STATISTICAL INFERENCE

Olga Vitek

College of Science

College of Computer and Information Science

Northeastern University

OUTLINE

- Describing the data
  - Center and variation
- Basic statistical inference
  - T-test and p-values
- P-values: a word of caution
  - Instability, multiplicity, alternative approaches

CENTER: MEAN

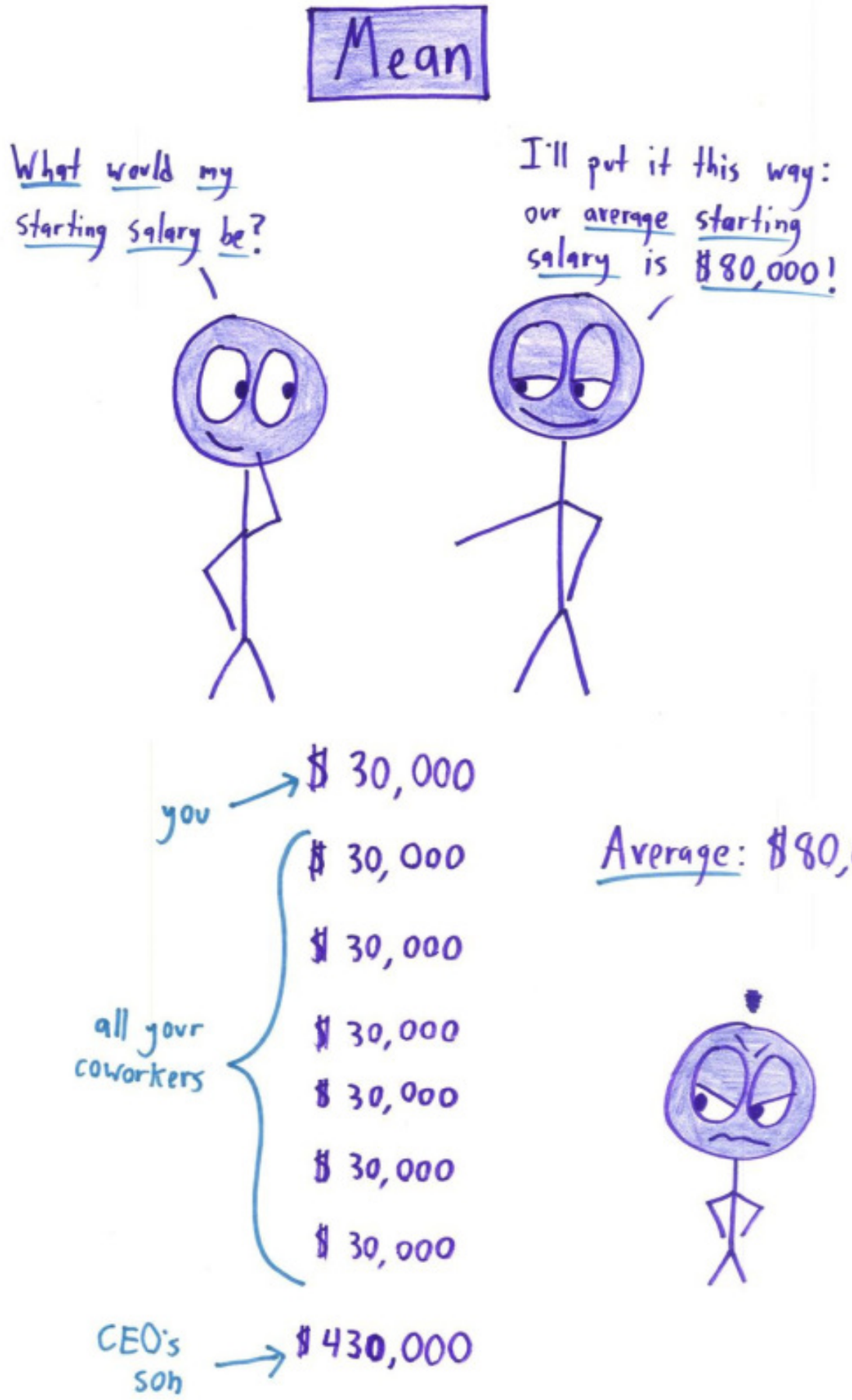
Mean

$$\bar{y} = \frac{1}{n}(y_1 + y_1 + \ldots + y_n)$$

$$= \frac{1}{n}\sum_{i=1}^{n} y_i$$

Weighted mean

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}$$

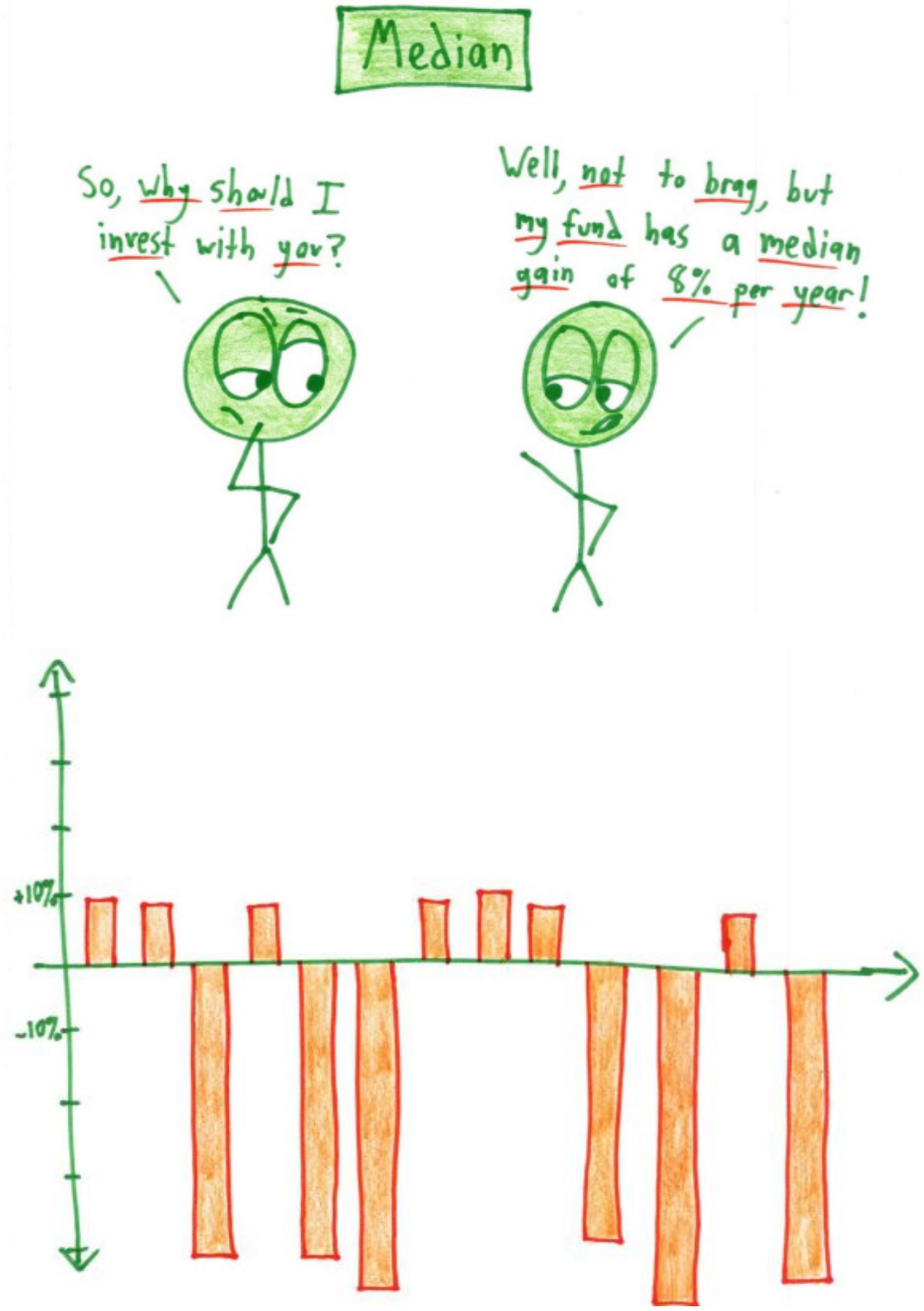


https://mathwithbaddrawings.com/2016/07/13/why-not-to-trust-statistics

CENTER: MEDIAN

A value such as 50% data points are smaller, and 50% are larger

- Symmetric distributions: median=mean
- Skewed to the right: median < mean
- Skewed to the left: median > mean

