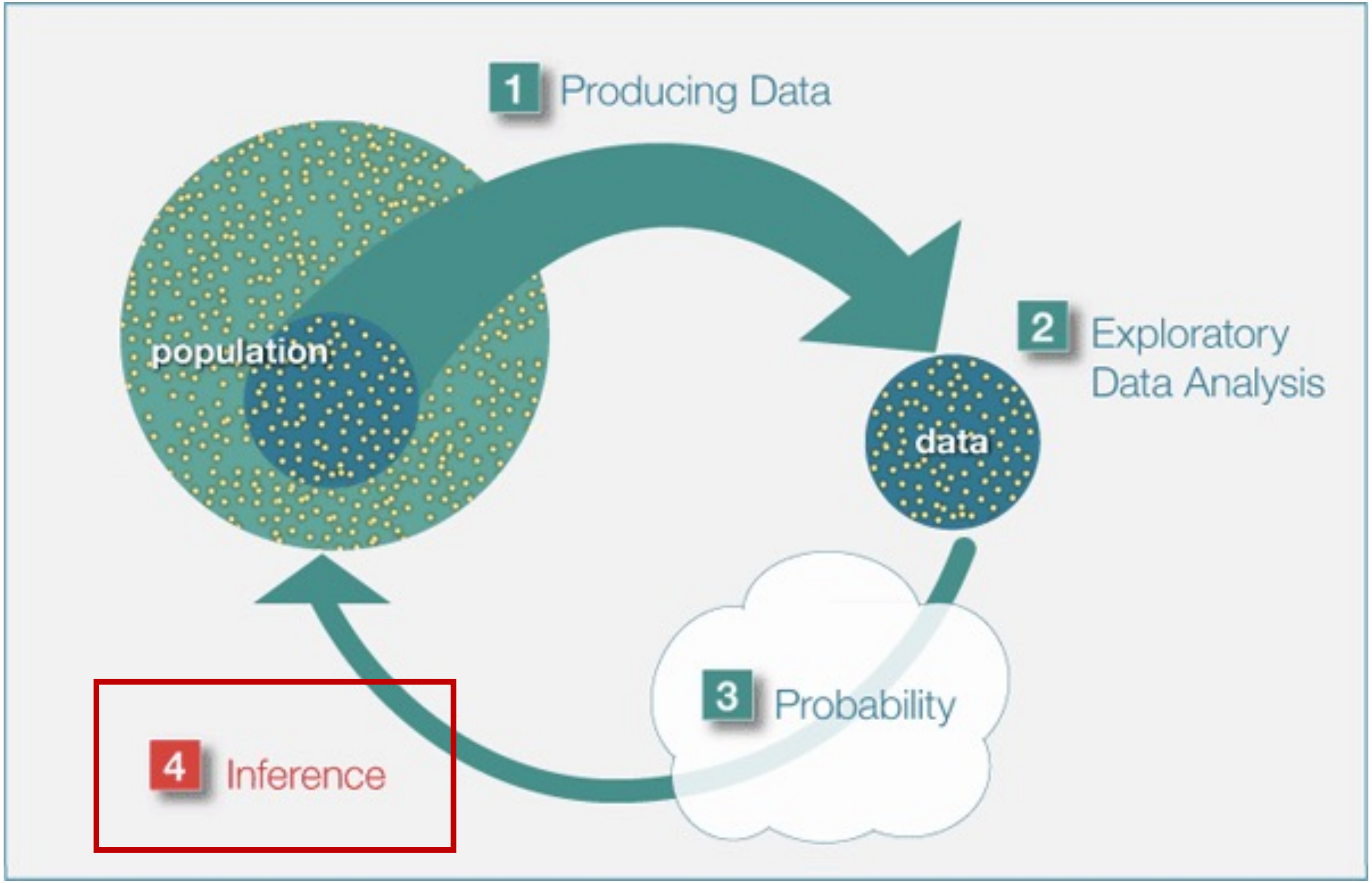
Special Topics in Biostatistics and Bioinformatics
Week VI

Ege Ülgen, M.D.

