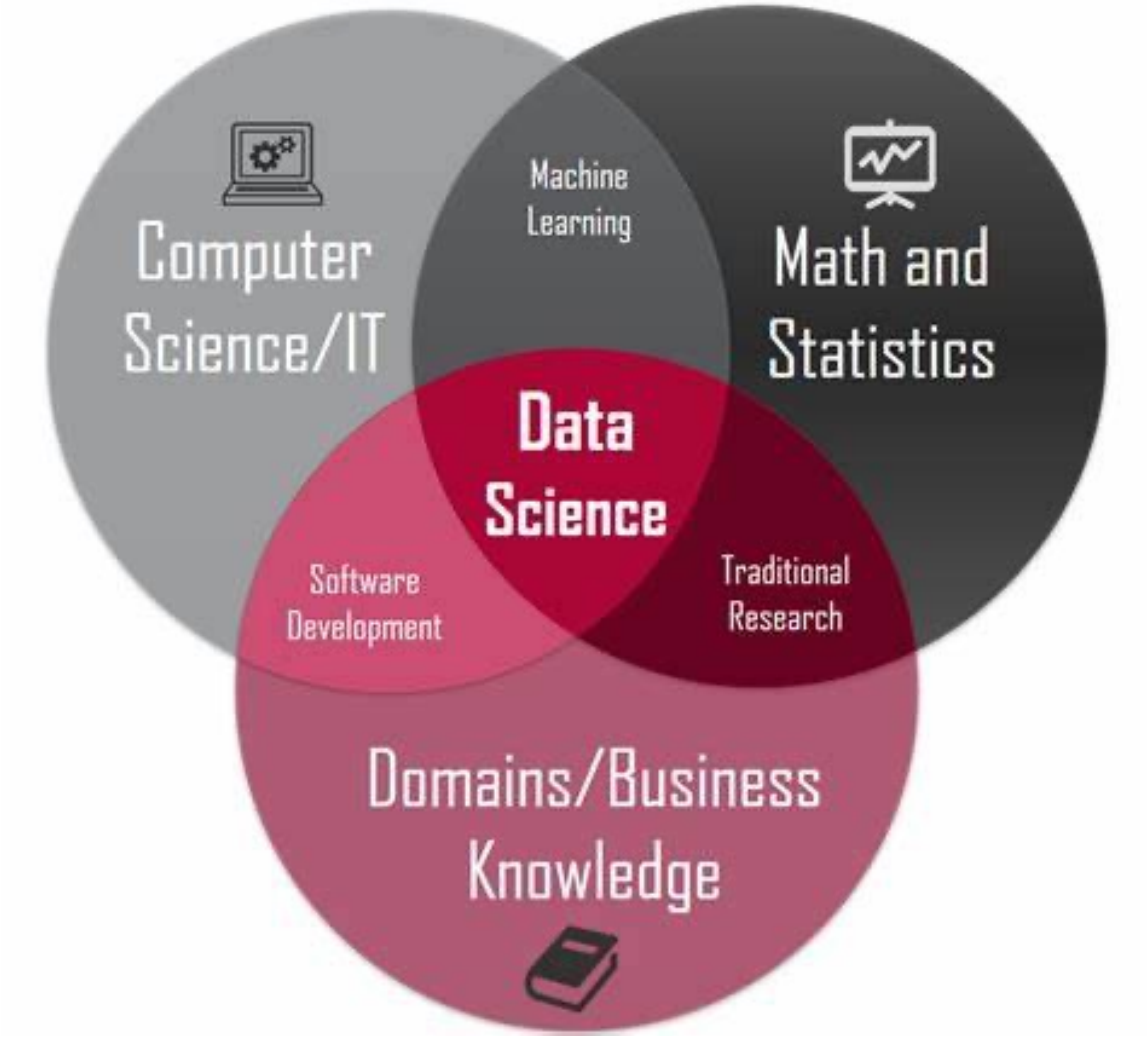
STATISTICS REVIEW

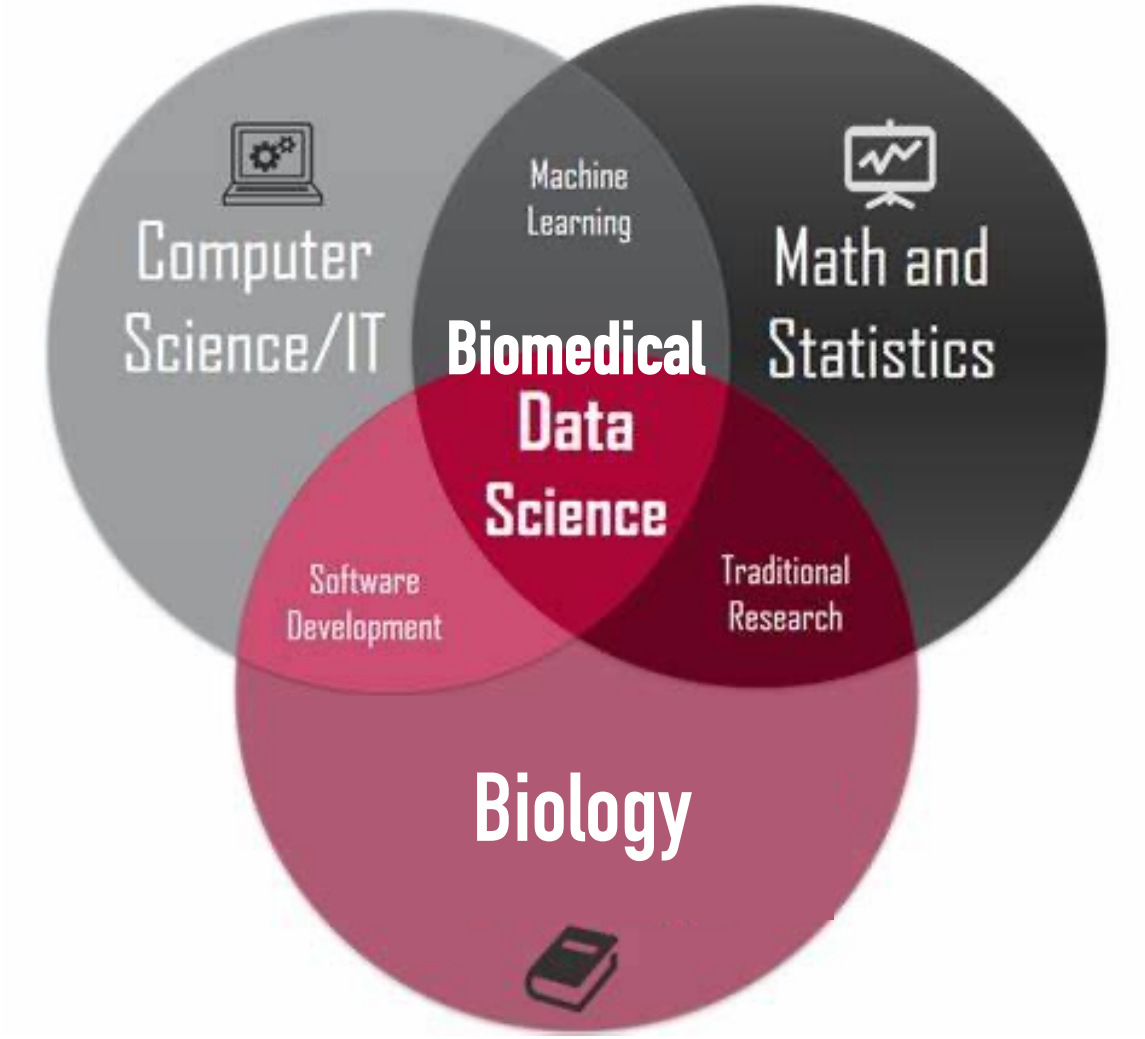
OUTLINE

- Random variable
- Probability distribution
- Central limit theorem
- Hypothesis testing
- P-value
- Multiple testing correction
- Type I and type II errors
- False discovery rate (FDR)

Data Science



Computational Biology



Why statistics is important

- Statistics is the theoretical foundation of machine learning and data science
- Statistics is the bridge between experiments and theories
- Statistics is the bridge between observations (data) and discoveries (science)

What is a model

- A model describes the relationship between quantities
  - Quantitative/mathematical models
- Statistical models
  - Relationship between random variables
- “All models are wrong; but some are useful.” (George Box)

