STATISTICS REVIEW II

OUTLINE

- Sampling Bias
- Simpson’s Paradox
- Type I and type II errors
- Frequentist vs. Bayesian
- A case study

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.

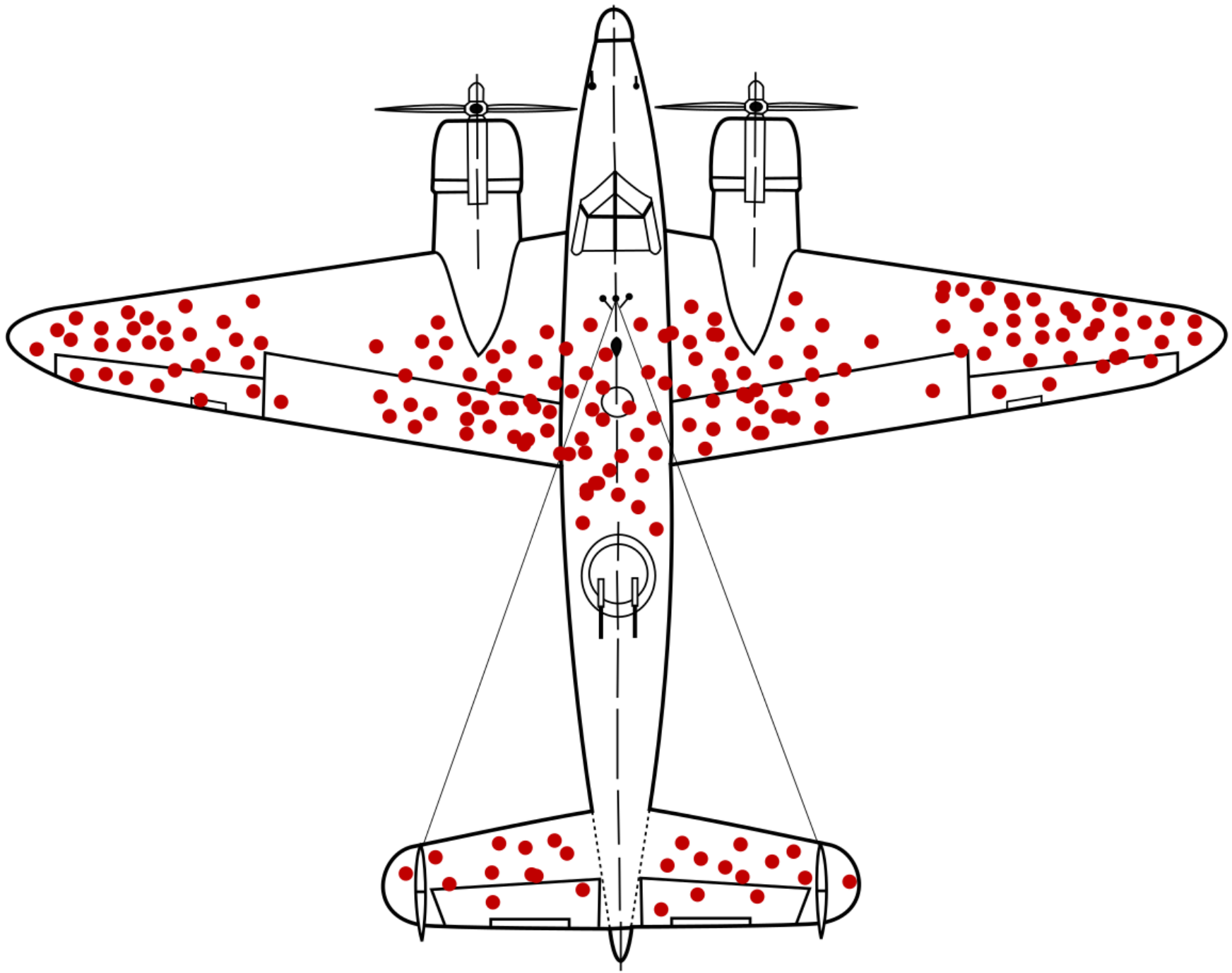
ASA statement on statistical significance and p-values

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.

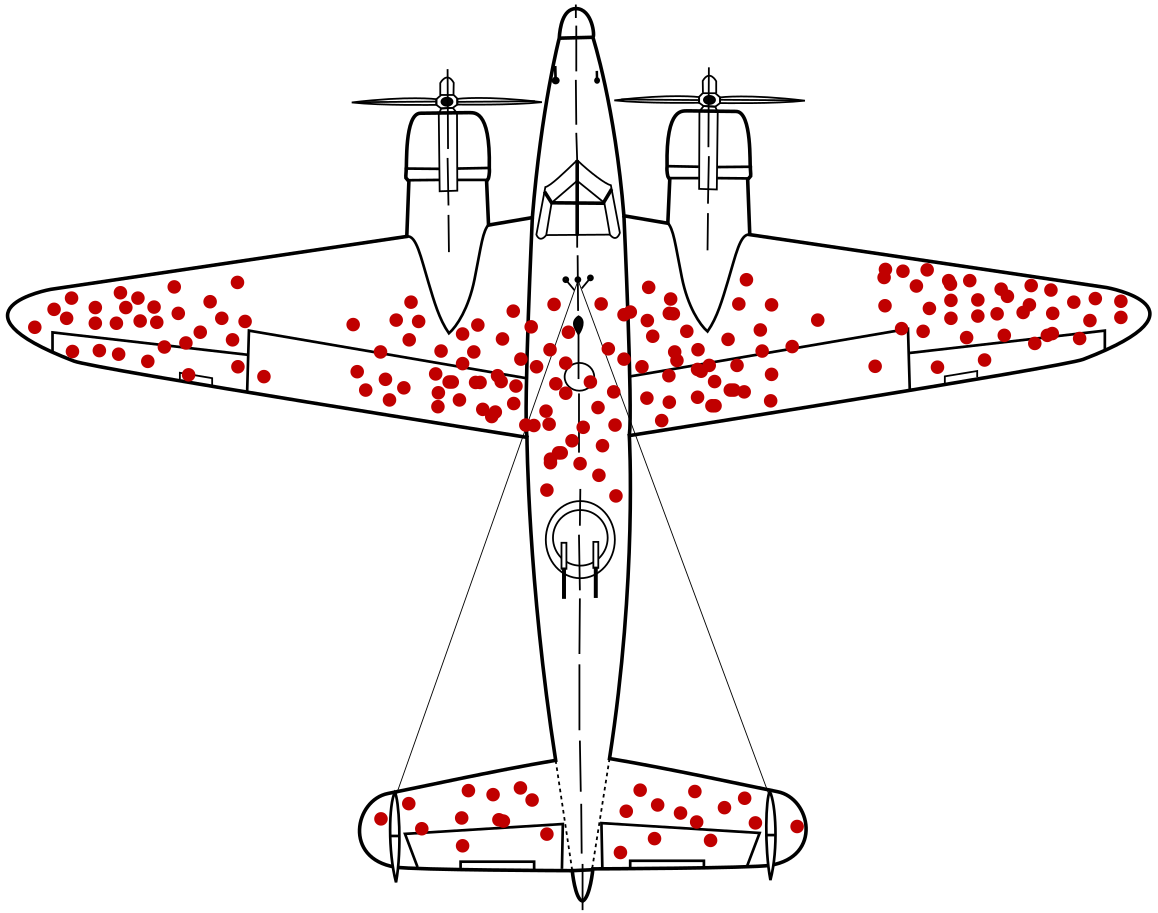
What is the control?
What is the null hypothesis?