7 April 2022

ACIBADEM
MEHMET ALİ AYDINLAR
ÜNİVERSİTESİ



Unit 4a: introduction to statistical inference [Internet]. [cited 2021 Sep 27]. Available from: https://bolt.mph.ufl.edu/6050-6052/unit-4/

Parameter vs. Statistic

- Parameters are characteristics of the population
- We estimate population parameters with data
  - But the estimation may change from sample to sample
- Goal: Estimate the parameter + How certain are we?

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$CI = \left(\bar{X} - c\,\frac{s_x}{\sqrt{n}}, \bar{X} + c\,\frac{s_x}{\sqrt{n}}\right)$$

Hypothesis Testing

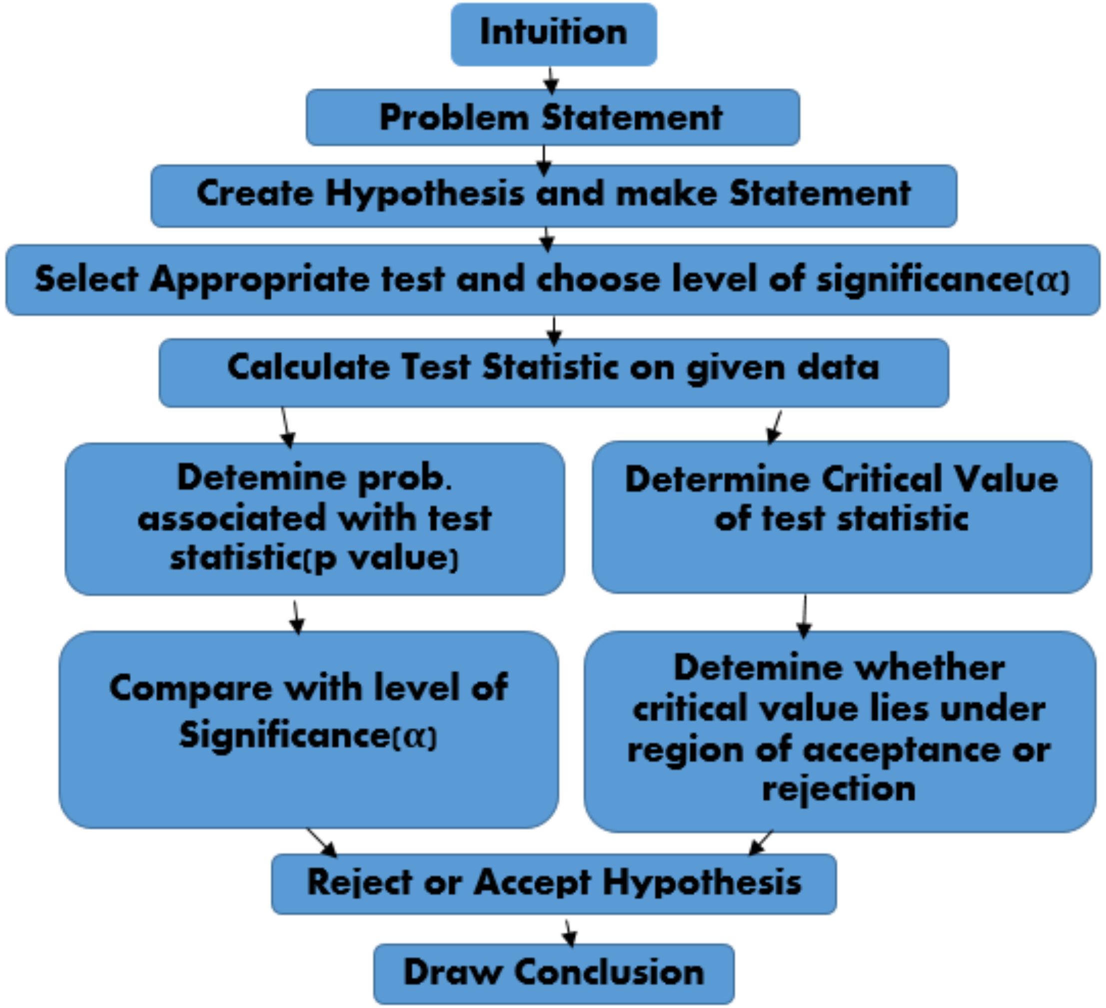
- A form of inference
- A decision rule

$$Expr = \beta_0 + \beta_1 Treatment + \varepsilon$$

- H_0: The relationship between treatment and expression is exactly zero
- H_1: The relationship between treatment and expression is not zero