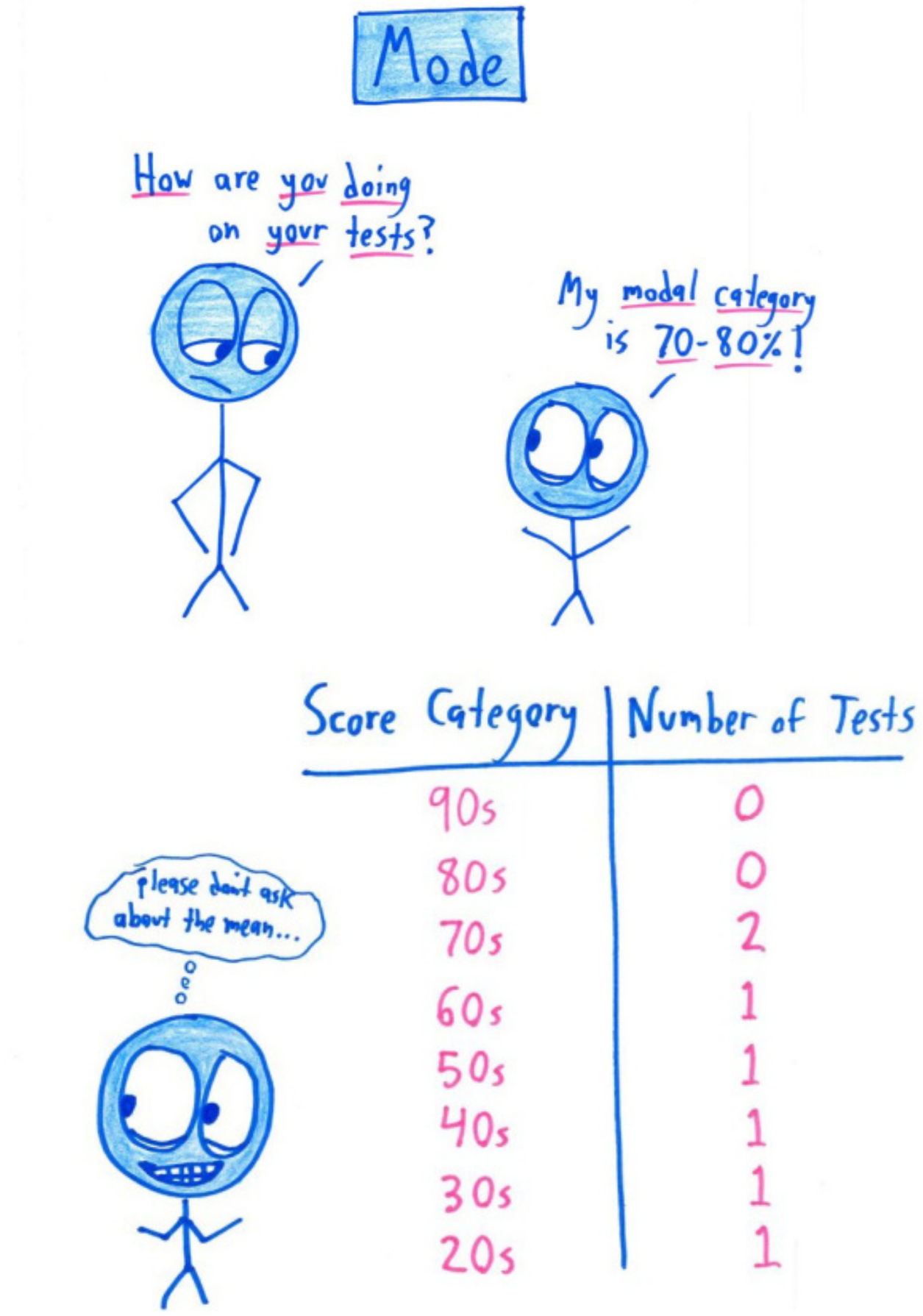


https://mathwithbaddrawings.com/2016/07/13/why-not-to-trust-statistics

CENTER: MODE

Most frequent value

- Best used with categorical data



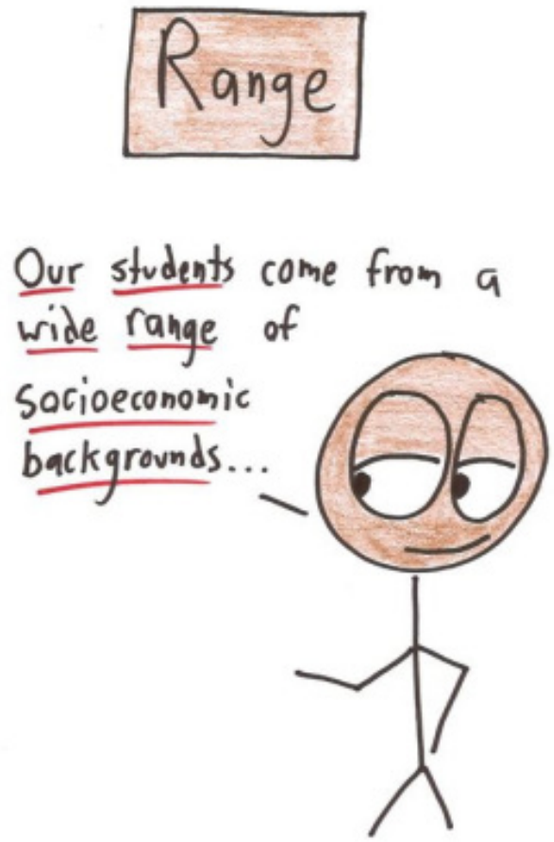
https://mathwithbaddrawings.com/2016/07/13/why-not-to-trust-statistics

