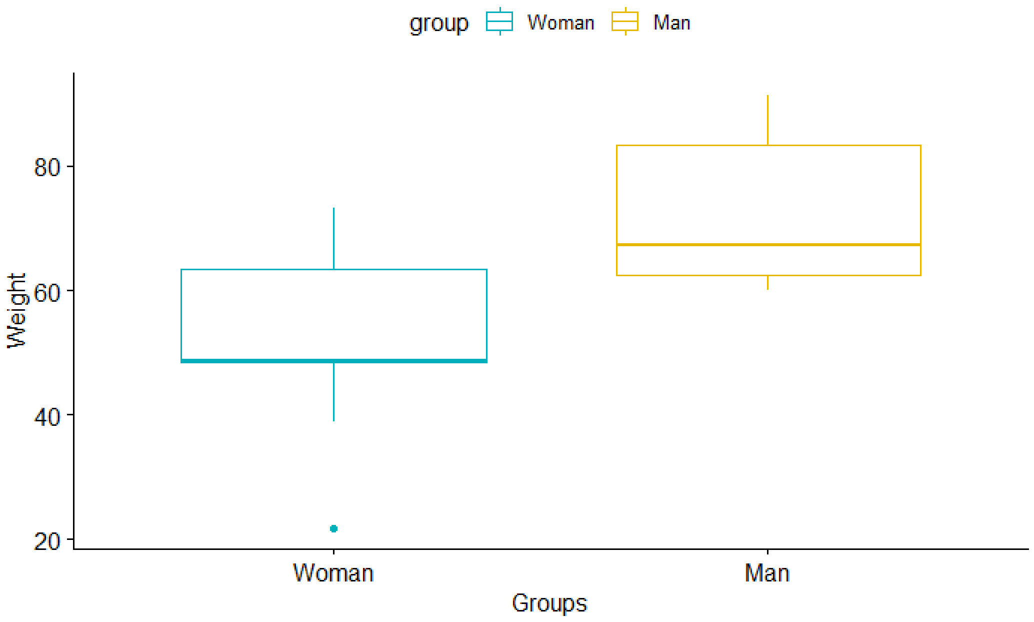
Mann Whitney test: Example 2

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

wilcox.test(weight~group)

