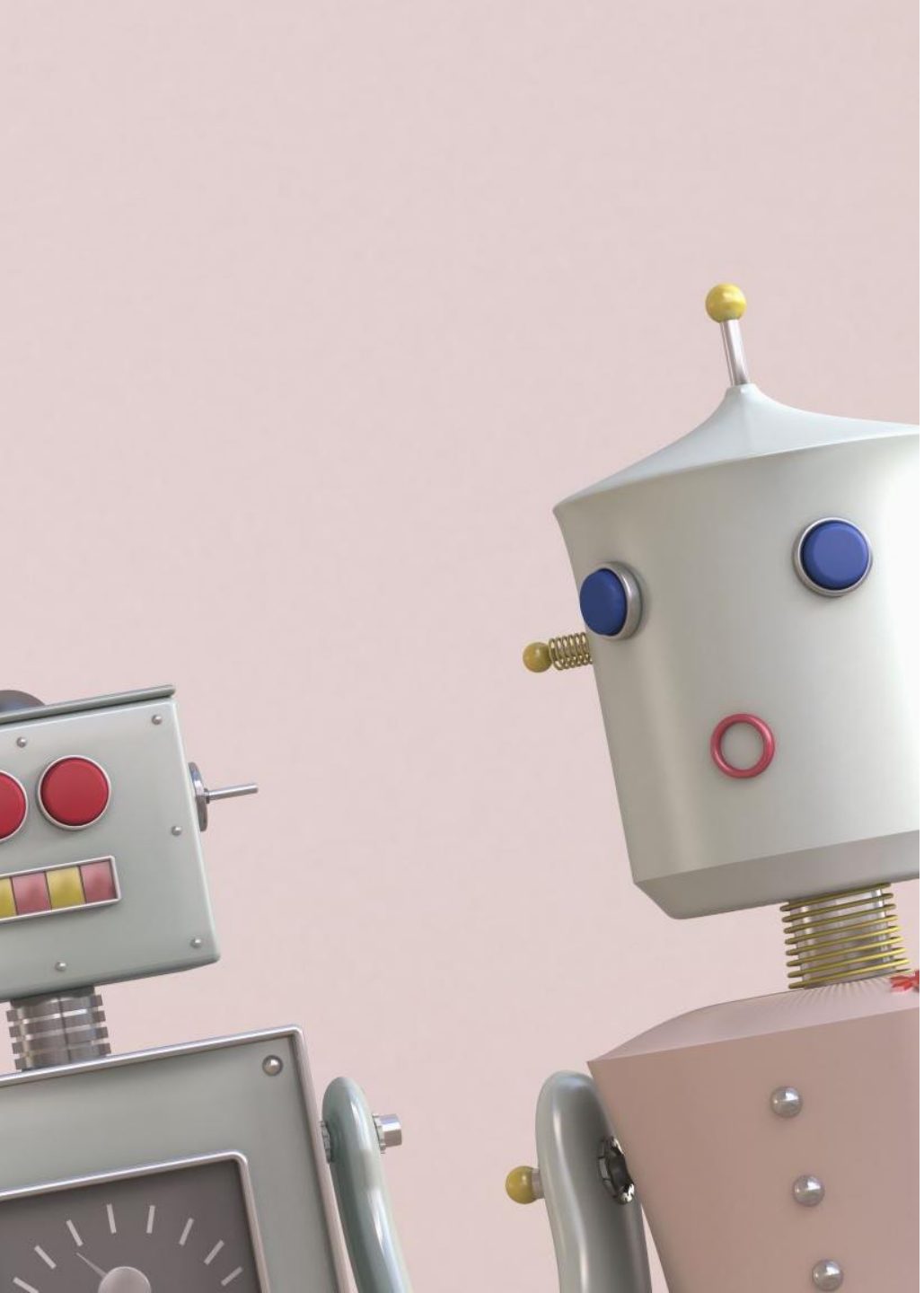
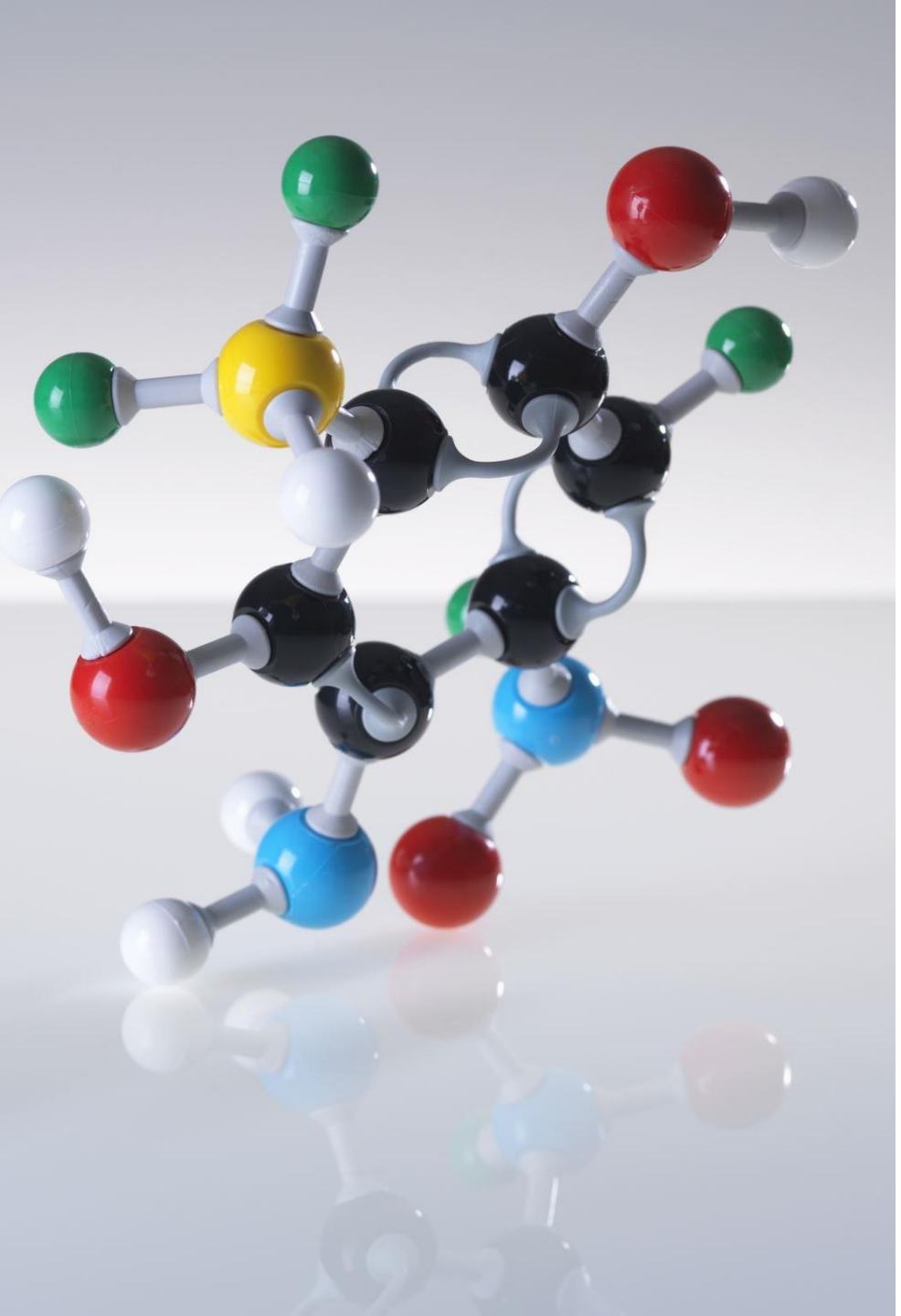


SOCIAL ROBOTS AND HUMAN-ROBOT INTERACTION

Week 3, Statistical Analysis

Ana Paiva

P2
2025/2026



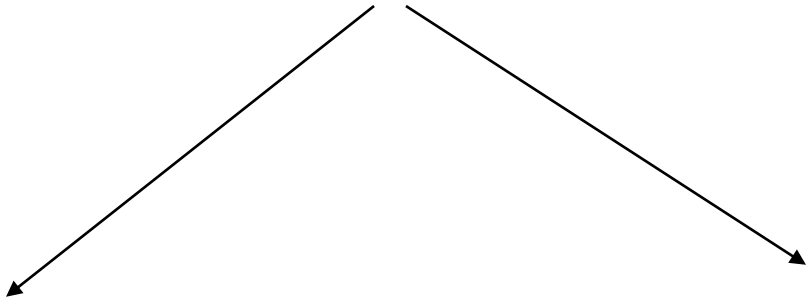
STRUCTURE OF LECTURE

- Summary of practical and methodological aspects of experimental design in HRI
- Obtaining results: statistical analysis
  - Why do we need it?
  - Descriptive and Inferential statistics
  - Types of test statistics (T-test and ANOVA)
  - Significance, sample size
  - Errors
- Reporting and writing up studies in HRI

SUMMARY:HYPOTHESES AND VARIABLES

To prove that an hypothesis we need to measure something

H1. A robot that commits errors during its interaction with humans, is perceived as more likeable than a robot that performs flawlessly.



Scenario: robot commits errors	Measure: degree of likeability

Summary: Dependent versus Independent Variables

In a study we have the Independent variables and Dependent variables:

The Independent variable (the "causes" is the variable that is purposely change It is the manipulated variable.

The *Dependent variable* (the "effects") changes in response to the independen variable. It is the responding variable.

Summary: Between-Group and Within Subjects Experimental Designs

Between groups (or independent measures)

- Use separate groups of users (participants) for each of the different conditions (associated with the manipulation done).
- Each participant is tested only once.

Within-subjects (or repeated measures)

- Participants are exposed to the different conditions of the experiment.

Hybrid design (mixed)

- Designs that involve a combination of a between groups and within subjects variables.

WHY DO WE NEED DATA ANALYSIS & STATISTICS?