What is the population/baseline?

- Aircrafts that had returned from missions

- All aircrafts that went to missions

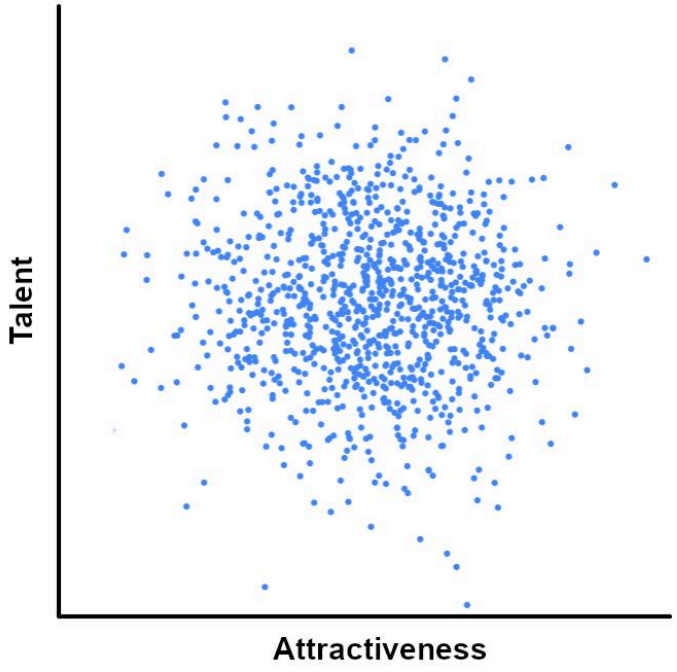


More examples about sampling bias

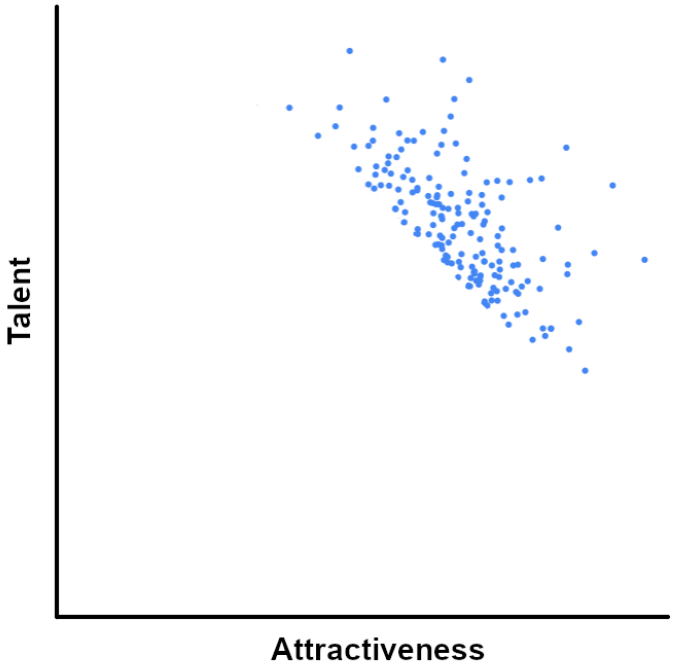
- A survey conducted at a healthcare provider found that 80% of its visitors were diagnosed with a disease.

More examples about sampling bias

- Are talent and attractiveness negatively correlated?



