# Statistical Foundations for Biomedical Research

**Mariana M S Santos, PhD**
Investigator Assistant
ERA Chair iSTARS

Faculdade de Medicina 1825

# Statistical Mindset

Beyond the P-value

**Statistics vs. Math**
Math is about certainty; Statistics is about **quantifying uncertainty**.

**The Workflow:** Population → Sample → Inference.

**Why it matters in BioMed:** Biological variation is huge. We need to distinguish "signal" (the treatment effect) from "noise" (natural variation).

## Descriptive Stats – The "First Look"

Don't Trust the Mean!

**Central Tendency:** Mean vs. Median.
- *Tip:* Use Median for skewed data (e.g., length of hospital stay).

**Dispersion:** Standard Deviation (SD) vs. Interquartile Range (IQR).

**The Visual Check:** Always plot your data (Histograms, Boxplots) before running tests. Anscombe's Quartet shows that different data patterns can have the same mean.

# Where to Start? The 4-Step "Health Check"

Before you analyze, "Listen" to your Data

1. **The Shape & Types (The Skeleton)**

- How many patients (N) vs. how many variables?
- Are the variables correctly typed? (e.g., Is "Gender" a category or a number? Is "Date" actually a date or just text?)
- Check if your variables are correctly recognized (Numeric, Categorical, Date). Ensure your Sample Size (N) matches your expectations.

# Where to Start? The 4-Step "Health Check"

Before you analyze, "Listen" to your Data

## 2. Missingness (The Gaps)

- Where are the holes? Is data missing at random, or is there a pattern? (e.g., Are older patients skipping the weight question?)

  *Rule:* If a variable has >30-40% missing values, it might be unusable.

- Identify missing values. Are they random or biased? *Tip:* Variables with >30%missing values may compromise your study.

# Where to Start? The 4-Step "Health Check"

Before you analyze, "Listen" to your Data

### 3. The "Sense" Test (Outliers & Sanity)

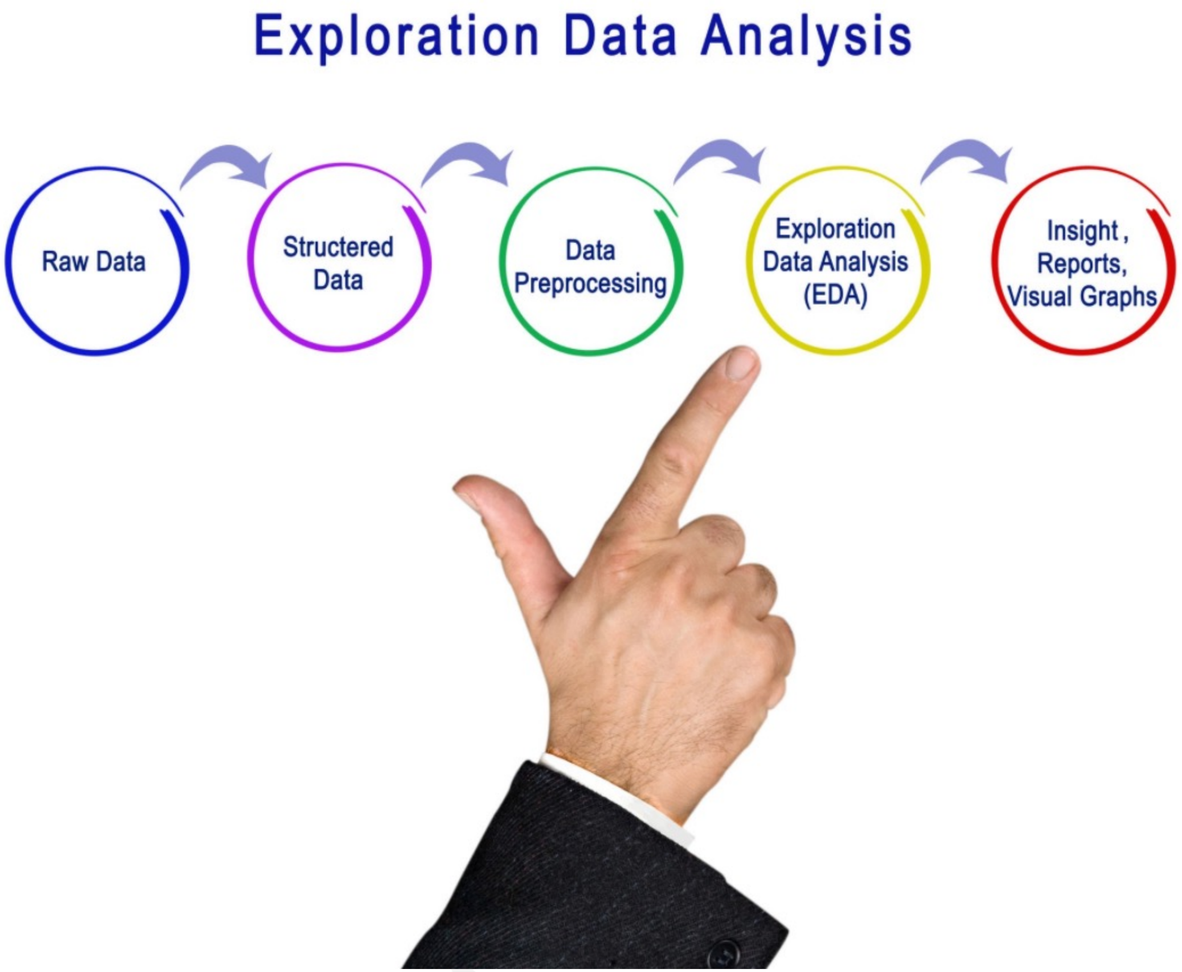
- Look for impossible values: A heart rate of 0 or a BMI of 150.
  “Are these typos or extreme clinical cases?”