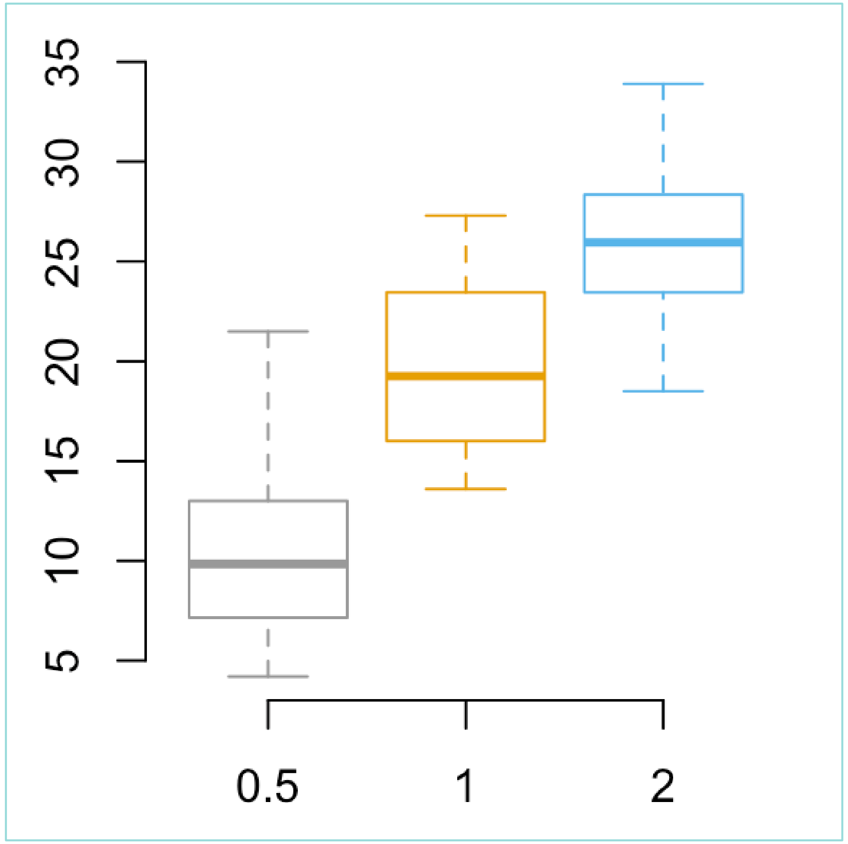
```
	Wilcoxon rank sum test with continuity correction

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.



Two or more independent samples

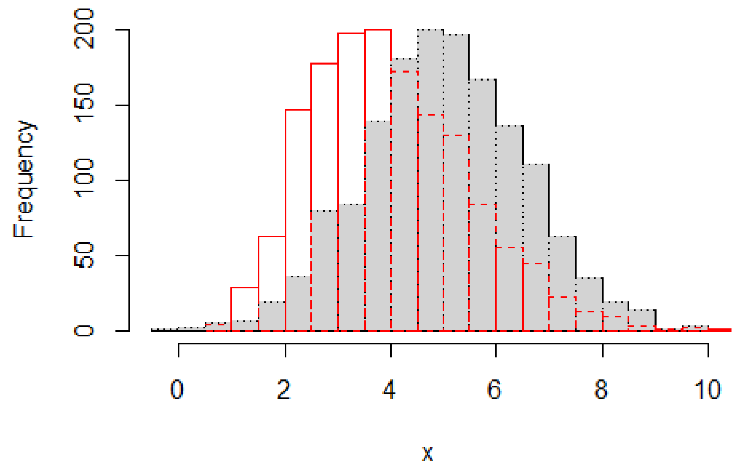
Kruskal-Wallis test

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

1. Hypothesis

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

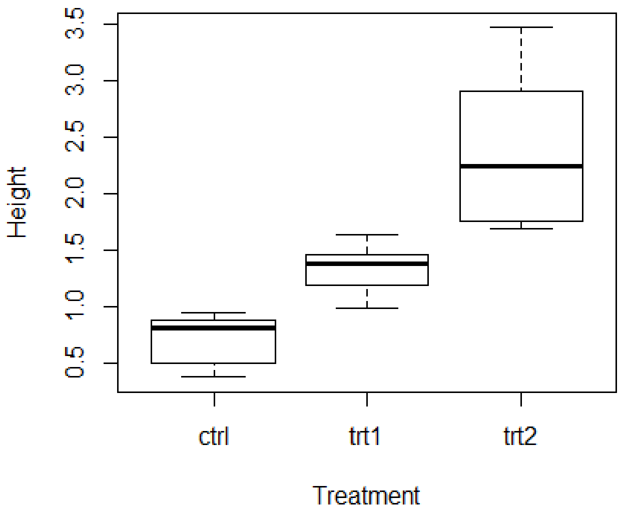
3. Decision based on p-value

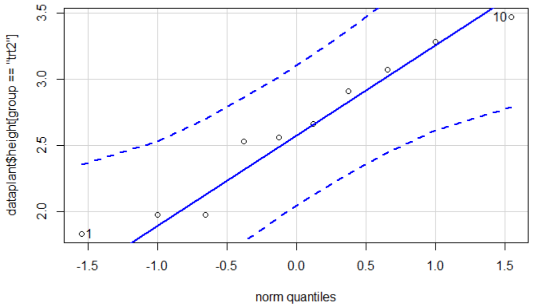
If $p<0.05$ reject the null hypothesis (Ho)

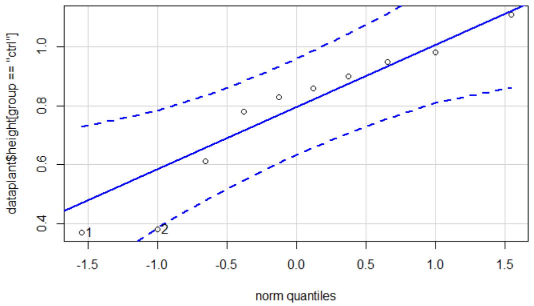
Example 3

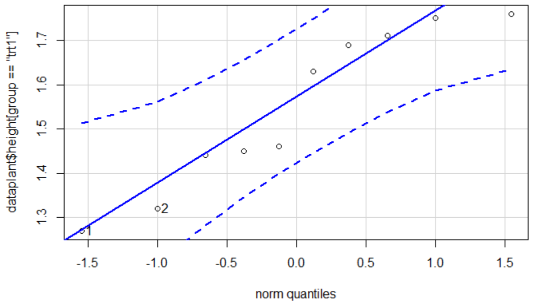
Data: Here, we'll use the data set named *PlantGrowth*. It contains the height of plants (cm) obtained under a control and two different treatment conditions.

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









Kruskal-Wallis test: Example 3

#In order to run the Kruskal-Wallis 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.

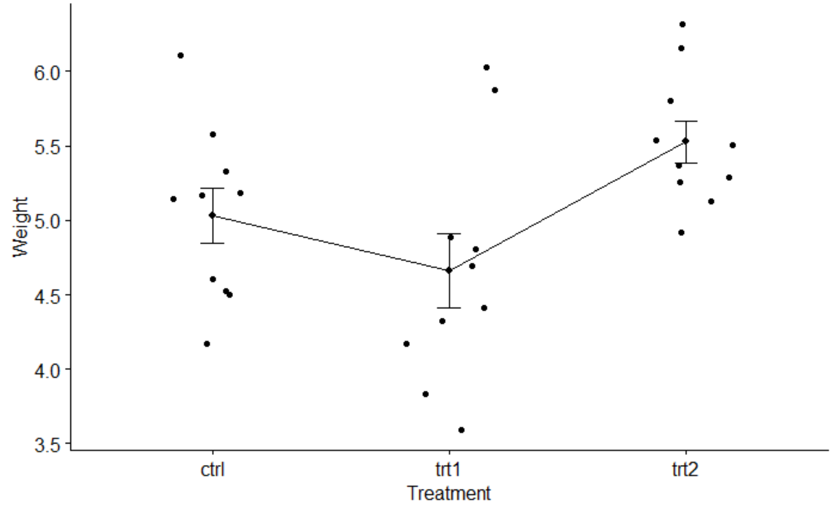
```
        Kruskal-Wallis rank sum test