- To answer our research question
- Thus, to enable us to ***test experimental hypotheses***

DESCRIPTIVE & INFERENTIAL STATISTICS

Descriptive Statistics

- Organize
- Summarize
- Simplify
- Presentation of data

Describing data

Inferential Statistics

- Generalize from samples to populations
- Hypothesis testing
- Relationships among variables

Make predictions

DESCRIPTIVE STATISTICS

Tell us about the "population and the distribution of the results".

What we need to look at is:

- **Mean** (M): the sum of all the scores divided by the number of scores;
- **Median** (Mdn): is the middle score of a distribution of scores, when they are ranked in order of magnitude.
- **Sum of the squared errors** (SS)
- Variance (s2): ***average squared deviation between the mean and the observed score***: tells us on average how much a given data point differs from the mean of all data points.
- **Standard Deviation** (SD) is the square root of the variance.

BUT…

We are interested not only in statistics that describe the population sample but also in statistics that allow us to **“infer”** things about the users in the groups (the conditions) interacting with our social robots:

we need **inferential** statistics.

STARTING POINT: CHARACTERIZING THE POPULATION

To answer general questions about the studies performed we need to look at **the population** that will test our robot.

- Ideally we would like for "everyone" to test it!

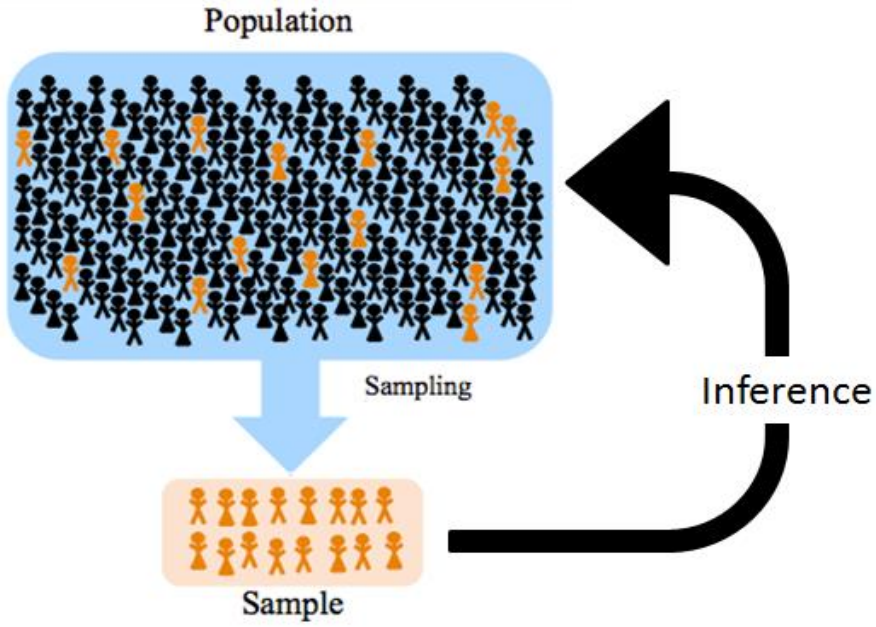
- Of course that is not possible…☺

So, we will rely on a

"**sample**" that we will assume "***represents***" the population.

WHAT IS INFERENTIAL STATISTICS?

Essentially, we want to make *inferences* on a *population,* after observing a certain phenomenon on a *sample*



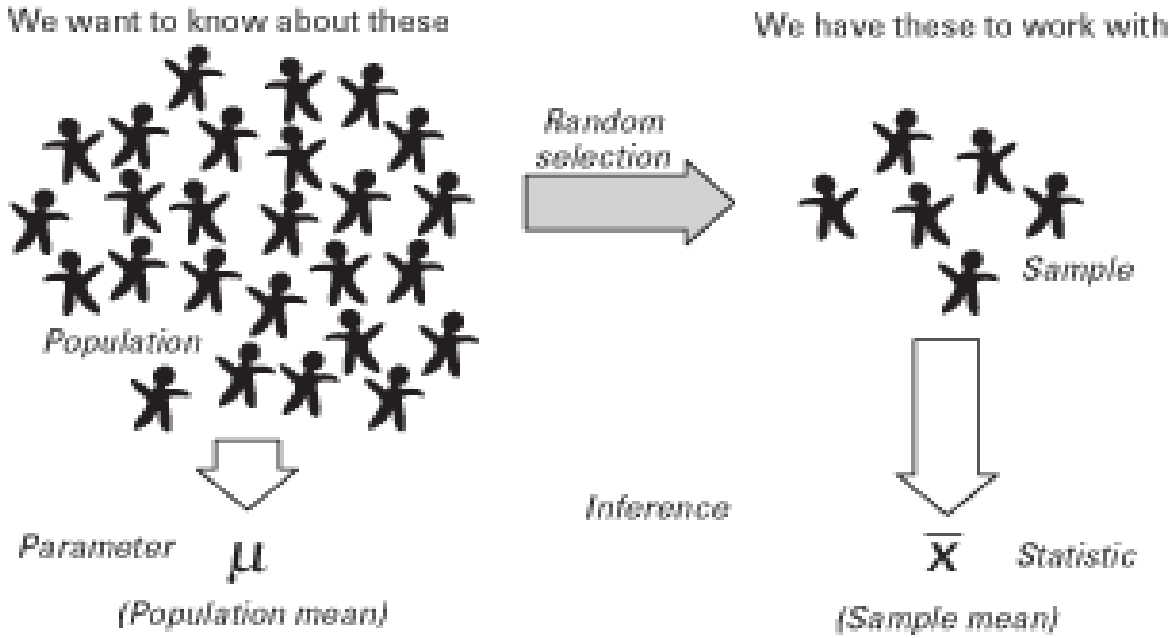
Sampling: how to sample the population?

- We sample all the time: when checking for salt, we don't eat the whole pot, just taste a little to see if there is enough salt
- If the sample is **representative** of the whole population, then our results on the sample should apply to the whole population (e.g. if the food is not stirred, and you taste from the bottom of the pot, the sample is not representative)

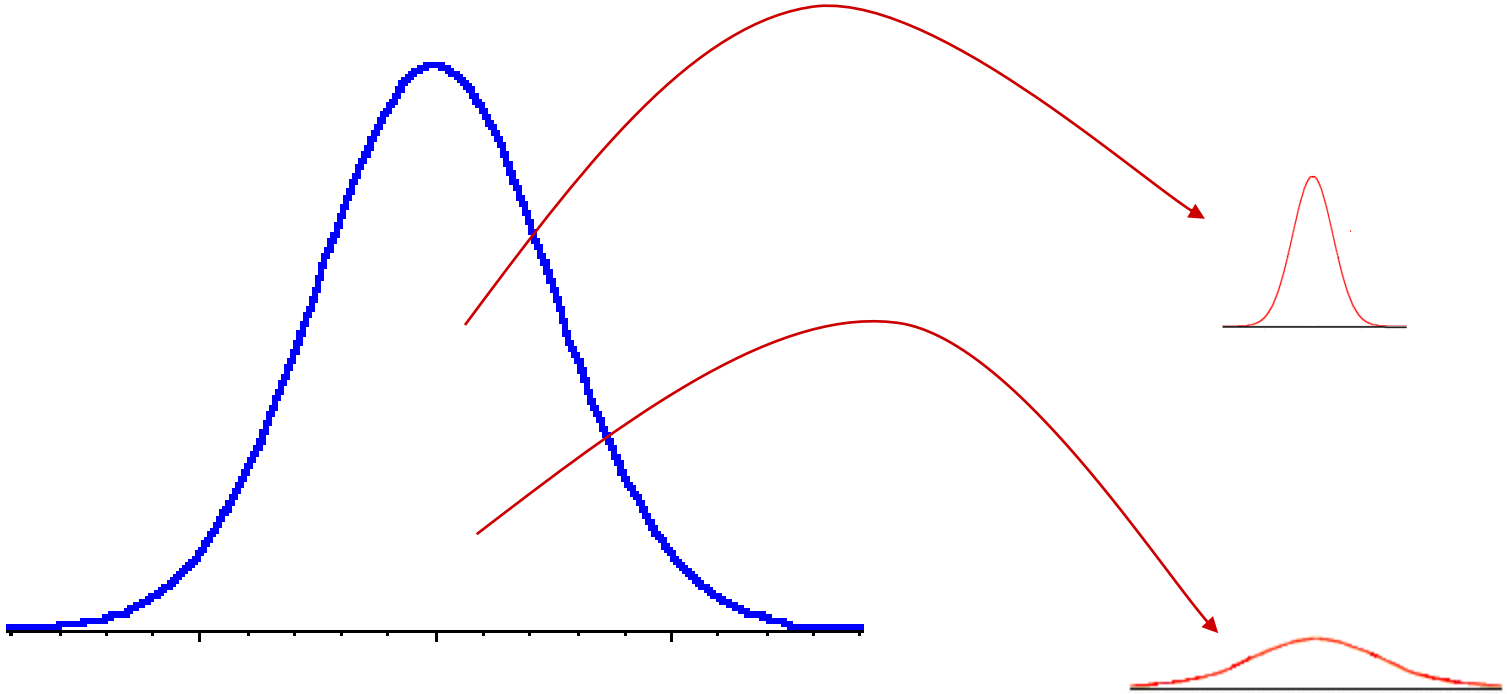
Sampling techniques examples: https://towardsdatascience.com/8-types-of-sampling-techniques-b21adcdd2124

CONCEPTS: PARAMETER VS STATISTIC

- Parameter: a value that describes a population (most often unknown)
- Statistic: a value that describes a sample (what we measure)



SAMPLING: ISSUES WHEN MAKING INFERENCES



The data collected represent only a single sample from a much larger population
So… if a different sample were used, then different results could have been obtained....

HYPOTHESIS TESTING

- To be used with an **experimental hypothesis** (that is, the prediction that the experiment manipulation will have an effect)
- Basically we need "**to prove that the results in the sample population that we observe of one condition is different from the other condition**".