Random variables

- Randomness
- All quantities obtained from experimental measurements can be considered as random variables
- Discrete vs. continuous random variables

- Probability distribution

Discrete probability distributions

- Bernoulli distribution

$$\Pr(X = 1) = p = 1 - \Pr(X = 0) = 1 - q.$$

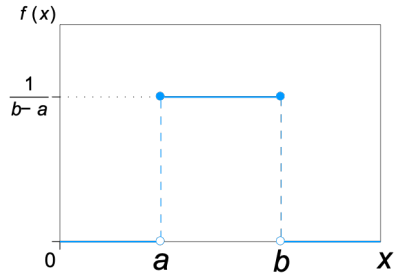
- Binomial distribution $\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

- Negative binomial distribution $\Pr(X = k) = \binom{k+r-1}{k} (1-p)^k p^r$

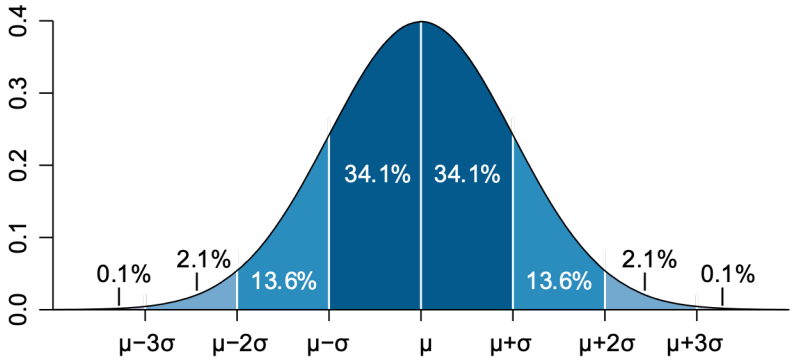
- Poisson distribution $\Pr(X{=}k) = \dfrac{\lambda^k e^{-\lambda}}{k!}$

Continuous probability distributions

- Uniform distribution
- Normal distribution (Gaussian distribution)

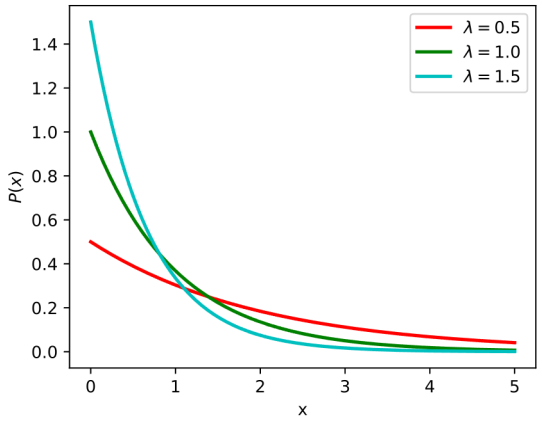


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



- Exponential distribution

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$



Use of probability distributions in computational biology

- Null models for hypothesis testing
- Noise or background models for signal detection and differential enrichment analysis
- Error models in regression/machine learning

Central limit theorem

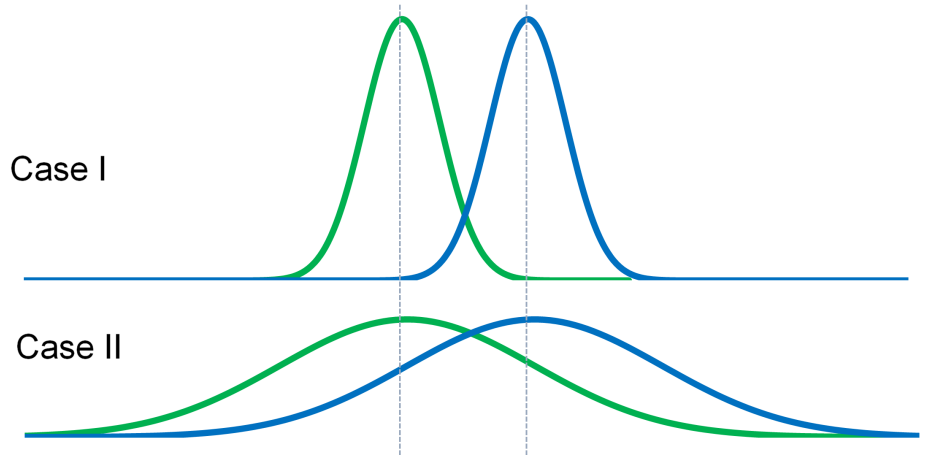
- Regardless of the population distribution, the sample mean will follow a standard normal distribution, if the samples are independent and equal size.

- Theoretical foundation of t-test

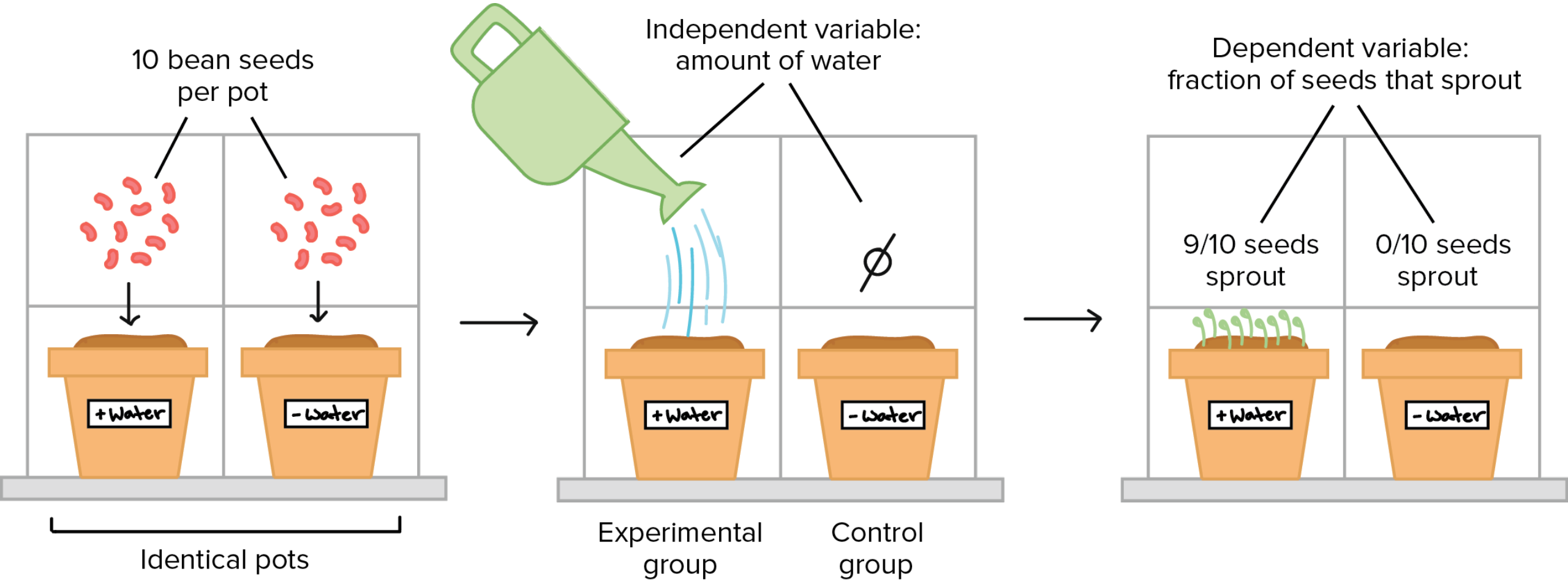
Hypothesis testing

- General thinking:
  - Are they different?
  - Is the difference “statistically significant”?

