HYPOTHESIS TEST

Bioinformatics Awareness Days @ TIGEM
July 11th, 2022

Eugenio Del Prete, M. Eng., Ph.D.
BIOINFORMATICS CORE
e.delprete@tigem.it

Bioinformatics Core: Tasks

- **STATISTICAL DATA ANALYSIS**
  Experimental Design, Hypothesis Testing, Differential Expression Analysis, Cluster Analysis, Time Series Data Analysis, Survival Analysis, Correlation Analysis
- **OMICS**
  Microarray Analysis, Gene Networks, Pathway Analysis, TFBS Identification, Gene Annotation, Integration, Protein Analysis, Drug Networks
- **NEXT GENERATION SEQUENCING**
  Whole Exome, Targeted Gene, RNA, miRNA, ChIP, Visualization, Interpretation
- **DATABASE AND SOFTWARE**
  DB Creation, DB Maintenance, Web Sites Creation, Web Service Support
- **BIOINFORMATICS AND (BIO)STATISTICS TRAINING**

Bioinformatics Core: People

DIEGO DI BERNARDO

https://www.tigem.it/research/facilities/core-facilities/bioinformatics

https://bioinformatics.tigem.it/

DIEGO CARRELLA

ROSSELLA DE CEGLI

SERGIO SARNATARO

EUGENIO DEL PRETE

Bioinformatics Core: Something about Me

- **TLC ENGINEER** @ UNIVERSITY OF ROME 'SAPIENZA'
  MAIN TOPICS: Signal Processing, Remote Sensing, Bioinformatics
  THESIS: miRNA Analysis, Genomic Data Mining, Consensus Analysis, PSSM Creation
- **BIOINFORMATICS RESEARCH FELLOW** @ INSTITUTE OF FOOD SCIENCES (CNR)
  Protein Prediction and Classification, Protein Analysis, Proteomic Mass Spectra Analysis, Sequence Alignment and Phylogenetic Tree, Docking
- **PHD IN APPLIED BIOLOGY** @ UNIVERSITY OF BASILICATA
  Celiac Disease and Comorbities, Microarray Data Analysis, Ontologies, Gene Set Enrichment Analysis, Semantic Similarity, Proteomic Mass Spectra Analysis
- **BIOINFORMATICS RESEARCH FELLOW** @ INSTITUTE OF APPLIED MATHEMATICS (CNR)
  Proteomic Mass Spectra Analysis, Metabolomic (Lipidomic) Data Analysis, Web Tools Developer, Hypothesis Tests, Omics Data Integration
- **BIOSTATISTICIAN AND DATA ANALYST** @ TIGEM

Outline

UNCERTAINTY AND VARIABILITY

- Descriptive Statistics
- Uncertainty and Variability
- Measurement

HYPOTHESIS TESTING

- Inferential Statistics
- Hypothesis Testing: What, How, Errors, Which
- Multiple Test Correction

EXAMPLES

- Example One
- Example Two
- Example Three

CONCLUSION

- Take Home Message
- Final Remarks

Not Only Aphorism...

Trilussa (1871 - 1950)
Carlo Alberto C. M. Salustri

Poet, Writer, Journalist

LA STATISTICA

Sai che d'è la statistica? È 'na cosa
che serve pé fa' un conto in generale
de la gente che nasce, che sta male,
che more, che va in carcere e che sposa.
Ma pé me la statistica curiosa
è dove c'entra la percentuale,
pé via che, lì, la media è sempre eguale
puro co' la persona bisognosa.
Me spiego: da li conti che se fanno
secondo le statistiche d'adesso
risurta che te tocca un pollo all'anno:
e, se nun entra ne le spese tue,
t'entra ne la statistica lo stesso
perché c'è un antro che ne magna due.

Statistics

Science

- Study of collective and measurable phenomena, with **quantifiable** data
- Answer to a **well-posed** question to find a solution, with a degree of **uncertainty**
- Application of mathematical principles and techniques to **learn** from data

