Concepts of Biostatistics & Biological Data

BioData Training School 2025

December 10-13, 2025

University of Tirana, Albania

Outline

What is Biostatistics?

Why biostatistics matters?

Key concepts

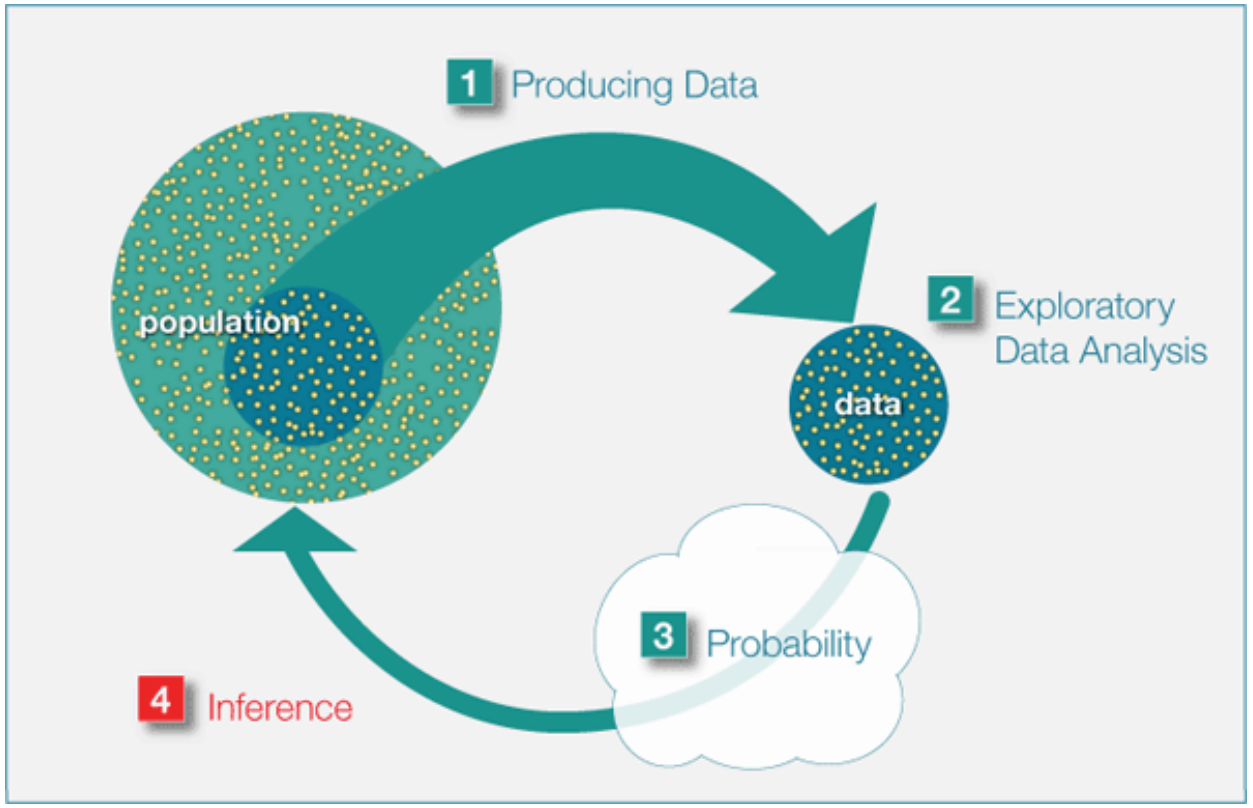
- Biological data types
- Descriptive statistics
- Probability & inference
- Sample size calculation
- Core statistical methods

Definition of Biostatistics

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

--- Much more statistics than biology and medicine, however, biostatisticians must learn some biological and medical concepts also.

Aim of Biostatistics

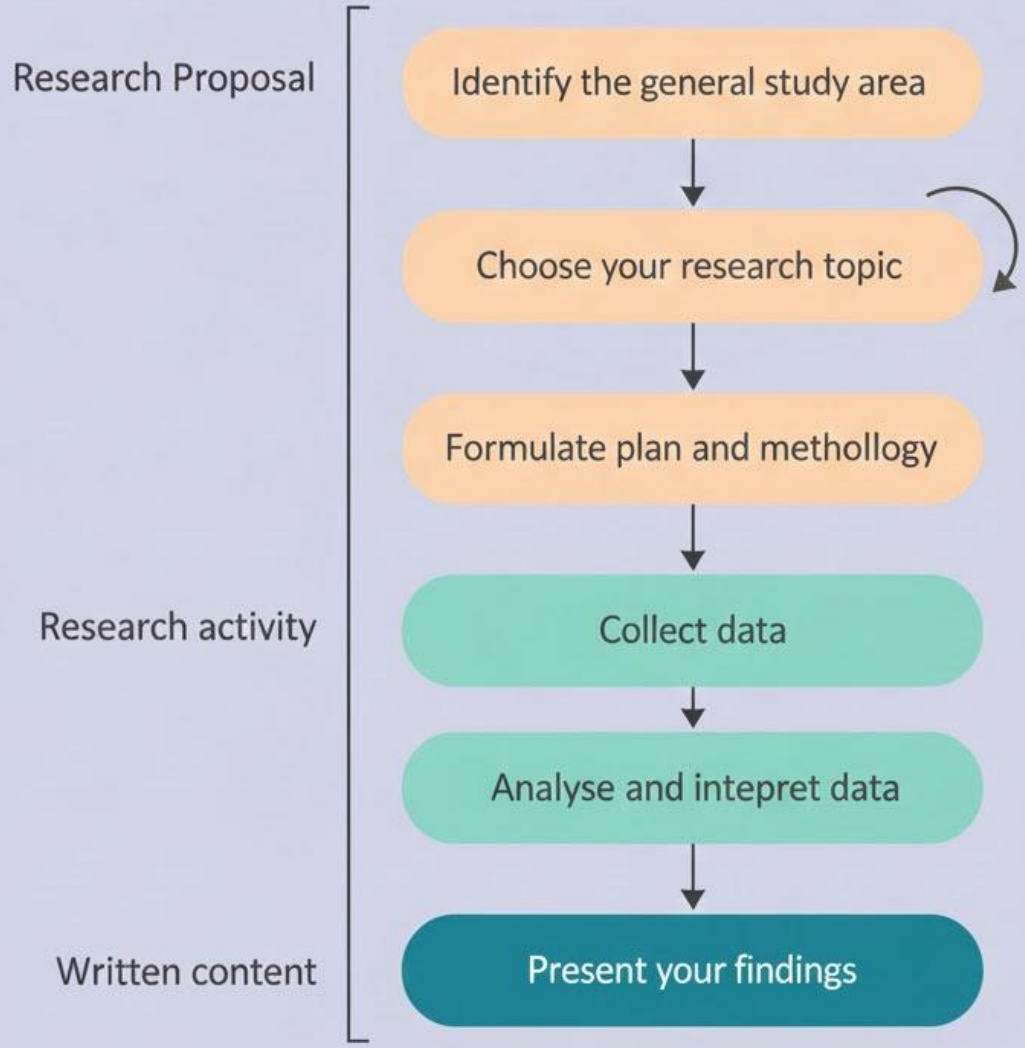


Source: UF-Open learning

- Draw conclusions about a population based on the data obtained from a sample chosen from it.