Population



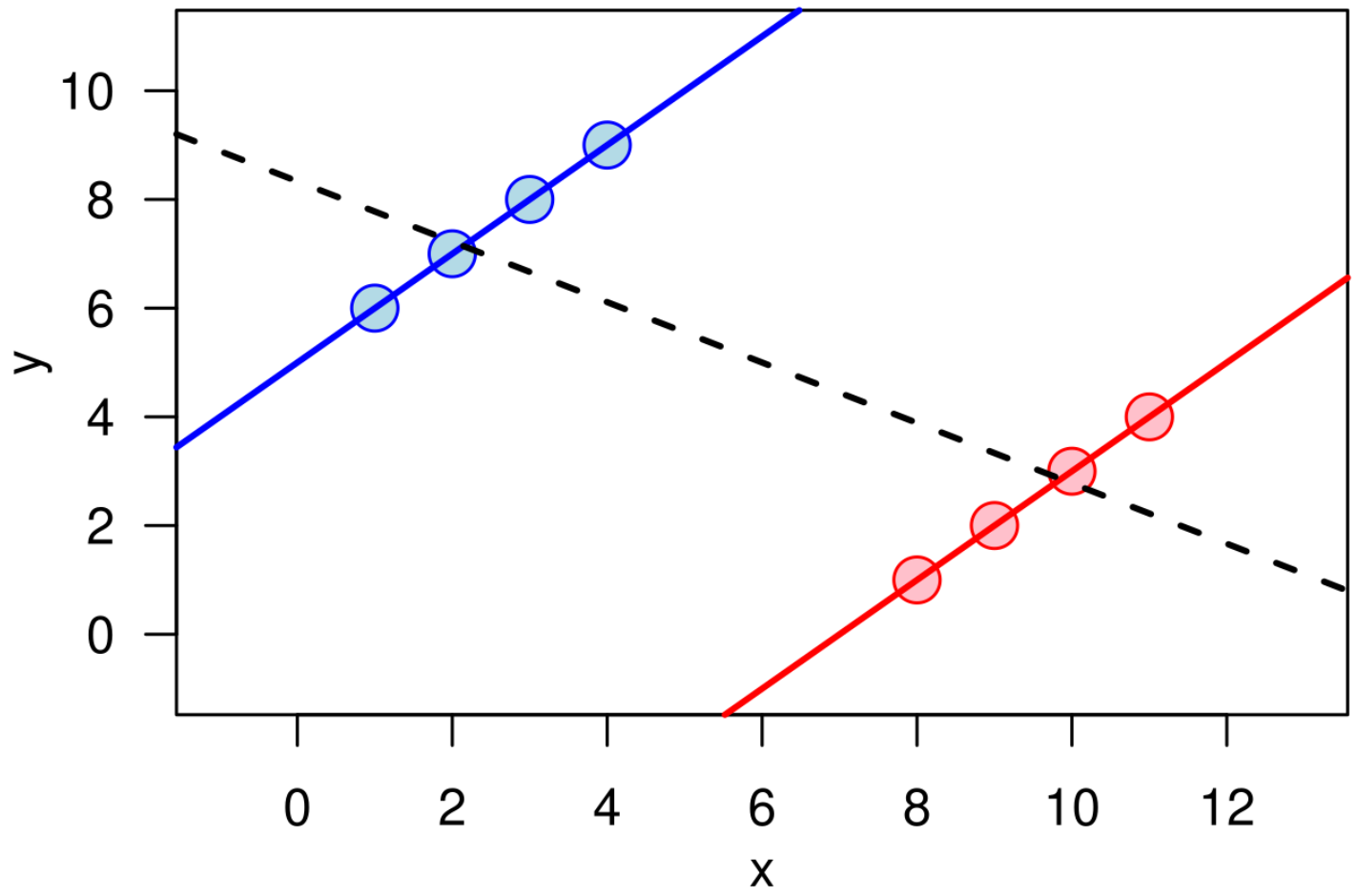
Celebrities

Simpson's Paradox

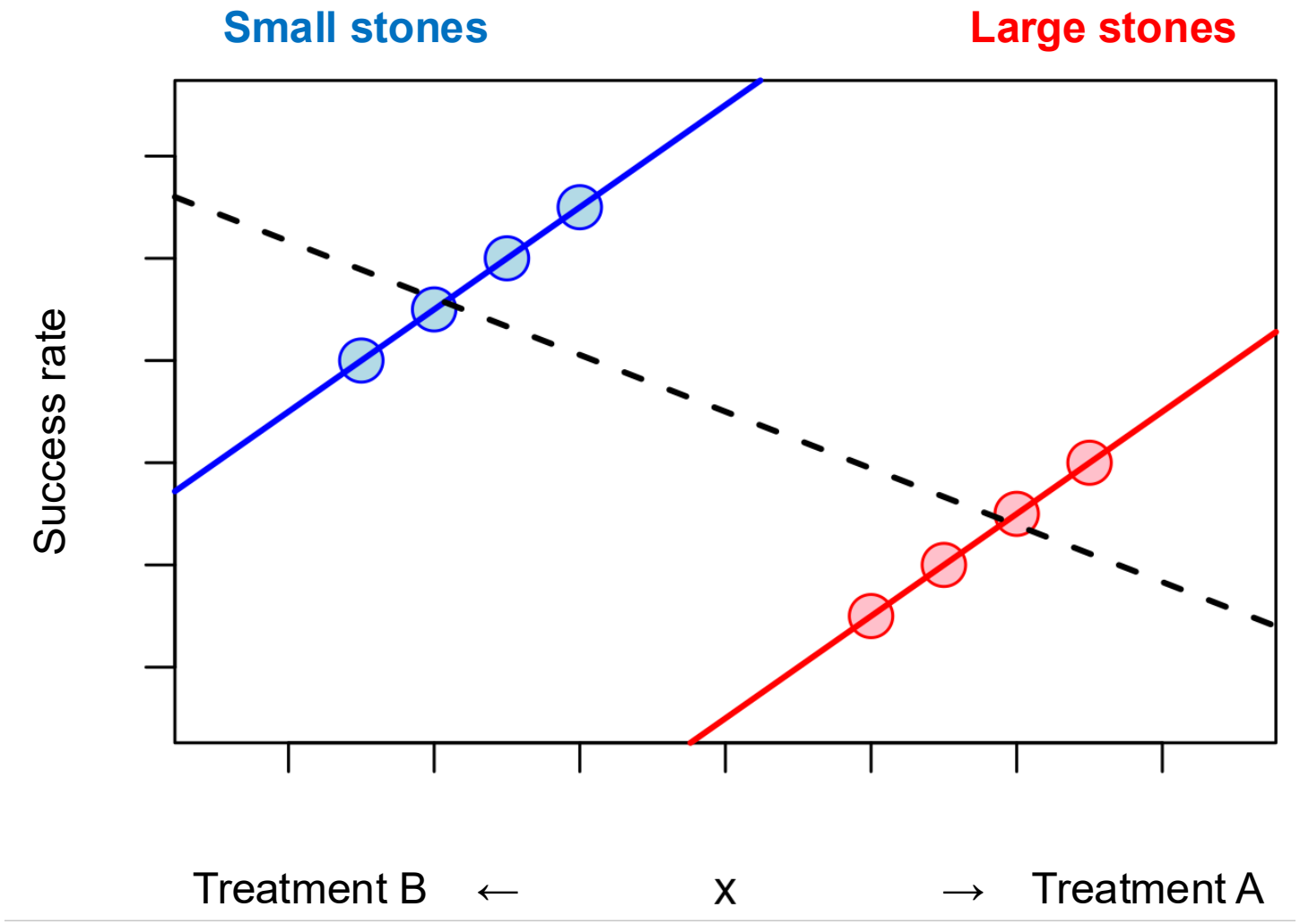
Kidney stone treatments' success rates

Treatment Stone size	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

Simpson’s Paradox



Simpson's Paradox



Simpson's Paradox: An scRNA-seq example

Proportion	Pre-treatment	Post-treatment
Subpopulation A	0.04	0.80
Subpopulation B	0.16	0.16
Subpopulation C	0.80	0.04
Total	1.00	1.00