Workflow

PLANNING → COLLECTING → ANALYZING → DEDUCING



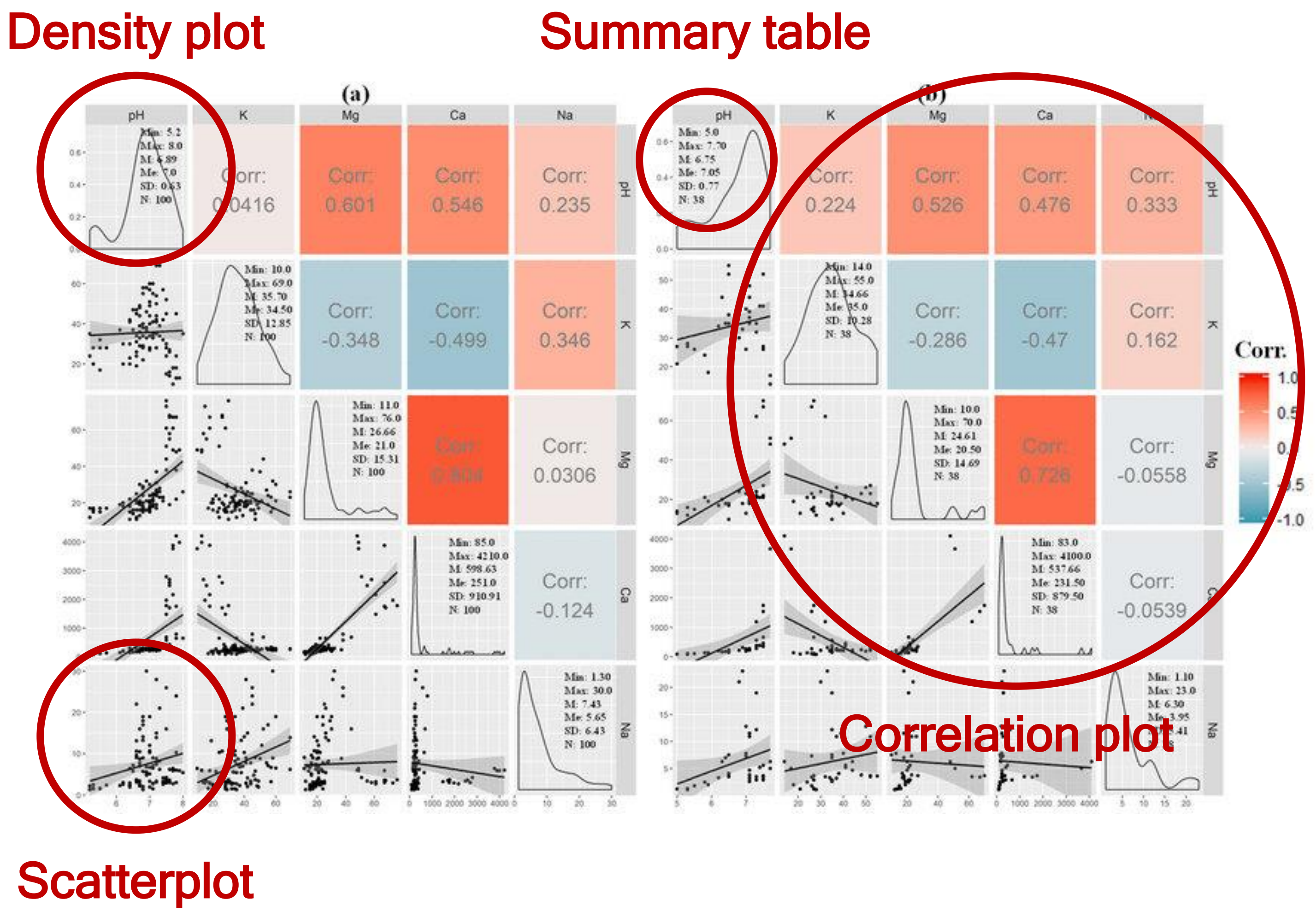

Descriptive Statistics

Descriptive Statistics

- **Description** of the features for a specific dataset
- **Summary** of the information from a specific dataset

Description Tools

- **Plots**: barplot, boxplot, pie chart, scatter plot, density plot, correlation plot
- **Tables**: descriptive table, summary table



Uncertainty

Uncertainty

- COMMON SENSE: not known beyond doubt, not having complete knowledge
- STATISTICAL: probability and repeatability

Example: Coin Flip

a) Flip the coin **10 times**: H, H, H, T, T, T, T, H, H, H
b) Calculate percentage: H 60%, T 40%
c) Flip the coin **1000 times** (1000 >>10)
d) Calculate percentage: H 54%, T 46 %

(Strong) Law of Large Numbers

I) $X_1, X_2, \ldots, X_n$ is an infinite sequence of independent and identical distributed random variables

II) Expected values $E(X_1) = E(X_2) = \ldots = E(X_n) = \mu$ and sample average $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$

then

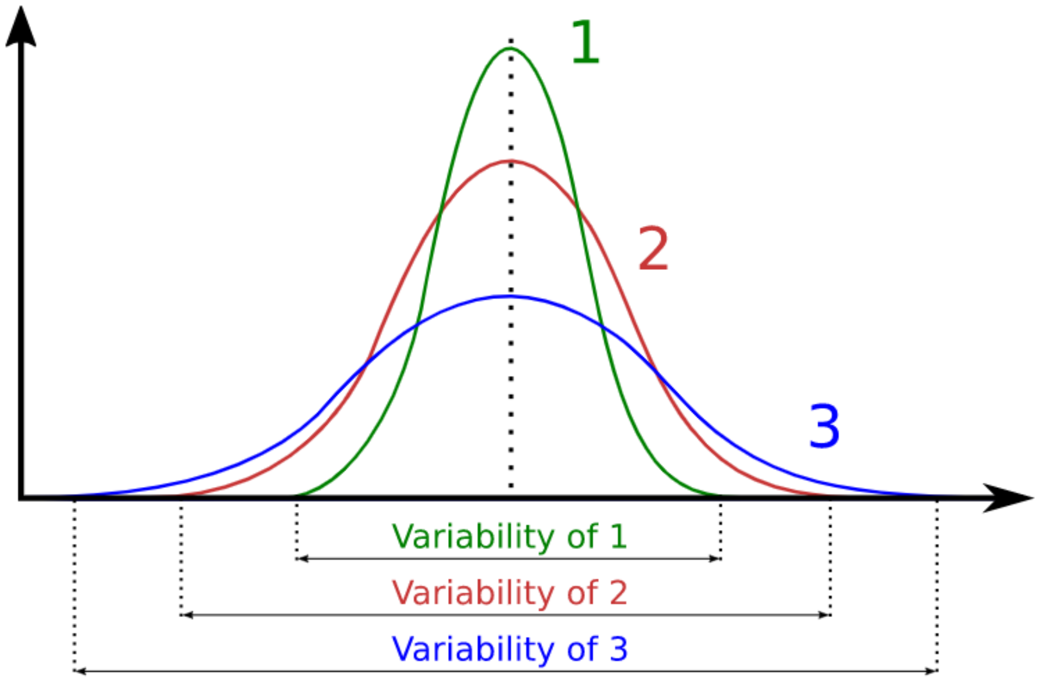
$$\bar{X}_n \to \mu \text{ when } n \to \infty$$

Variability

Variability

- COMMON SENSE: different values in a particular condition
- STATISTICAL: divergence of data from its mean value (spread, dispersion)

Normal distribution



Sample Mean: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_n$

Sample Variance: $\sigma_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

Sample Error: $\sigma_{\bar{X}} = \frac{\sigma_n}{\sqrt{n}}$

WHAT ABOUT σ_n?
WHAT ABOUT n?