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

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*

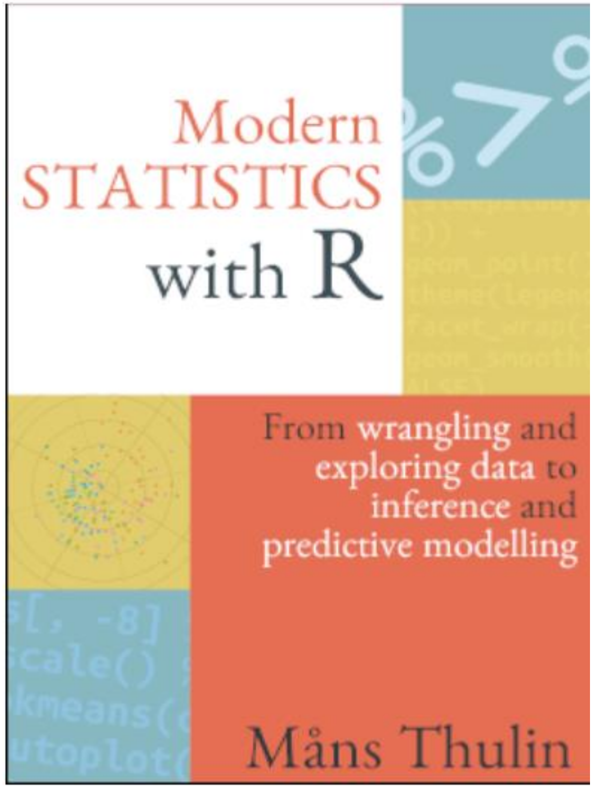
Pairwise Wilcoxon signed rank test: Example 3