Barai A. Hypothesis Testing | All a beginner needs to know [Internet]. Medium. 2020 [cited 2021 Oct 4]. Available from: https://medium.com/analytics-vidhya/hypothesis-testing-all-a-beginner-needs-to-know-dd1bdeebdbf1

Hypothesis Testing - Steps

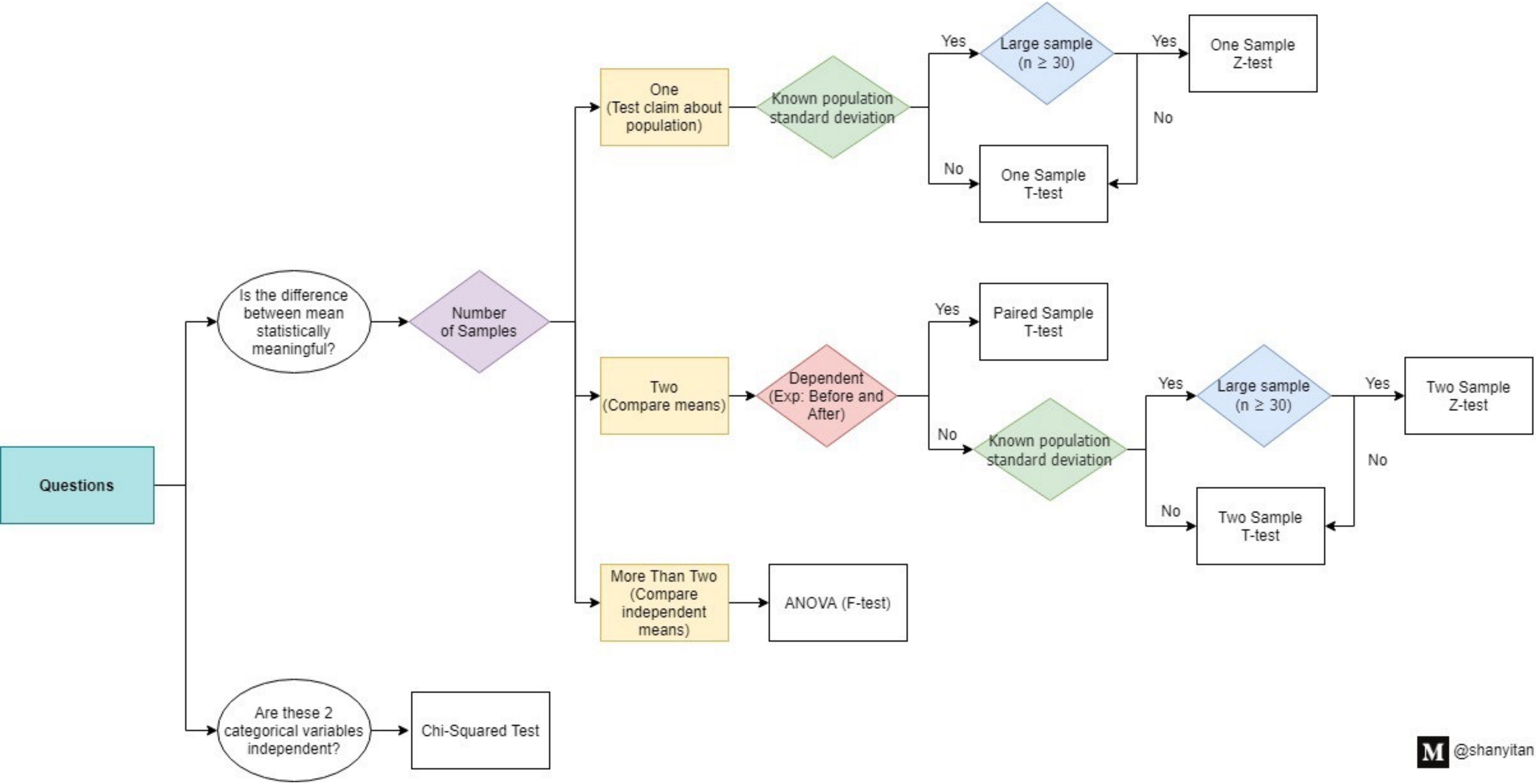


The most popular statistic: t-statistic

$$t = \frac{\hat{\beta}}{se(\hat{\beta})}$$

Moderated t-statistic (Remove bias due to small variability):