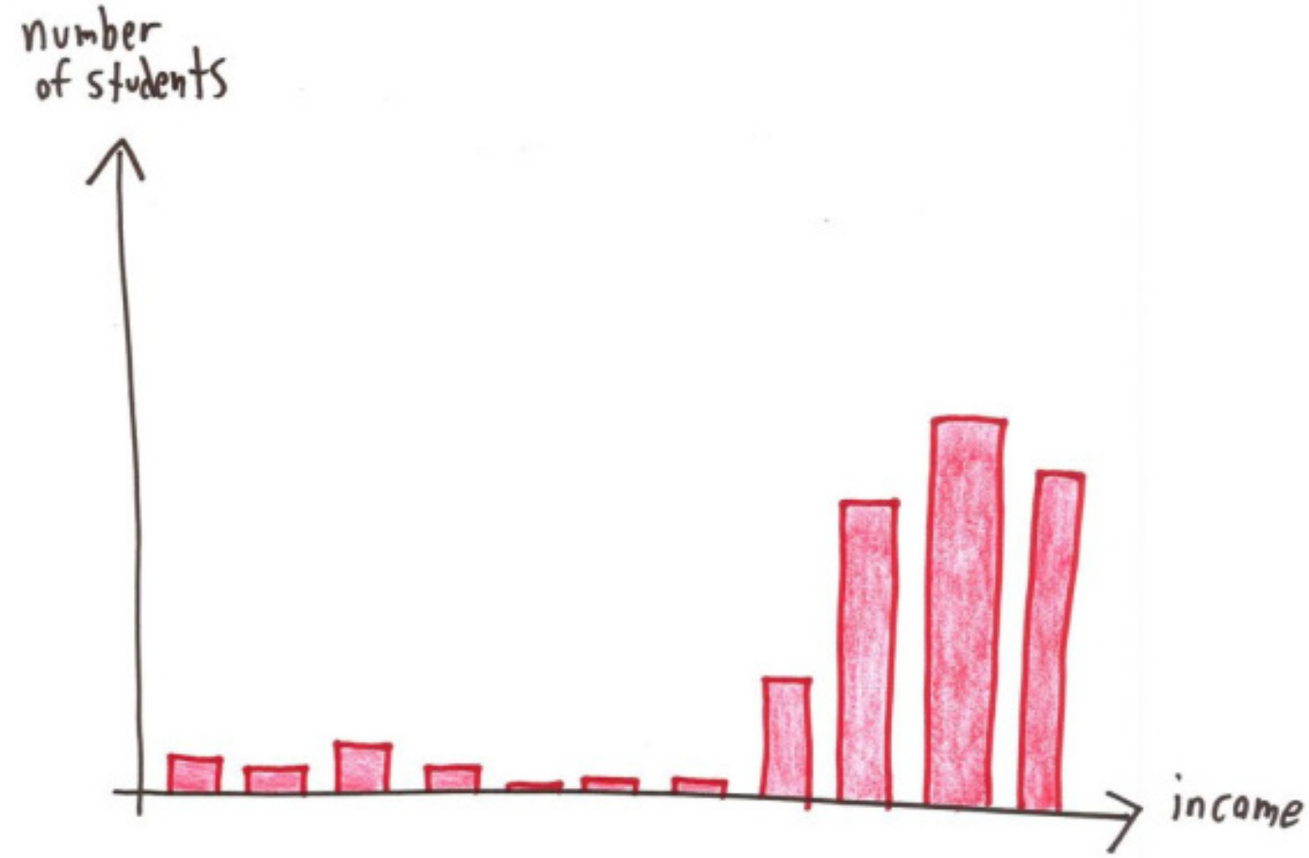
VARIATION: RANGE

Difference between largest and smallest value

- Deviation: difference between a value and the mean

$$y_i - \bar{y}$$





https://mathwithbaddrawings.com/2016/07/13/why-not-to-trust-statistics

VARIANCE

Sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

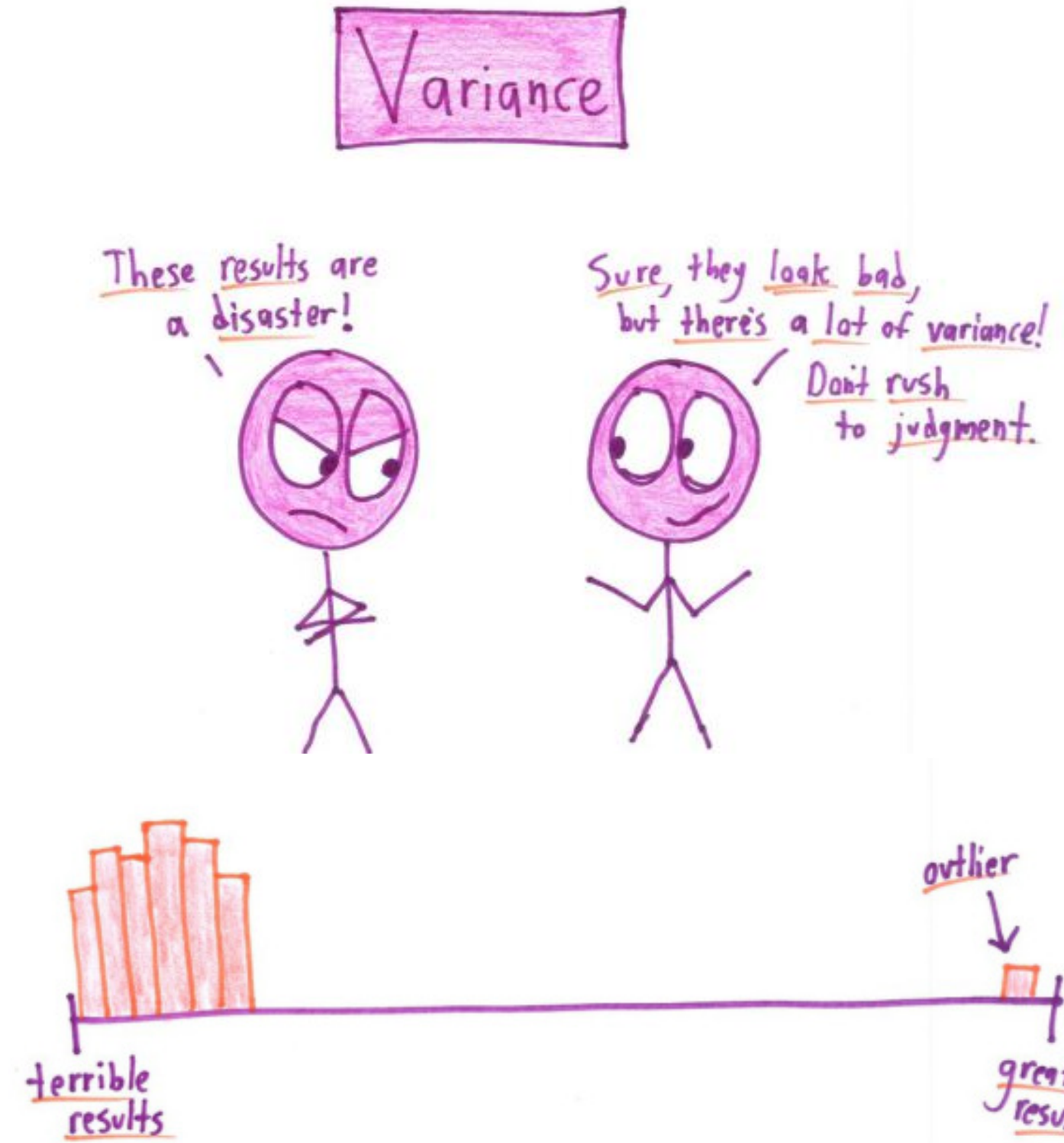
$$= \frac{\text{sum of squared deviations}}{\text{sample size - 1}}$$

Standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

$$= \sqrt{\frac{\text{sum of squared deviations}}{\text{sample size - 1}}}$$

- Translates deviations^2 to original scale
- $n - 1$ is degrees of freedom



https://mathwithbaddrawings.com/2016/07/13/why-not-to-trust-statistics

INTERPRETING VARIATION

- If the data are approximately bell shaped
  - $\approx 68\%$ of the observations fall between $\bar{y} - s$ and $\bar{y} + s$
  - $\approx 95\%$ of the observations fall between $\bar{y} - 2 \cdot s$ and $\bar{y} + 2 \cdot s$
  - Nearly all observations fall between $\bar{y} - 3 \cdot s$ and $\bar{y} + 3 \cdot s$

- Coefficient of variation for a variable Y is

$$CV = \frac{s}{\bar{y}} \cdot 100\%$$

  - standard deviation expressed as a % of mean

- Z-score of an observation y_i is

$$z = \frac{y_i - \bar{y}}{s} = \frac{\text{observation - mean}}{\text{standard deviation}}$$

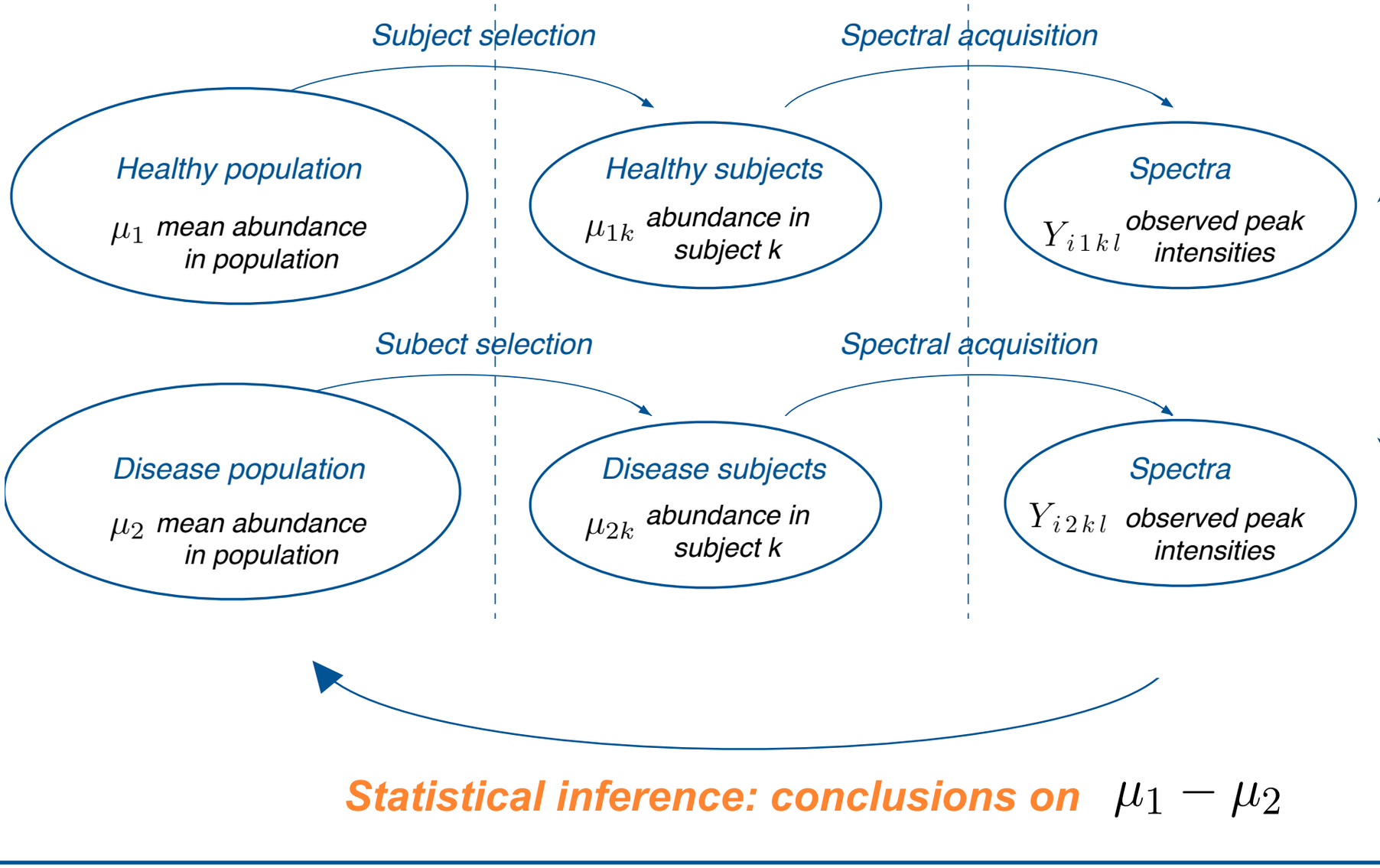
Used to describe the magnitude of the individual values, in relationship to the mean

OUTLINE

- Describing the data
  - Center and variation
- Basic statistical inference
  - T-test and p-values
- P-values: a word of caution
  - Instability, multiplicity, alternative approaches

COMPARE DESIGNS

In terms of bias and (in)-efficiency

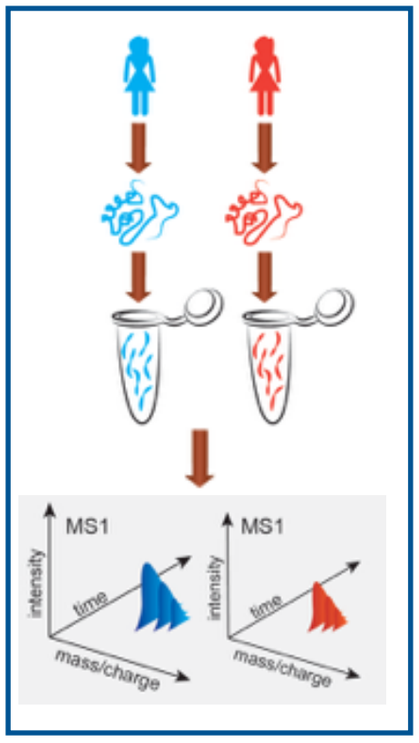


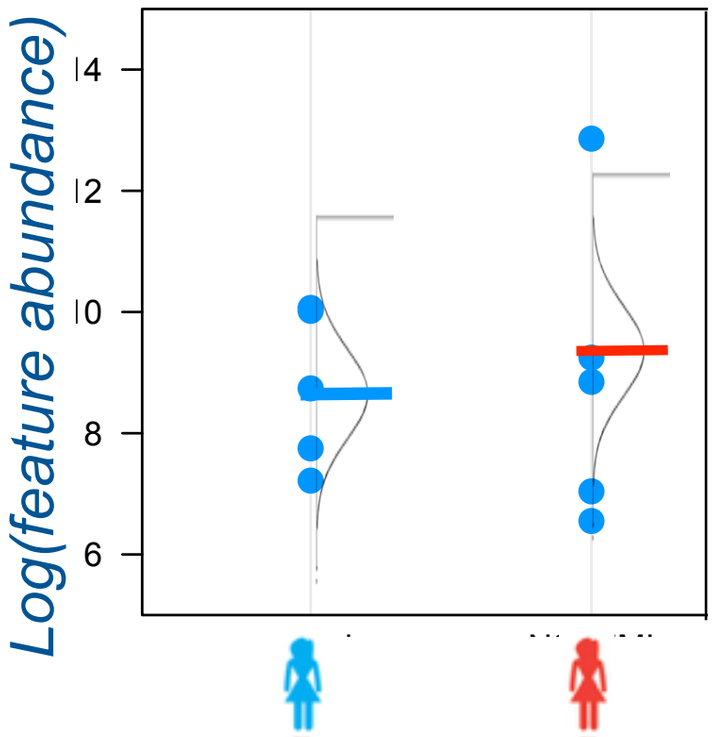
Bias: $\bar{Y}_{.1..} - \bar{Y}_{.2..}$ *systematically different from* $\mu_{1k} - \mu_{2k}$