Reporting the results:

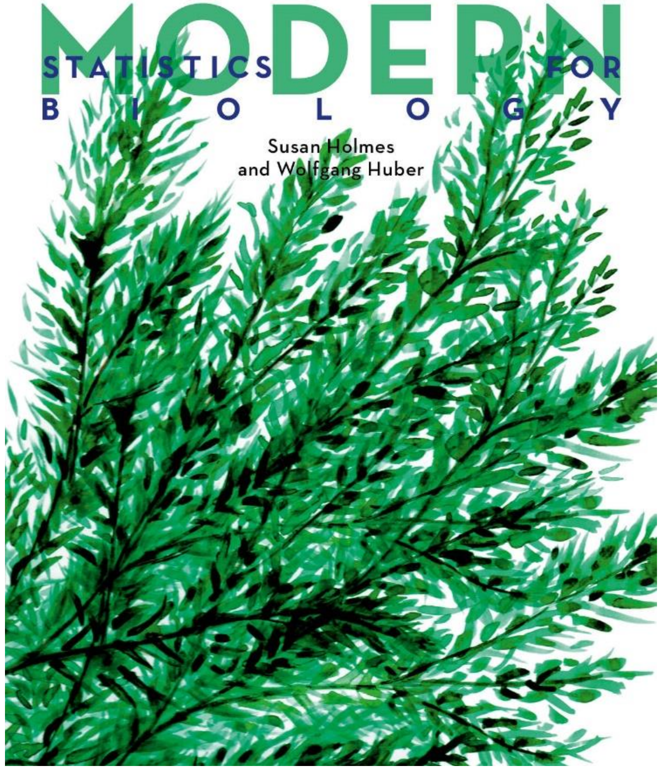
A Kruskal-Wallis test was carried out to compare plant height after two treatments and control (no treatment). There was very strong evidence of a difference (p-value < 0.001) between the mean ranks of at least one pair of groups. Wilcoxon signed rank pairwise tests were carried out for the three pairs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