Gene X expression	Pre-treatment	Post-treatment	Log2 Fold Change
Subpopulation A	0.10	0.30	+1.58
Subpopulation B	1.50	1.80	+0.26
Subpopulation C	3.00	3.50	+0.22
Population Average	2.64	0.67	-1.98

Credit: Jean Fan, JEFworks

Confusion Matrix

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP (Type I Error)
	Negative	FN (Type II Error)	TN

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

Source: Wikimedia. Author: Lavender888000

Example: Rare disease screening

Suppose:

- Disease prevalence: **1 in 10,000** (0.01%)
- Test sensitivity: **99%** (correctly detects 99% of cases)
- Test specificity: **99%** (correctly rules out 99% of non-cases)

Now test **1,000,000** people:

- **True cases** = 100
  - True positives = 99 (99% of 100)
  - False negatives = 1
- **Non-cases** = 999,900
  - False positives = 9,999 (1% of 999,900)
  - True negatives = 989,901
- So total positives reported by the test = **99 + 9,999 = 10,098**.
  Only **99** of those are real.
- The **Positive Predictive Value (PPV)** = 99 / 10,098 ≈ **0.98%**.
  That means **99.02% of the "positive" results are false alarms**.

Why screening does not work well for rare diseases with imperfect tests?

- **Key issue**: Even if a test has "good" accuracy (say, 99% sensitivity and 99% specificity), when the disease is rare, most positive results will actually be **false positives** rather than **true positives**.
- This is because the **prevalence** (base rate) of the disease is very low, so the number of healthy individuals vastly outnumbers the true cases.
- **Example**: **COVID antibody tests** in the early stage of the COVID pandemic: When prevalence was <5% in most populations, even tests with 95% specificity yielded more false positives than true positives.

The Statistics Behind It

- The key relationship is given by **Bayes' theorem**:

$$PPV = \frac{(\text{sensitivity}) \times (\text{prevalence})}{(\text{sensitivity} \times \text{prevalence}) + (1 - \text{specificity}) \times (1 - \text{prevalence})}$$

- When prevalence is very small, the denominator is dominated by false positives (the (1−specificity)×(1−prevalence) term).
- This drives PPV close to zero, even for high-quality tests.

Summary

- Population-wide screening for rare diseases with tests of “ordinary” accuracy does not work because the **false positives overwhelm the true positives**.
- Instead, targeted screening of **higher-risk subgroups** (increasing effective prevalence) may dramatically improve predictive value.

Bayes' Theorem

LIKELIHOOD
The probability of "B" being True, given "A" is True

PRIOR
The probability "A" being True. This is the knowledge.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

POSTERIOR
The probability of "A" being True, given "B" is True

MARGINALIZATION
The probability "B" being True.

Frequentist vs. Bayesian

Frequentist

- P-value
- Confidence
- Maximum Likelihood Estimation (MLE)

$$\mathcal{L}_n(\theta) = \mathcal{L}_n(\theta; \mathbf{y}) = f_n(\mathbf{y}; \theta)$$