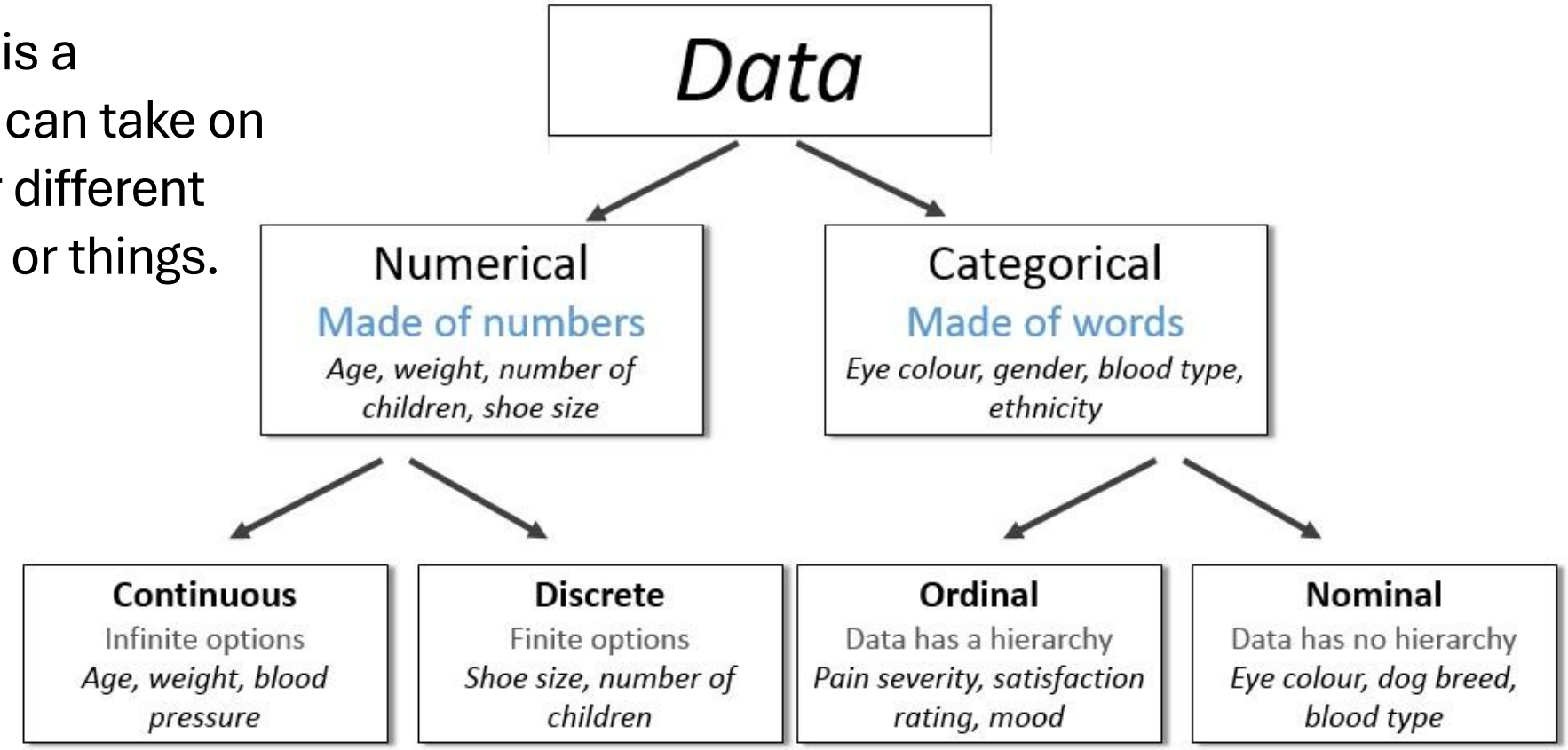
Biostatistics in research

- 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 all steps**.

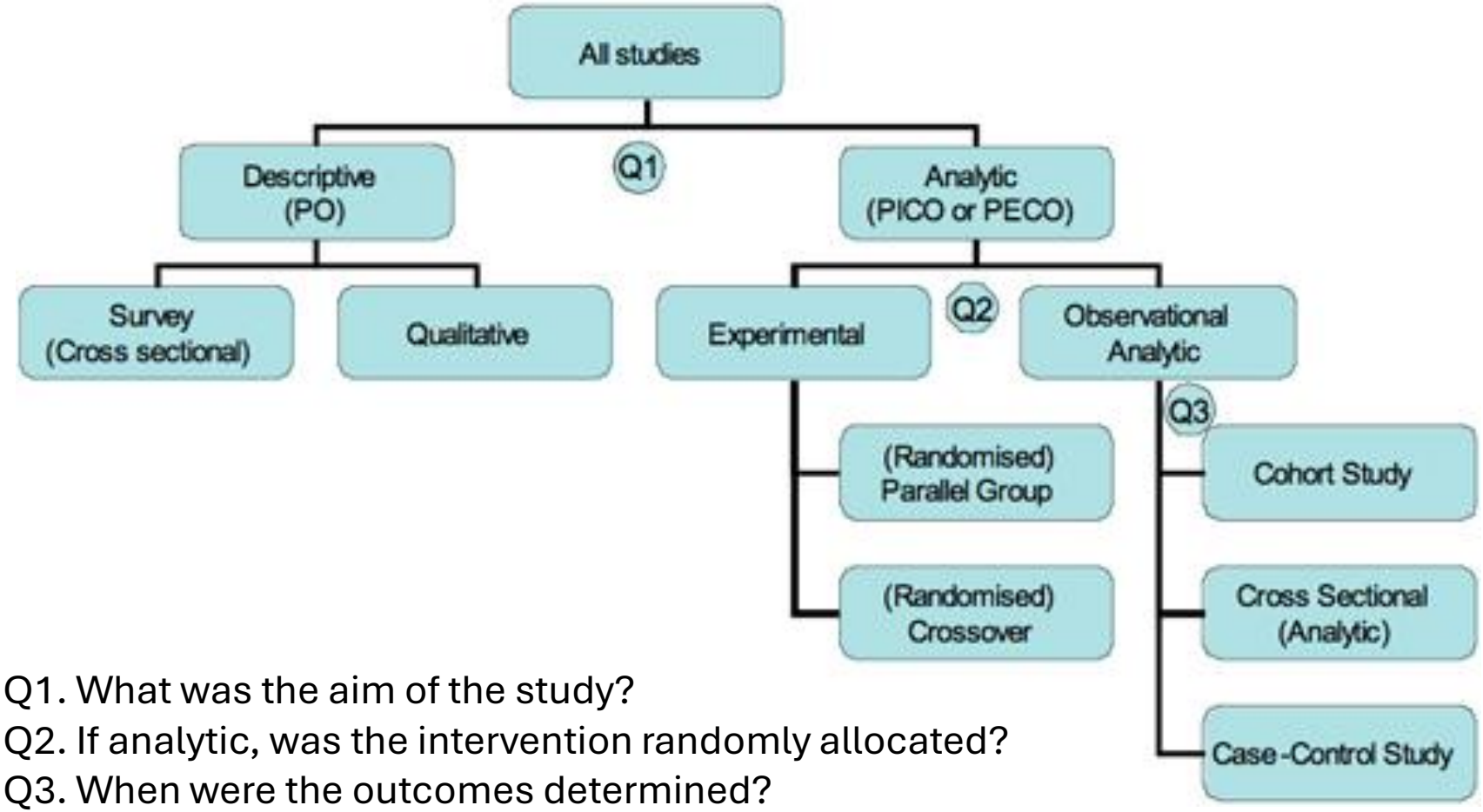


Types of variables in biostatistics

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



Data collection/Study design



Q1. What was the aim of the study?
Q2. If analytic, was the intervention randomly allocated?
Q3. When were the outcomes determined?

https://www.cebm.ox.ac.uk/resources/ebm-tools/study-designs

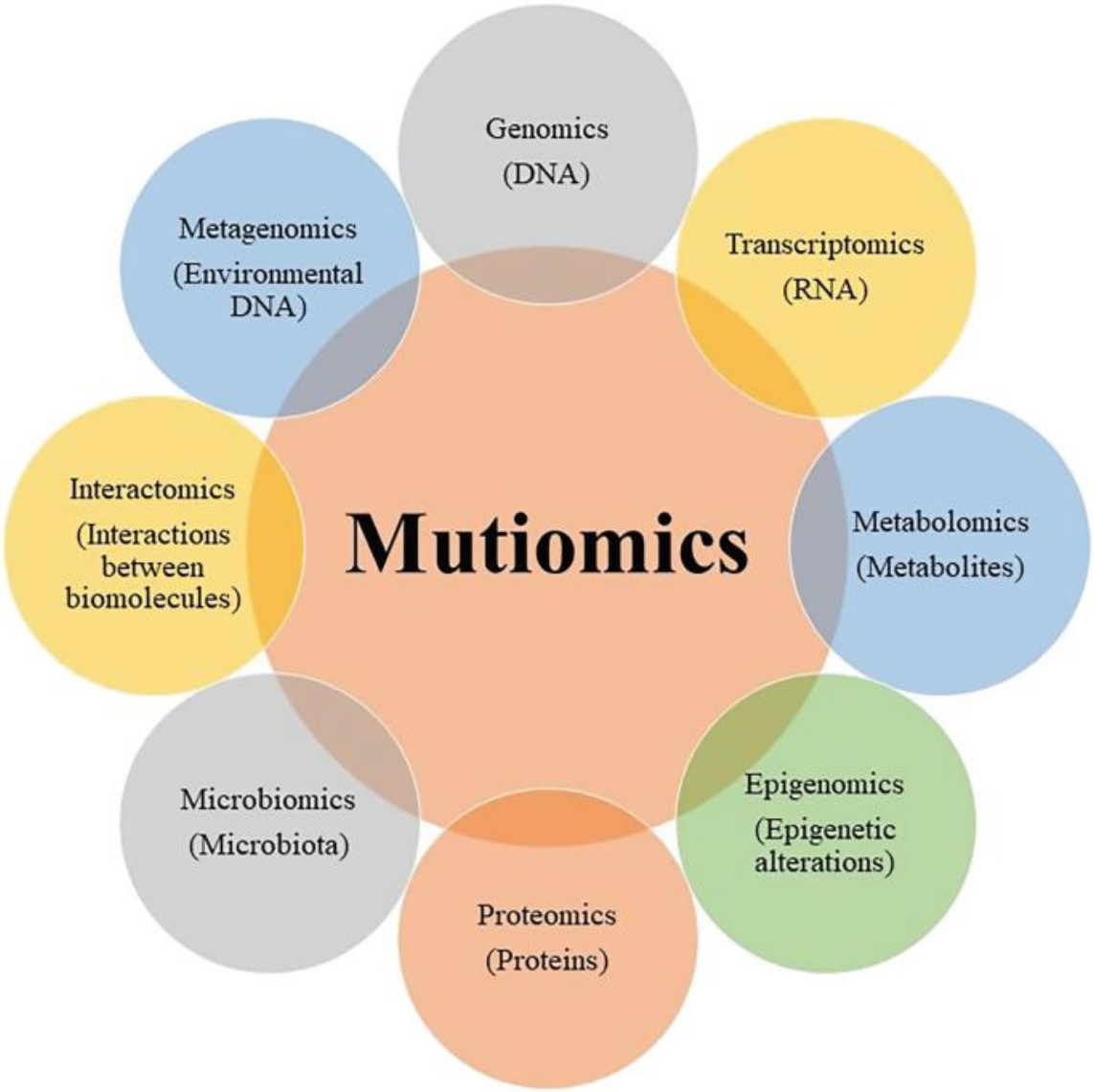
Biological & Health data

Clinical & epidemiological data (blood pressure, BMI, disease status, incidence rates, etc)