Inefficiency: *Large* $Var(\bar{Y}_{.1..} - \bar{Y}_{.2..})$

TWO-SAMPLE T-TEST

Simple example: label-free experiment, one feature/protein





$$FoldChange = \frac{'typical'\ value\ in\ group\ 1}{'typical'\ value\ in\ group\ 2}$$

$$log_2(FoldChange) = \\ = log_2('typical'\ value\ in\ group\ 1) \\ - log_2('typical'\ value\ in\ group\ 2)$$

- $\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}$ = estimates log-fold change

$$\frac{1}{n_1}\sum_j Y_{1j} - \frac{1}{n_2}\sum_j Y_{2j} = \frac{1}{n_1}\sum_j log_2\, X_{1j} - \frac{1}{n_2}\sum_j log_2 X_{2j} =$$

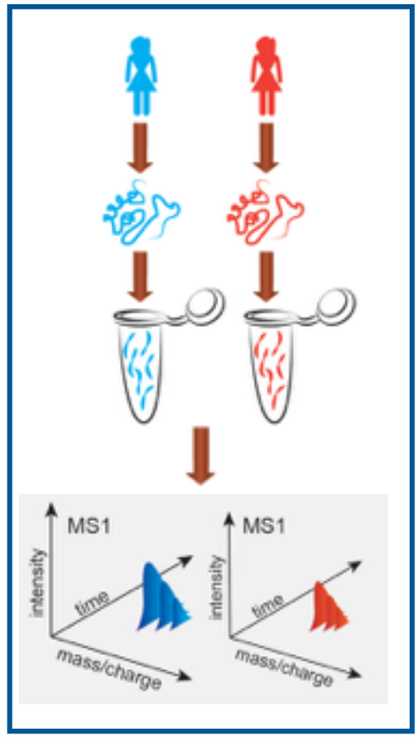
$$log_2\left(\prod_j X_{1j}\right)^{\frac{1}{n_1}} - log_2\left(\prod_j X_{2j}\right)^{\frac{1}{n_2}} = log_2 \frac{\left(\prod_j X_{1j}\right)^{\frac{1}{n_1}}}{\left(\prod_j X_{2j}\right)^{\frac{1}{n_2}}}$$

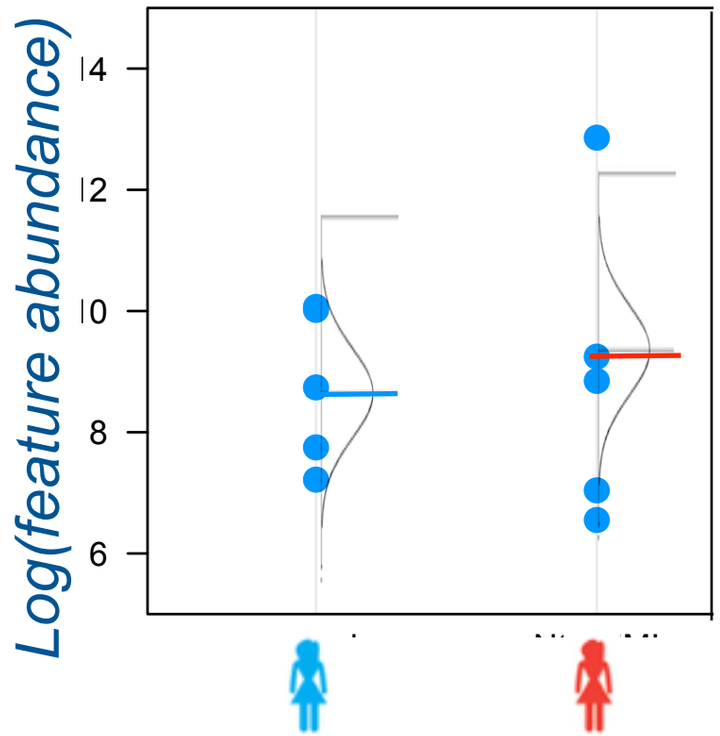
Conclusion:

On log scale, estimates of FC are ratios of geometric means

TWO-SAMPLE T-TEST

Simple example: label-free experiment, one feature/protein





Sample means in each group

H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

$$\textbf{observed t} = \frac{\textbf{difference of group means}}{\textbf{estimate of variation}} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

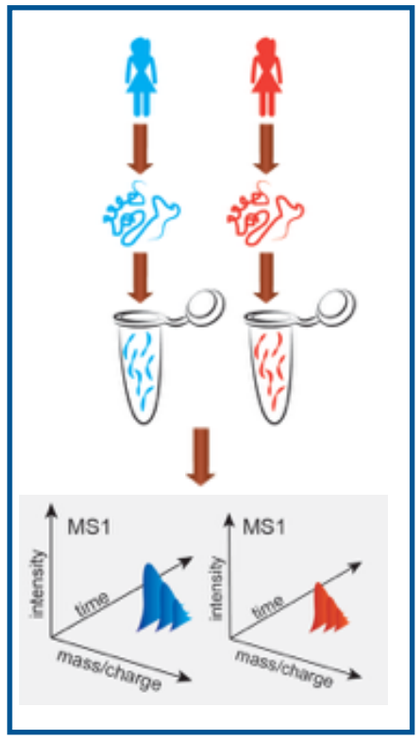
Number of replicates

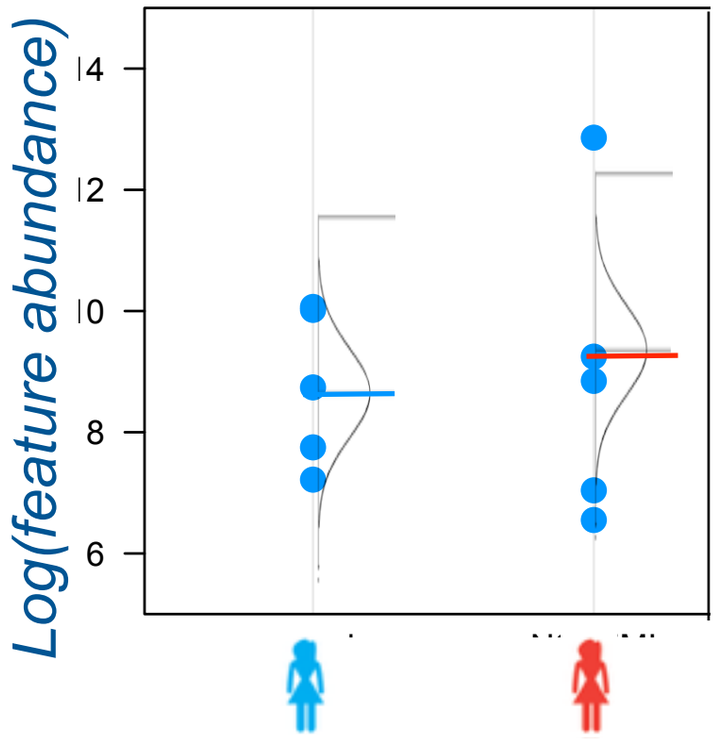
$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_{1\cdot})^2$$

Sample variance

TWO-SAMPLE T-TEST

Simple example: label-free experiment, one feature/protein





H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

$$\textbf{observed t} = \frac{\textbf{difference of group means}}{\textbf{estimate of variation}} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_{1\cdot})^2$$

Properties of the means

$$\frac{s_1^2}{n_1}$$

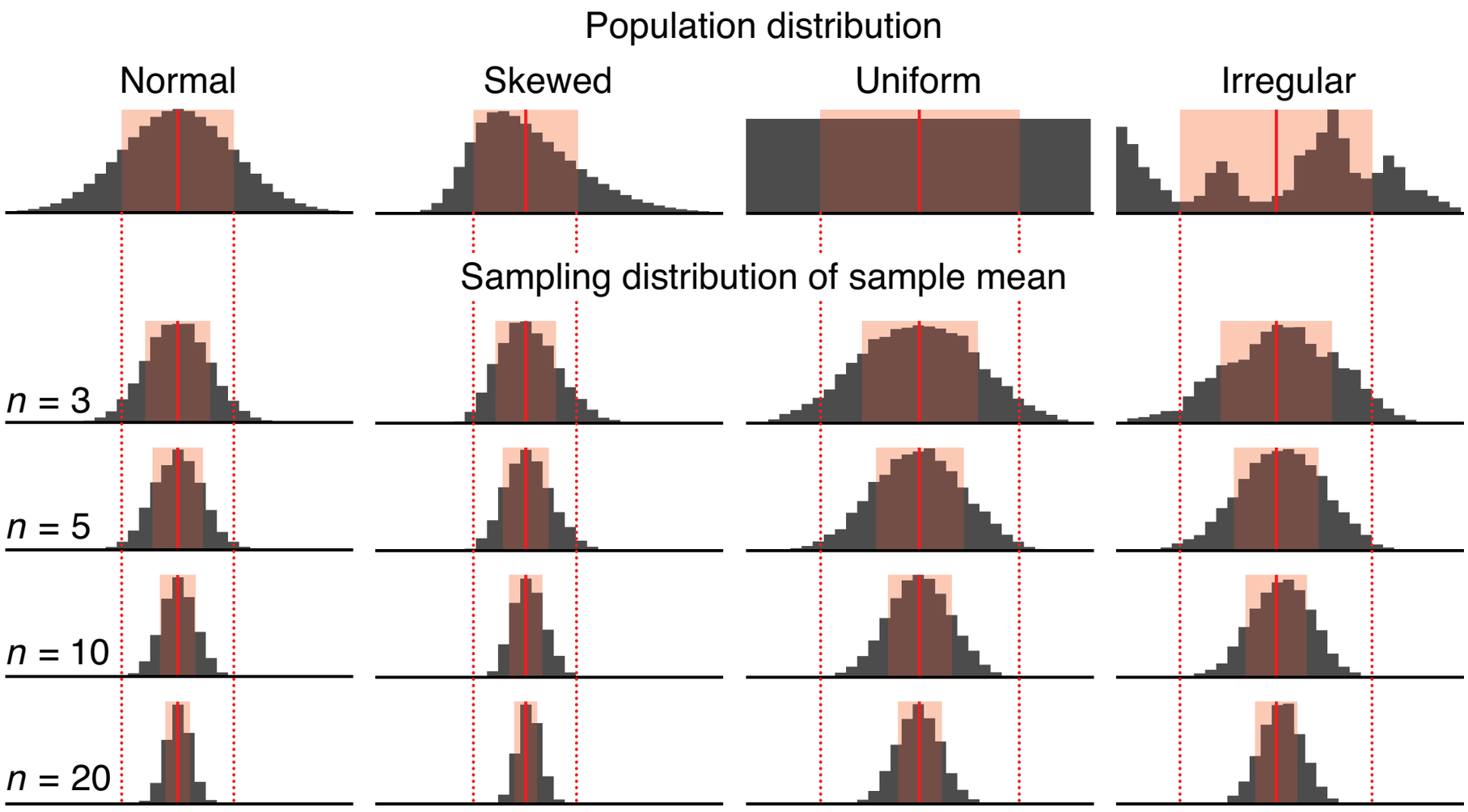
Variance of the sampling distribution of first mean