Null hypothesis (H0): There is no difference	**Alternative hypothesis (H1):** There is a difference

- We will calculate the "**probability**" that the results obtained are a "chance result"- and as this result decreases we become more confident that the experimental hypothesis is correct and the null hypothesis can be rejected.

USE OF A "TEST" STATISTICS

- A test statistics is ***a statistic that has known properties*** (we know its frequency distribution). We calculate the ***systematic variation*** (the variation due us manipulating something in one condition and not in the other) and the ***unsystematic variation*** which is due to the natural differences between people in different samples (such as differences in intelligence or motivation)
- The **test statistic** forms a ***"ratio comparing the obtained difference between the sample mean and the hypothesized population mean"*** versus the amount of difference we would expect without any interaction effect (the standard error).

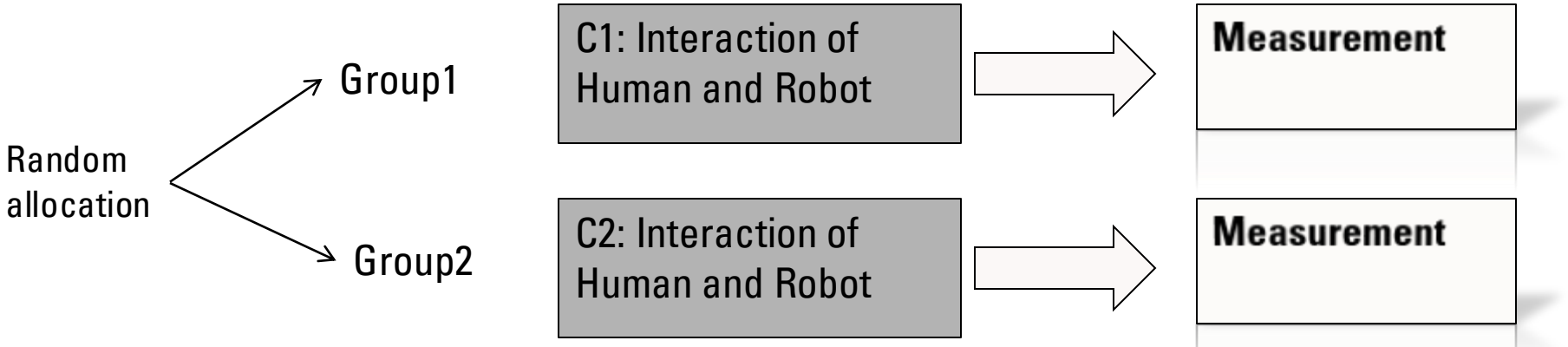
$$TestStatistic = \frac{SystematicVariance}{UnsystematicVariance}$$

EXAMPLES OF DIFFERENT TEST STATISTICS

$$\textbf{Test } \textit{statistic} = \frac{Effect}{Noise}$$

Effect = what you observe (difference between means, relationship between variables…)
Noise = variability expected under the null hypothesis
This ratio tells whether the observed effect is big enough to be unlikely under the null.

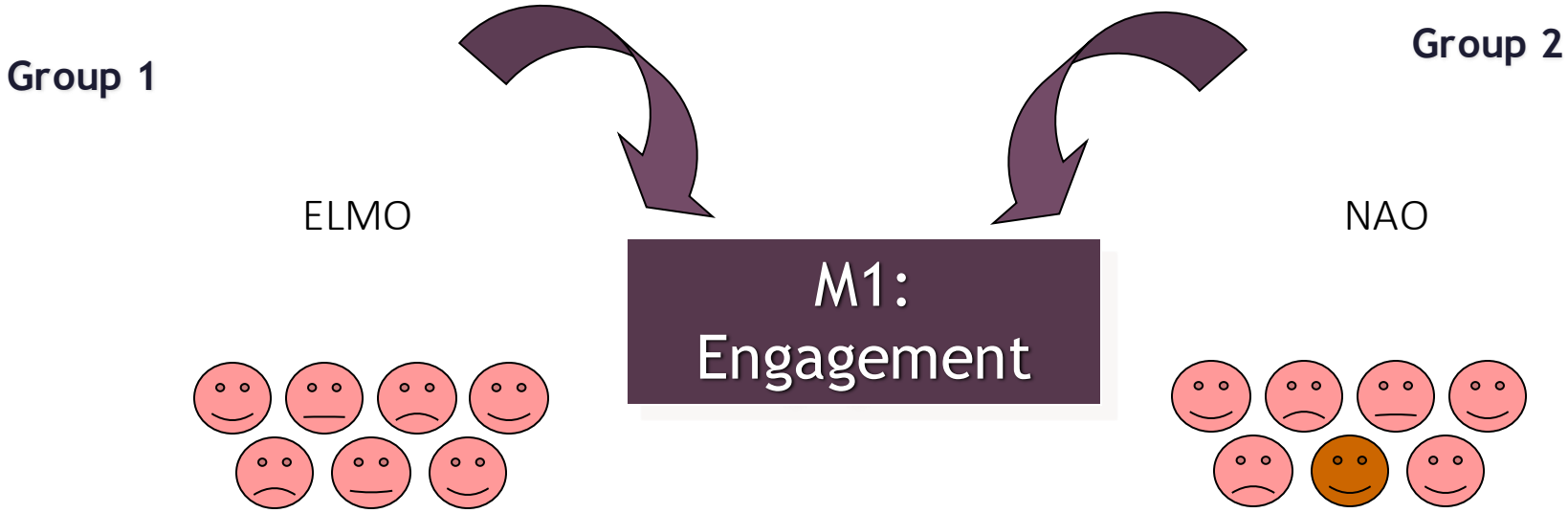
SIMPLE EXPERIMENTS STATISTICAL ANALYSIS
(BETWEEN GROUPS)



The simplest form of experiment that can be done is one with only one independent variable that is manipulated in only two ways and only one outcome is measured.

- More often than not the manipulation of the independent variable involves having an experimental condition and a control one.

EXAMPLE



- Variance created by our manipulation
  - Interaction with NAO or with ELMO
- Variance created by unknown factors
  - E.g. Differences in ability (unsystematic variance)

→ Independent T-test

T-TEST

- The t-test assesses whether the means of **two groups** are statistically different from each other.
- This analysis is appropriate whenever you want to compare the means of **two groups**

RATIONAL FOR THE *T*-TEST

Two samples of data are collected and the **sample means** calculated. These means might differ by either a little or a lot.

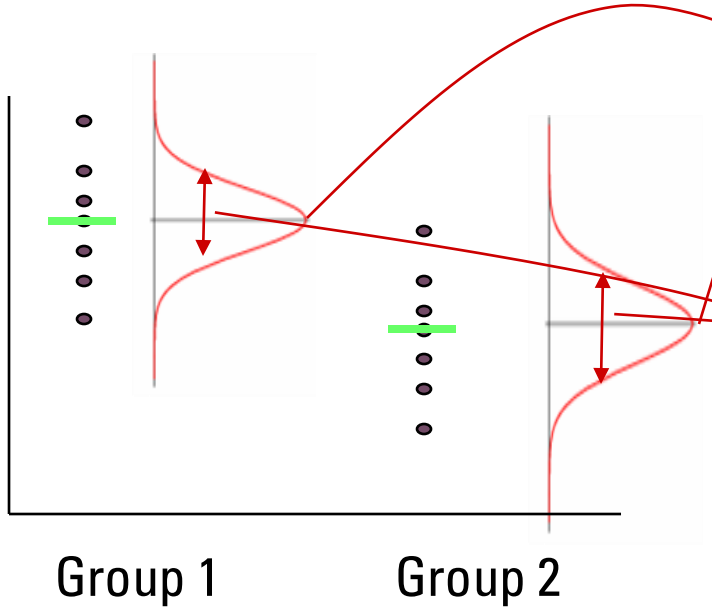
If the samples come from the same population, then we expect their means to be roughly equal.

Although it is possible for their means to differ by chance alone, we would expect large differences between sample means to occur very infrequently.

We compare the difference between the sample means that we collected to the difference between the sample means that we would expect to obtain if there were no effect (i.e. if the null hypothesis were true).

CALCULATING T

Difference between the means divided by the pooled standard error of the mean



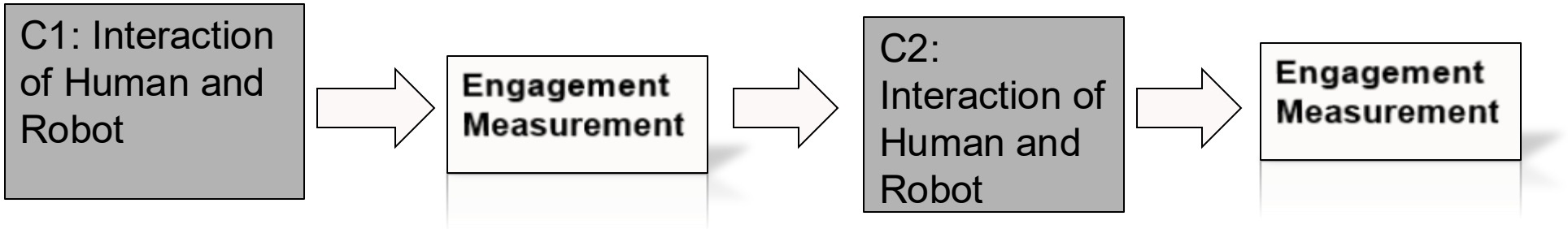
$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

← sample means

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^{\,2}}{n_1} + \frac{s_2^{\,2}}{n_2}}$$