- Statistical thinking:
  - Null hypothesis
  - Alternative hypothesis



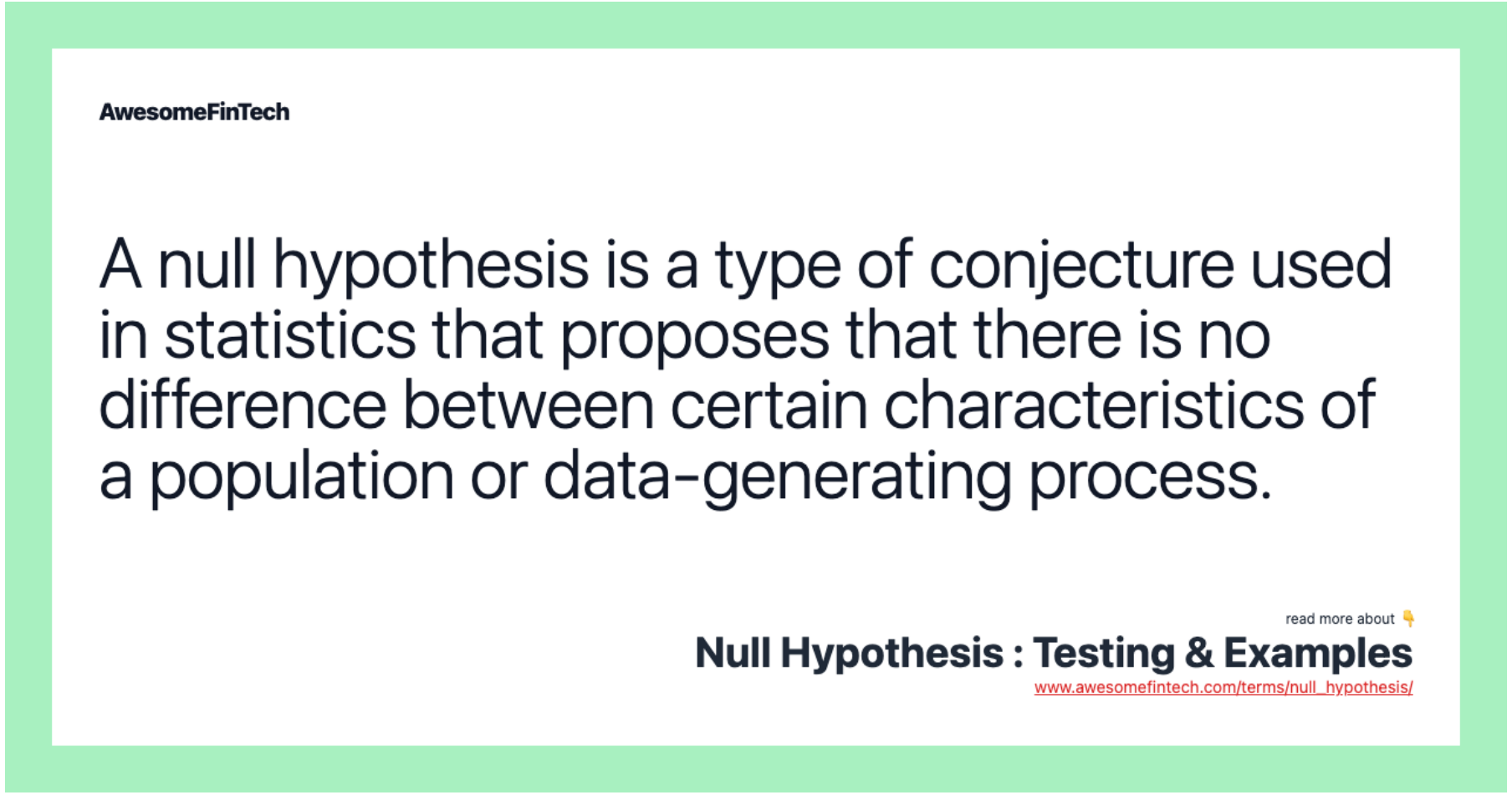
Negative control in experiments



Source: https://www.khanacademy.org/science/biology/intro-to-biology/science-of-biology/a/experiments-and-observations