$$\sqrt{\frac{s_1^2}{n_1}}$$

Standard error of the first mean

ASSUMPTION: NORMAL DISTRIBUTION

The Central Limit Theorem



Probability distribution of the data

Repeatedly selecting n data points and calculating means

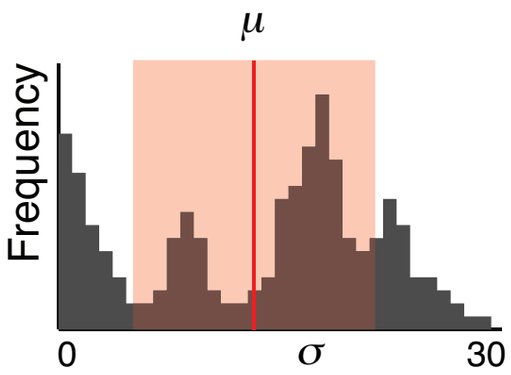
Conclusion:

As n increases, the mean is less variable and more Normal

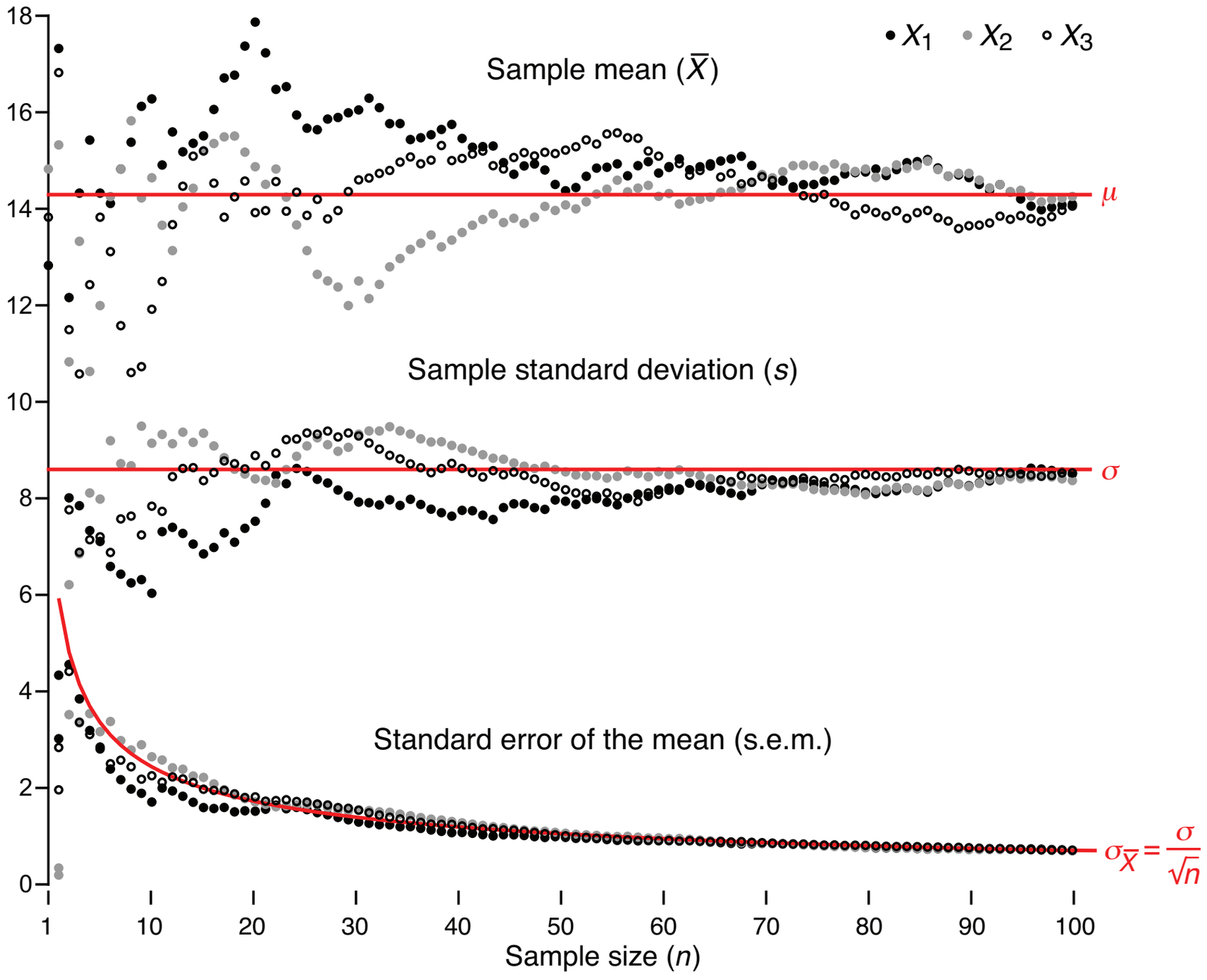
Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