$$t = \frac{\hat{\beta}}{se\left(\hat{\beta}\right) + c}$$


Questions
Is the difference between mean statistically meaningful?
Number of Samples
One (Test claim about population)
Known population standard deviation
Yes
Large sample (n ≥ 30)
Yes
One Sample Z-test
No
No
One Sample T-test
Two (Compare means)
Dependent (Exp: Before and After)
Yes
Paired Sample T-test
No
Known population standard deviation
Yes
Large sample (n ≥ 30)
Yes
Two Sample Z-test
No
No
Two Sample T-test
More Than Two (Compare independent means)
ANOVA (F-test)
Are these 2 categorical variables independent?
Chi-Squared Test
@shanyitan

Non-parametric Tests

- Often used when assumptions of parametric tests are not met
- **Robust with respect to the distribution of data**
- **Less assumptions**
  - e.g., they do not depend on the assumption of normality

- **Less statistical power** compared to parametric tests
  - Higher risk of type II errors (e.g., high probability of accepting there is no difference between the groups where there is a difference)

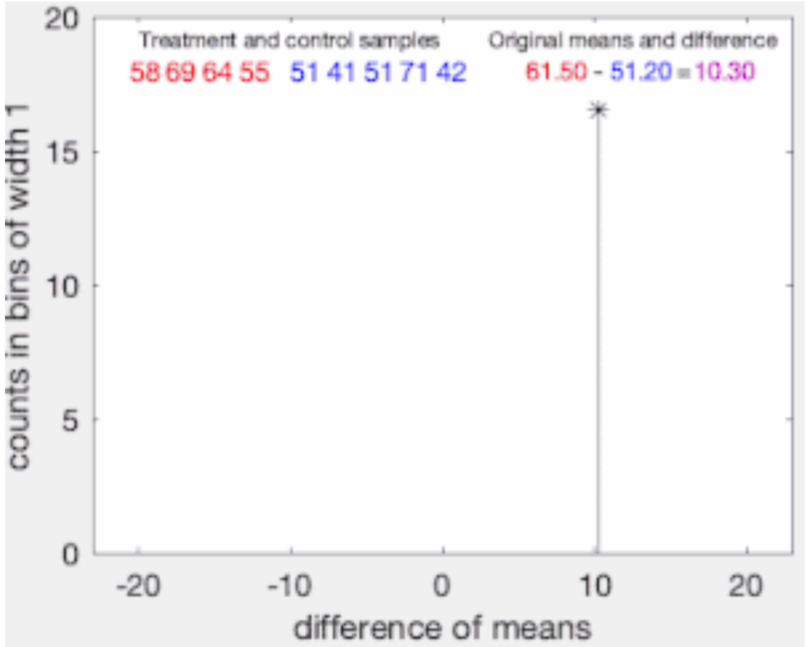
Non-parametric Tests

- **χ^2 test**
- **Wilcoxon rank-sum test (Mann–Whitney U test)** ~ Independent samples t-test
- **Kruskal-Wallis test** ~ one-way ANOVA
- **Mood's Median Test** ~ one-way ANOVA
- **Friedman test** ~ two-way ANOVA
- **Spearman's rank correlation test** ~ Pearson correlation test
- …

Permutation Test

- Permute labels to "break any relationship"
- the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under all possible rearrangements of the observed data points

https://www.jwilber.me/permutationtest/



H_0	Decision	
	Fail to reject	Reject
True	Correct decision	**Type I Error** α
False	**Type II Error** **ß**	Correct decision