Standard error of the difference between the means

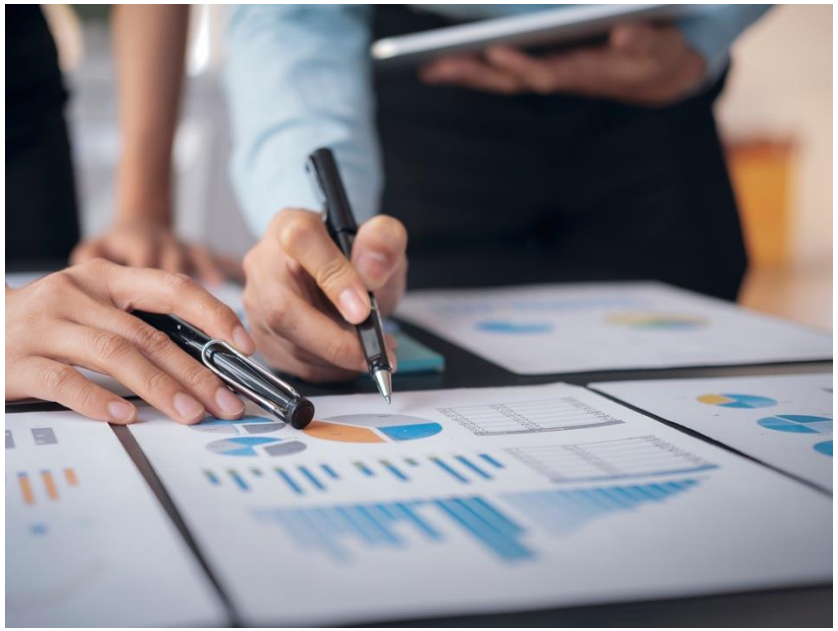
AND REPEATED MEASURES? DEPENDENT T-TEST



Dependent t-test (or paired t-test or repeated-measures t-test)

- Compares two means based on related data.
- Compares participants to themselves (using difference scores $D = X_1 - X_2$)
- E.g., Data from the same users measured at different times (people interact with both robots).
- Advantage: reduces noise by controlling for individual variability

SUMMARY: T-TEST



- Independent *t*-test
  - Compares two means based on independent data
  - E.g., data from different groups of users
- Dependent (repeated measured) *t*-test
  - Compares two means based on related data
  - E.g., Data from the same users measured at different times (people interact with both robots)

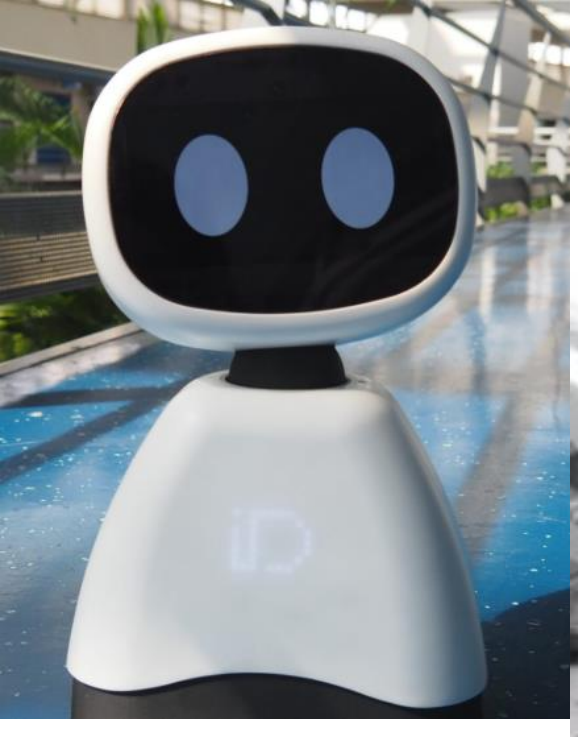
IMPORTANT: ASSUMPTIONS OF THE T-TEST

- Both the independent *t*-test and the dependent *t*-test are ***parametric tests*** based on the **normal distribution**.

Therefore, they assume:

  - The sampling distribution is normally distributed.
  - Data are measured at least at the interval level.

- The independent *t*-test, because it is used to test different groups of people, also assumes:
  - Variances in these populations are roughly equal (*homogeneity of variance*).
  - Scores in different treatment conditions are independent (because they come from different people).



LIMITATIONS!

The t-test is **limited to situations where there only two levels of the independent variable.**

However, sometimes we have experiments (given the types of experiments with our robots) that have three or four levels of the independent variable.

SO….WHAT IF THERE ARE MORE THAN TWO FACTOR LEVELS?

The *t*-test does not directly apply!!!!

- The **analysis of variance** (ANOVA) is the appropriate analysis "engine" for these types of experiments
- The name "analysis of variance" stems from a **partitioning** of the total variability in the response variable into components that are consistent with a **model** for the experiment
- The ANOVA was developed by Fisher in the early 1920s, and initially applied to agricultural experiments…..

Used extensively today for many experiments

ANOVA

ANOVA compares the means of three or more groups simultaneously.
At its simplest form ANOVA tests the following **hypotheses**:

- H0: The means of all the groups are equal
- H1: Not all the means are equal

It does not say how, or which groups differ (this can be found by follow up with “multiple comparisons”)

Examples: Comparing friendliness of 3 types of robots

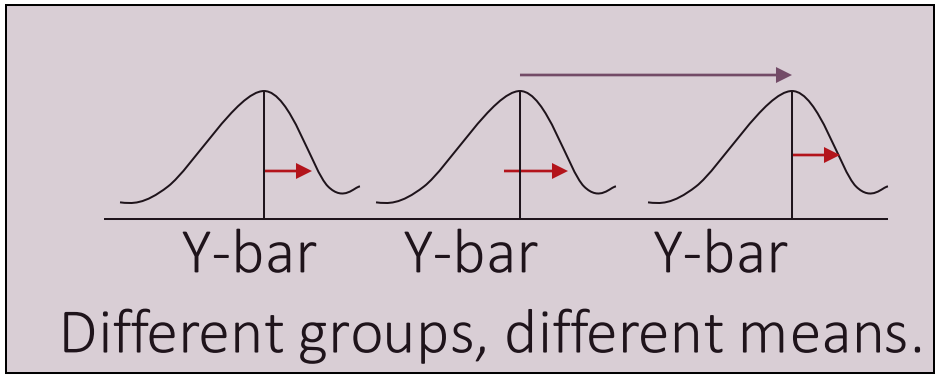
Note: you can use ANOVA for only 2 groups, but it becomes equivalent to a t-test.

ASSUMPTIONS OF ANOVA

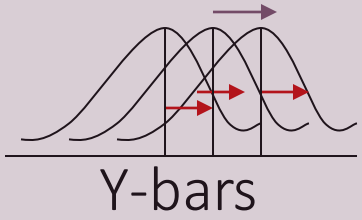
- each group is approximately normal
  - check this by looking at histograms and/or normal quantile plots, or use assumptions
  - can handle some nonnormality, but not severe outliers
- standard deviations of each group are approximately equal
  - rule of thumb: ratio of largest to smallest sample st. dev. must be less than 2:1

ANOVA

Rational: when the groups have little variation within themselves, but large variation between them, it would appear that they are distinct and that their means are different.



Different groups, different means.



Similar groups, similar means.

F-statistic

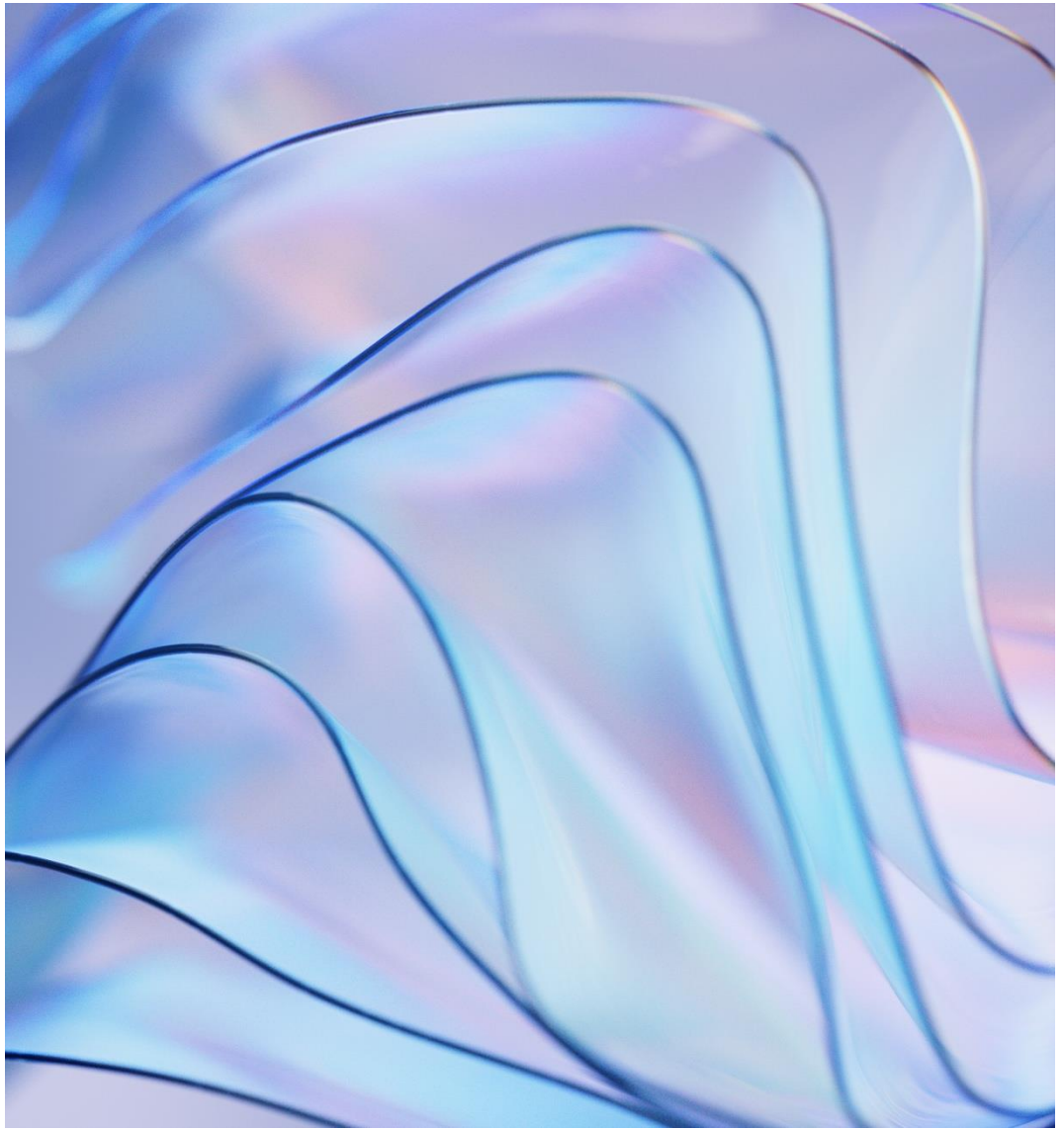
$$F = \frac{Variance\ between\ groups}{Variance\ within\ groups}$$

If the variance *between* groups is much larger than the variance *within* groups

→ groups differ.