(See GALTON'S BOARD)

Measure of Central Tendency

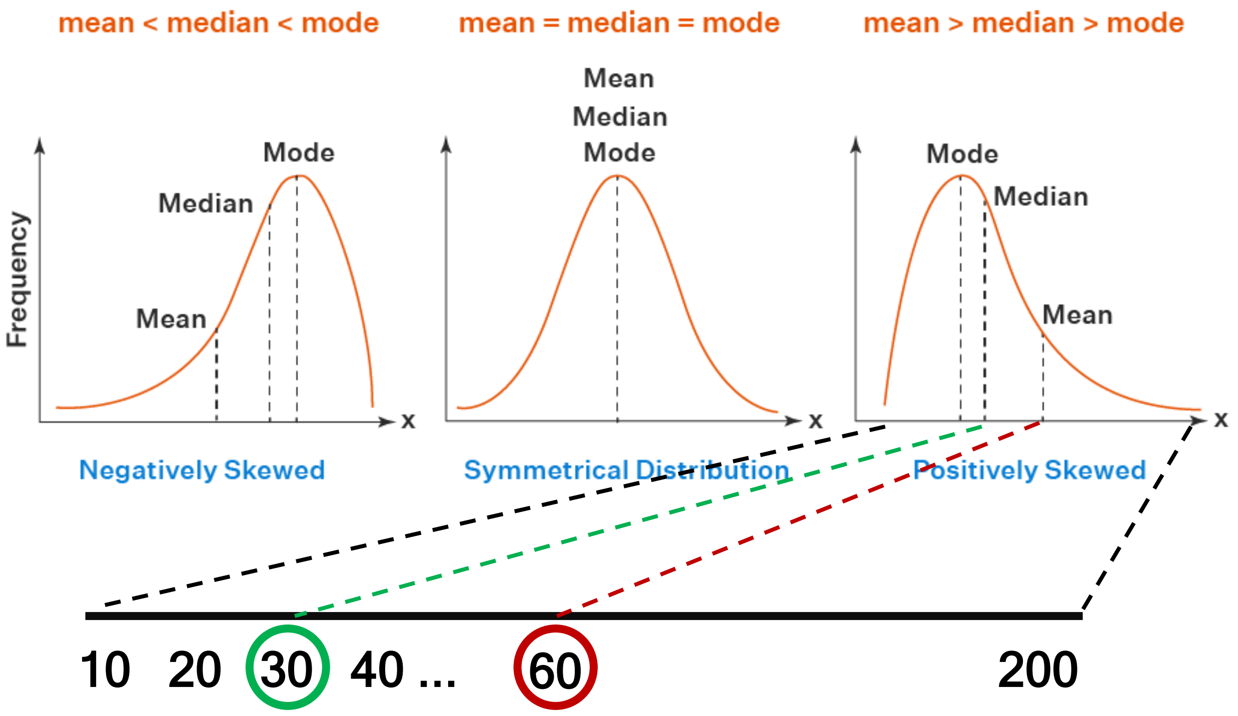
- **Mode**
  Most frequent value in the data set

  (nominal data)

- **(Arithmetic) Mean**
  Sum of all measurements divided by the number of observations in the data set

- **Median**
  Middle value that separates the higher half from the lower half of the data set

  (ordinal data)