Laboratory measurements (enzyme activity, hormone levels, cell counts, etc)

Omics data (Genomics, Transcriptomics, Metabolomics)

Imaging data (MRI scans, microscopy images, CT images)

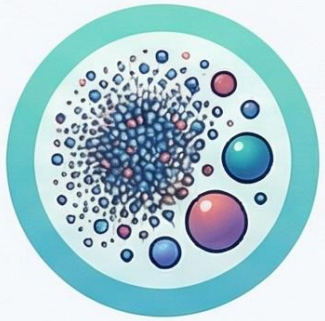


https://doi.org/10.1007/978-981-97-1844-3_16

DATA-CENTRIC HURDLES

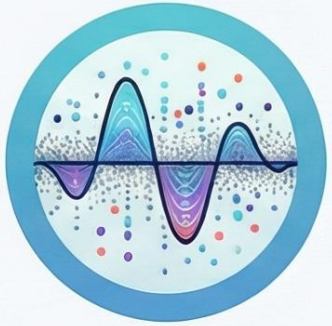
High Dimensionality vs. Low Sample Size

Thousands of variables (genes, biomarkers) often overwhelm a small number of samples.



Inherent Variability & Noise

Data differs across individuals and time, complicated by batch effects and measurement errors.



Heterogeneous & Incomplete Sources

Data comes from varied sources (omics, clinical) with different formats and missing values.

PROCESS & RESOURCE BARRIERS

Ethical, Privacy & Access Constraints

Sensitive health information is restricted by regulations, data sharing policies, and consent.

Lack of Standardization & Reproducibility

Inconsistent analysis pipelines and preprocessing steps make it difficult to reproduce published results.

Intensive Computational Demands

Large datasets require high-performance computing, optimized workflows, and scalable algorithms.

Biological Data: Key Challenges

The Data Analysis Process

5 Main Steps

- Data preparation includes editing, coding, and data entry in forms that are appropriate for analysis.

Data exploration

Nominal variables

- Frequency
  - Count (How frequent values occur)
  - Relative (The % of observations with a specific value)

Numerical variables

- Descriptive statistics
  - Mean, Mode, Median
  - Variance, Standard deviation, SE
  - Range, Percentiles

Data exploration: Descriptive statistics

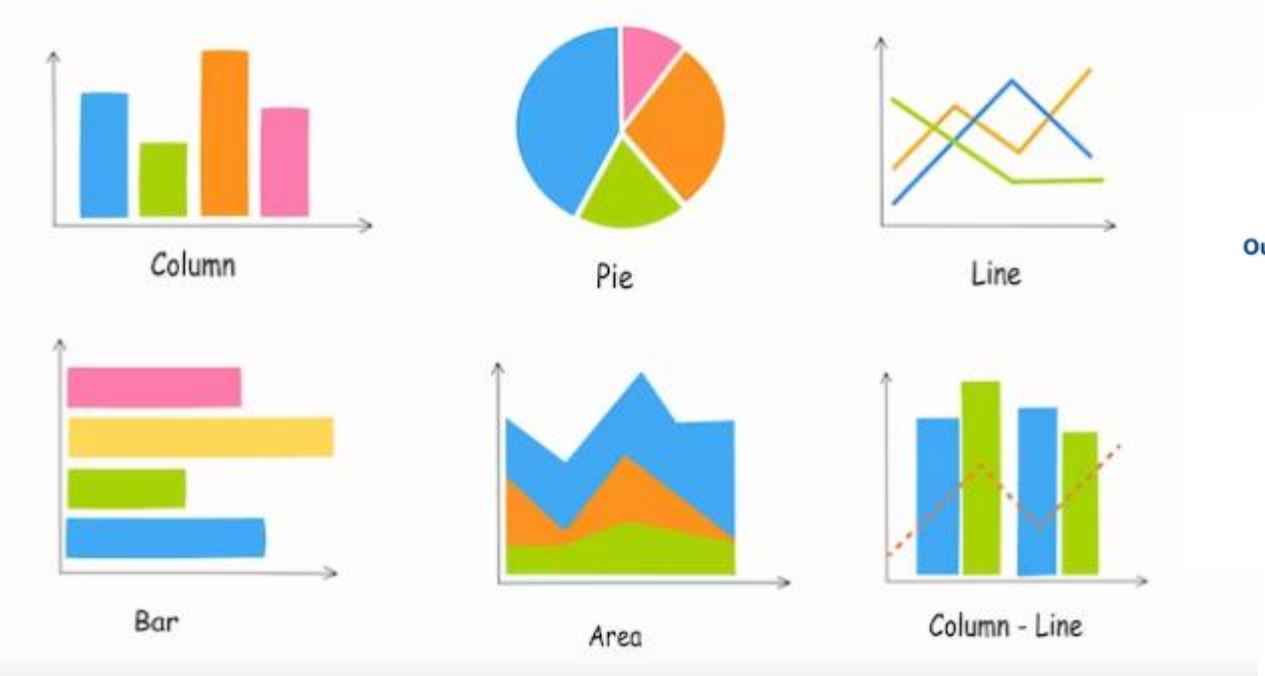
1. Arithmetic mean $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

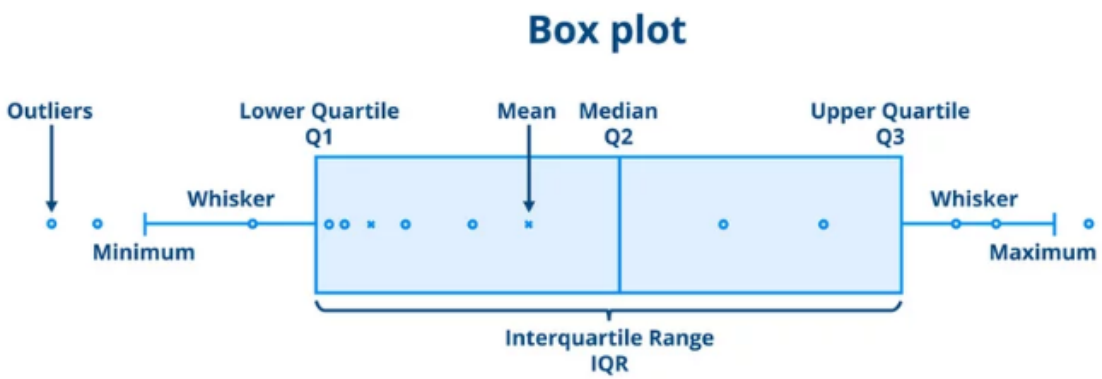
2. Median
   (1) The $\left(\frac{n+1}{2}\right)$th largest observation if n is odd
   (2) The average of the $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th largest observations if n is even

3. Variance $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

4. Standard deviation $$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\text{sample variance}}$$

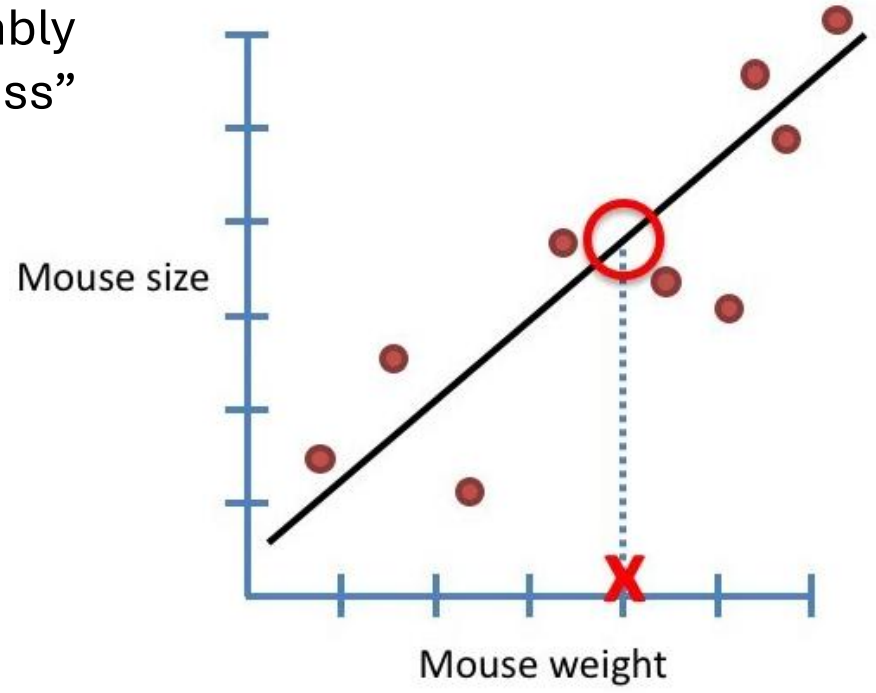
Data exploration: Graphical visualizations





Statistical modeling

- “A statistical model is usually specified as a mathematical relationship between one or more random variables and other non-random variable and represents, often in considerably idealized form, the data-generating process” (Wikipedia).
- Includes
  - T-tests
  - ANOVA
  - Linear Regression
  - Generalized linear models
  - And many more…



Draw conclusions

A framework for critically evaluating your analysis before drawing conclusions.