ANOVA: SUMMARY

So, essentially, ANOVA uses your sample to tell in the population if you **have overlapping group distributions** (no difference between means) or fairly distinct group distributions (differences between means).



ANOVA FOLLOW UP COMPARISONS

The ANOVA only indicates that the groups do not all have the same means…

So, to understand the differences we can compare them two by two using the 2-sample t test.

Example: Imagine we have 4 groups.

- The null Hypothesis H0 is that the groups are all the same.
- The ANOVA test tells us that there is a difference.
- We want to see how the 4 groups differ between themselves (pairwise comparisons- cond 1 against cond2, etc)

ANOVA FOLLOW UP COMPARISONS

Problems with Ad-hoc pairwise comparisons

- Need to be careful to avoid the **"multiple comparisons problem"**
  - the more comparisons are made, the more likely that at least one comparison will be significant merely due to random error (due to false positives)
- Use **corrections**, like the Bonferroni correction, or other tests, such as Tukey's HSD (honestly significant difference) test

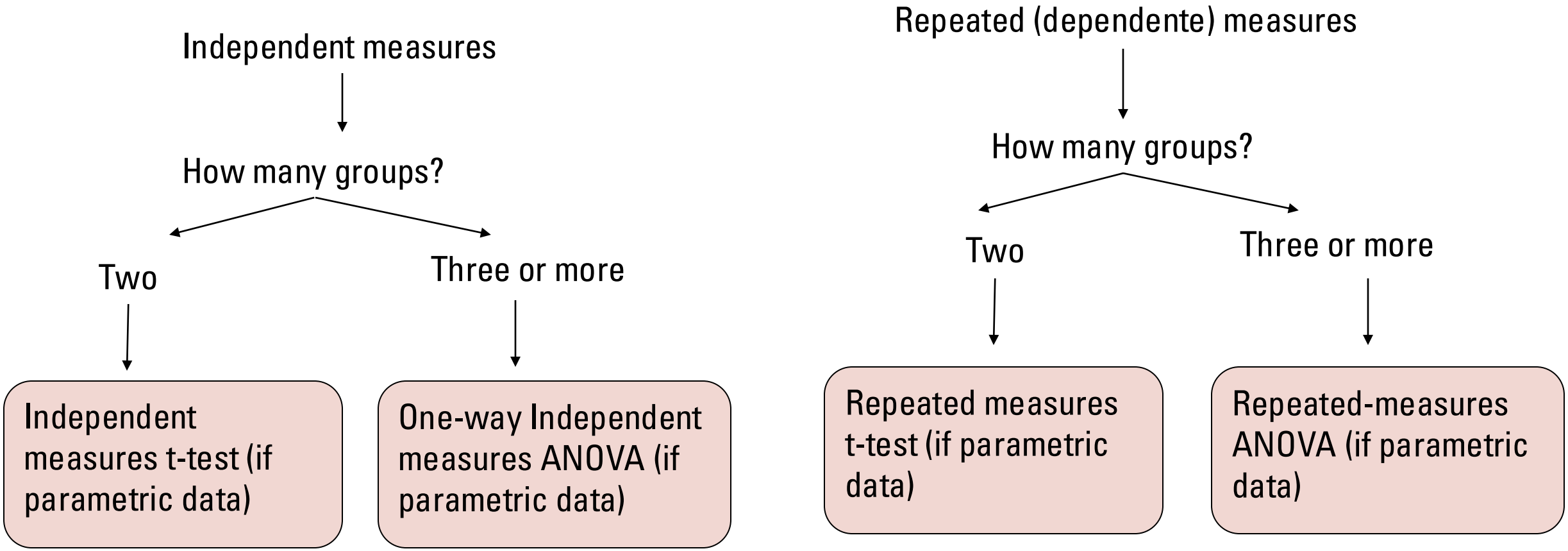
DIFFERENT TYPES OF ANOVA

- **one-way ANOVA** is the multi-group equivalent of an **independent** t –test
- **repeated-measures ANOVA** for the case of **dependent** observations

And:

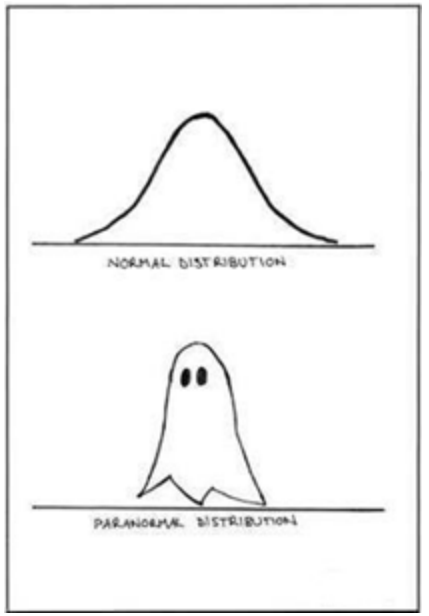
- **MANOVA i**f you want to test **more than one dependent variable**

SUMMARY TESTS



AND WHEN THE DATA IS NOT NORMALLY DISTRIBUTED?

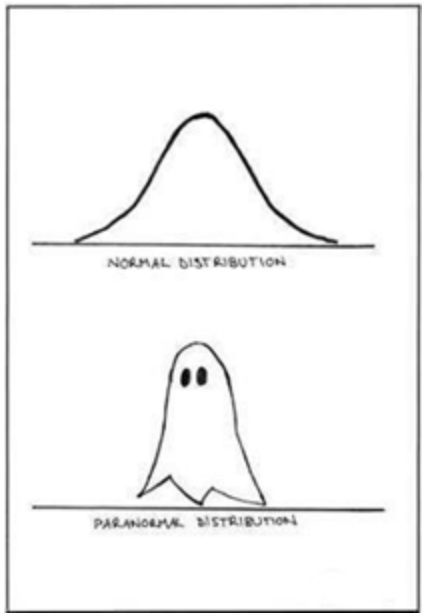
All these previous tests work under the assumptions that the data **is normally (or almost) distributed**, and that variance is roughly equal (homogeneity of variance)



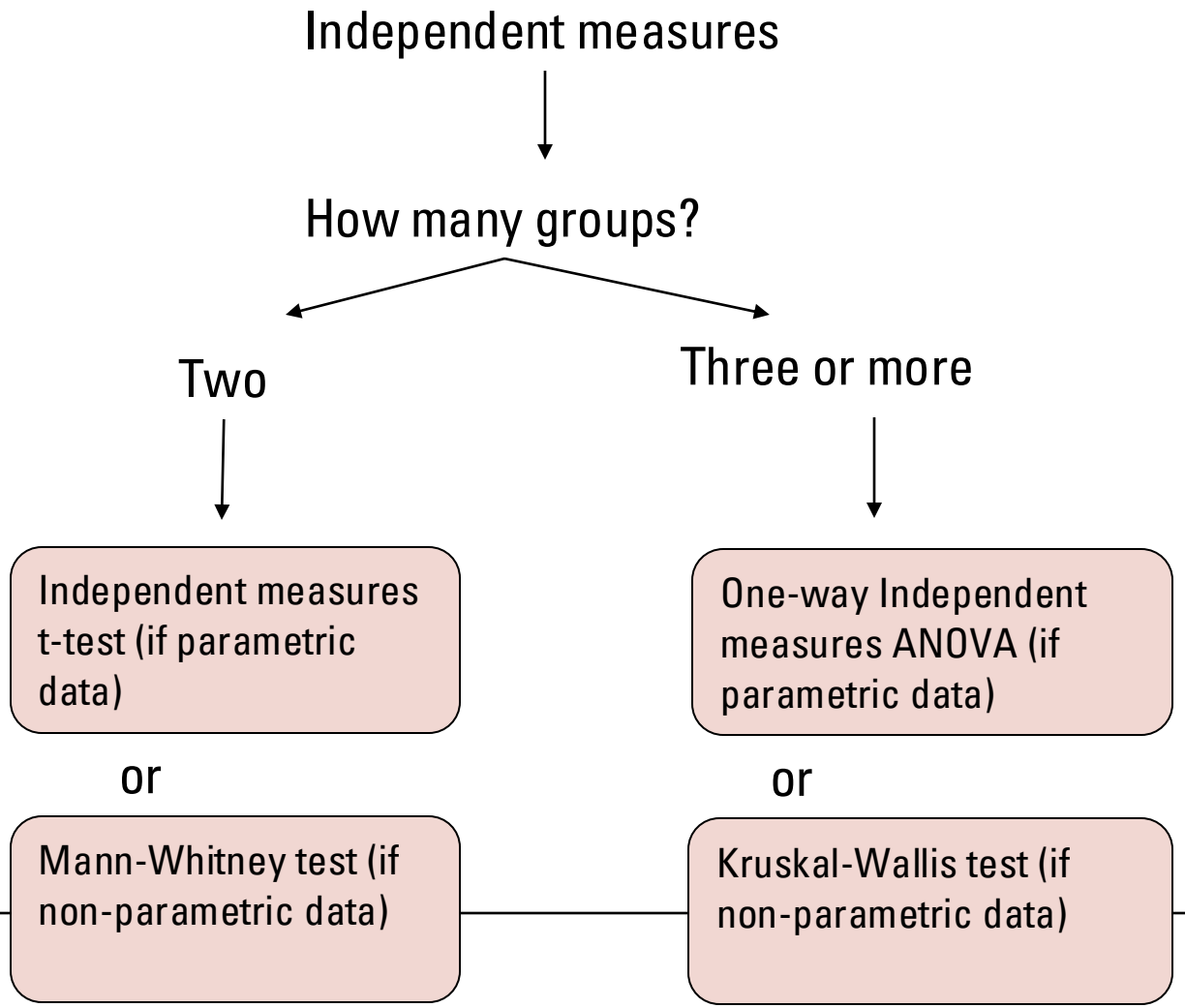
- You can test for these assumptions with:
  - Kolmogorov–Smirnov or Shapiro–Wilk tests of normality
  - Levene's and Barlett's tests of homogeneity

WHAT TO DO WHEN THE DATA IS NOT NORMALLY DISTRIBUTED?