EFFECT OF SAMPLE SIZE

As n increases, the estimates stabilize



Probability distribution of the data

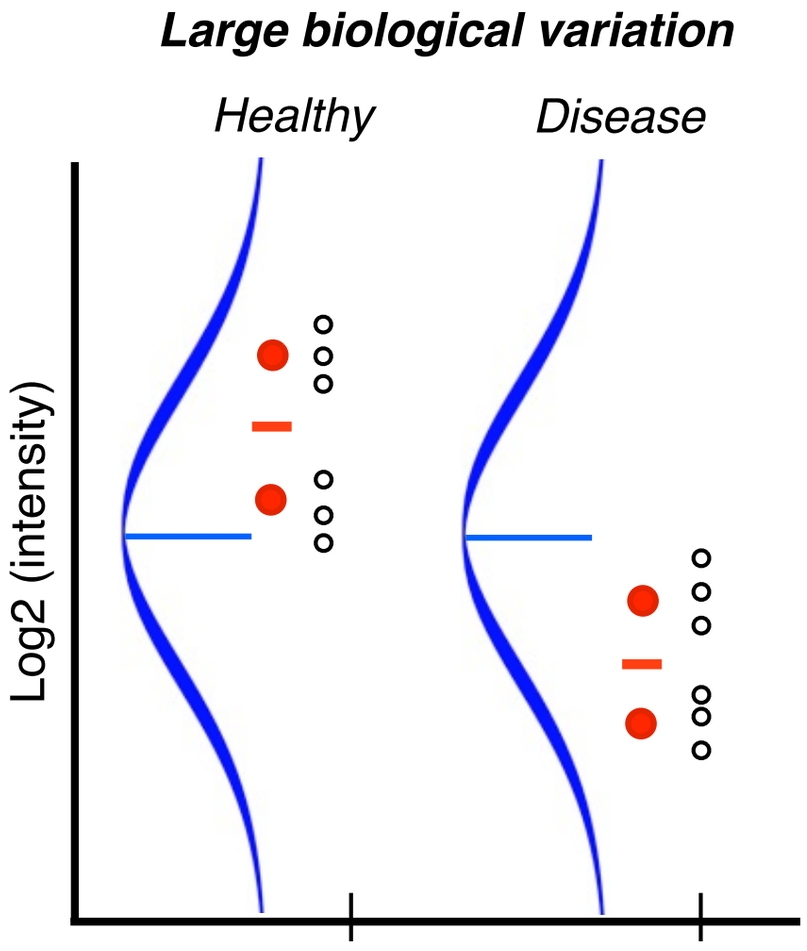


Simulated example

Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

EFFECT OF SAMPLE SIZE

As n increases, the estimates stabilize



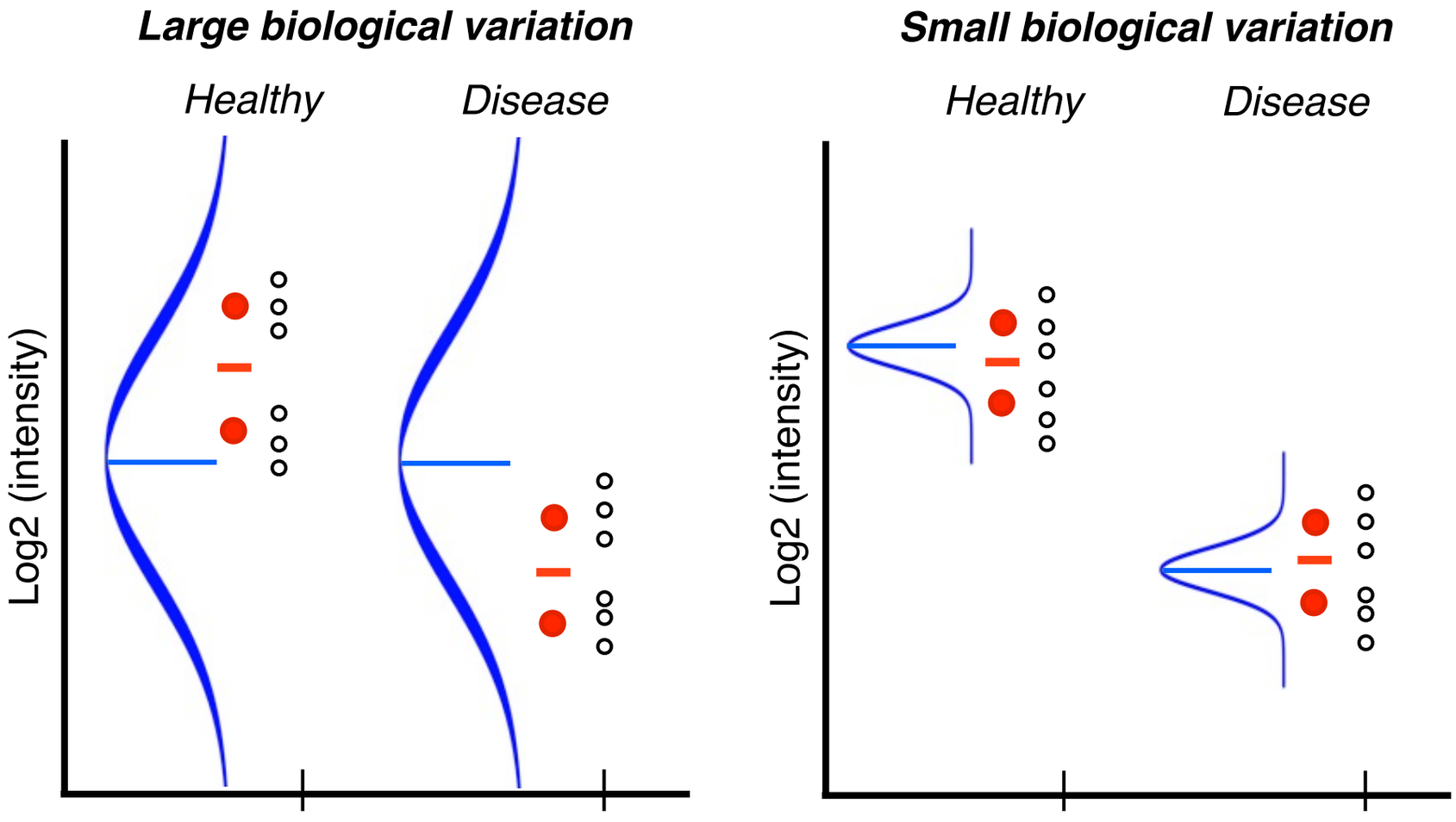
- *Protein abundance in the population*
- *Average abundance in the population*
- *Average abundance in the selected subjects*
- *Abundance in the individual subject*
- *Observed log-intensities of L*

When biological variance is large, more biological replicates are needed to accurately estimate the variance

Chang et al, *Molecular and Cellular Proteomics,* 11:M111.014662, 2012

EFFECT OF SAMPLE SIZE

As n increases, the estimates stabilize

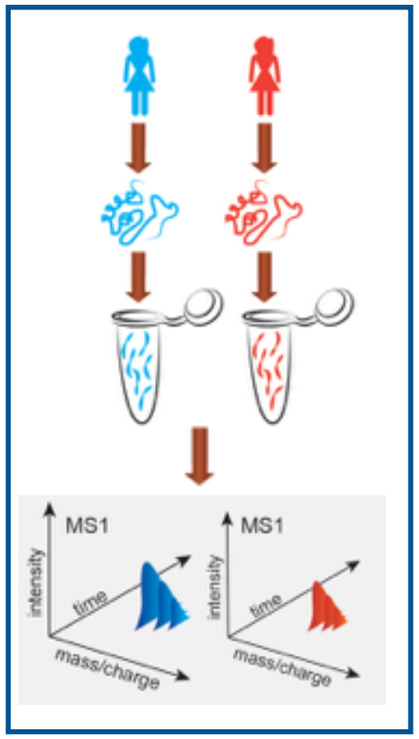


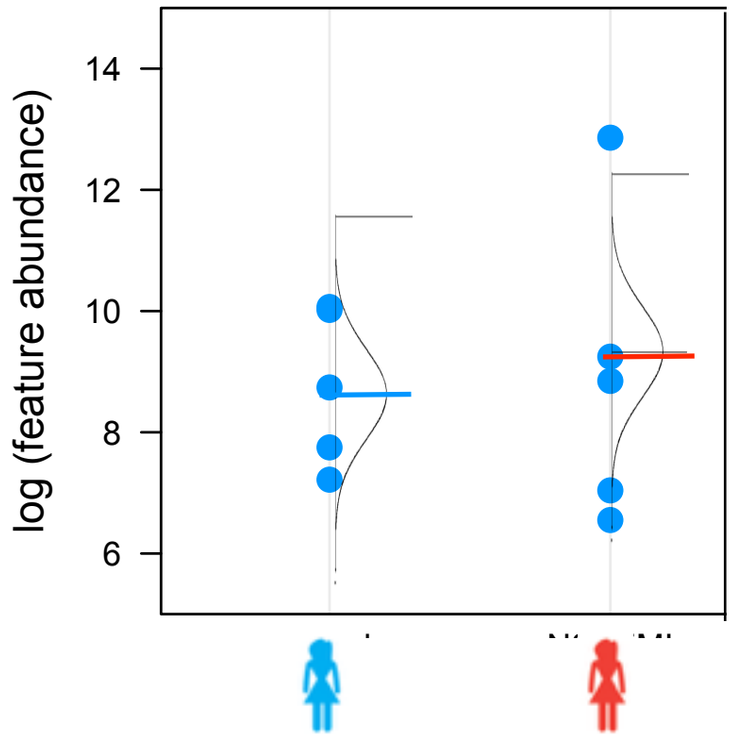
When biological variance is large, more biological replicates are needed to accurately estimate the variance

Chang et al, *Molecular and Cellular Proteomics,* 11:M111.014662, 2012

FINDING DIFFERENTIALLY ABUNDANT PROTEINS

False positive rate





H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

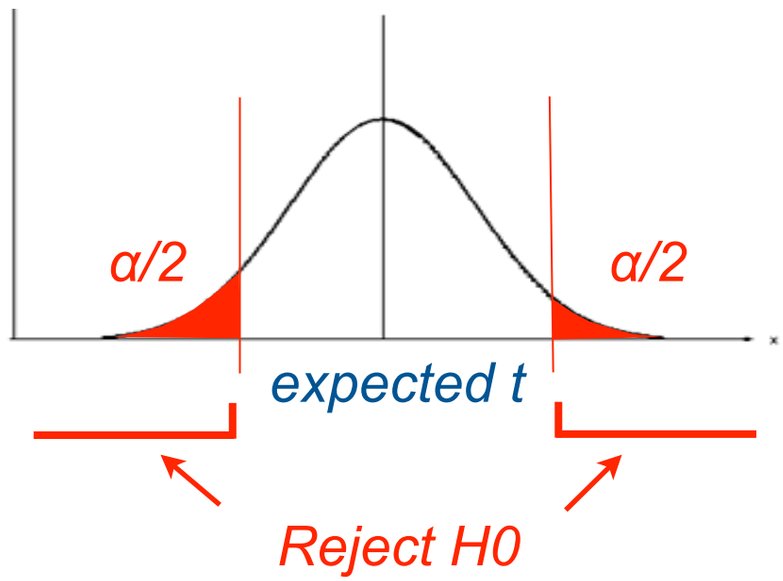
Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

$$\text{observed t} = \frac{\text{difference of group means}}{\text{estimate of variation}}$$