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta; \mathbf{y})$$

$$p(\text{data}|\theta)$$

Bayesian

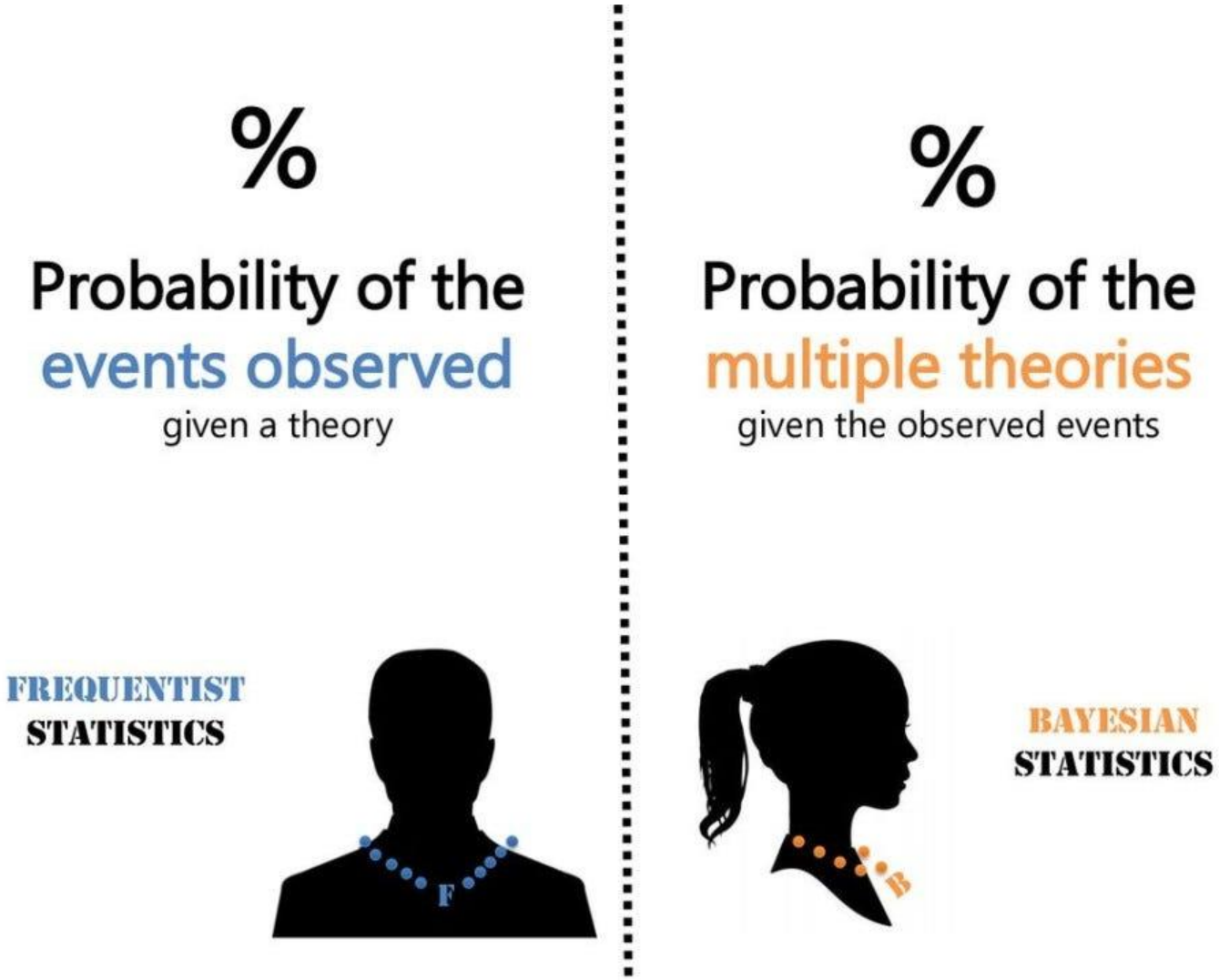
- Bayes' Theorem

$$\underbrace{p(\theta \mid \text{data})}_{\text{Posterior}} = \frac{\overbrace{p(\text{data} \mid \theta)}^{\text{Likelihood}} \cdot \overbrace{p(\theta)}^{\text{Prior}}}{\underbrace{p(\text{data})}_{\text{Normalization}}}$$

Frequentist vs. Bayesian

	Frequentist	Bayesian
Hypothesis test	*p* value (null hypothesis significance test)	Bayes factor
Estimation with uncertainty	maximum likelihood estimate with confidence interval (The "New Statistics")	posterior distribution with highest density interval

Frequentist vs. Bayesian



A case study

Article

Spatial transcriptomics reveal neuron–astrocyte synergy in long-term memory

https://doi.org/10.1038/s41586-023-07011-6

Received: 16 March 2023

Accepted: 21 December 2023

Published online: 7 February 2024

Open access

Check for updates

Wenfei Sun[1,2,6], Zhihui Liu[2,3,6], Xian Jiang[2], Michelle B. Chen[1], Hua Dong[4], Jonathan Liu[5], Thomas C. Südhof[2,3] ✉ & Stephen R. Quake[1,5] ✉

Memory encodes past experiences, thereby enabling future plans. The basolateral amygdala is a centre of salience networks that underlie emotional experiences and thus has a key role in long-term fear memory formation[1]. Here we used spatial and single-cell transcriptomics to illuminate the cellular and molecular architecture of the role of the basolateral amygdala in long-term memory. We identified transcriptional signatures in subpopulations of neurons and astrocytes that were memory-specific and persisted for weeks. These transcriptional signatures implicate neuropeptide

Article

Spatial transcriptomics reveal neuron–astrocyte synergy in long-term memory

https://doi.org/10.1038/s41586-023-07011-6

Received: 16 March 2023

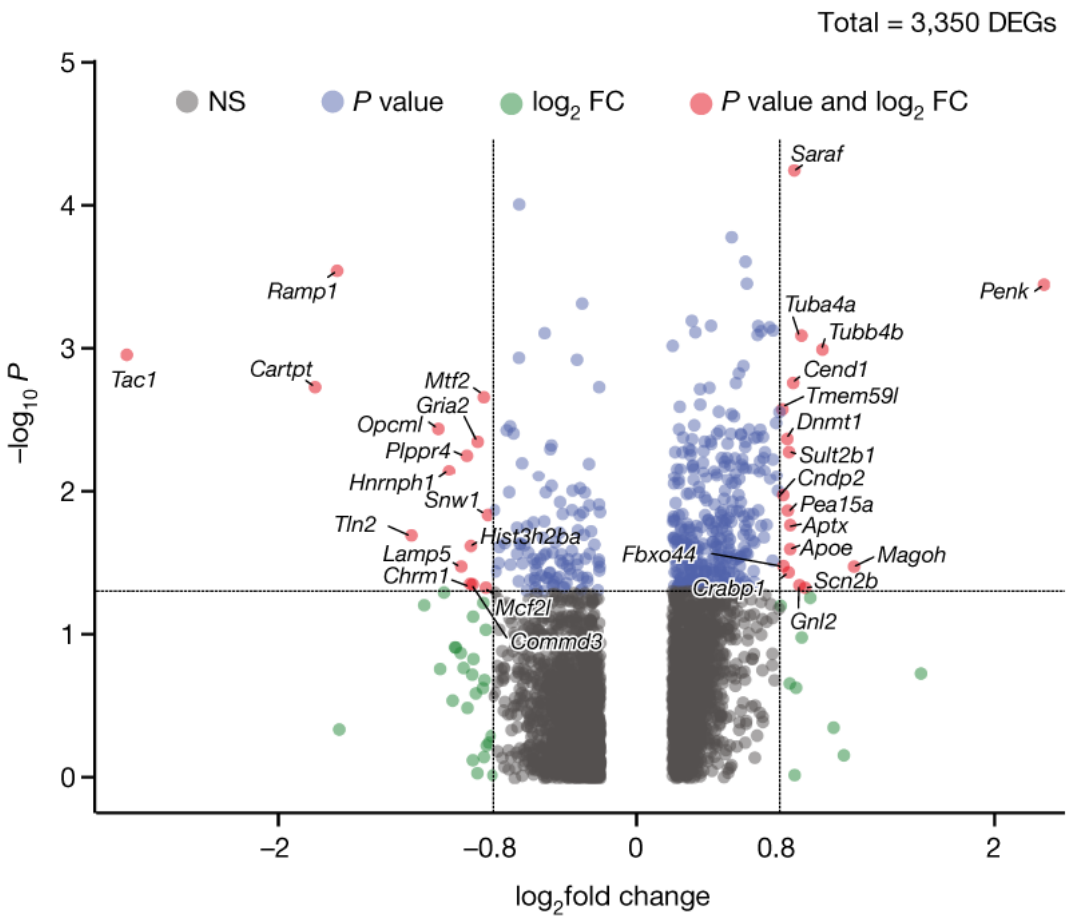
Accepted: 21 December 2023

Published online: 7 February 2024

Open access

Check for updates

Wenfei Sun[1,2,6], Zhihui Liu[2,3,6], Xian Jiang[2], Michelle B. Chen[1], Hua Dong[4], Jonathan Liu[5], Thomas C. Südhof[2,3]✉ & Stephen R. Quake[1,5]✉