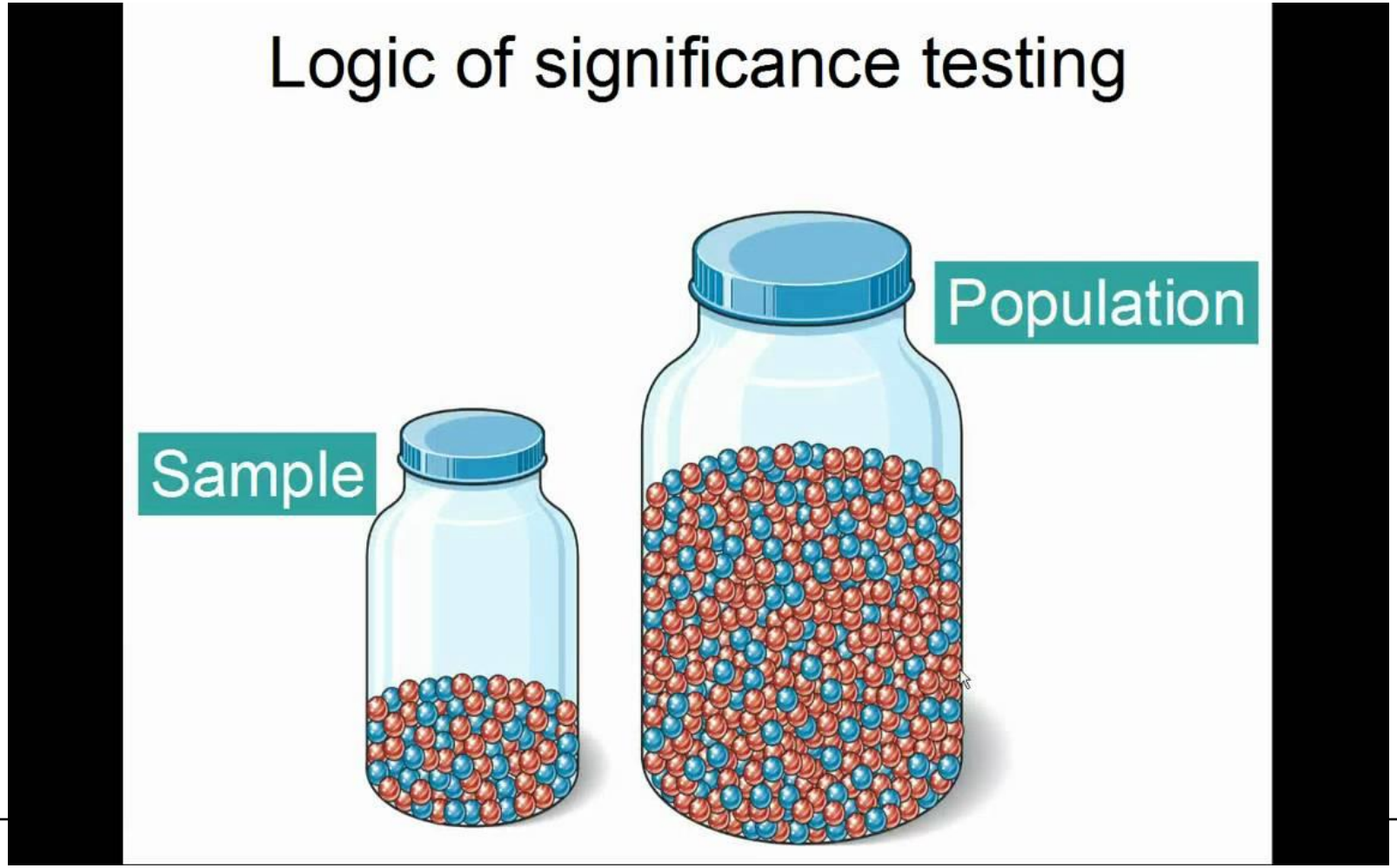
- For **non-normal distributions**, there are many non-parametric tests (e.g. **Kruskal-Wallis, Mann-Whitney, Wilcoxon…)**



TESTS SUMMARY



HOW DO I KNOW HOW SIGNIFICANT ARE MY RESULTS?



STATISTICALLY SIGNIFICANT?

- 1) The **null hypothesis**, H0, always states that the "condition" has no effect (no change, no difference).
- 2) **Significant** = probably not caused by chance for a given significance level, but rather caused by a factor of interest (your manipulation)

P-VALUES

So… we want to find out the probability that the results was caused by chance for a given significance level (and not caused by our manipulation)

Null Hypothesis (H_0): "There is no effect," "No difference exists," or "Nothing is happening."

P values = the probability that the observed result was obtained by chance

So, we ran our experiment, and choose our test (T-test, ANOVA).. and calculate the test statistics *(T-statistic, F-ratio, etc).* Then calculate p-value as the probability that the test statistic is bigger than the observed value, when H0 is true)

NOTE: to calculate the p-value use SPSS, R, or other software

P-VALUES

P value: why the **< 0.05** the Gold Standard?

P values = the **probability** that the observed result was obtained by chance

- i.e. when the null hypothesis is true

- We usually consider that is significant when the probability is low (Usually 0.05)
- If p < = 0.05 we reject the null hypothesis and accept the experimental hypothesis
  - 95% certain that our experimental effect is genuine
- If however, p > 0.05 level then we reject the experimental hypothesis and accept the null hypothesis

The 0.05 threshold is a convention, not a magic number!

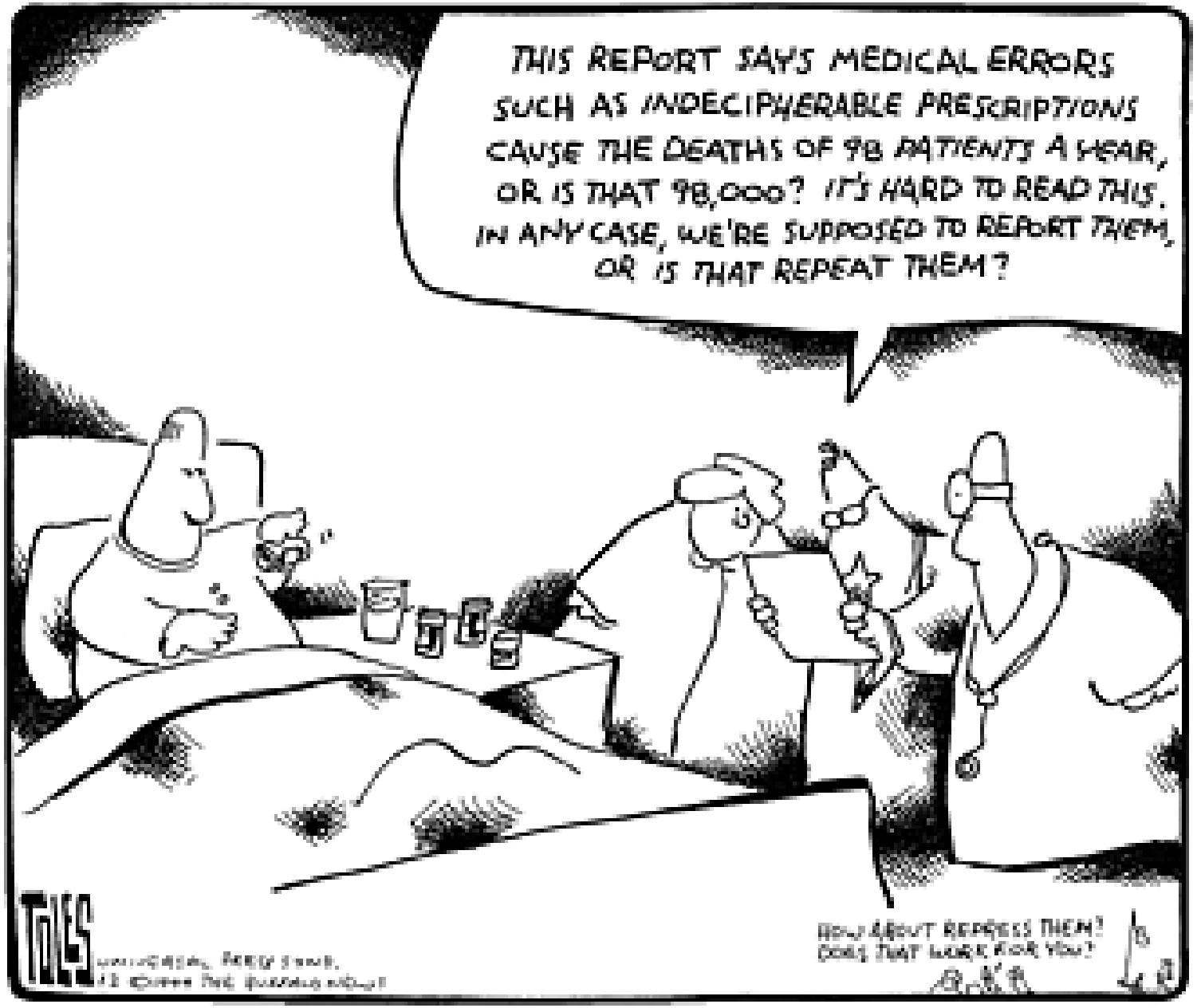
SAMPLE SIZE

"How many participants should I test?"