1. Does the analysis answer the research question?

Confirm that your results directly address the original query you set out to investigate.

2. Are there limitations affecting the conclusions?

Identify any constraints or weaknesses in your analysis that might influence the results.

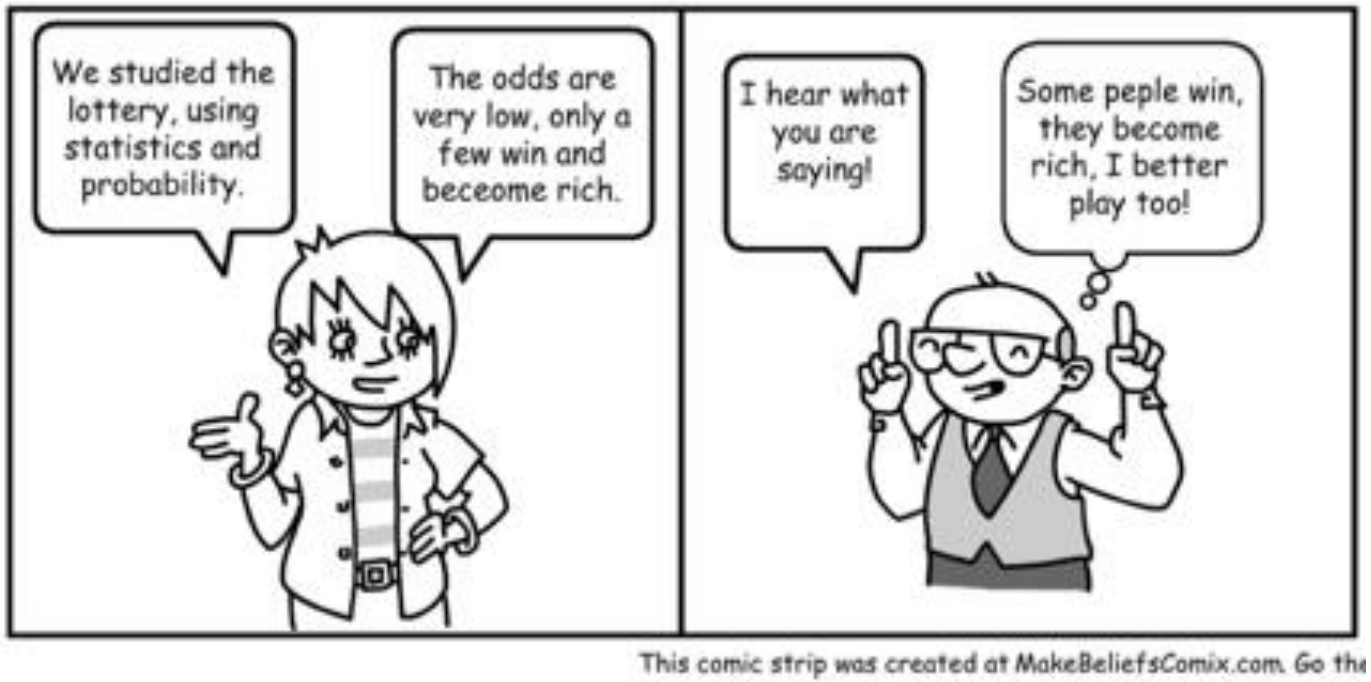
3. Is the analysis sufficient for decision-making?

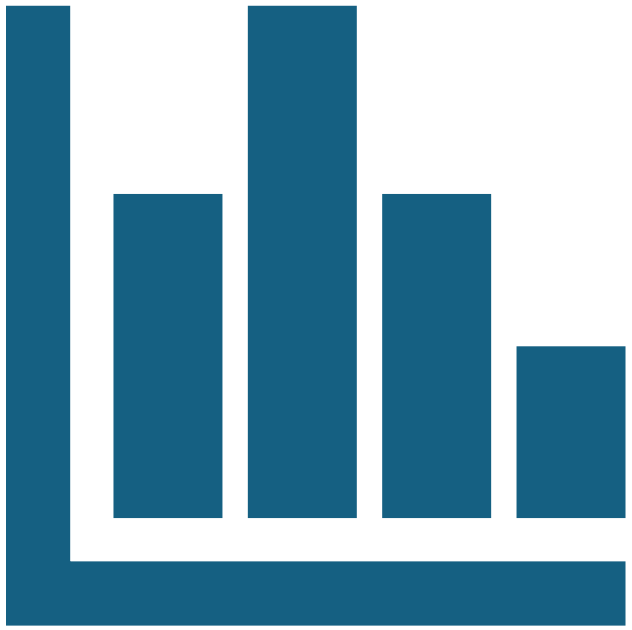
Determine if the findings provide enough clear insight to support and inform future actions.

Communicate the results

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

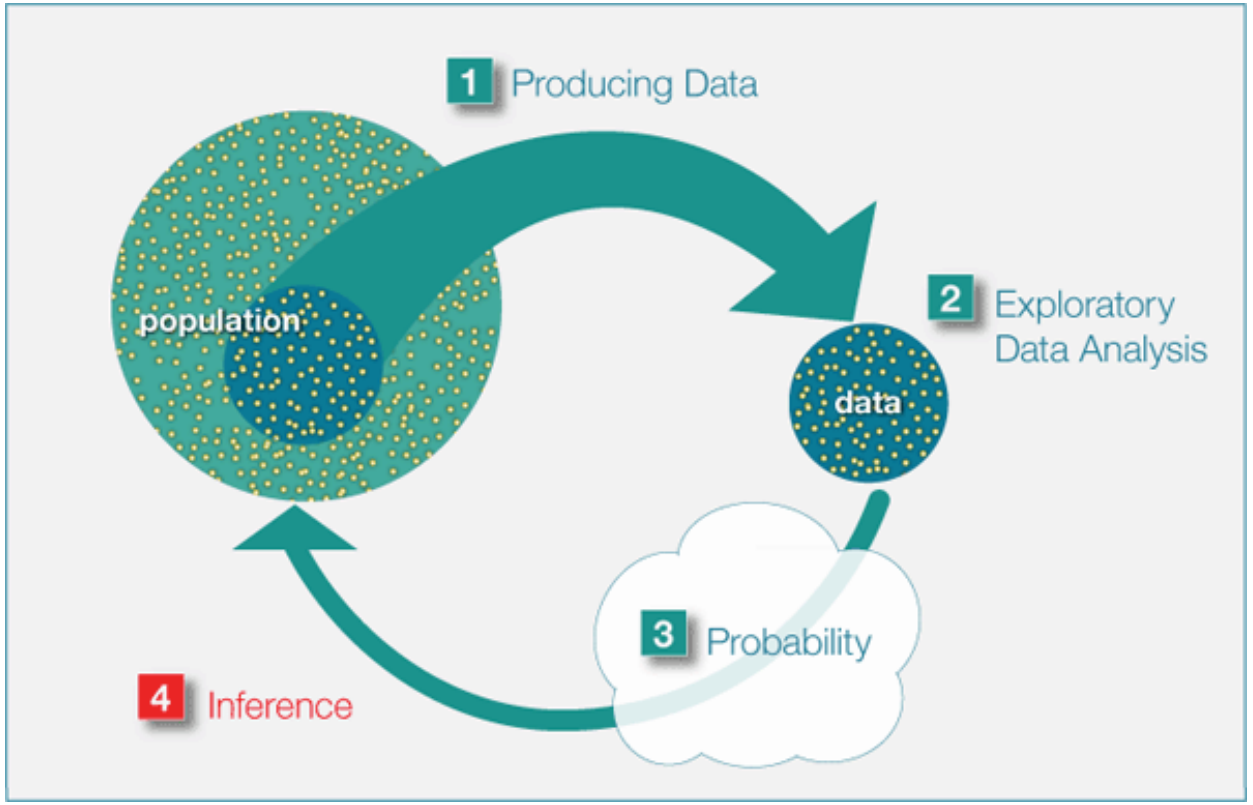
- Extra skills needed at this stage
  - Writing
  - Presenting
  - Communication