Matters arising

False positives in study of memory-related gene expression

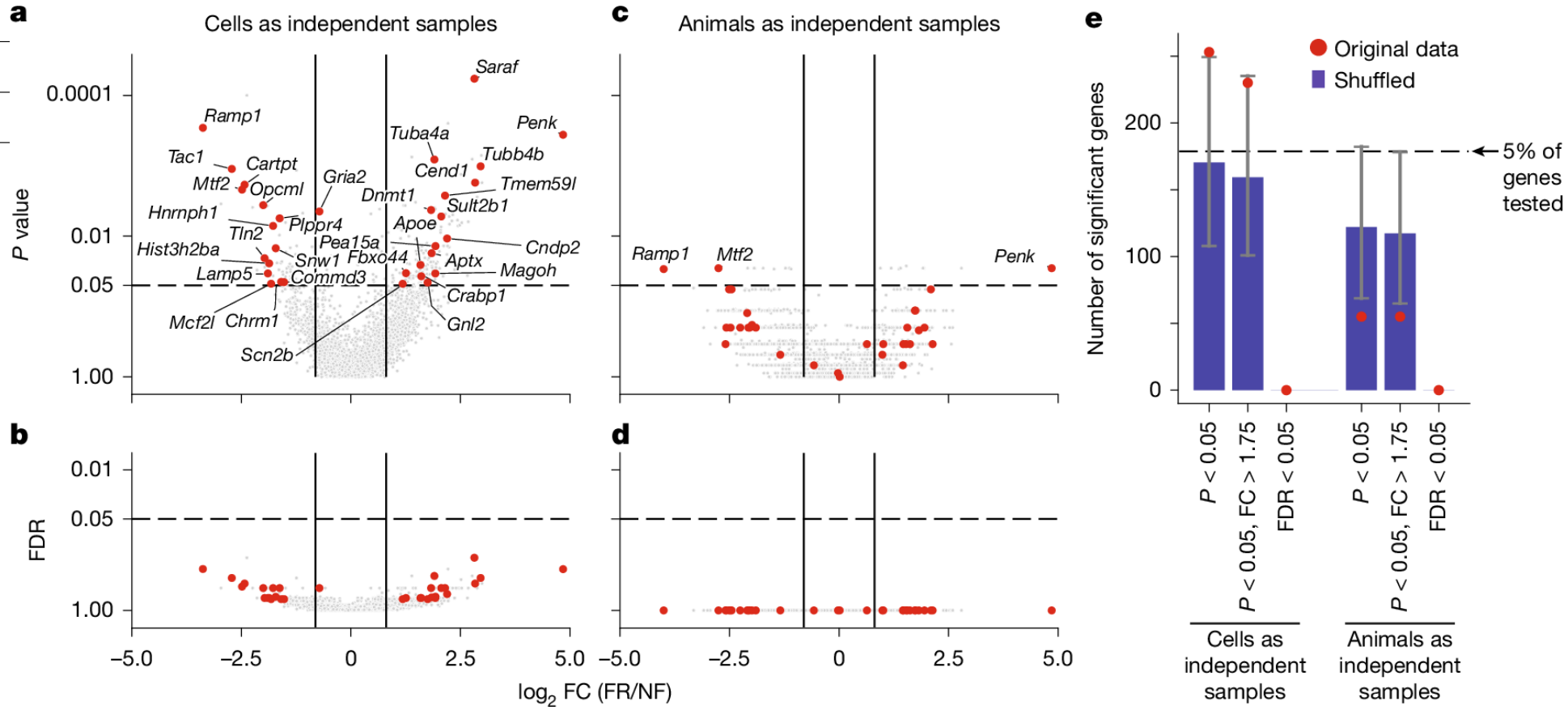
https://doi.org/10.1038/s41586-025-08988-y

Received: 18 July 2024

Accepted: 7 April 2025

Published online: 4 June 2025

Eran A. Mukamel[1]✉ & Zhaoxia Yu[2]