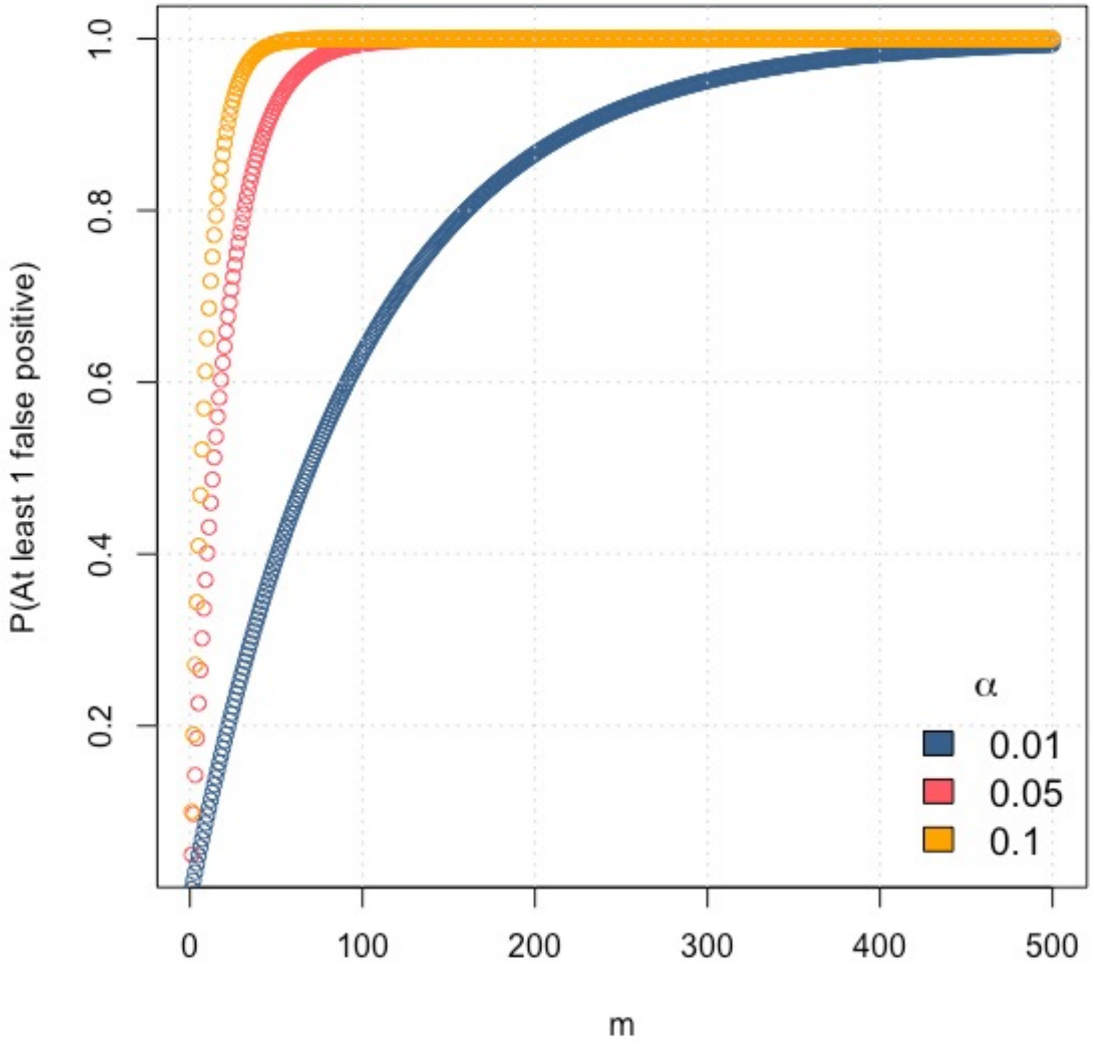
Multiple Testing - Example

- A typical microarray experiment might result in performing 10000 separate hypothesis tests
- If we use a standard p-value cut-off of 0.05, we'd expect ***500*** genes to be deemed "significant" by chance

Multiple Testing

- P(making a type I error) = α
- P(not making a type I error) = $1 - \alpha$
- P(not making a type I error in m tests) = $(1 - \alpha)^m$
- P(making at least 1 type I error in m tests) = $1 - (1 - \alpha)^m$

Multiple Testing



Approaches to Control Type I Error Rate (V)

- Per comparison error rate (PCER)
- Per-family error rate (PFER)
- **Family-wise error rate (FWER) – Bonferroni, Holm**
- **False discovery rate (FDR) - Benjamini & Hochberg**
- Positive false discovery rate (pFDR)