Statistical Inference

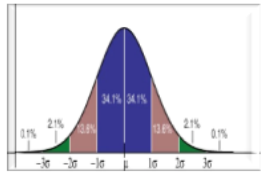
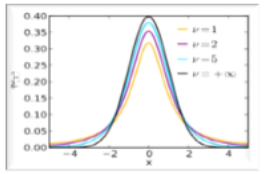
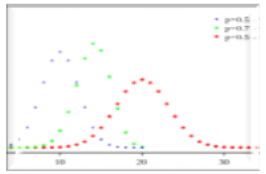
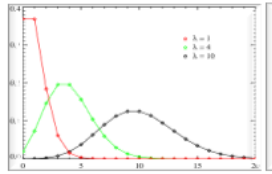
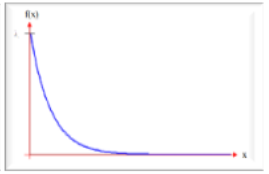
Aim and applications



- 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

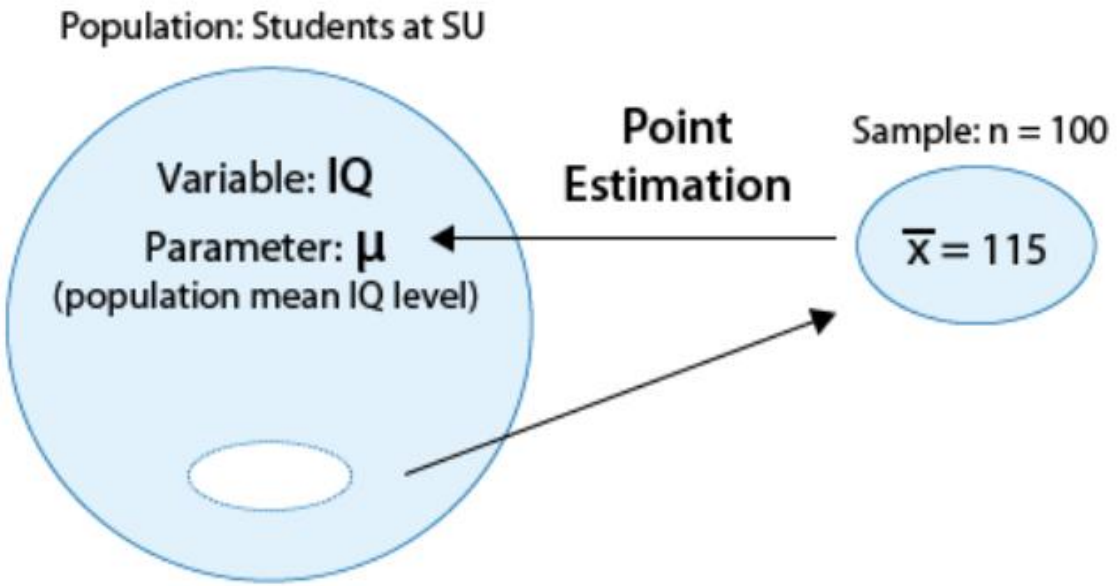
Statistical inference is based on probability distributions

	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.
  - 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



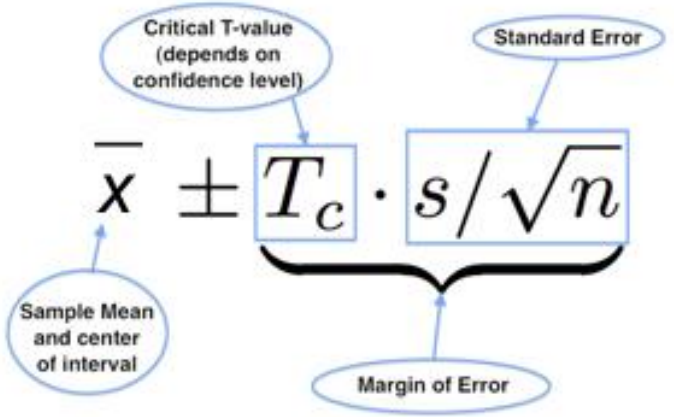
- In **point estimation**, we estimate an unknown parameter using a single number that is calculated from the sample data.

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

More on: UF-Open learning

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)

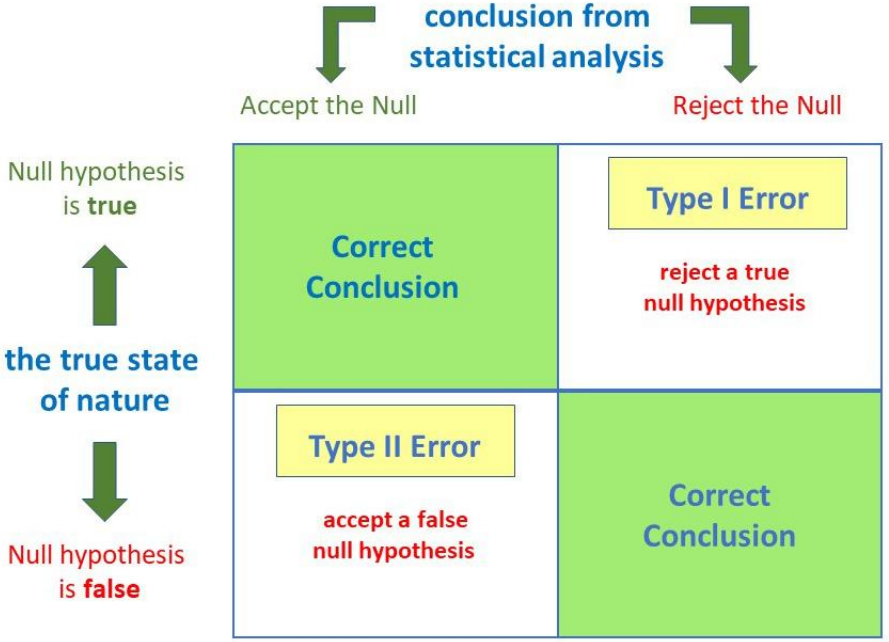
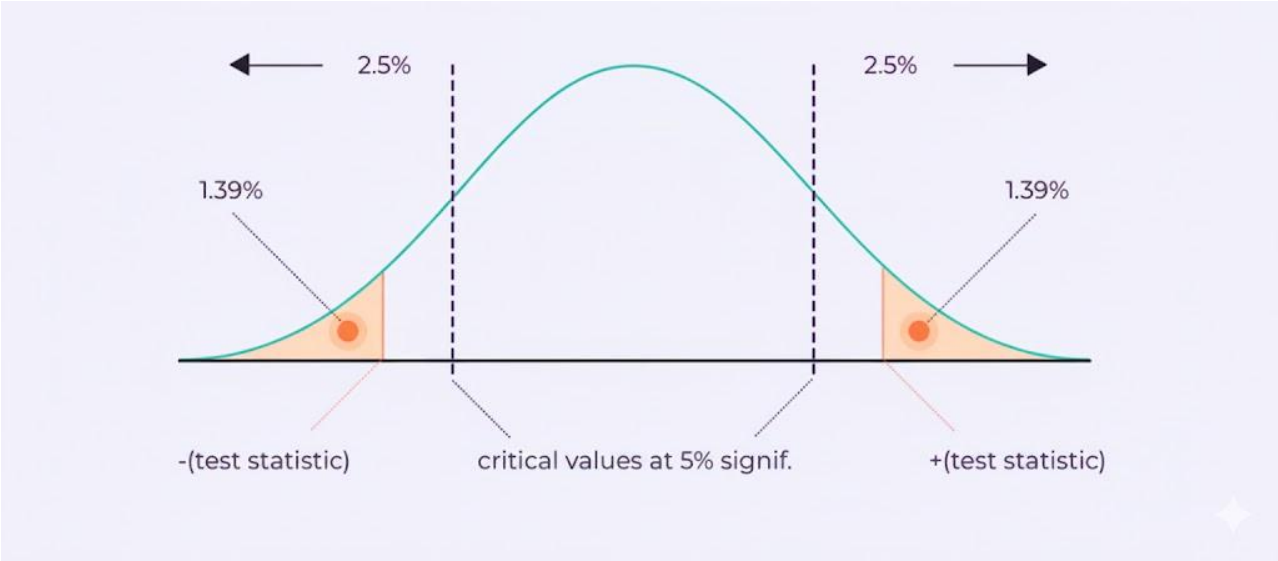


Image source: simplypsychology.org

P-value

- A p-value (probability value) in statistics measures the likelihood of observing your data (or more extreme data) if the null hypothesis (no real effect/difference) were true.