$$\overset{\text{no difference}}{\sim} \textbf{Student distribution}$$

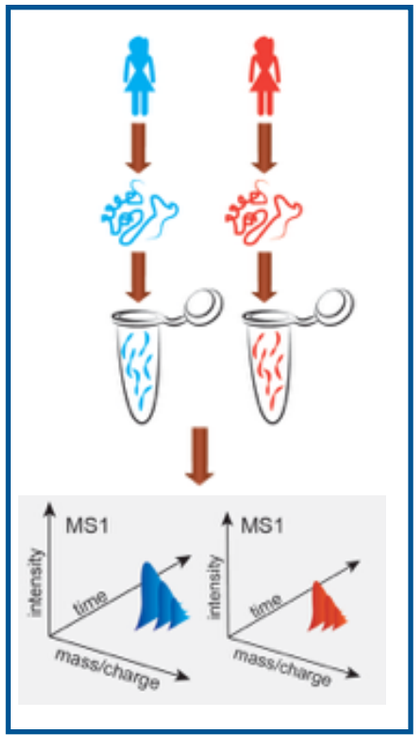
Distribution of the score if H0 is true

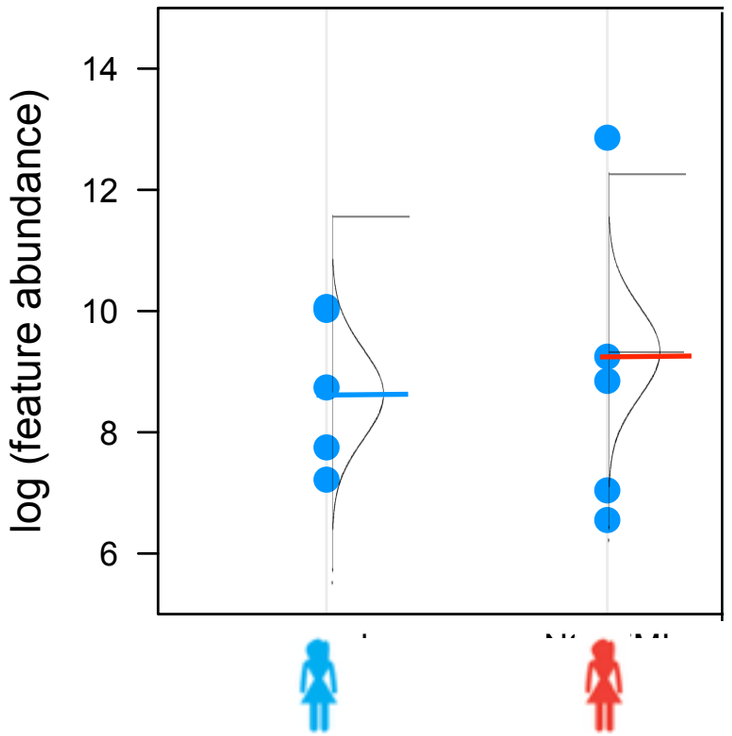
α = False Positive Rate



FINDING DIFFERENTIALLY ABUNDANT PROTEINS

P-value





H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

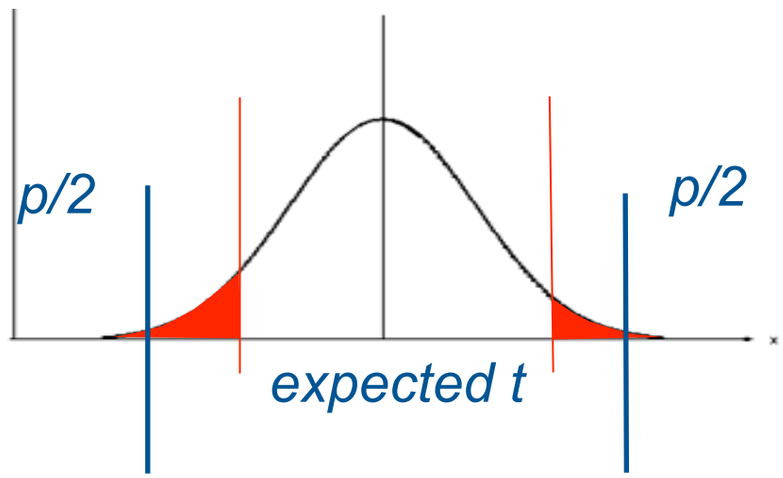
Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

$$\text{observed t} = \frac{\text{difference of group means}}{\text{estimate of variation}}$$

$$\overset{\text{no difference}}{\sim} \text{Student distribution}$$

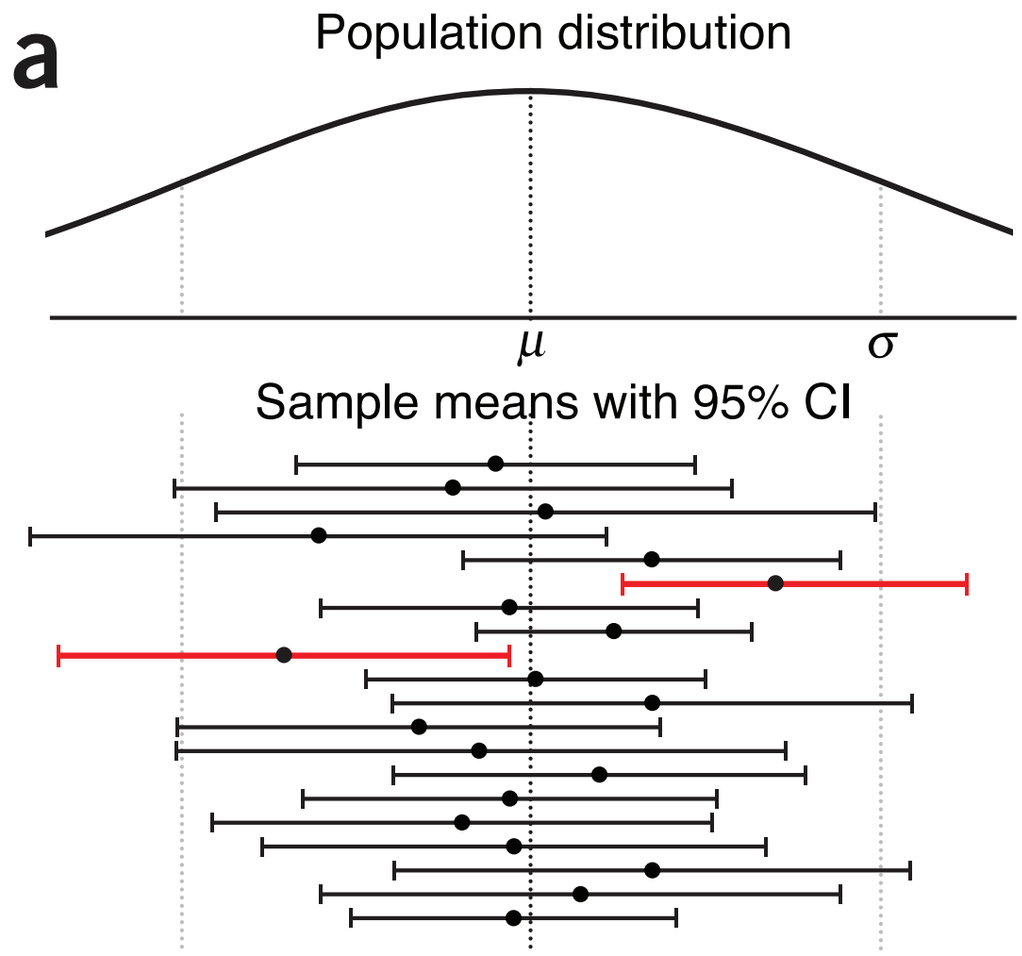
Distribution of the score if H0 is true

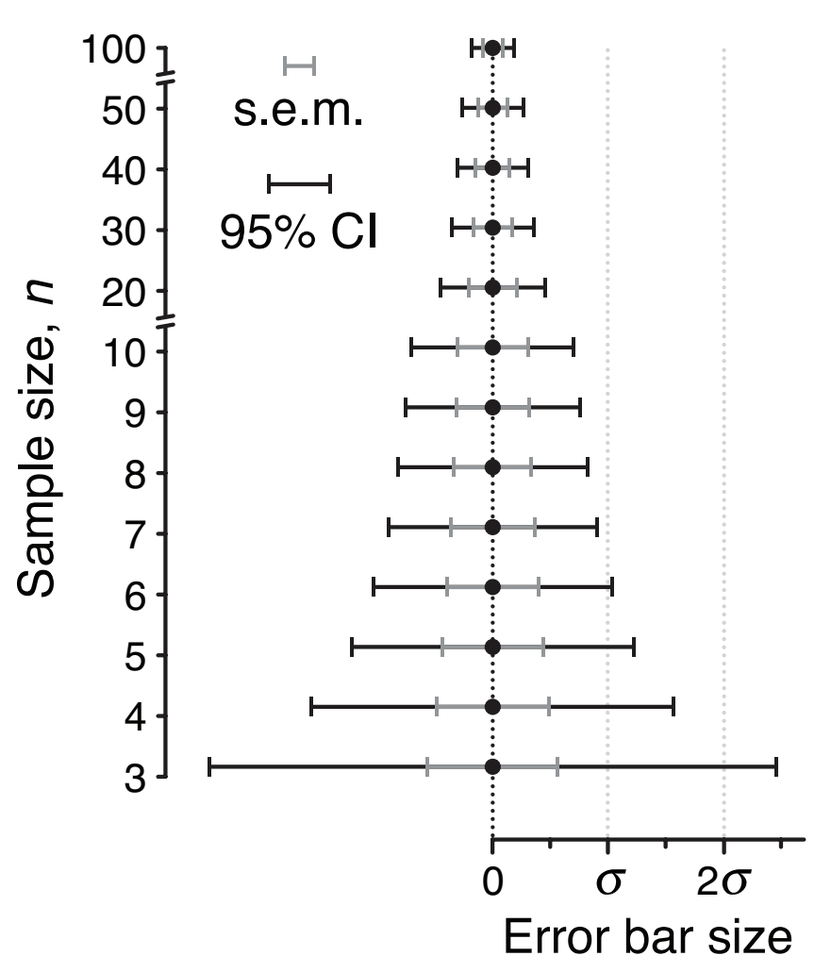
p = p-value



ALTERNATIVE TO TESTING: CONFIDENCE INTERVALS

Not all error bars are made equal





A 95% CI: if we repeatedly collect data and draw confidence intervals, then 95% of them will contain the true mean

CI are wider than bars indicating standard error of the mean!

$$\left[(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \;,\; (\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