  Decide how to handle them **before** you start.

### 4. Distribution & Balance (The Heartbeat)

- Is your outcome variable balanced? (e.g., In a rare disease study, do you have 100 cases vs. 1,000,000 controls?)

- Is the data normally distributed? This dictates if you use a T-test or a Non-parametric test.


Exploration Data Analysis
Raw Data
Structered Data
Data Preprocessing
Exploration Data Analysis (EDA)
Insight , Reports, Visual Graphs

# Hypothesis Testing (The Logic)

Guilty until proven Innocent

- ❑ **Null Hypothesis (H0):** No effect, no difference.
- ❑ **Alternative Hypothesis (H1):** There is an effect.
- ❑ **The P-value:** The probability of seeing your results (or more extreme) **if the null hypothesis is true**.

***Warning!* A p-value < 0.05 is not a "seal of truth"; it's just a threshold of surprise.**

## Choosing the Right Test

| Data Type | Comparison | Test Example |
| --- | --- | --- |
| **Continuous (Normal)** | 2 Groups | T-test |
| **Continuous (Non-normal)** | 2 Groups | Mann-Whitney U |
| **Categorical** | Frequencies | Chi-square / Fisher’s Exact |
| **Continuous** | Relationship | Pearson/Spearman Correlation |

## Clinical vs. Statistical Significance

Do large samples reveal everything?

**The Sample Size Trap:** With 1,000,000 patients, even a 0.1 mmHg drop in blood pressure becomes "statistically significant" ($p<0.001$), but it's **clinically irrelevant**.

**Effect Size:** Tell us *how much* it matters (e.g., Cohen's d, Odds Ratio).

**Confidence Intervals (CI):** as much useful as p-values. They show the range of the likely true effect.

***Ok, it is statistically significant, but is it clinical relevant?***

# Correlation vs. Causation

The Data Scientist’s Golden Rule

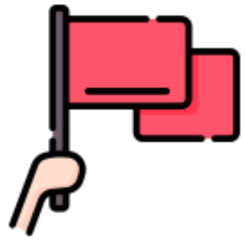
<u>**The Data Scientist’s Golden Rule**</u>

- **Correlation:** Two variables move together.
- **Causation:** One variable *causes* the other to move.
- **Confounding Factors:** The "Hidden Variable" (e.g., Ice cream sales and drowning deaths both increase in summer; the cause is the Heat, not the ice cream).

**Avoid Data Dredging: do not torture your data.**

## The "Bullshit Detector" – Top 3 Statistical Red Flags

Don't be Fooled by Numbers



1

### P-Hacking (The "Fishing Expedition")

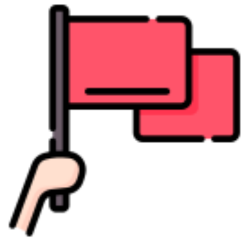
**The Trap:** Testing 50 different variables until one finally hits $p<0.05$ by pure chance.

**The Red Flag:** A study that reports a significant result for a very specific sub-group that wasn't the original focus.

**The Fix:** Always define your hypothesis *before* looking at the data.

## The "Bullshit Detector" – Top 3 Statistical Red Flags

Don't be Fooled by Numbers



2

### Missing Confidence Intervals (CI)

**The Trap:** Reporting only the p-value (p=0.04) to claim success.

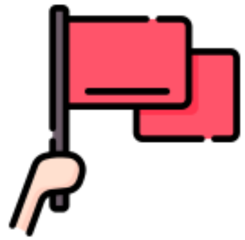
**The Red Flag:** A "significant" result where the Confidence Interval is huge (e.g., an Odds Ratio of 1.5 but with a CI of [1.01 to 25.0]).

**The Fix:** If the CI is too wide, the estimate is imprecise. The "p" tells you it's there; the "CI" tells you if it's reliable.

Faculdade de Medicina 1825

## The "Bullshit Detector" – Top 3 Statistical Red Flags

Don't be Fooled by Numbers



3

### Hidden Axes & Relative Risk Bias

**The Trap:** Saying "Treatment X reduces risk by 50%"

**The Red Flag:** If the absolute risk drops from 2% to 1%, it is a 50% relative reduction, but only a **1% absolute difference**.

**The Fix:** Always look at the **Absolute Risk Reduction (ARR)**. Don't let the Y-axis of a graph start at 40% to make a small difference look like a mountain.

U LISBOA UNIVERSIDADE DE LISBOA

Feitos de outra matéria

Faculdade 1825
de Medicina

# Comments & Questions

**Mariana M S Santos**
marianamsantos@medicina.ulisboa.pt

LISBOA
UNIVERSIDADE DE LISBOA

Feitos de outra matéria