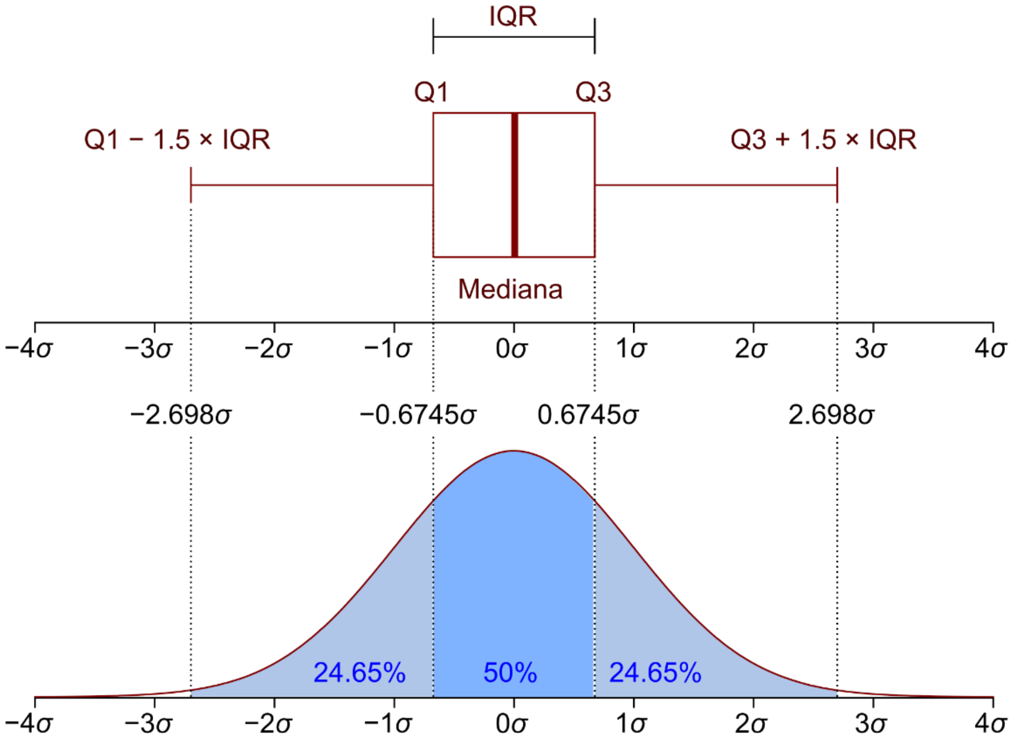
Measure of Variability

- **Range**
  Difference between the smallest and the largest value in the data set
- **Standard Deviation (SD)**
  How data is spread out going from the mean
- **Coefficient of Variation (CV)**
  Relative dispersion of data around the mean

$$c_v = \frac{\sigma}{\mu}\ (x\ 100)$$

- **InterQuartile Range (IQR)**
  How widespread the interval is, in which the middle 50 % of all the values lie

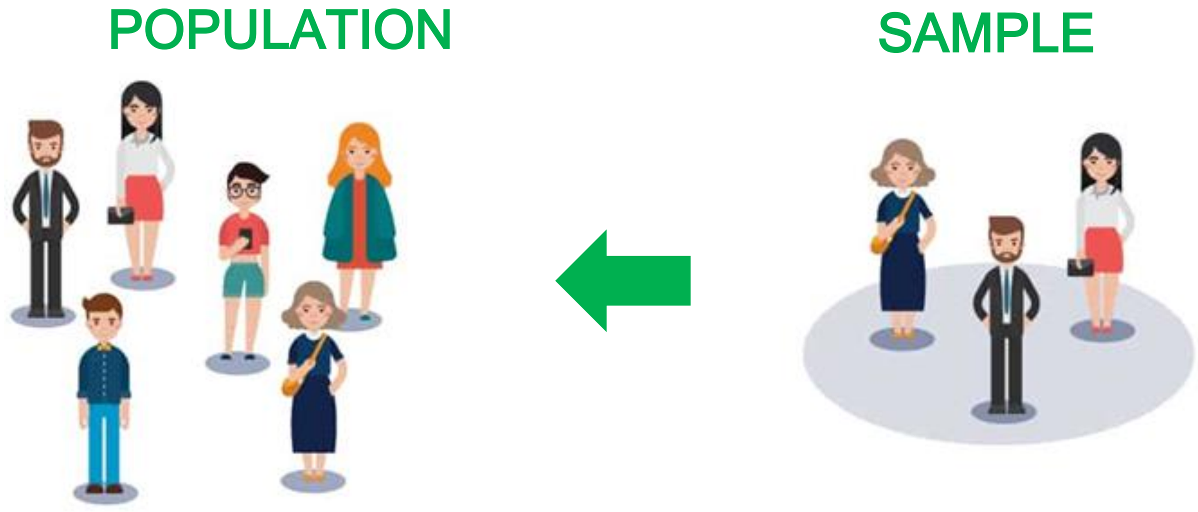
- SD is the square root of sample variance
- CV is a normalization (dimensionless)



Inferential Statistics

Inferential Statistics

- **Assumption** from the features of a specific dataset and **validation**
- Statistical methods for inferring the characteristics of a population (**parameter**) from a sample (**statistic**)



Estimation

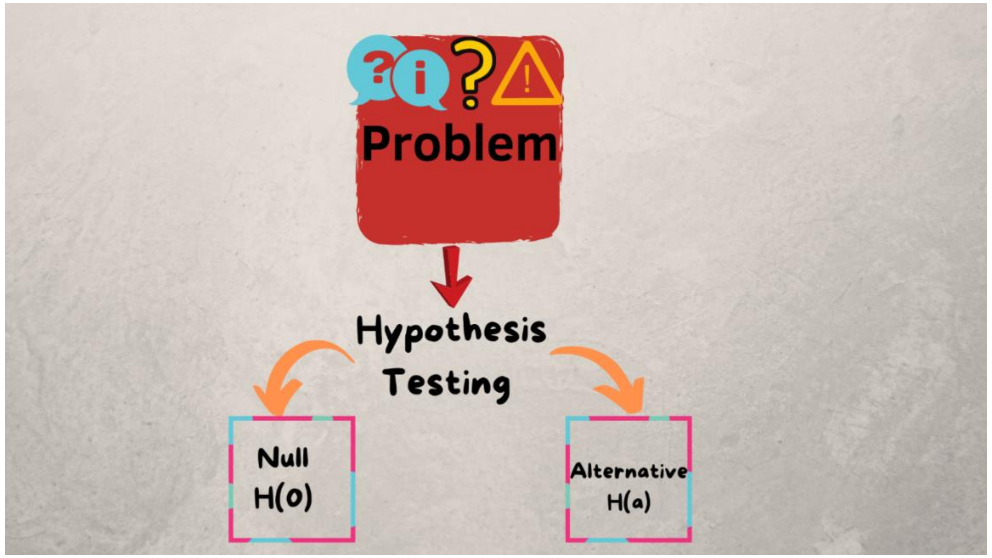
- Measure a statistic from the sample
- Generalize to the population:

a) approximate estimation (**margin of error**)
b) sample ≠ population (**probability of error**)