Slide from Jessica Li

Null hypothesis in statistical hypothesis testing

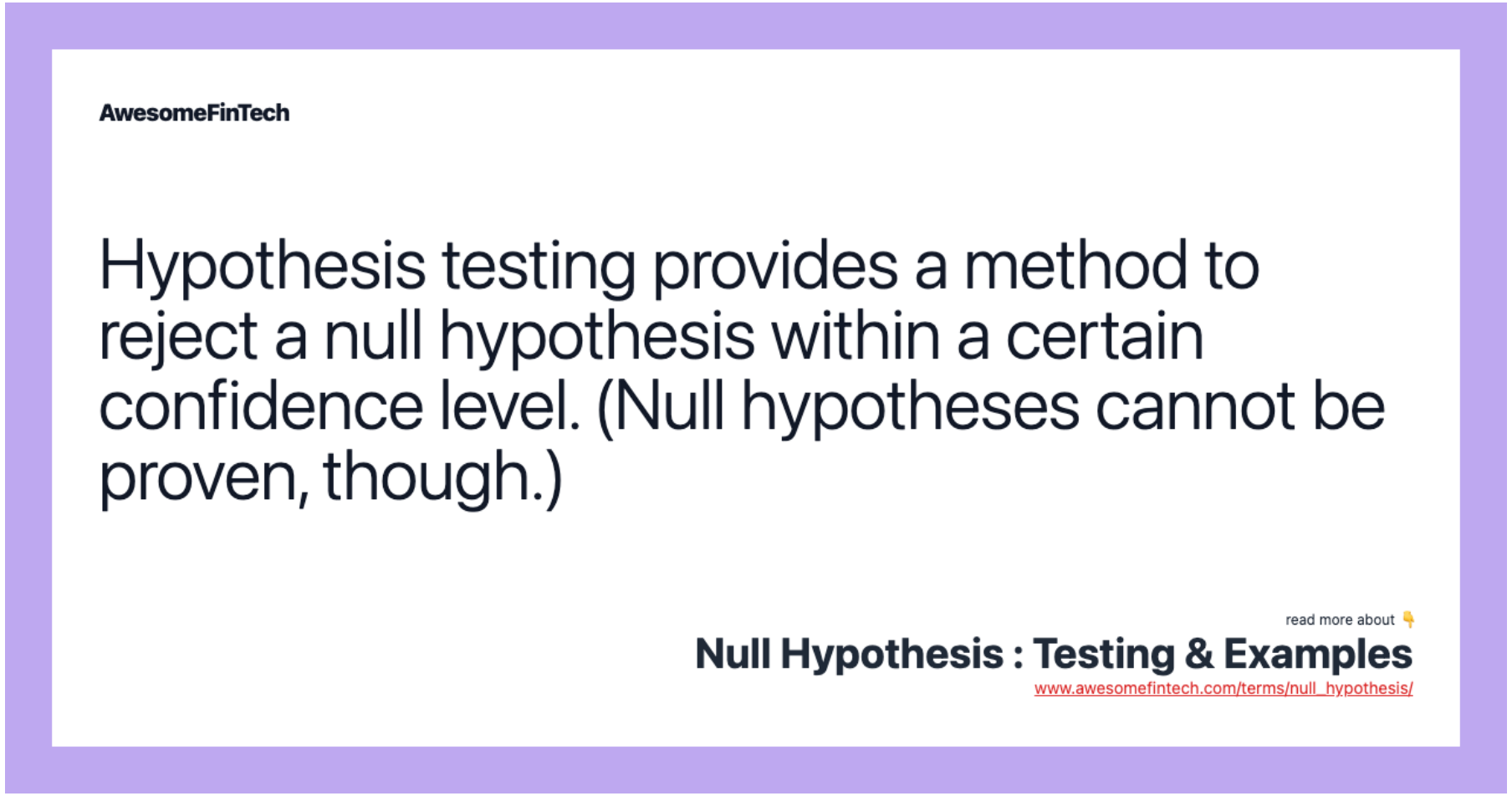


The null hypothesis does not depend on a test procedure

Source: https://www.awesomefintech.com/term/null_hypothesis/

Slide from Jessica Li

Null hypothesis in statistical hypothesis testing

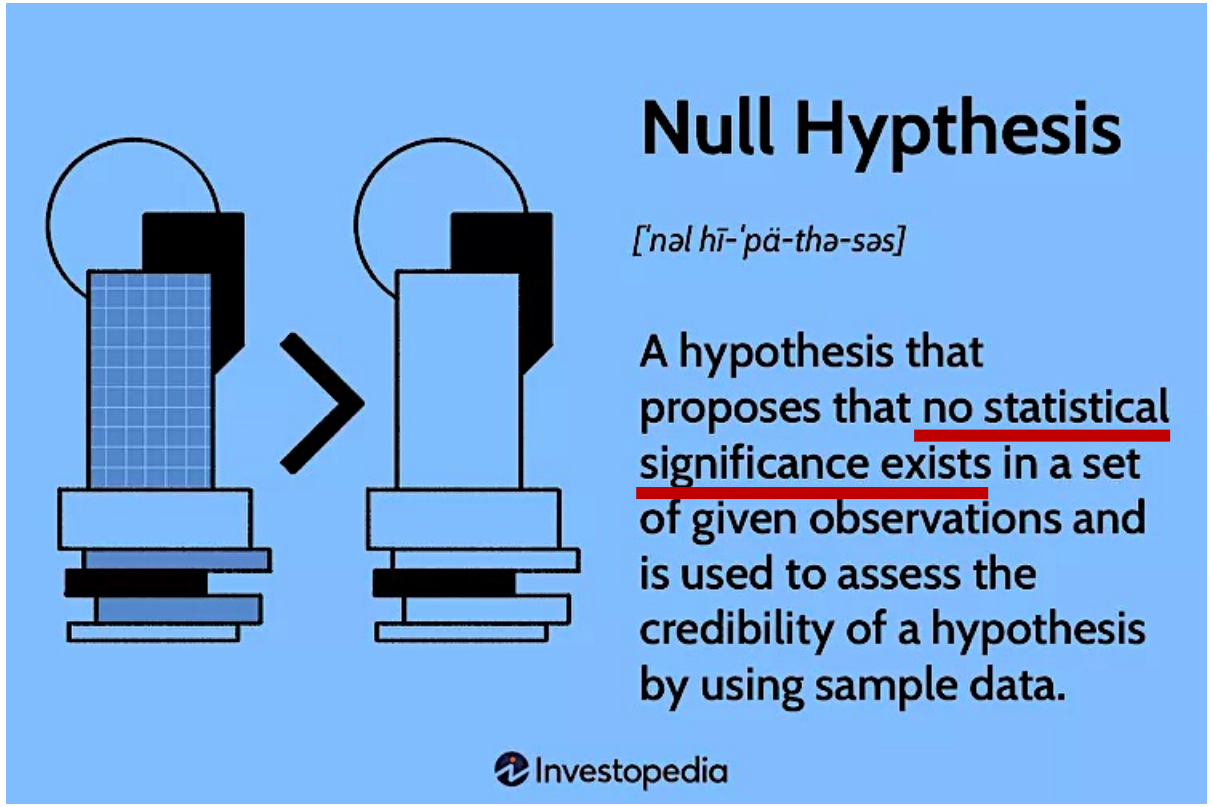


Statistics cannot tell us everything b/c data is random

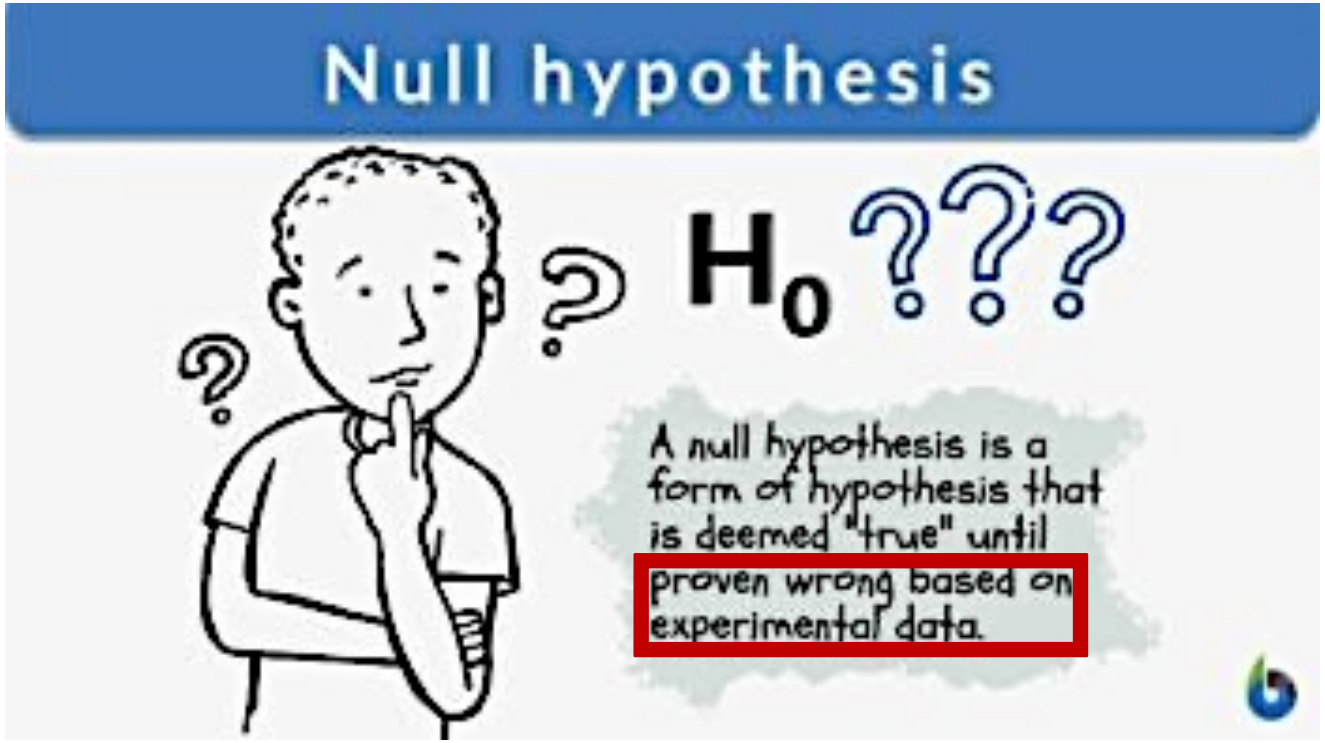
Source: https://www.awesomefintech.com/term/null_hypothesis/

Slide from Jessica Li

Null hypothesis is often misunderstood



Source: https://www.investopedia.com/terms/n/null_hypothesis.asp



Source: https://www.biologyonline.com/dictionary/null-hypothesis

Slide from Jessica Li

Common hypothesis tests

- Student's t-test (parametric)
  - Null (2-sample): The sample means are equal.
- Fisher's exact test
  - Null: The two groups are equally likely for an event/feature.
- Wilcoxon (rank-sum) test
  - Null: The two samples X and Y, $P(X>Y) = P(X<Y) = 0.5$
- Kolmogorov-Smirnov test (K-S test)
  - Null: The two samples have the same cumulative distribution.

- Hypothesis test statistic, p-value

Which of the following statements about p-values is true?

A. P-values measure how big the difference is between the datasets compared.
B. P-value is the probability of observing the data by random chance.
C. P-value is the least probability of observing the data under the assumption that the null hypothesis is true.

ASA statement on statistical significance and p-values

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

ASA statement on statistical significance and p-values

4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Multiple testing correction

- High-throughput experiments
  - RNA-seq: 500 Differentially Expressed genes from 20K genes
  - ChIP-seq: 10,000 TF binding sites from the genome

Bonferroni correction

- Controls family-wise error rate
- N tests
- Adjusted P-value = N * P
- Very strict

Carlo Emilio Bonferroni (1892-1960)