Power analysis / sequential analysis

- The hypothesis test is influenced not only by the **size of the manipulation effect** but also by the **size of the sample**.
- Thus, even a very small effect can be significant if it is observed in a very large sample.

ERRORS



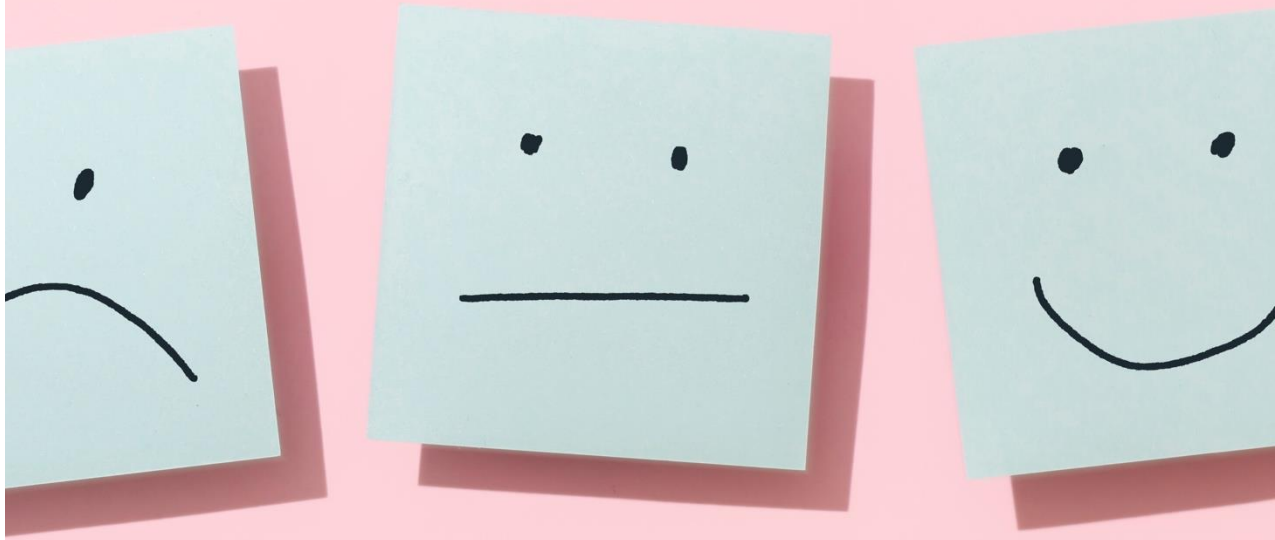
TWO TYPES OF ERRORS

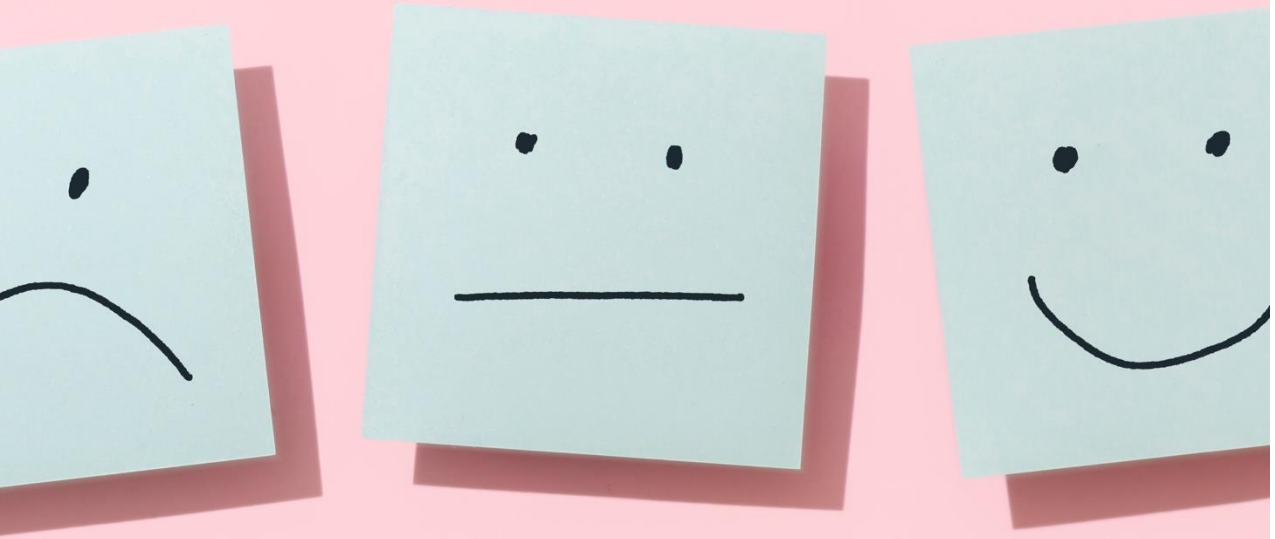
- Type I error = false positive
  - α level of 0.05 means that there is 5% risk that a type I error will be encountered
- Type II error = false negative

"I have a positive attitude…
but it might be a false positive."

TYPE I ERRORS: FALSE POSITIVE

- A **Type I error** occurs when the sample data appear to show a treatment effect when, in fact….there is none. !!!!
- In this case we reject the null hypothesis but falsely conclude that the condition has an effect.
- Type I errors are caused by **unusual, unrepresentative samples**. Just by chance the researcher selects an extreme sample with the result that the sample falls in the critical region even though the treatment has no effect.
- Yet, the hypothesis test is structured so that Type I errors are very unlikely; specifically, the probability of a Type I error is equal to the alpha level.



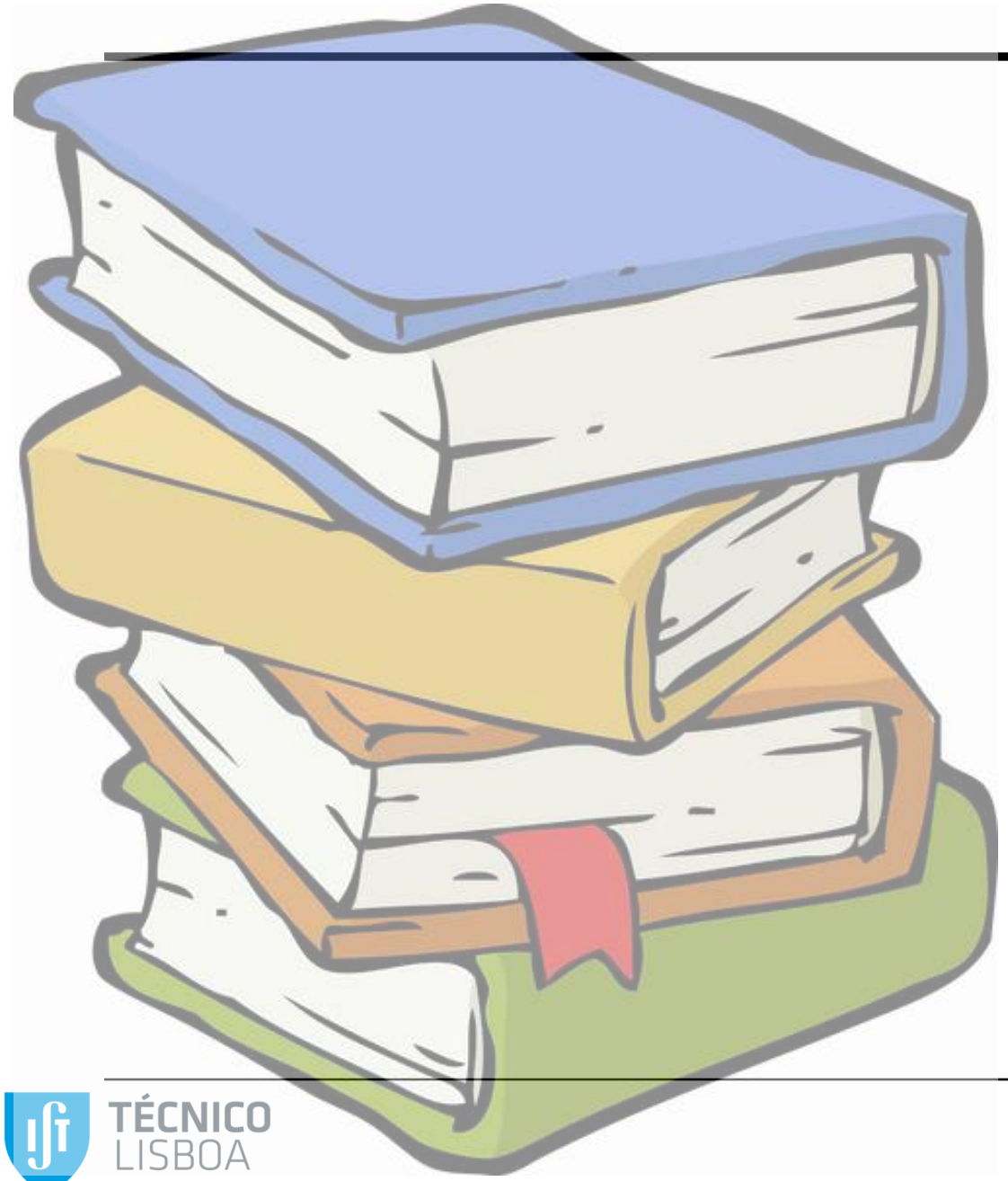


TYPE II ERRORS: FALSE NEGATIVES

- A **Type II error** occurs when the sample does not appear to have been affected by the manipulation when, in fact, it does have an effect.
- In this case, the researcher will fail to reject the null hypothesis and falsely conclude that the manipulation does not have an effect.
- **Type II errors are commonly the result of a very small manipulation effect**. Although the manipulation does have an effect, it is not large enough to show up in the research study.