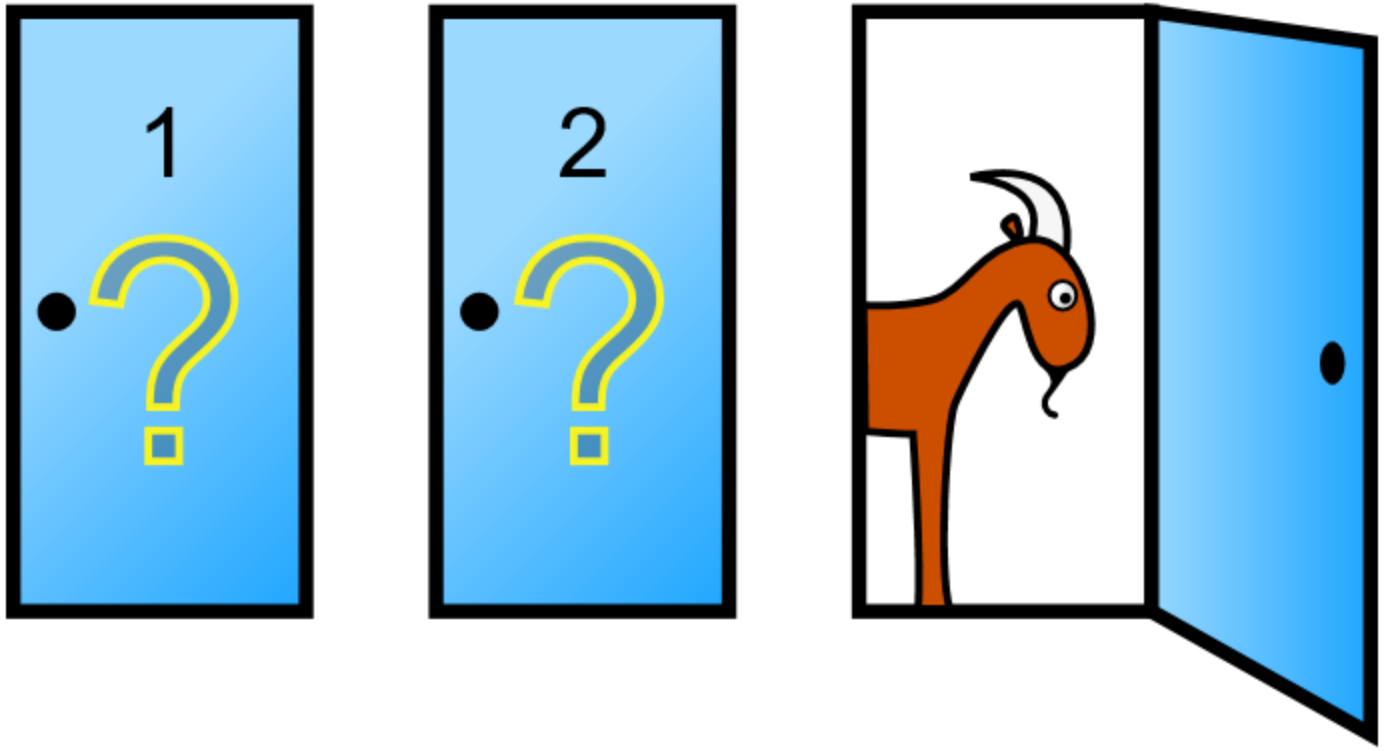


Matters arising

False positives in study of memory-related gene expression

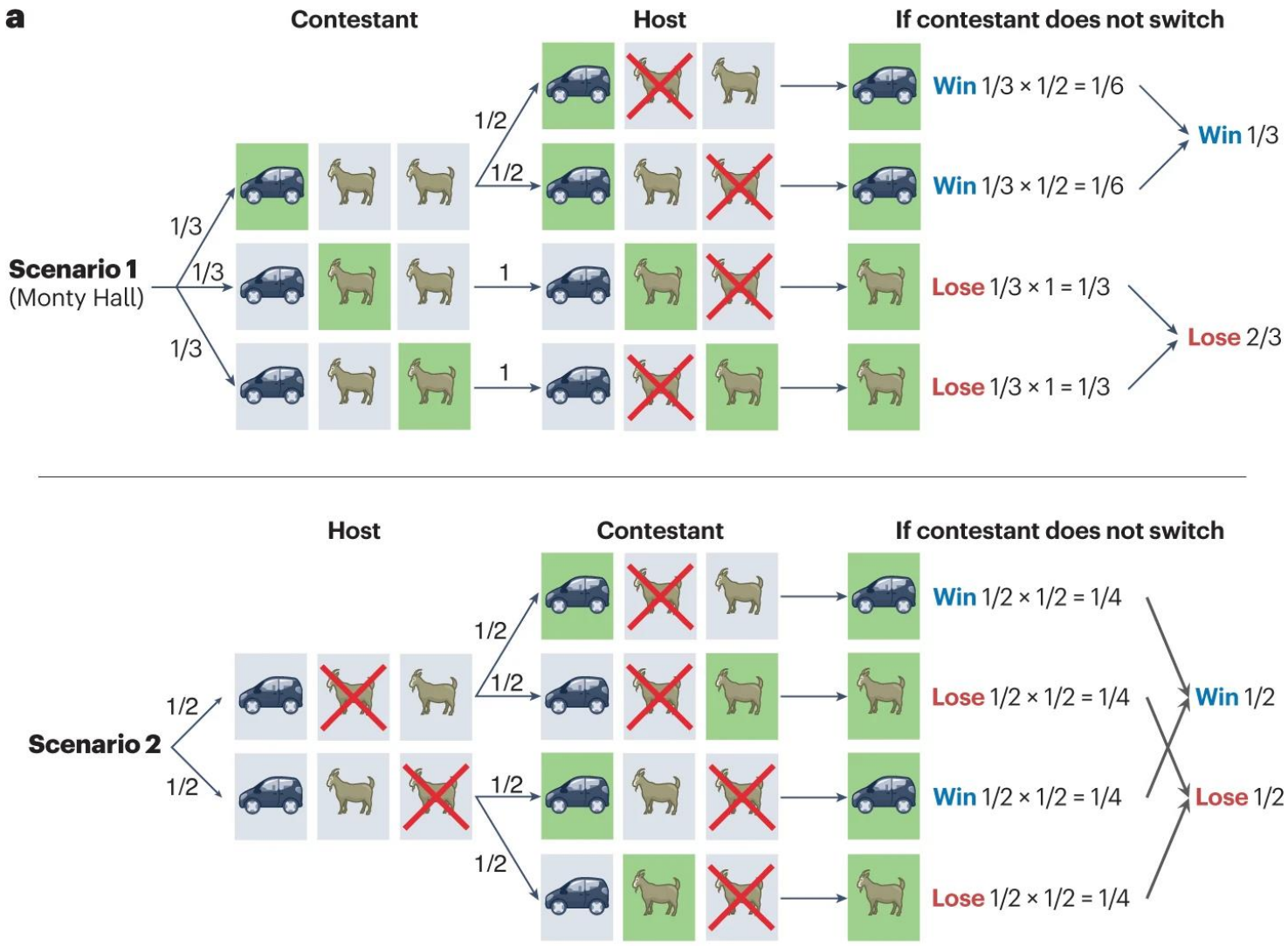
- Multiple testing correction (FDR) was not applied correctly.
  - Sun et al. “used a series of criteria to pre-select 56 candidate genes of interest, thus reducing the burden of multiple hypothesis testing.”
  - This is double-dipping!
- Fail to use animal as sample.
  - “Treatment of individual cells as independent samples.”
  - Cells correlate from the same animal/sample.