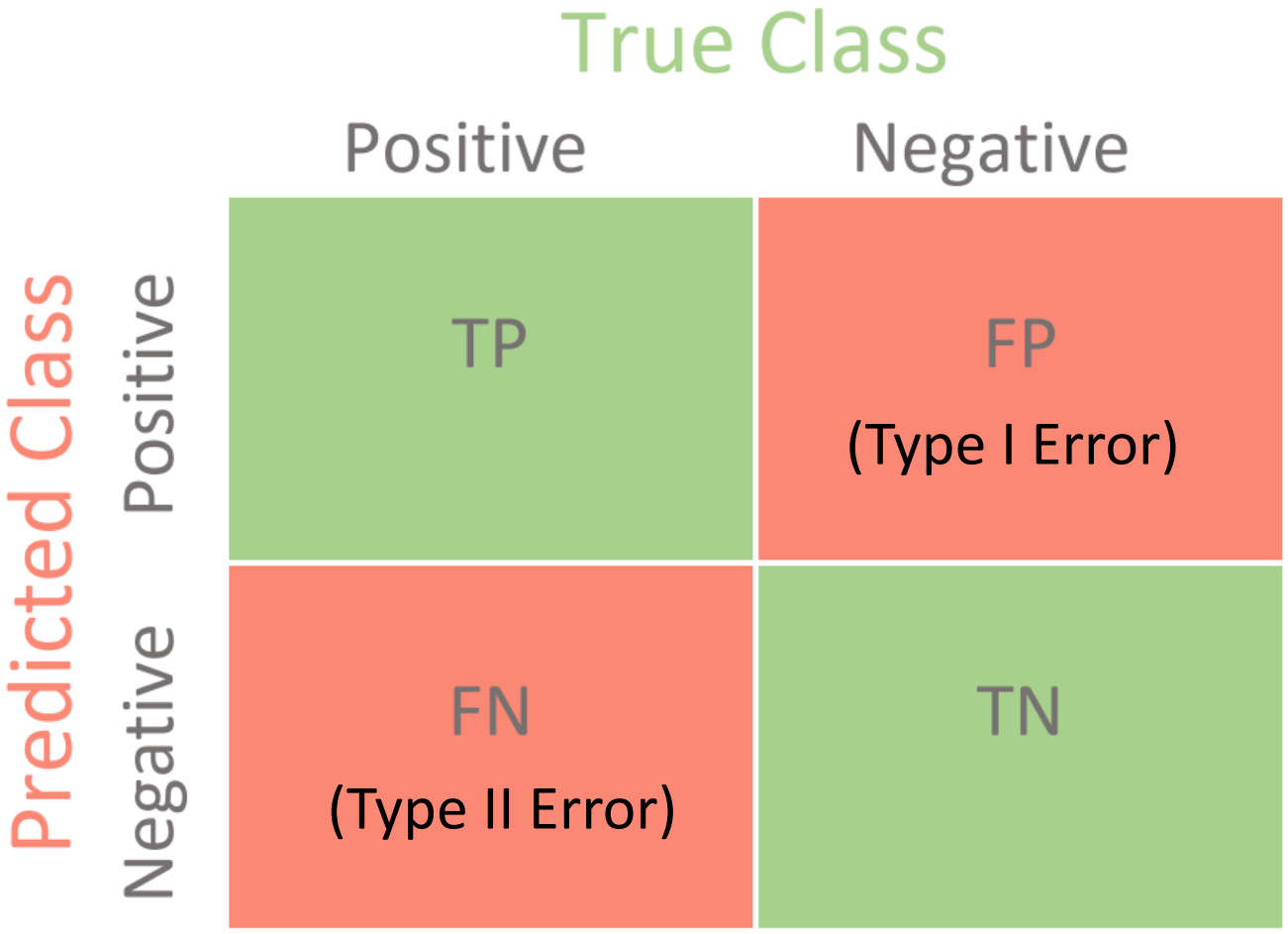
Benjamini-Hochberg (B-H) Correction

- N observations w/ various p-values
- Rank N
- Adjusted P = P * N / R
- Moderate

Yoav Benjamini (1949-)

and Yosef Hochberg

Confusion Matrix



Source: https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

Summary

		CONDITION determined by "Gold Standard"			
	TOTAL POPULATION	CONDITION POS	CONDITION NEG	PREVALENCE $\frac{\text{CONDITION POS}}{\text{TOTAL POPULATION}}$	
TEST OUT-COME	TEST POS	True Pos TP	*Type I Error* False Pos FP	*Precision* Pos Predictive Value PPV = $\frac{TP}{TEST\ P}$	False Discovery Rate FDR = $\frac{FP}{TEST\ P}$
	TEST NEG	*Type II Error* False Neg FN	True Neg TN	False Omission Rate FOR = $\frac{FN}{TEST\ N}$	Neg Predictive Value NPV = $\frac{TN}{TEST\ N}$
	ACCURACY ACC ACC = $\frac{TP + TN}{TOT\ POP}$	*Sensitivity (SN), Recall* Total Pos Rate TPR TPR = $\frac{TP}{CONDITION\ POS}$	*Fall-Out* False Pos Rate FPR FPR = $\frac{FP}{CONDITION\ NEG}$	Pos Likelihood Ratio LR + LR + = $\frac{TPR}{FPR}$	Diagnostic Odds Ratio DOR DOR = $\frac{LR+}{LR-}$
		Miss Rate False Neg Rate FNR FNR = $\frac{FN}{CONDITION\ POS}$	*Specificity (SPC)* True Neg Rate TNR TNR = $\frac{TN}{CONDITION\ NEG}$	Neg Likelihood Ratio LR - LR - = $\frac{TNR}{FNR}$	

Source: Wikimedia. Author: Lavender888000

Li *et al. Genome Biology* (2022) 23:79
https://doi.org/10.1186/s13059-022-02648-4

Genome Biology

SHORT REPORT Open Access

Exaggerated false positives by popular differential expression methods when analyzing human population samples

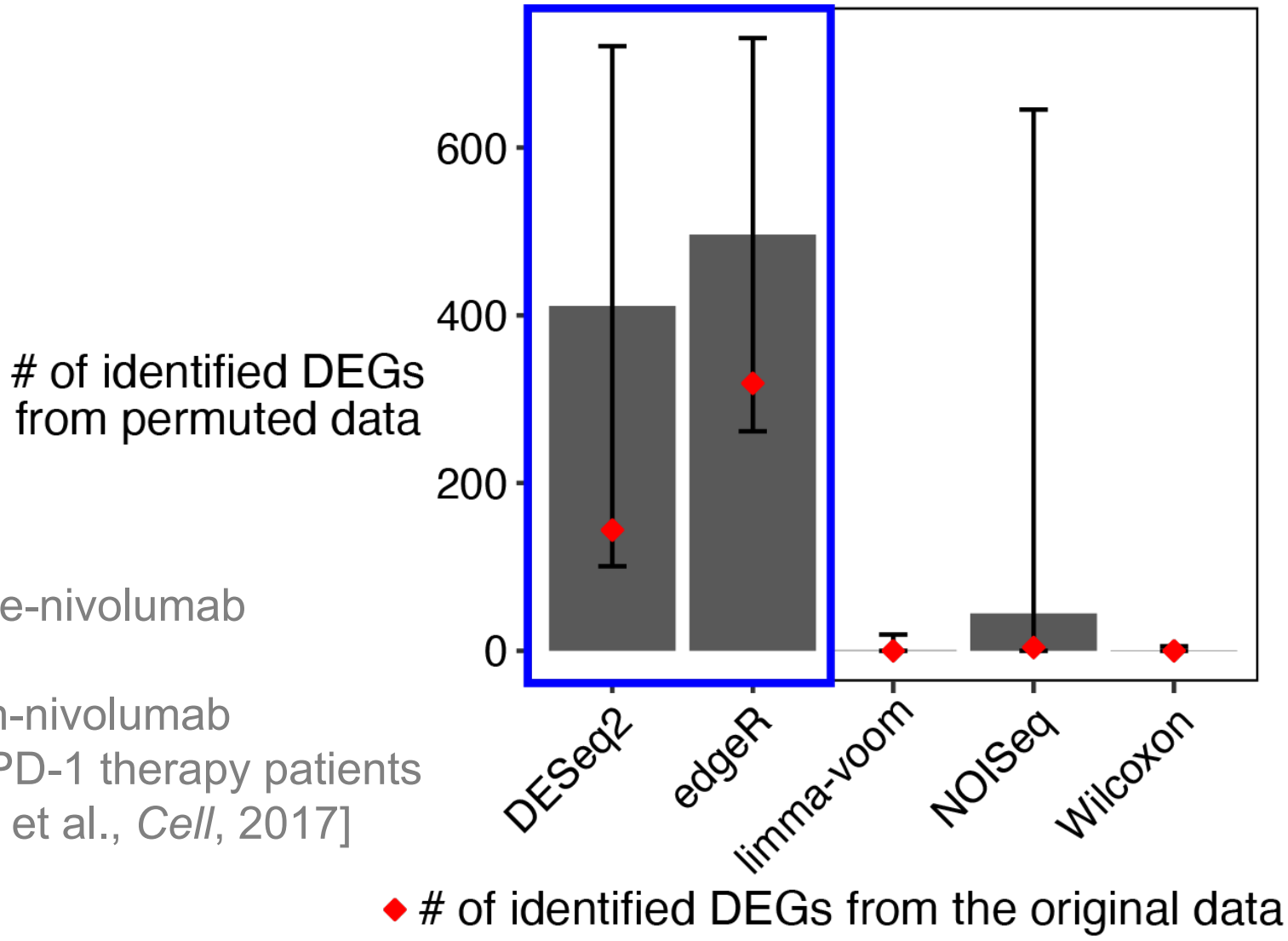
Yumei Li[1†], Xinzhou Ge[2†], Fanglue Peng[3], Wei Li[1*] and Jingyi Jessica Li[2,4,5,6,7*]