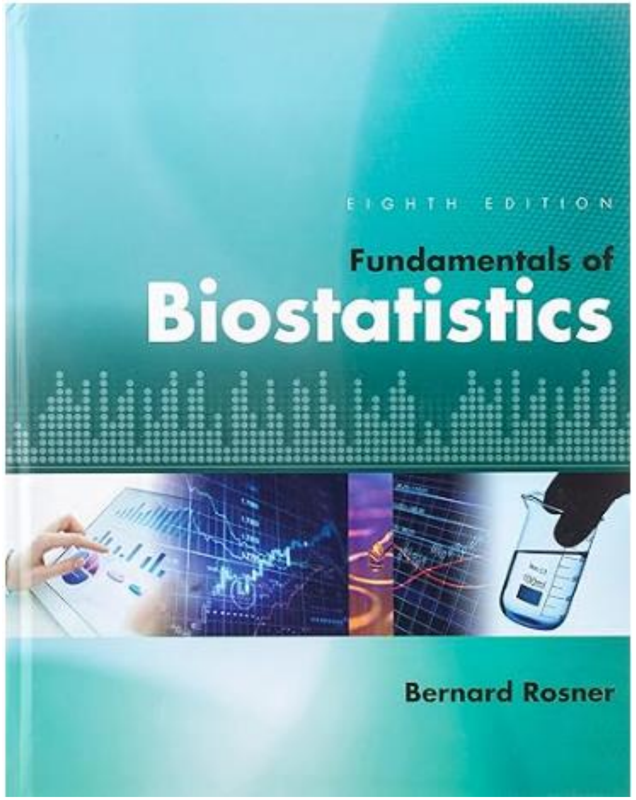
Recommended books



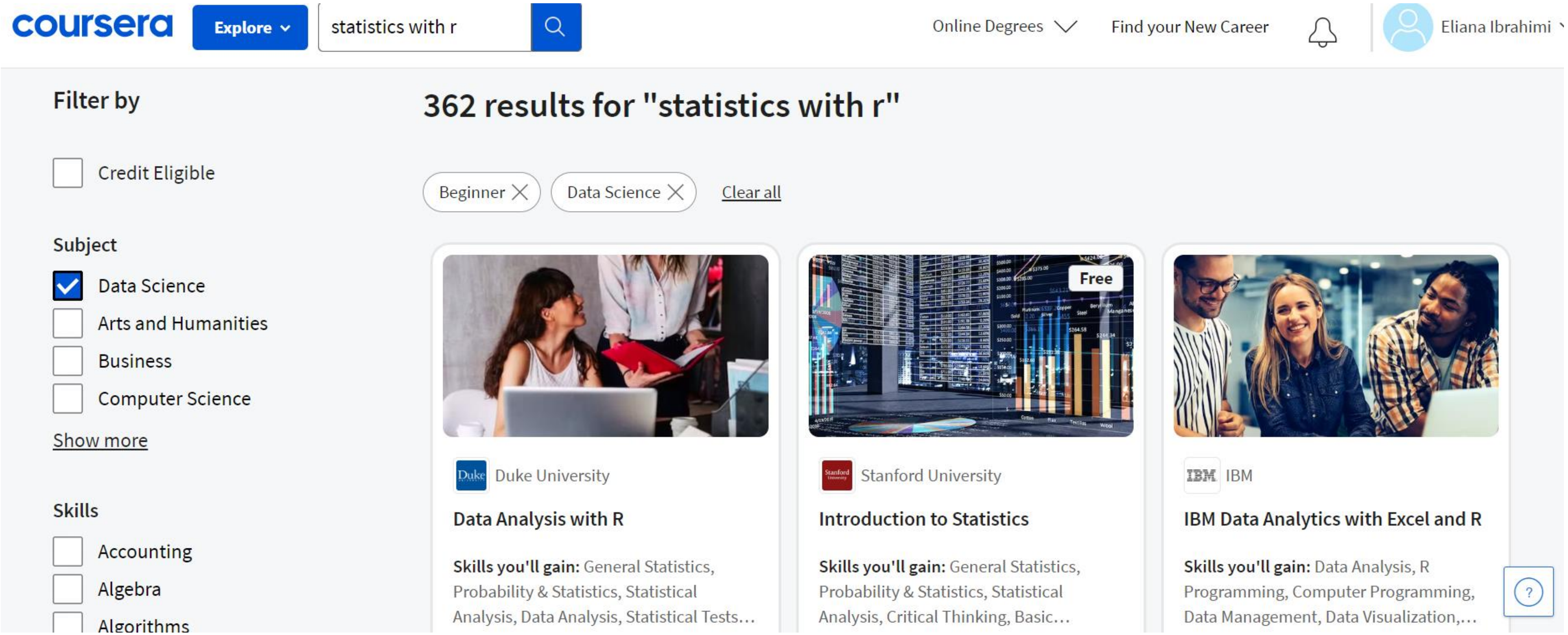
https://www.modernstatisticswithr.com/



https://web.stanford.edu/class/bios221/book/00-chap.html



Online courses



References/Useful links

1. Rosner, Bernard. Fundamentals Of Biostatistics. Cengage Learning, 2011.
2. Pezzullo, John. Biostatistics For Dummies. Wiley, 2013.
3. Kloke, J., & McKean, J.W. (2014). Nonparametric Statistical Methods Using R (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b17501
4. https://bolt.mph.ufl.edu/6050-6052
5. http://www.biostathandbook.com/HandbookBioStatThird.pdf
6. https://data-flair.training/blogs/why-learn-r/
7. https://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf


Any questions?