Hypothesis Testing: What and How

Hypothesis Testing

- An analyst **tests** an assumption regarding a population parameter
- The methodology employed depends:
  - a) on the **nature of the data** used
  - b) on the **reason for the analysis**



How to test a hypothesis

1. **State null hypothesis H0**
   Children who take vitamin C are no less likely to become ill during flu season
2. **State alternative hypothesis H1**
   Children who take vitamin C are less likely to become ill during flu season
3. **Determine significance level α**
   Percentage of error be willing to accept (5%)
4. **Calculate H0 probability p-value**
   One group with vitamin C during flu season and the other with a placebo. Collecting a p-value of 0.1
5. **Reject or not H0**
   P-value > α, H0 cannot be rejected

Hypothesis Testing: Types of Error

- H0: LEMUR (NOT SLOTH)

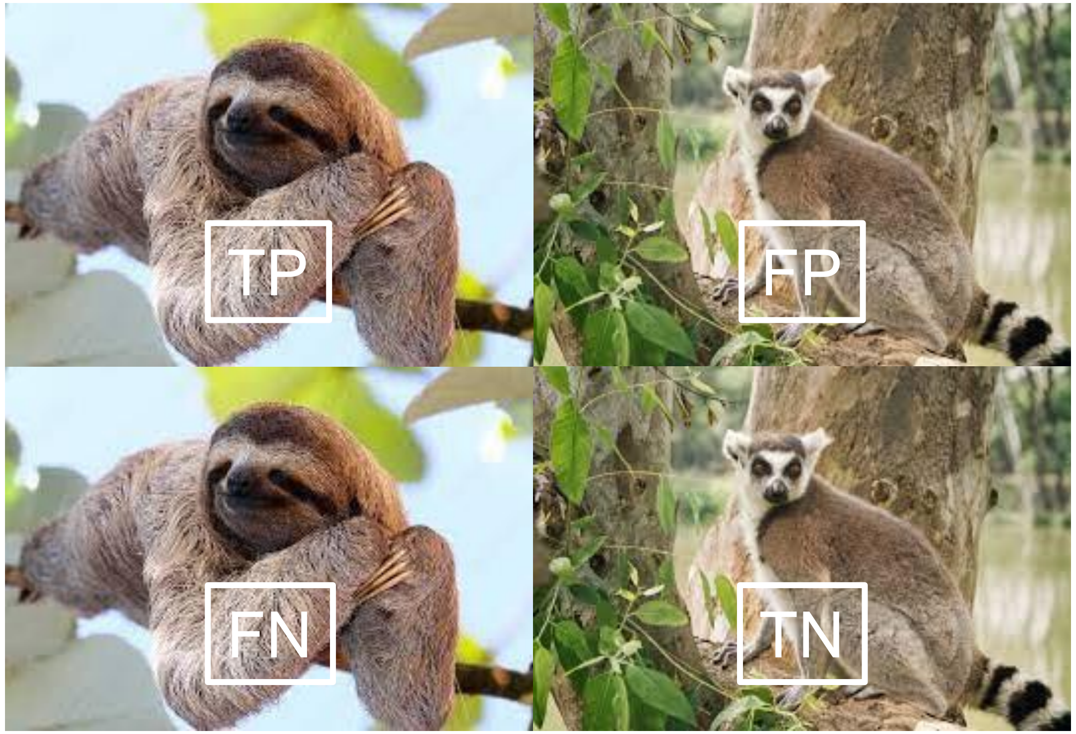
REALITY

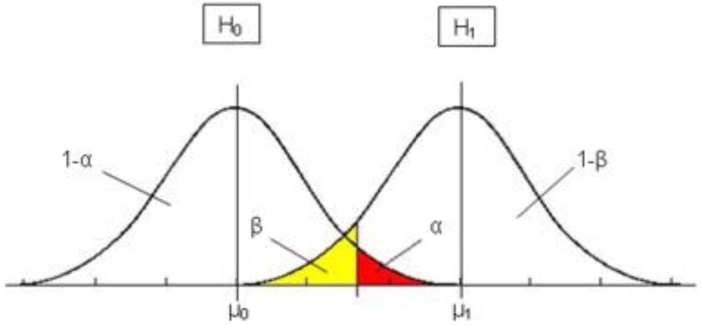
H0 FALSE | H0 TRUE

STUDY FINDINGS

REJECT H0

ACCEPT H0





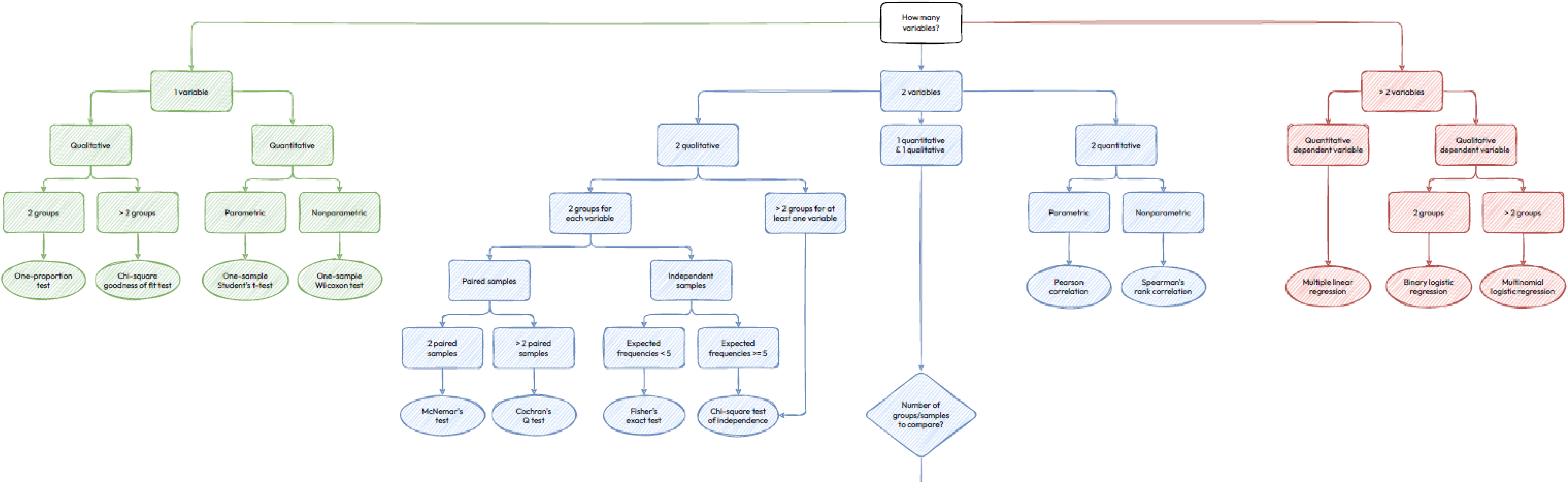
TRUE POSITIVE (TP) - POWER (1-β)
Probability to REJECT H0 when H0 is FALSE

FALSE POSITIVE (FP) - TYPE I ERROR, α
Probability to REJECT H0 when H0 is TRUE

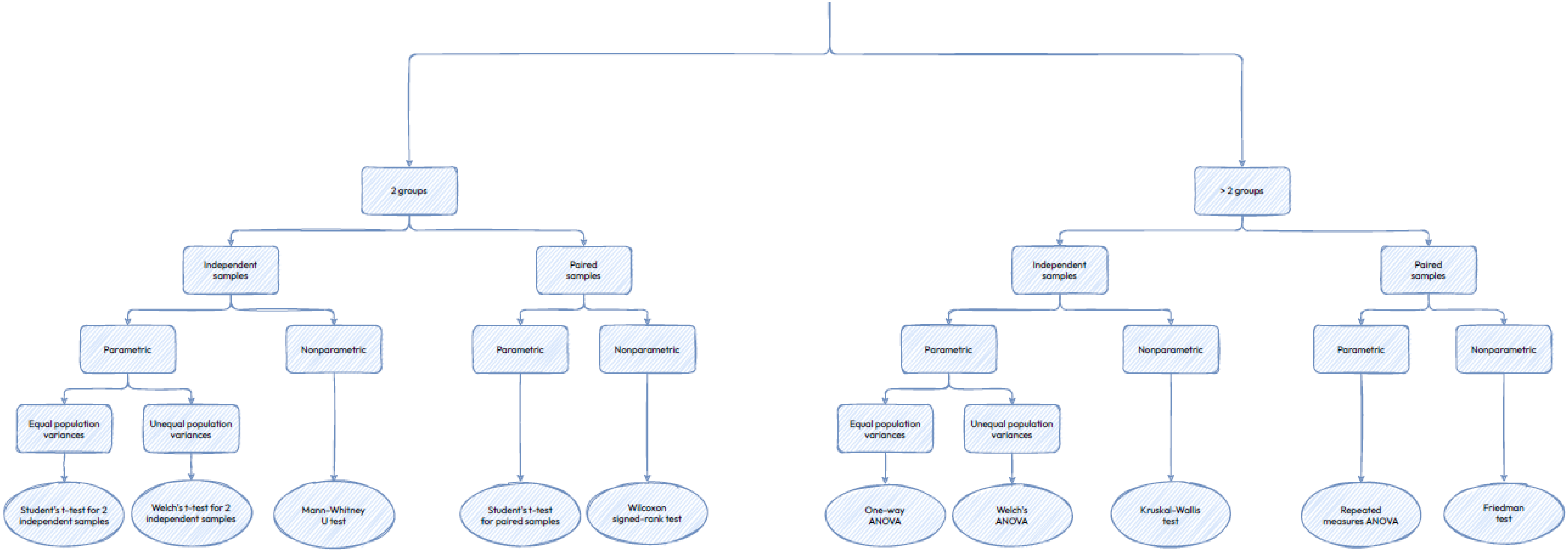
FALSE NEGATIVE (FN) - TYPE II ERROR, β
Probability to ACCEPT H0 when H0 is FALSE