*Correspondence: wei.li@uci.edu; lijy03@g.ucla.edu
†Yumei Li and Xinzhou Ge contributed equally to this work.
[1] Division of Computational Biomedicine, Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, CA 92697, USA
[2] Department of Statistics, University of California, Los

Abstract

When identifying differentially expressed genes between two conditions using human population RNA-seq samples, we found a phenomenon by permutation analysis: two popular bioinformatics methods, DESeq2 and edgeR, have unexpectedly high false discovery rates. Expanding the analysis to limma-voom, NOISeq, dearseq, and Wilcoxon rank-sum test, we found that FDR control is often failed except for the Wilcoxon rank-sum test. Particularly, the actual FDRs of DESeq2 and edgeR sometimes exceed 20% when the target FDR is 5%. Based on these results, for population-level RNA-seq studies with large sample sizes, we recommend the Wilcoxon rank-sum test.

Why there are many DE genes identified from permuted data?



51 pre-nivolumab
vs.
58 on-nivolumab
anti-PD-1 therapy patients
[Riaz et al., *Cell*, 2017]

Li *et al*. *Genome Biol* 2022

Important assumption in DESeq2 and edgeR

Both DESeq2 and edgeR assume a **negative binomial (NB)** distribution per gene and condition

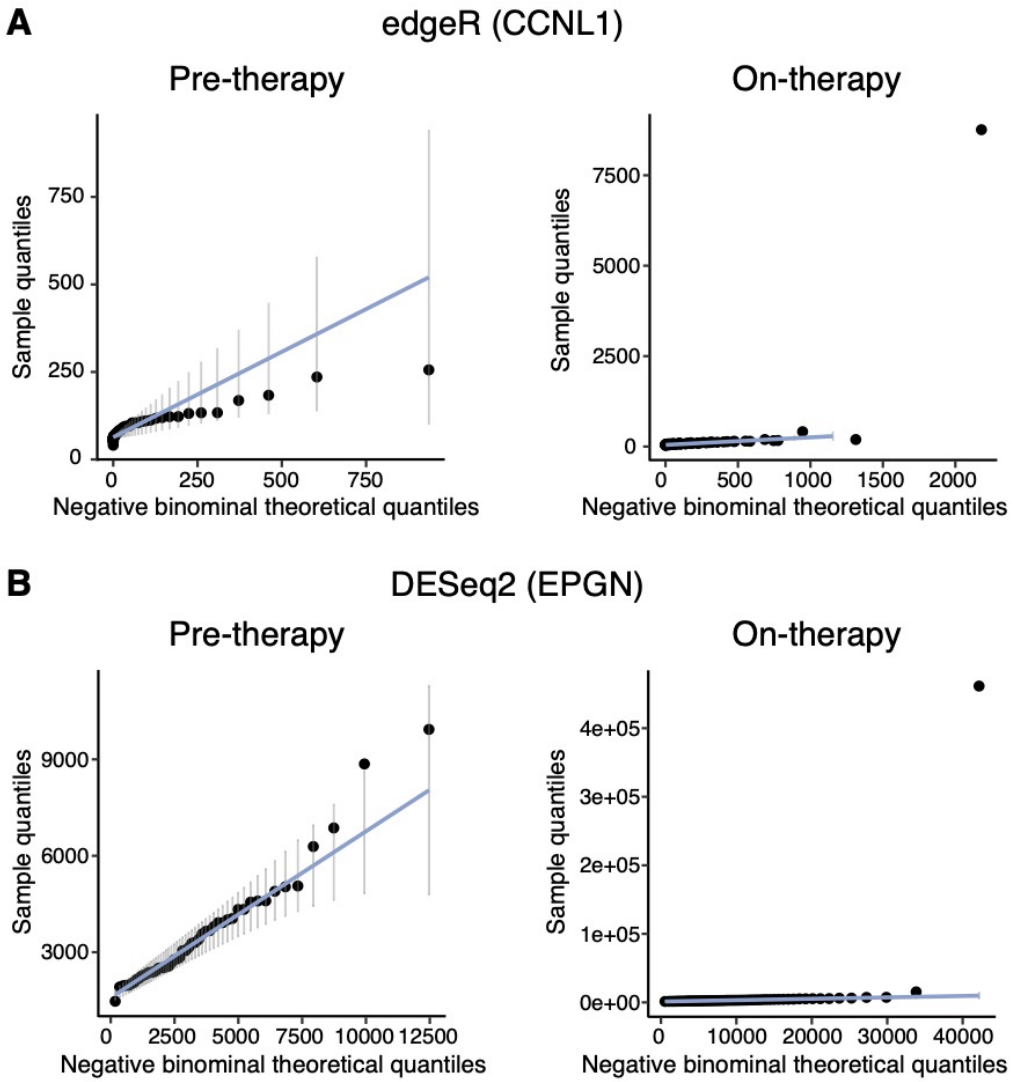
For each gene,

- Condition 1: $X_i \overset{\text{ind}}{\sim} \text{NB}(\mu_1 s_i, \sigma_1),\ i = 1, \ldots, n$
- Condition 2: $Y_j \overset{\text{ind}}{\sim} \text{NB}(\mu_2 s_j, \sigma_2),\ j = 1, \ldots, m$

Null hypothesis $\boldsymbol{H_0}: \mu_1 = \mu_2$

appropriate ***only if*** the NB assumption is reasonable

Gene expression can deviate from NB distribution



Li *et al*. *Genome Biol* 2022

Why does Wilcoxon test work in this scenario?

For each gene, the normalized counts

Condition 1: $\widetilde{X}_i,\ i = 1, \ldots, n$

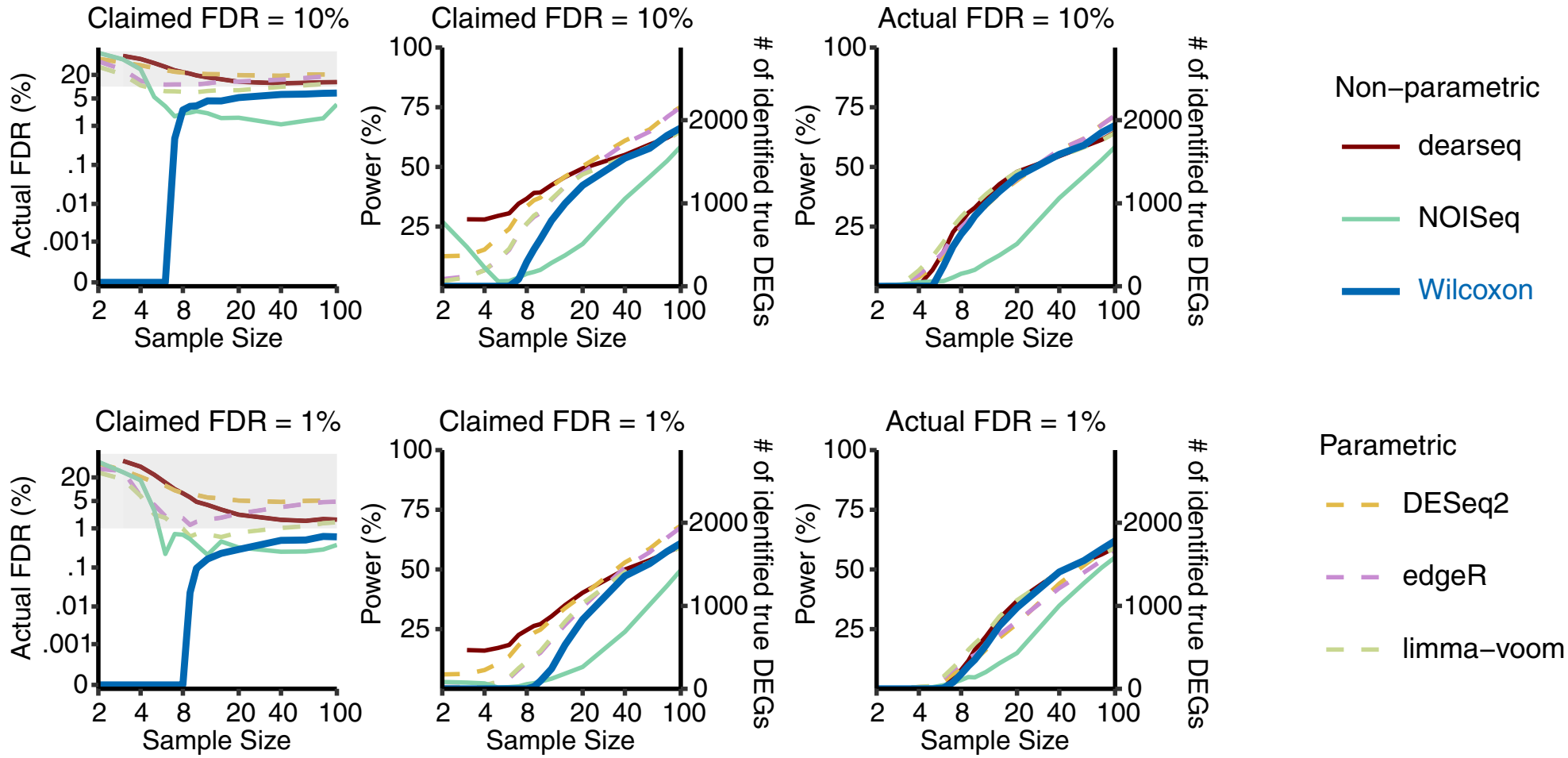
Condition 2: $\widetilde{Y}_j,\ j = 1, \ldots, m$