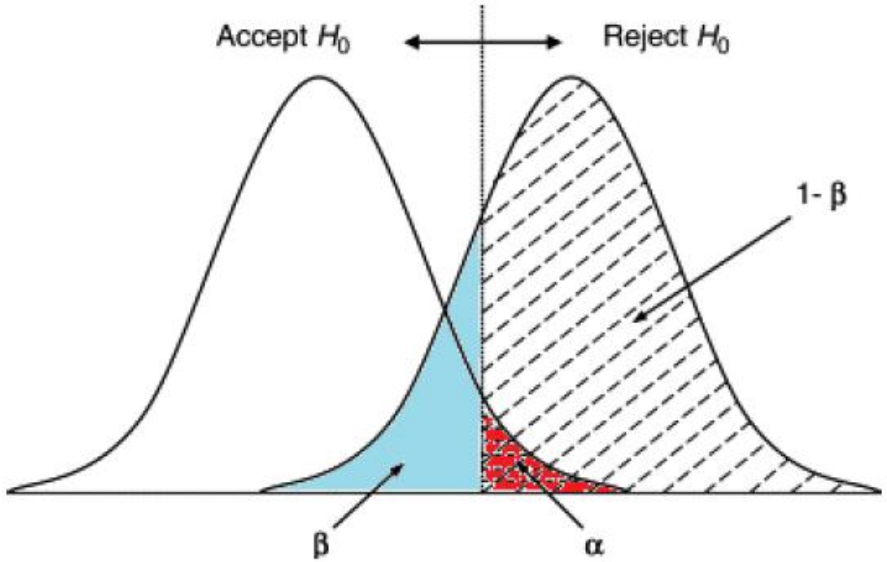


- A small p-value (e.g., < 0.05) suggests strong evidence against the null, implying the results aren't just random chance, while a large p-value means the data is consistent with the null hypothesis.

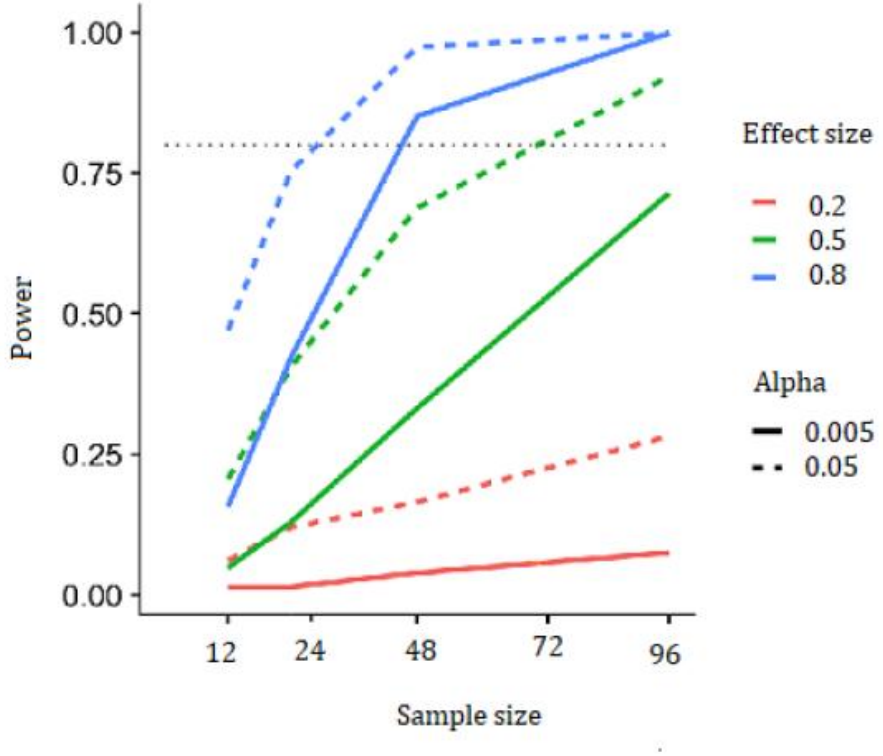
Power and sample size calculation

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-β



Power and sample size calculation

Sample Size Calculator

Home | About | Support

Standard tests

- Two-sample t-Test
- Paired t-Test
- Analysis of variance
- Wilcoxon Test
- Single proportion
- Chi-squared Test
- Fisher's exact Test
- McNemar's Test

Sample size for a two-sample t-test

Input and calculation

Mean difference: 1

Standard deviation: 2

Alpha two-sided: 0.05

Power: 0.8

Calculate

Press the Calculate button to calculate the sample size.

Copy result statement to clipboard

Options

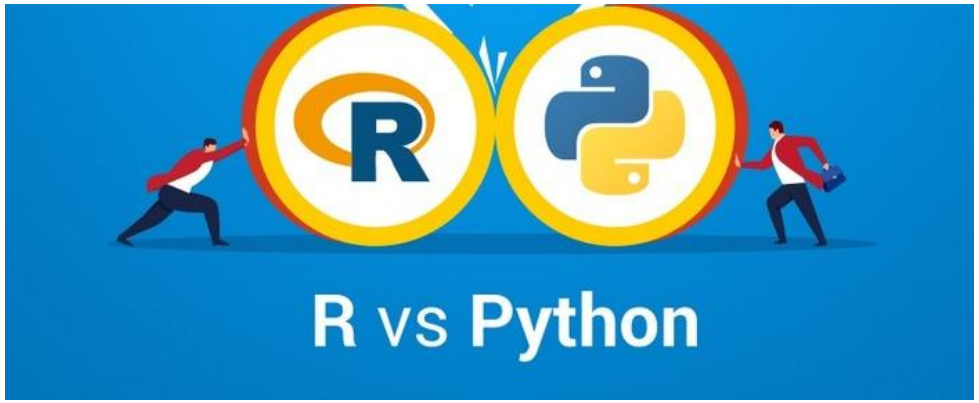
Calculate
- (•) Sample size
- ( ) Power (output decimal places: 4)

Specific options
- [ ] Unequal variances
- [ ] Calculate mean difference from group means

Advanced
- [ ] Unequal sample sizes
- [ ] Account for drop-outs
- [ ] Bonferroni correction

https://homepage.univie.ac.at/robin.ristl/samplesize.php

Software



Basic Statistical Tests

Type of data?

Numerical

How numerical data are organized

Two groups

Data distribution assumptions?

Parametric

t-test

Non-parametric

Mann-Whitney U Test

Paired data

Data distribution assumptions?

Parametric

Paired t-test

Non-parametric

Wilcoxon signed-rank test

More than two groups

Data distribution assumptions?

Parametric

ANOVA

Non-parametric

Kruskal-Wallis test

Relationship between multiple variables

Data distribution assumptions?

Parametric

Pearson correlation
Regression analysis

Non-parametric

Spearman correlation

Categorical

Target task?