Monty Hall Problem



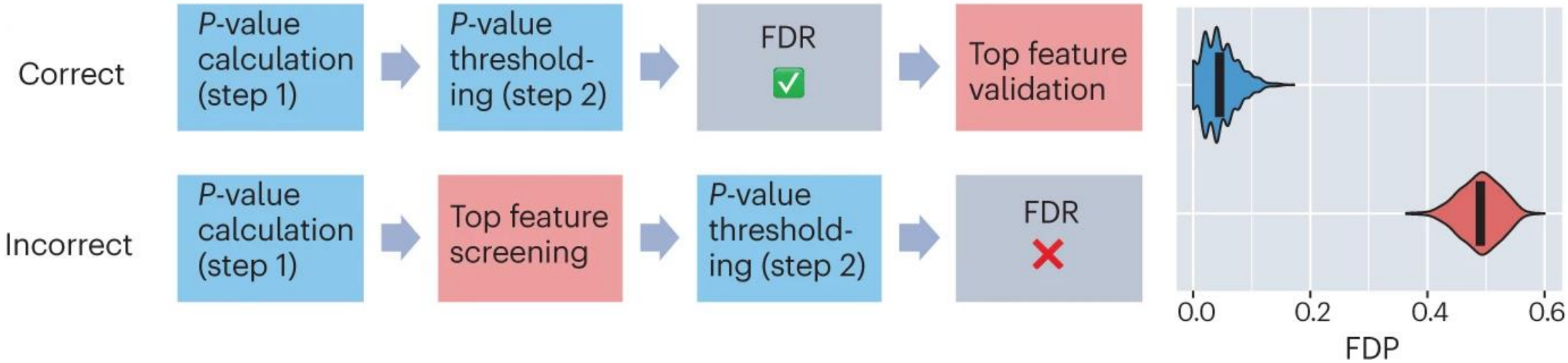
Source: Wikipedia

Monty Hall Problem



Li *Nat Biotech* 2023. PMID: 37198440

The order of action matters

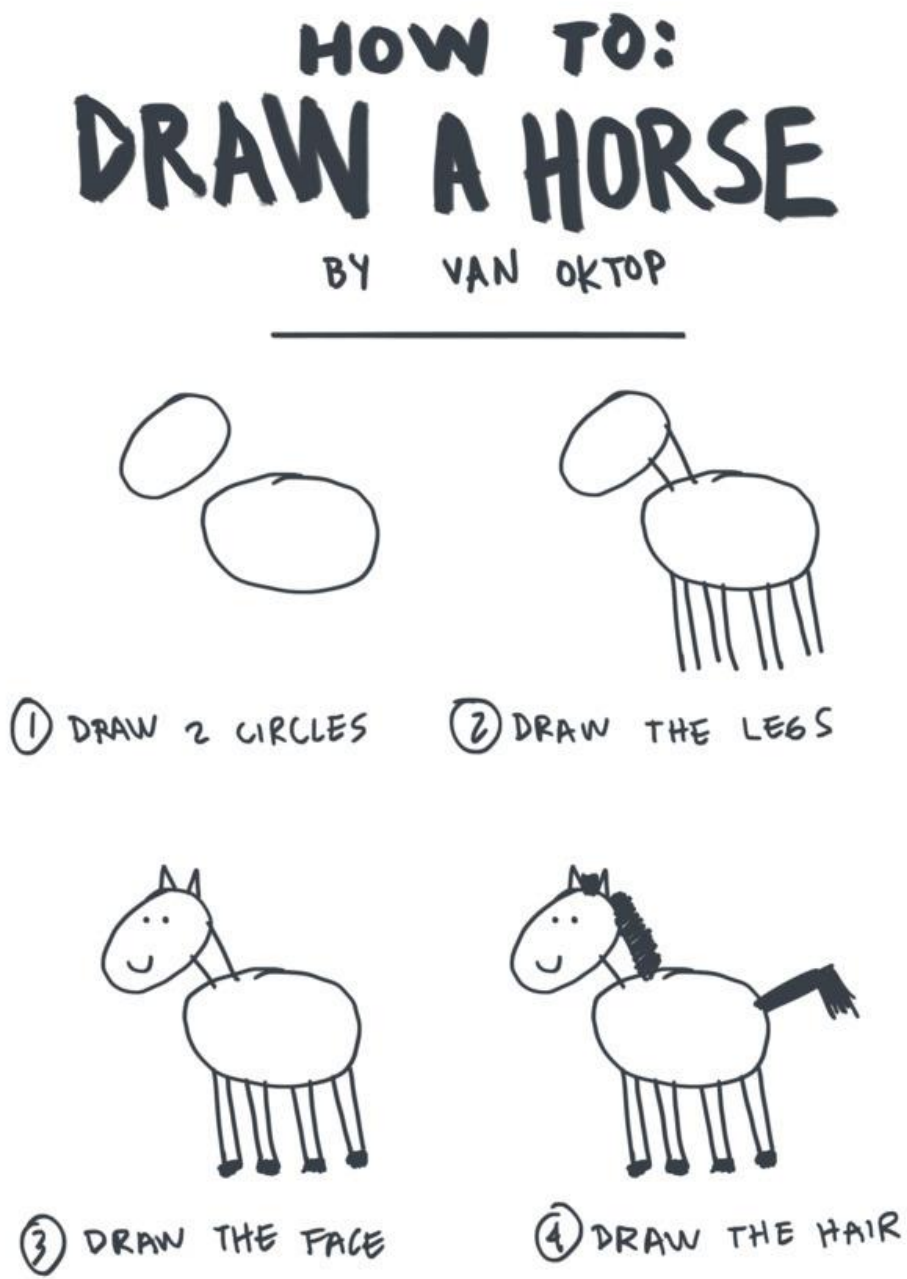


Li *Nat Biotech* 2023. PMID: 37198440

SUMMARY

1. Avoid sampling bias.
2. Carefully plan the study design.
3. Beware Simpson’s paradox.
4. Think before you analyze.
5. Statistical analysis is more than a set of computations.

https://www.youtube.com/watch?v=cUqoHQDinCM



Record procedure details!