HOW TO PROCEED?

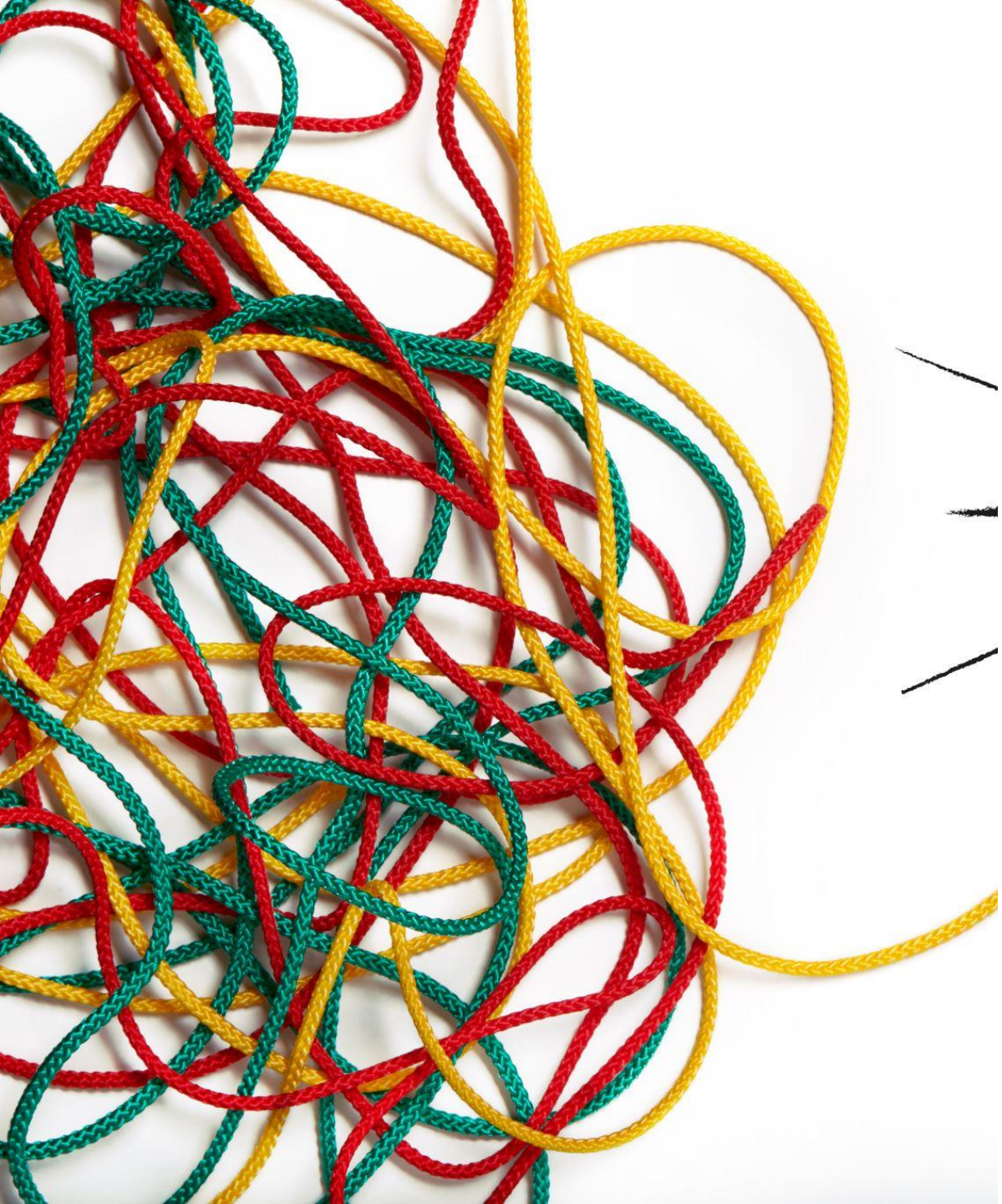
- Bibliography
  - Lectures by Andy Field on youtube https://www.youtube.com/channel/UCakigkjm3vBzEHpFzECDXQQ
  - Book: How to design and Report Experiments, by Andy Field and Graham Hole
- SPSS (Statistical Package for the Social Sciences)- can be obtained in IST, fenix
  - https://www.slideshare.net/sspink/seminar-on-spss
  - https://www.youtube.com/watch?v=ADDR3_Ng5CA
- R (https://www.r-project.org/)



HOW TO REPORT A STUDY?

STRUCTURE

- Motivation and Research Question
  - Includes the hypothesis
- Methods
- Results
- Discussion



MOTIVATION

- Motivation: which research question did you set out to answer? Why?
- What was your expected answer or assumptions about the outcome of this investigation?
  - Hypothesis?
  - Designed to prove?
- Relate assumptions to findings

METHODS

- How was the experiment designed?
- Subjects and Materials
  - On what subjects or materials was the experiment performed?
  - How were the subjects/materials prepared?
- Setting
  - What machinery/equipment was used?
- What sequence of events did you follow as you handled the subjects/materials or as you recorded the data?

NOTE:

- Must be an accurate and complete account of what was done and what materials (robots characteristics etc) used
- Usually a chronological structure
- Write in the past tense

RESULTS

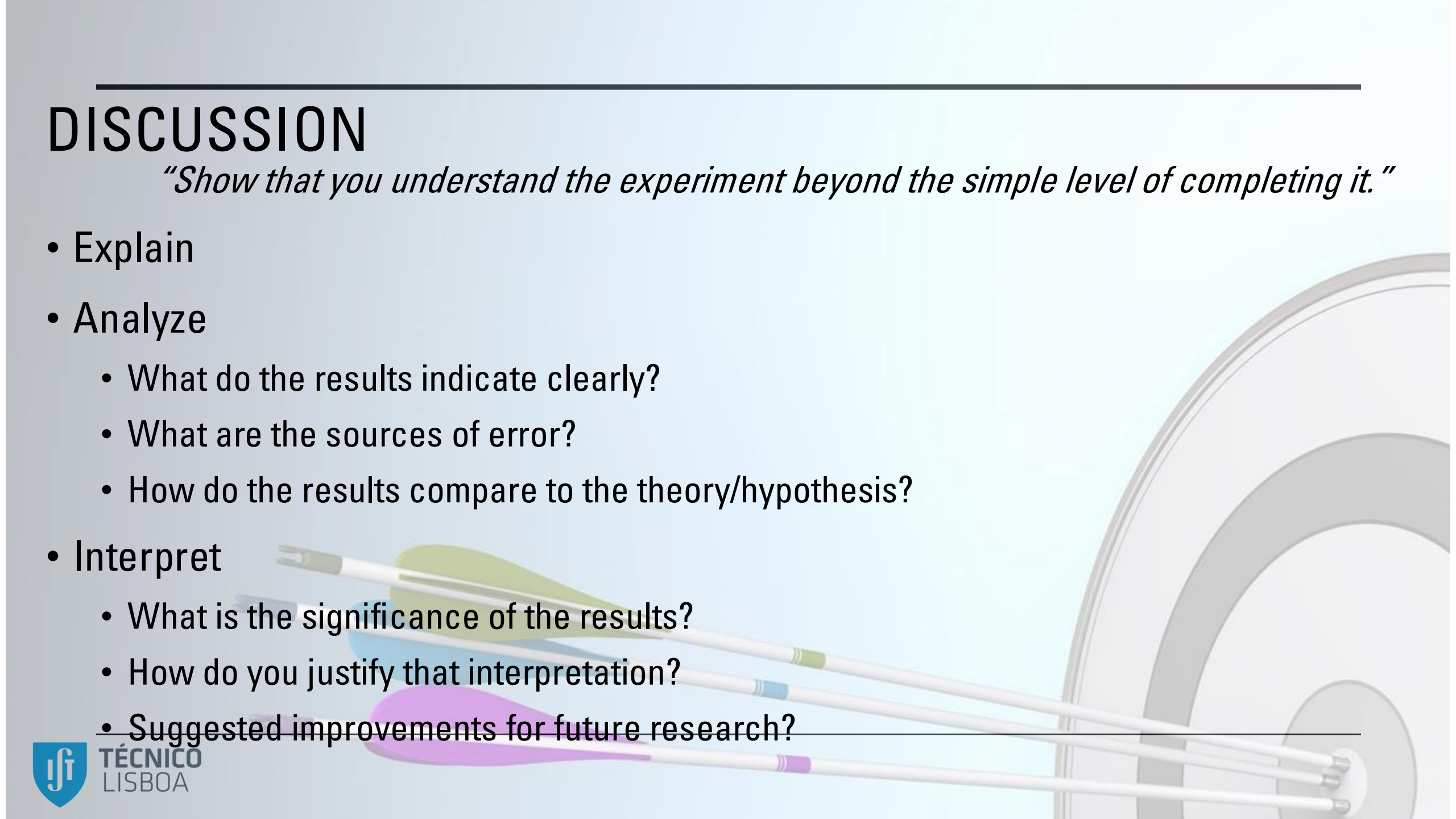
- Present data (descriptive statistics and inferential statistics)
- State in verbal form as well as visual (provide graphs and figures)
- Use sentence to draw attention to key points in graphs, figures, etc.
- Number and title tables and graphs
- Use appendix for raw data or complex calculations (if needed)

- What are the main results?
- Is the data presented so results are clear, logical and self-explanatory?
- What is the main point – what ties results together?

DISCUSSION

"Show that you understand the experiment beyond the simple level of completing it."

- Explain
- Analyze
  - What do the results indicate clearly?
  - What are the sources of error?
  - How do the results compare to the theory/hypothesis?
- Interpret
  - What is the significance of the results?
  - How do you justify that interpretation?
  - Suggested improvements for future research?



TÉCNICO LISBOA

FURTHER READING

- *Book: How to design and Report Experiments, by Andy Field and Graham Hole, SAGE, 2003*