Compare two variables

Chi-Square test of independence

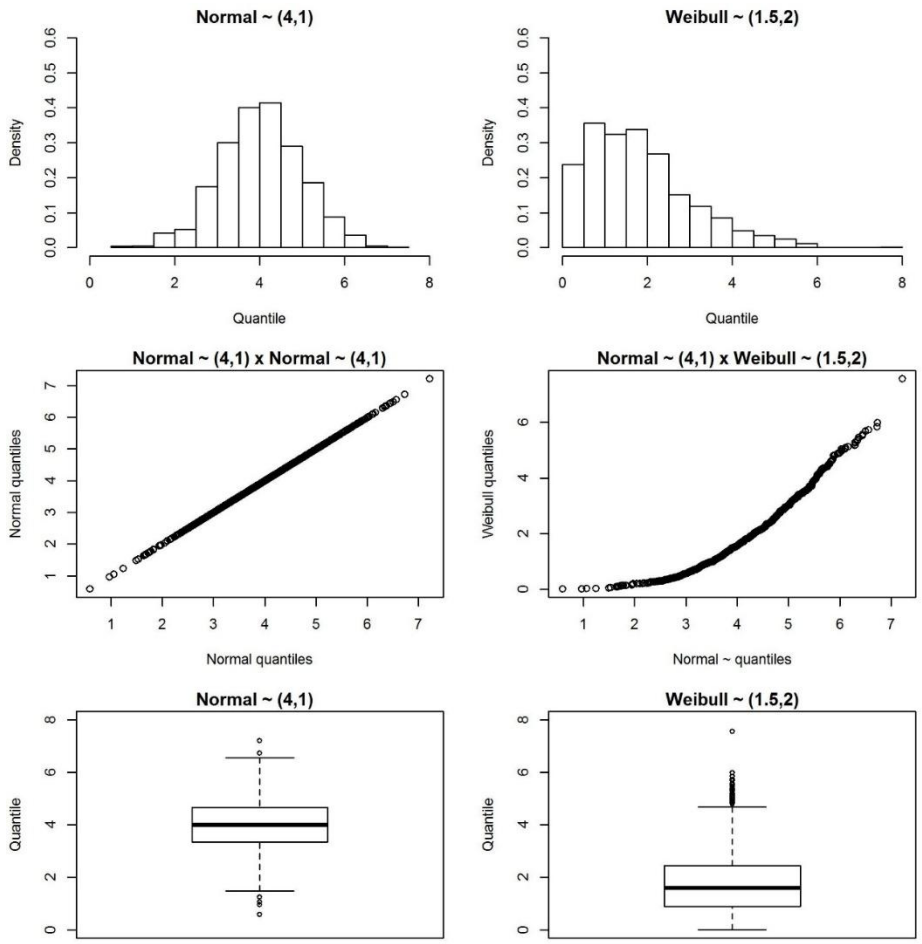
Test proportions

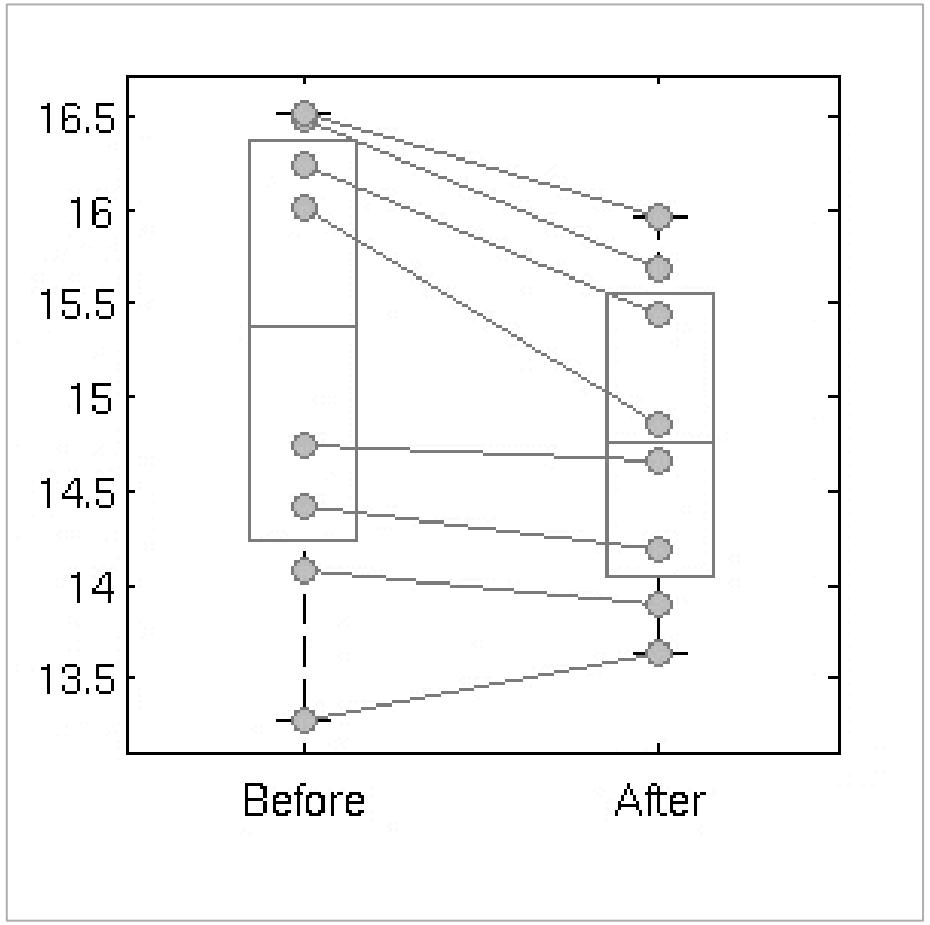
Z-test for proportions
Binomial test

https://www.statology.org/

How to test for normality?

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





Two related samples

Paired Samples t-test (Parametric)

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

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

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 reject H_0

Wilcoxon signed rank test (nonparametric)

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.

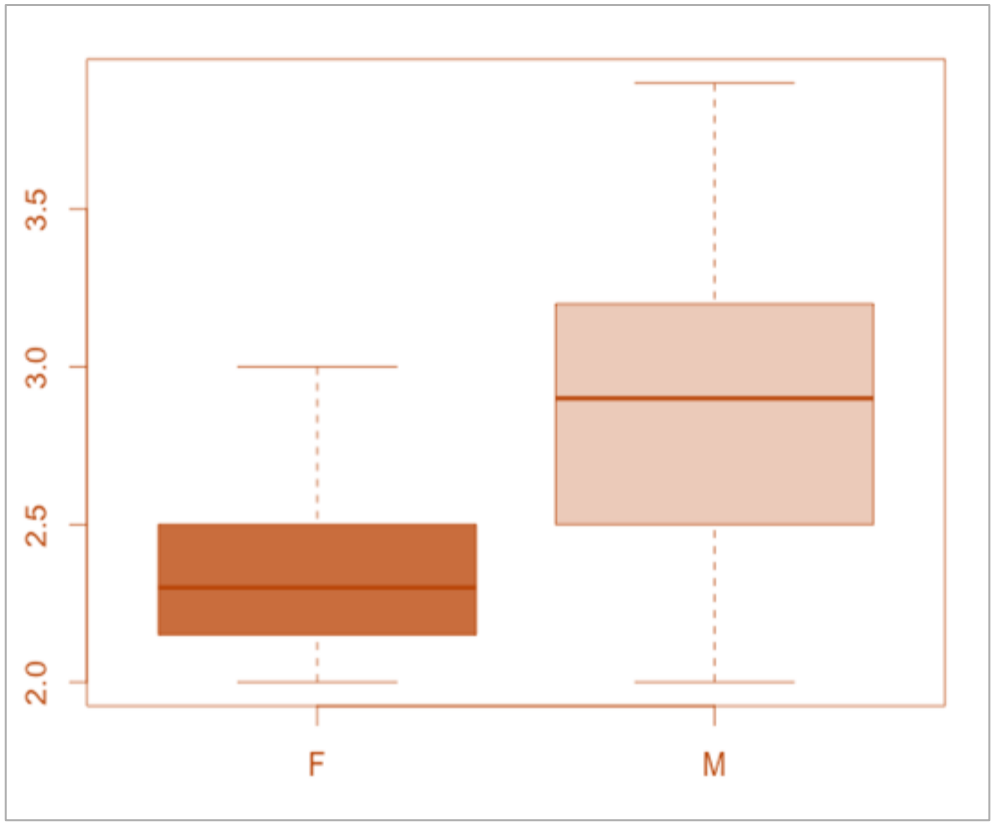
2. Compute the test statistics (in R)

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

3. Decision based on p-value

If $p<0.05$ reject Ho

If the difference is not symmetric around zero then there is difference between groups.



Two independent samples

Independent Samples t-test (parametric)

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

1. Hypotheses

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

2. Test statistics

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

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

- **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 with Levene test.

3. P-value < 0.05 reject H_0

Independent Samples t-test: Unequal variances

1. Hypotheses

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

2. Test statistics

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

Compute the approximate degrees of freedom d', where

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

3. Decision

P-value < 0.05 reject H_0

- The populations means are significantly different from each other.

Mann Whitney U test

1. Hypothesis

H_0: the distributions of both populations are equal

H_1: the distributions are not equal

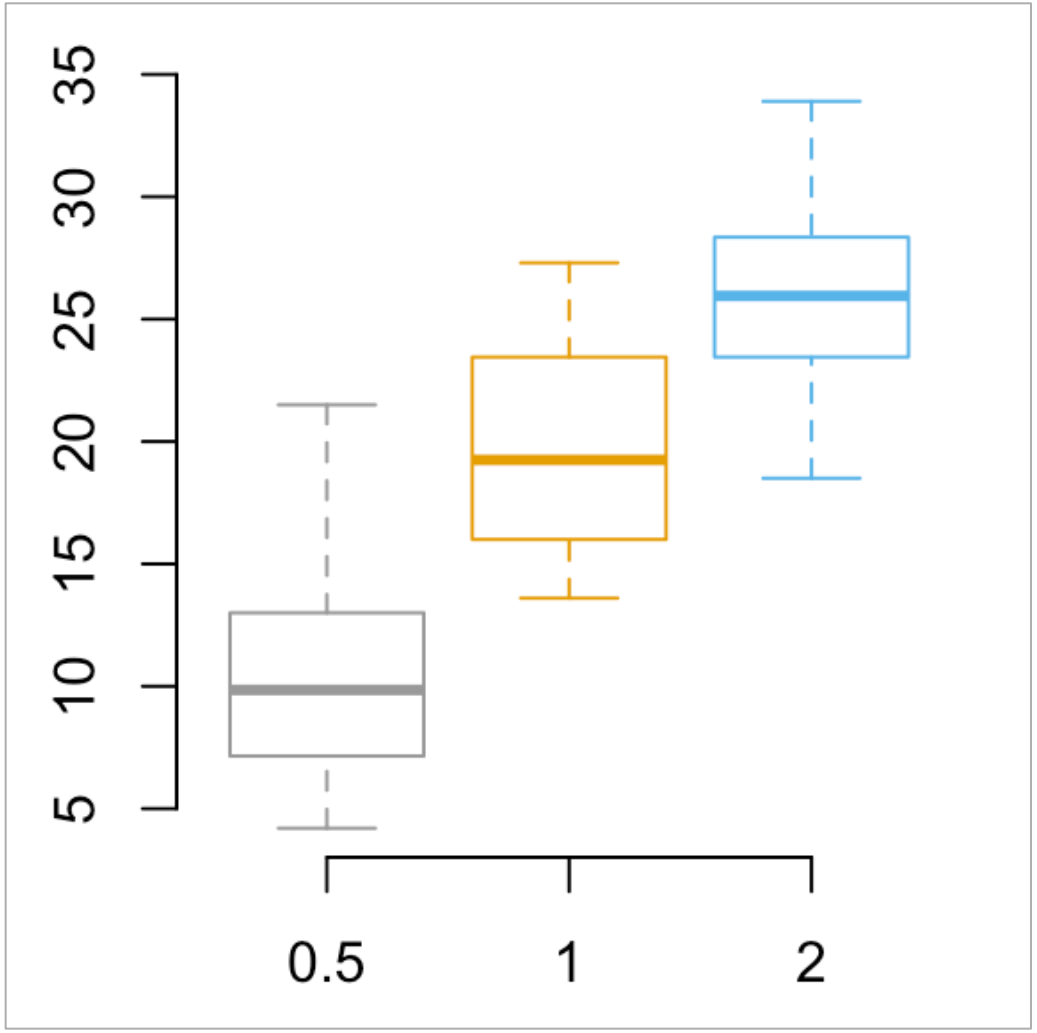
2. Compute the test statistics (in R)

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

3. Decision based on p-value

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

Two or more independent samples



One-way ANOVA (parametric)

- 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 (parametric)

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.