TRUE NEGATIVE (TN)
Probability to ACCEPT H0 when H0 is TRUE

Hypothesis Testing: Which



Hypothesis Testing: Which



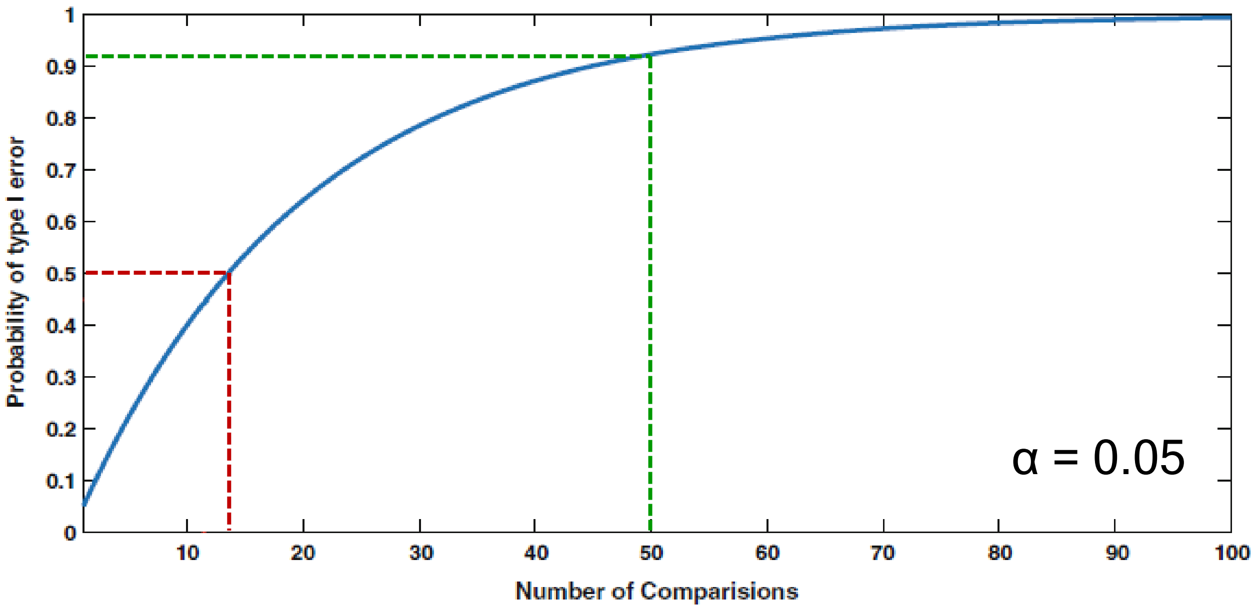
Hypothesis Testing: Which

- QUESTION 1: Which kinds of variable? CONTINUOUS, DISCRETE, CATEGORICAL
- QUESTION 2: How many groups per variable? 1 GROUP, 2 GROUPS, > 2 GROUPS
- QUESTION 3: Are the samples paired? UNPAIRED, PAIRED
- QUESTION 4: Are the distributions normal? PARAMETRIC, NON-PARAMETRIC
- QUESTION 5: Have the distributions the same variance? HOMOSCEDASTICITY

Hypothesis Testing: Multiple Test Correction

Probability of At Least One Type I Error

$$Pr(\alpha|m) = 1 - (1-\alpha)^m$$



α = 0.05, m = 1 → $Pr(\alpha = 0.05|m = 1) = 0.05$

α = 0.05, m = 13 → $\boldsymbol{Pr(\alpha = 0.05|m = 13) = 0.49}$

α = 0.05, m = 50 → $\boldsymbol{Pr(\alpha = 0.05|m = 50) = 0.92}$

Two Methods

Hard Correction: BONFERRONI

- $p_B = p_j m < \alpha$
- Control **Family-Wise Error Rate**
- Loss of power due to large number of tests

Soft Correction: BENJAMINI-HOCHBERG

- $p_1 < p_2 < p_j \ldots < p_m \rightarrow p_{BH} = \frac{p_j m}{j} < \alpha$
- Control **False Discovery Rate**
- Flexible procedure

Example: Identification of Differential Expressed Genes from RNASeq Data

Example One

- Suppose to test if the **average weight** of **10 mice** differs from **25 mg**

Dataset	
Name	Weight
M_1	20.6
M_2	20.0
M_3	20.4
M_4	22.0
M_5	19.9
M_6	20.7
M_7	18.8
M_8	20.5
M_9	20.4
M_10	23.3

- QUESTIONS
  a) Which kinds of variable?
  b) How many groups per variable?
  c) Are the samples paired?
  d) Are the distributions normal?
  e) Have the distributions the same variance?
  f) Multiple test correction?

ANSWERS
a) One continuous variable
b) One group
c) No (one measurement)
d) Yes
e) Not relevant
f) No (one test)

d) D'Agostino-Pearson test (0.1139), Shapiro-Wilk test (0.1634) → Answer: Yes

- TEST: **One sample Student's t-test** (< 0.0001, ****) → Answer: Yes

Example Two

- Suppose to test if **two different treatments** affect the weight of the mice

- QUESTIONS
  a) Which kinds of variable?
  b) How many groups per variable?
  c) Are the samples paired?
  d) Are the distributions normal?
  e) Have the distributions same variance?
  f) Multiple test correction?

ANSWERS
a) Two variables: one continuous, one categorical
b) Three groups for categorical
c) No (experimental design)
d) Yes
e) Yes
f) Yes (three test)

d) D'Agostino-Pearson test (0.8898, 0.6164, 0.6025), Shapiro-Wilk test (0.9567, 0.9304, 0.9410) → Answer: Yes

e) Brown-Forsythe test (0.3412), Bartlett's test (0.2371) → Answer: Yes

- TEST: **One way ANOVA** (0.0159, *)
  CTRL vs TRT1 (0.3909) → No, CTRL vs TRT2 (0.1980) → No,
  TRT1 vs TRT2 (0.0120) → Yes

Dataset

ID	Weight	Group
M_1	24.17	CTRL
M_2	25.58	CTRL
M_3	25.18	CTRL
M_4	26.11	CTRL
M_5	24.50	CTRL
M_6	24.61	CTRL
M_7	25.17	CTRL
M_8	24.53	CTRL
M_9	25.33	CTRL
M_10	25.14	CTRL
M_11	24.81	TRT1
M_12	24.17	TRT1
M_13	24.41	TRT1
M_14	23.59	TRT1
M_15	25.87	TRT1
M_16	23.83	TRT1
M_17	26.03	TRT1
M_18	24.89	TRT1
M_19	24.32	TRT1
M_20	24.69	TRT1
M_21	26.31	TRT2
M_22	25.12	TRT2
M_23	25.54	TRT2
M_24	25.50	TRT2
M_25	25.37	TRT2
M_26	25.29	TRT2
M_27	24.92	TRT2
M_28	26.15	TRT2
M_29	25.80	TRT2
M_30	25.26	TRT2

Example Three

- Suppose to test if **one treatment** affect the weight of the (**same**) mice

- QUESTIONS
  - a) Which kinds of variable?
  - b) How many groups per variable?
  - c) Are the samples paired?
  - d) Are the distributions normal?
  - e) Have the distributions same variance?
  - f) Multiple test correction?