Null hypothesis (approximate, ignoring ties):

$$\boldsymbol{H_0} : \mathbb{P}(\widetilde{X}_i > \widetilde{Y}_j) = 0.5,\ \text{for all } i, j$$

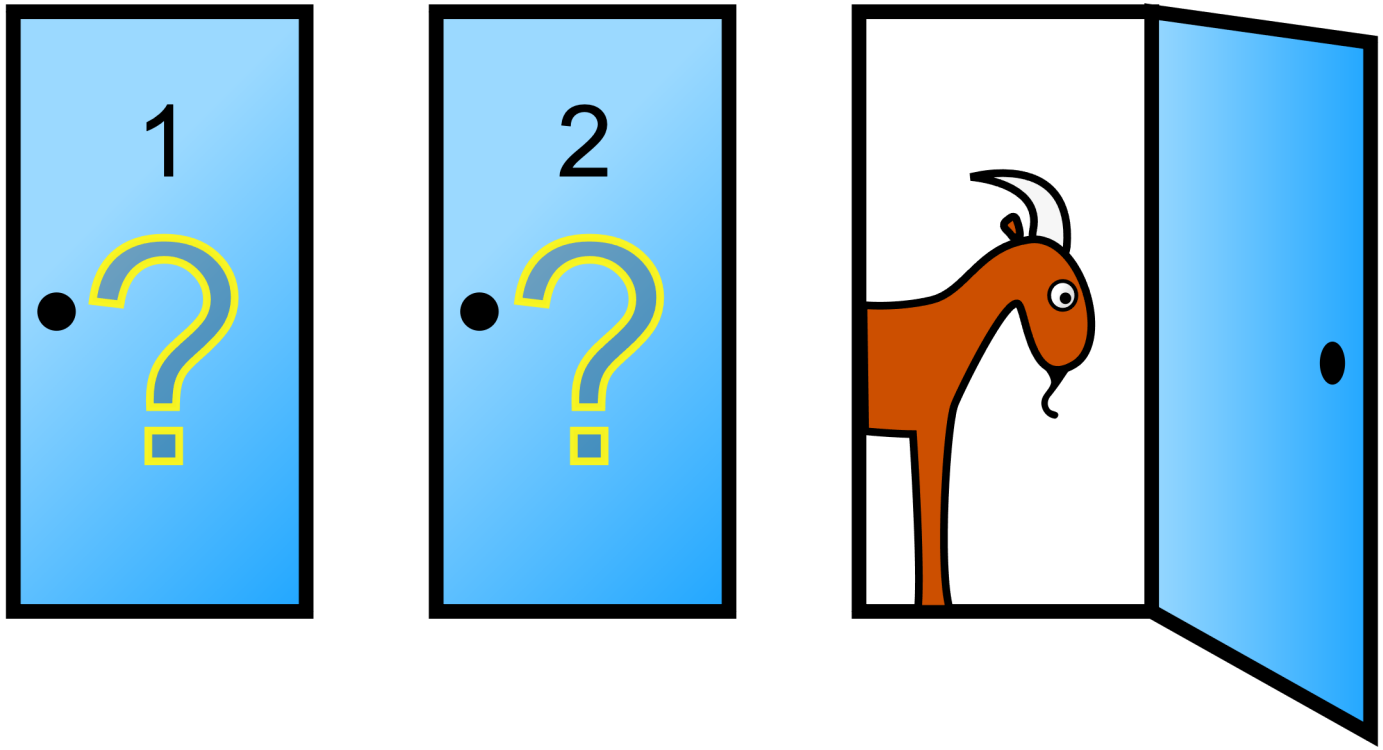
which does NOT have the NB assumption

Wilcoxon test is better when sample size > 8



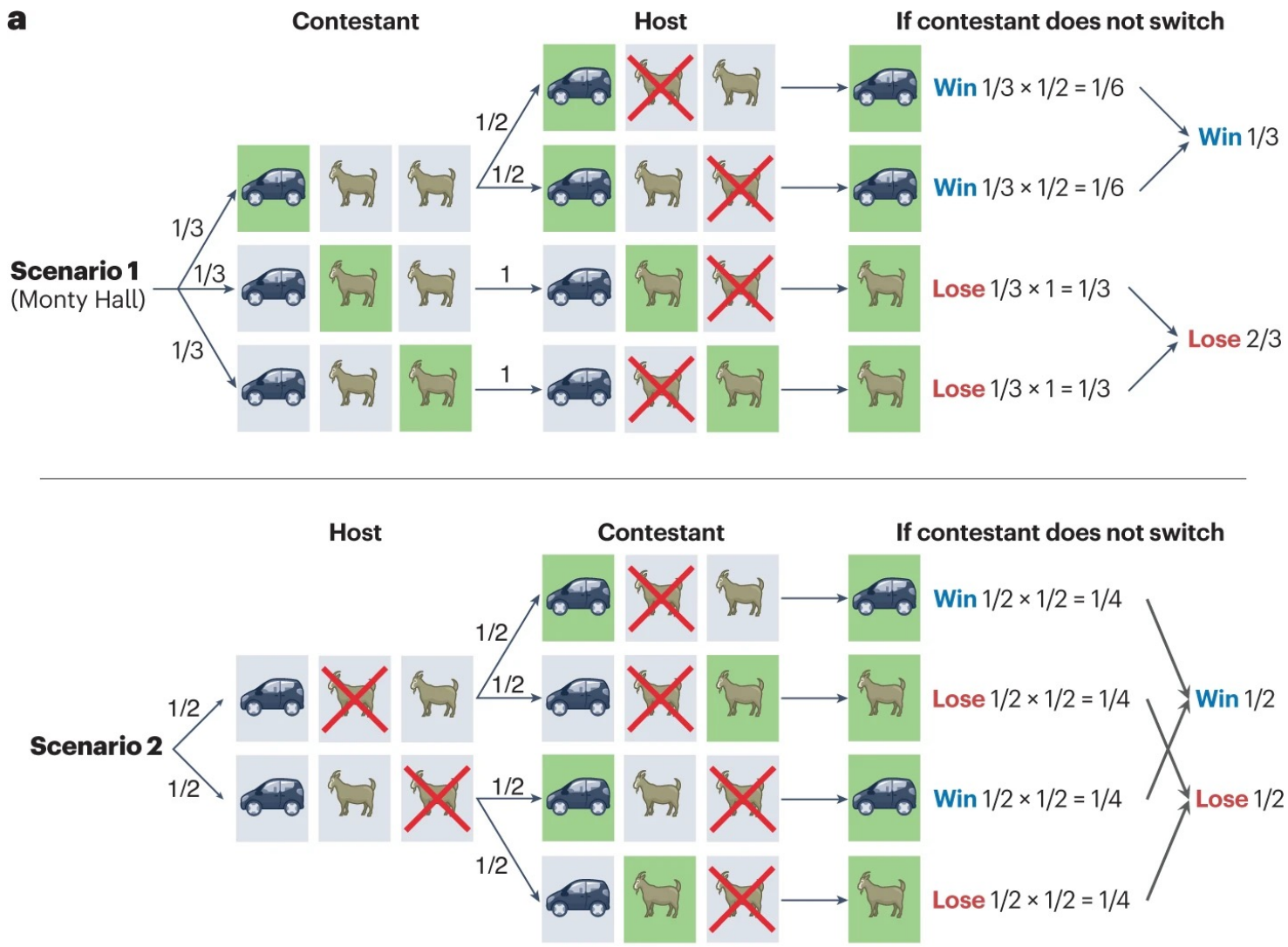
Li *et al*. *Genome Biol* 2022

Monty Hall Problem



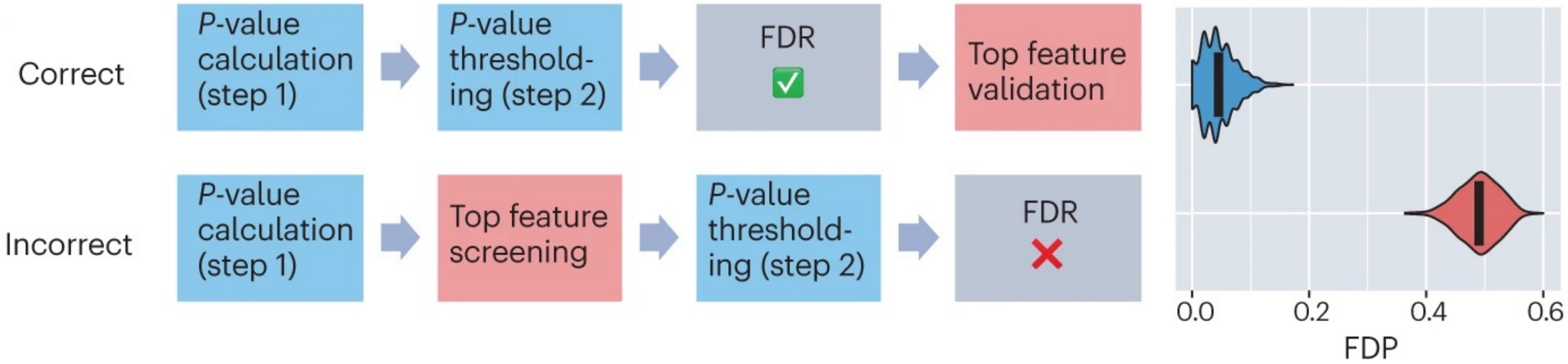
Source: Wikipedia

Monty Hall Problem



Li *Nat Biotech* 2023

The order of action matters



Li *Nat Biotech* 2023

PLOS | COMPUTATIONAL BIOLOGY

EDITORIAL

Ten Simple Rules for Effective Statistical Practice

Robert E. Kass[1], Brian S. Caffo[2], Marie Davidian[3], Xiao-Li Meng[4], Bin Yu[5], Nancy Reid[6]*

1 Department of Statistics, Machine Learning Department, and Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America, **2** Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States of America, **3** Department of Statistics, North Carolina State University, Raleigh, North Carolina, United States of America, **4** Department of