Kruskal-Wallis test

1. Hypothesis

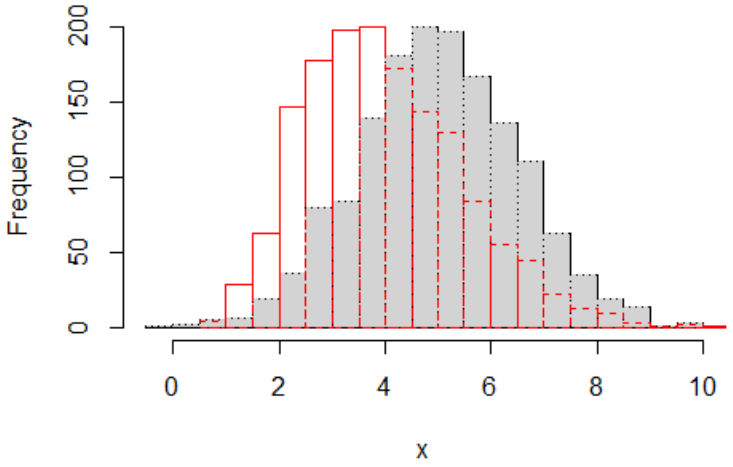
   Ho: the medians of all groups are equal, and

   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)

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

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



3. Decision based on p-value

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

Chi-square test for independence

- The chi-square test for independence, also called Pearson's chi-square test or the chi-square test of association, is used to discover if there is a relationship between two categorical variables.

1. Hypotheses

H_0: Variables are independent

H_1: Variables are related

2. Test statistics $X^2 = (O_{11} - E_{11})^2 / E_{11} + (O_{12} - E_{12})^2 / E_{12} + \cdots + (O_{RC} - E_{RC})^2 / E_{RC}$

- $\chi^2_{(R-1)(C-1)}$ *distribution*

- **Assumptions**

1. Your two variables should be measured at a nominal level (i.e., categorical data).
2. Your two variables should consist of two or more categorical, independent groups.
3. The expected counts should be larger than 5 in more than 75% of cases.

3. Decision

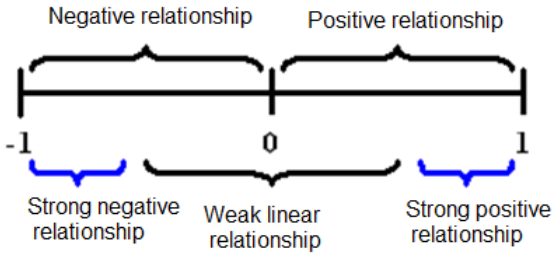
P-value < 0.05 reject H_0

- Variables are related.

Pearson Correlation

- The Pearson product-moment correlation coefficient (Pearson's correlation, for short) is a measure of the strength and direction of association that exists between two variables measured on at least an interval scale.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$



Cohen (1988):

|r| <0.3 Weak

0.3 ≤|r| < 0.5 Medium

|r| ≥0.5 Strong

- **Assumptions**

1. Your two variables should be measured at the interval or ratio level (i.e., they are continuous).
2. There is a linear relationship between your two variables. You can check by creating a scatterplot.
3. There should be no significant outliers.
4. Your variables should be approximately normally distributed

Spearman Correlation (nonparametric)

1. When normality assumption is violated use Spearman correlation.

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.

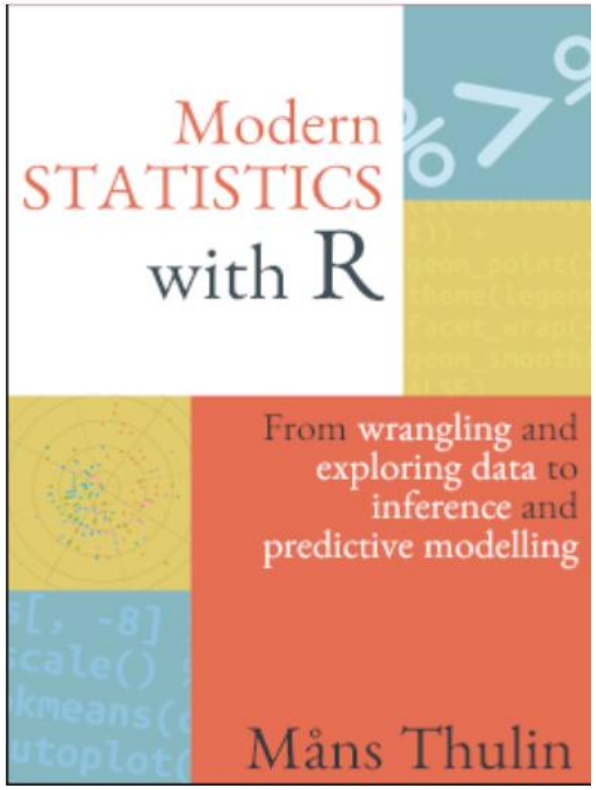
R application

Check R Notebooks:

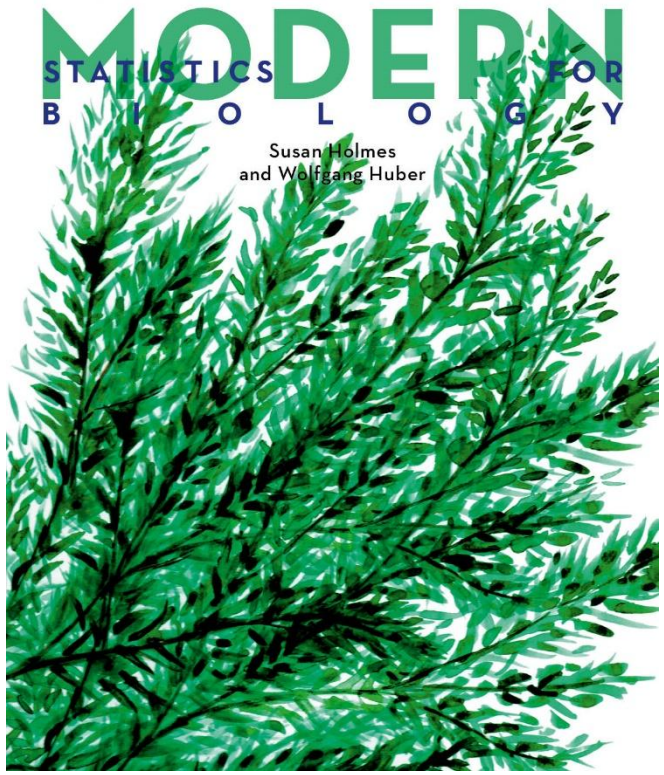
Parametric tests

Nonparametric tests

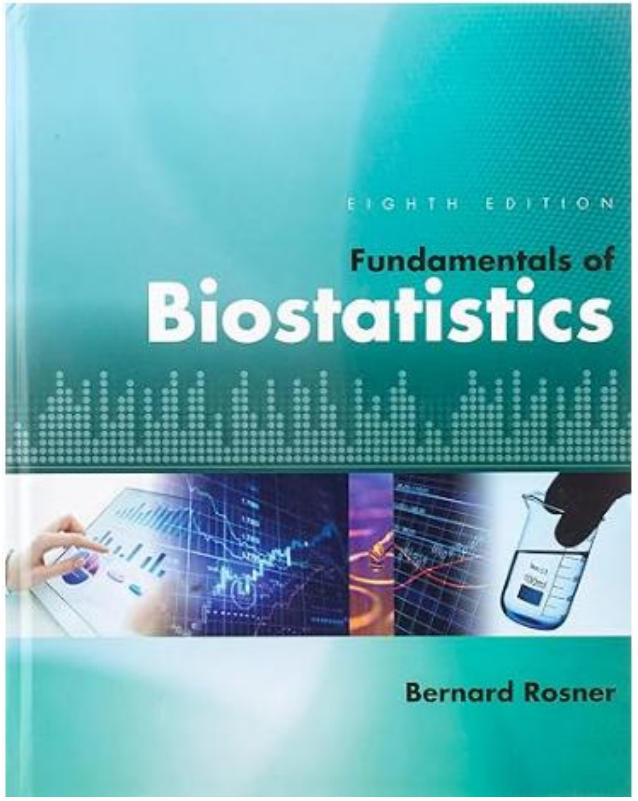
Recommended books



https://www.modernstatisticswithr.com/



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



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. http://www.biostathandbook.com/HandbookBioStatThird.pdf