ANSWERS
- a) Two variables: one continuous, one categorical
- b) Two groups for categorical
- c) Yes (experimental design)
- d) No
- e) Not relevant
- f) No (one test)

Dataset

ID	Before	After
M_1	20.01	39.29
M_2	19.09	39.32
M_3	19.27	34.51
M_4	21.30	39.30
M_5	24.14	43.40
M_6	19.69	42.79
M_7	17.22	42.20
M_8	18.55	38.39
M_9	20.52	39.23
M_10	19.37	35.22

d) D'Agostino-Pearson test (0.0445, 0.8714), Shapiro-Wilk test (0.2768, 0.2894) → Answer: No

- TEST: **Paired-sample Wilcoxon test** (0.002, **) → Answer: Yes

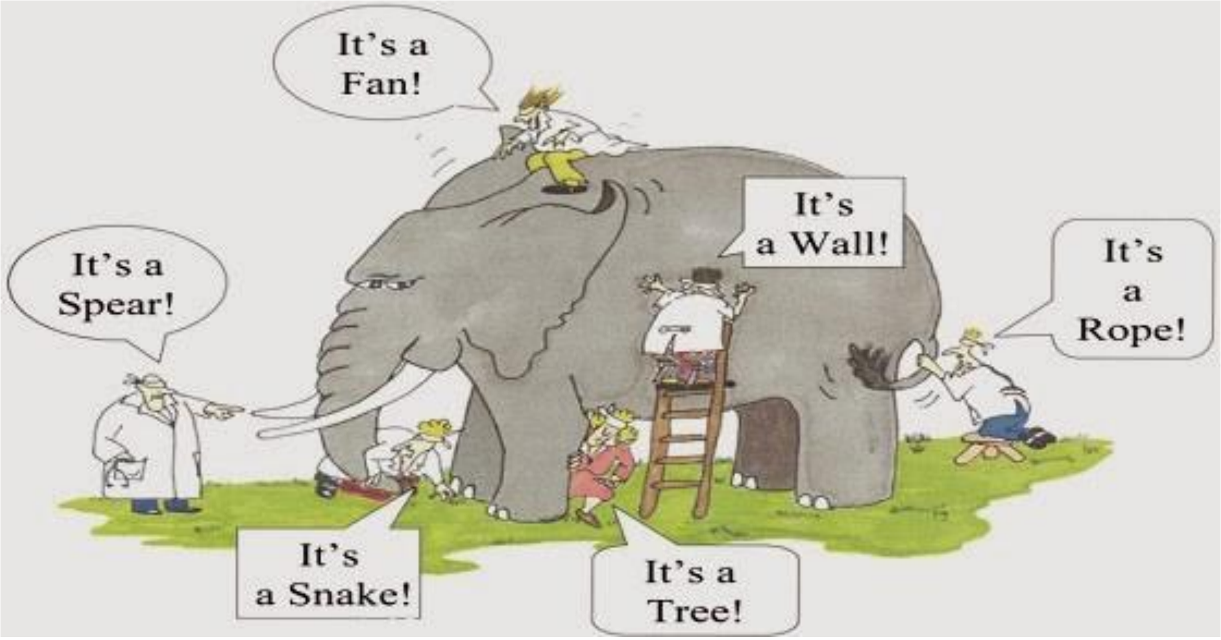
Take Home Message

- **Descriptive statistics** lend **inferential statistics** the quantities of interest
- Inferential statistics is correlated with the concept of **error**, because a sample **approximates the population**
- Type I error (α) and type II error (β) have a reverse trend: if it is possible, **increment the sample size**
- Select the **hypothesis test** corresponding to the actual experimental design, and correct for **multiple comparisons**
- Check the assumptions for selecting a **parametric** or **non-parametric test**

Final Remarks

To consult the statistician after an experiment is finished is often merely to ask him to conduct a postmortem examination. He can perhaps say what the experiment died of.

Sir R. A. Fisher
First Session of the Indian Statistical Conference, Calcutta, 1938



Eugenio Del Prete, M.Eng., Ph.D.
Biostatistician and Data Analyst
Telethon Institute of Genetics and Medicine (TIGEM)
Pozzuoli (NA), Italy
e-mail: e.delprete@tigem.it

References

[1] Emmert-Streib, F. **Understanding Statistical Hypothesis Testing: The Logic of Statistical Inference**, Machine Learning and Knowledge Extraction (2019).
[2] Banerjee, A. **Hypothesis testing, type I and type II errors**, Ind. Psychiatry J. (2009).
[3] Greenland, S. **Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations**, Eur J Epidemiol. (2016).

[h1] *https://bookdown.org/jgscott/DSGI/*
[h2] *https://statsandr.com/*
[h3] *https://youtu.be/EvHiee7gs9Y*