Statistics, Harvard University, Cambridge, Massachusetts, United States of America, **5** Department of Statistics and Department of Electrical Engineering and Computer Science, University of California Berkeley, Berkeley, California, United States of America, **6** Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

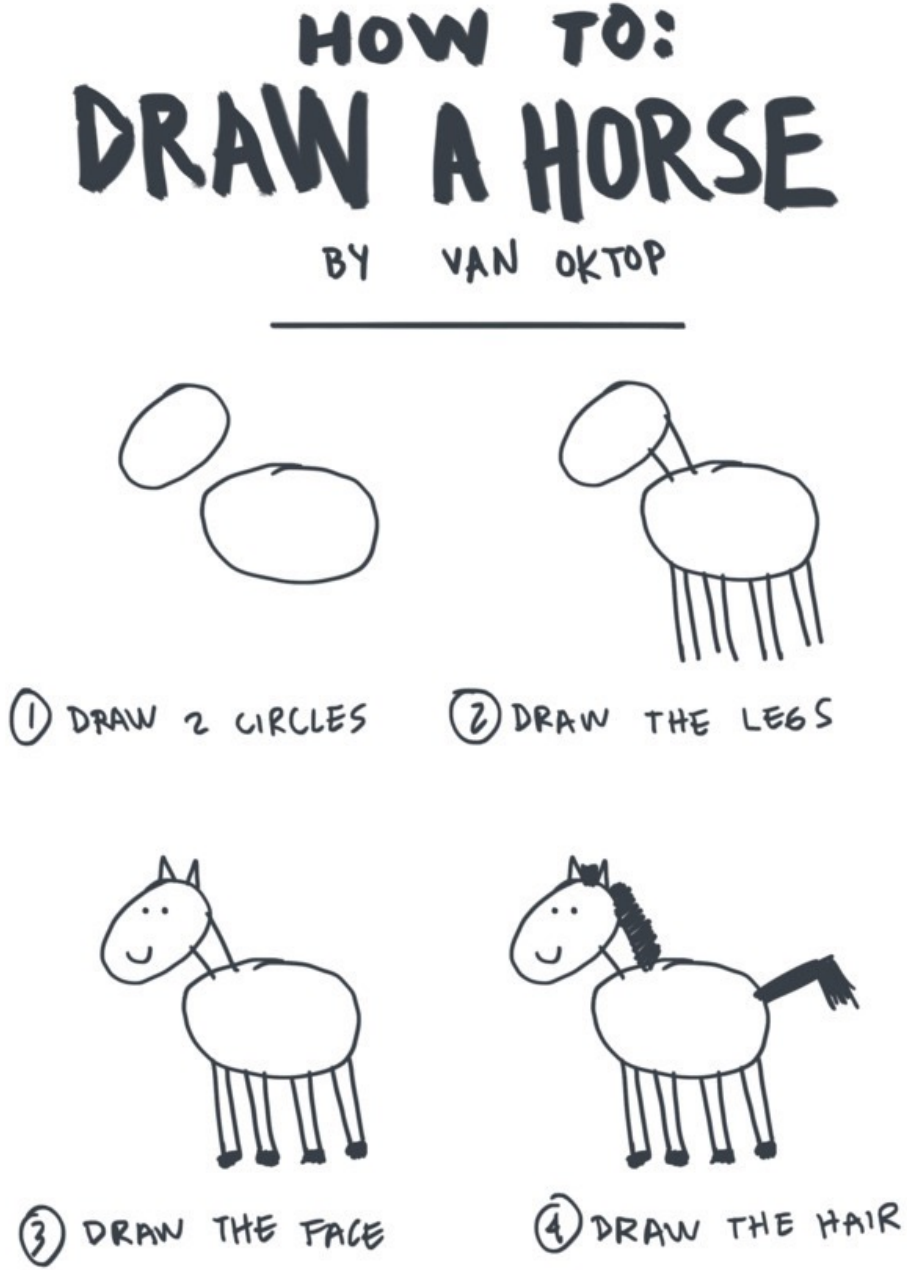
* reid@utstat.utoronto.ca

Ten Simple Rules for Effective Statistical Practice

1. Statistical methods should enable data to answer scientific questions.
2. Signals always come with noise.
3. Plan ahead, really ahead.
4. Worry about data quality.
5. Statistical analysis is more than a set of computations.

Ten Simple Rules for Effective Statistical Practice

6. Keep it simple.
7. Provide assessments of variability.
8. Check your assumptions.
9. When possible, replicate!
10. Make your analysis reproducible.



Record procedure details!

About Assignment 1 (Section B)

- Record all code and results
- Submit in any format (RMD preferred)

- Due Feb 19, 2024.
- Assignment 1 can be emailed to zang@virginia.edu