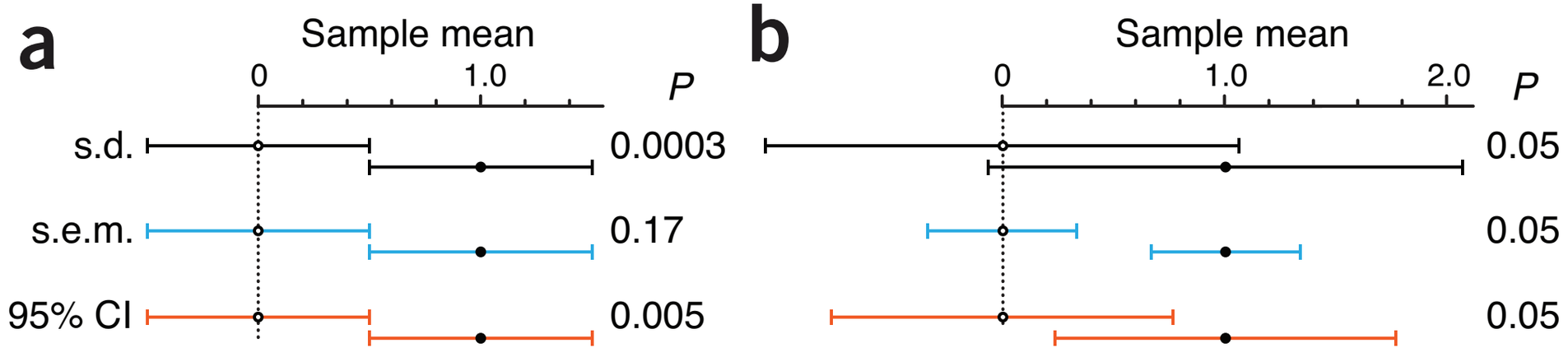
Width of the intervals depends on the sample size

Simulated example

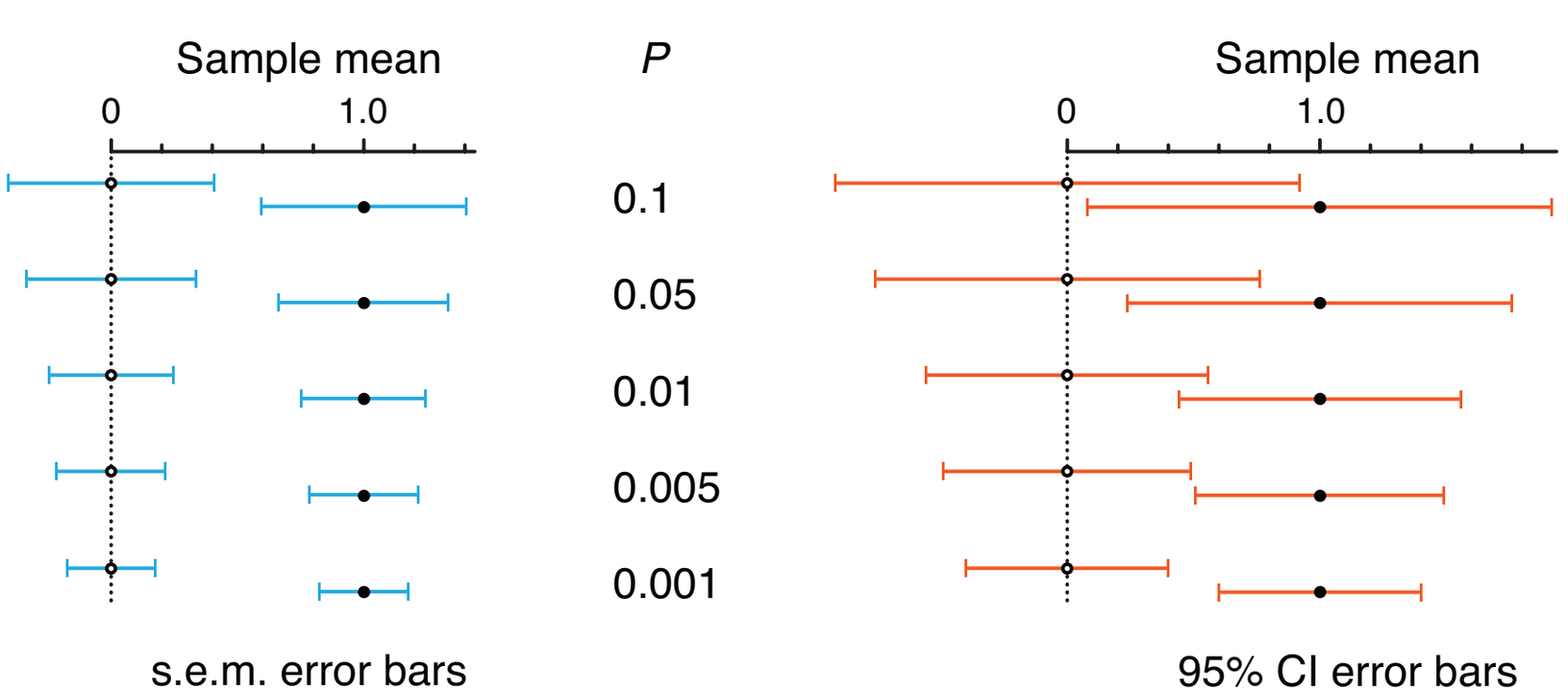
Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

ERROR BARS PROVIDE DIFFERENT INSIGHT

Absence of overlap does not always mean stat. significance



n=10



s.e.m. error bars

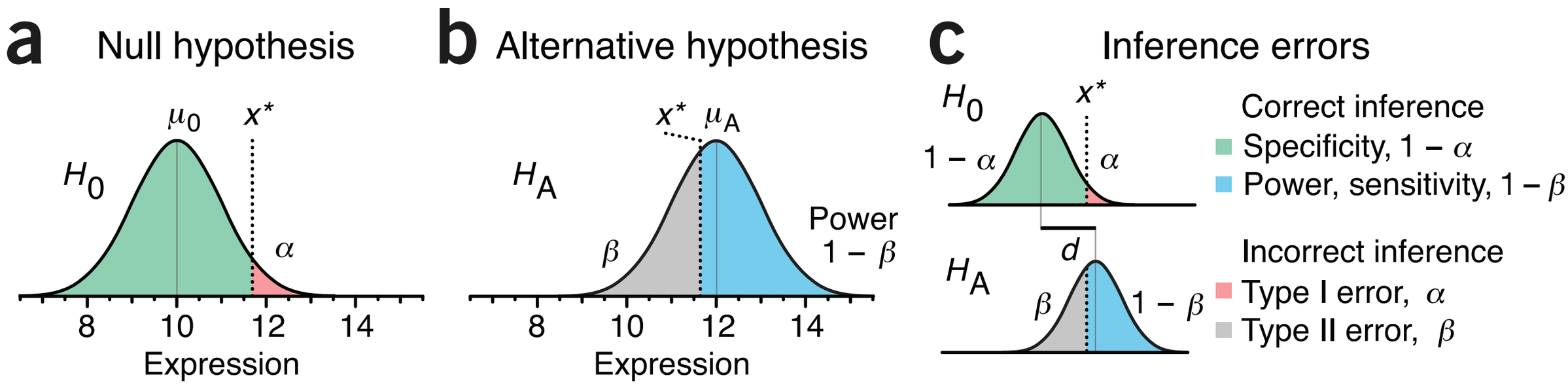
95% CI error bars

Simulated example

Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

STATISTICAL POWER

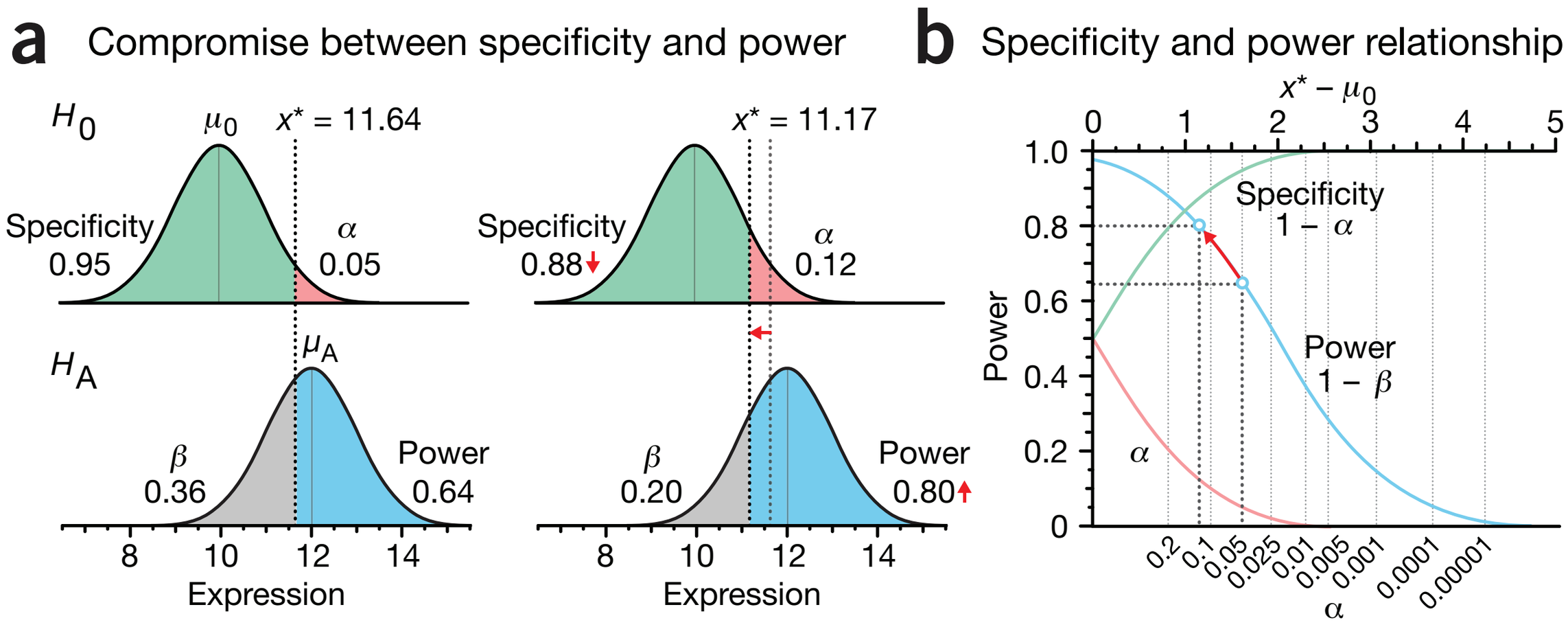
Probability to detect a difference when it exists



Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

STATISTICAL POWER

Probability to detect a difference when it exists

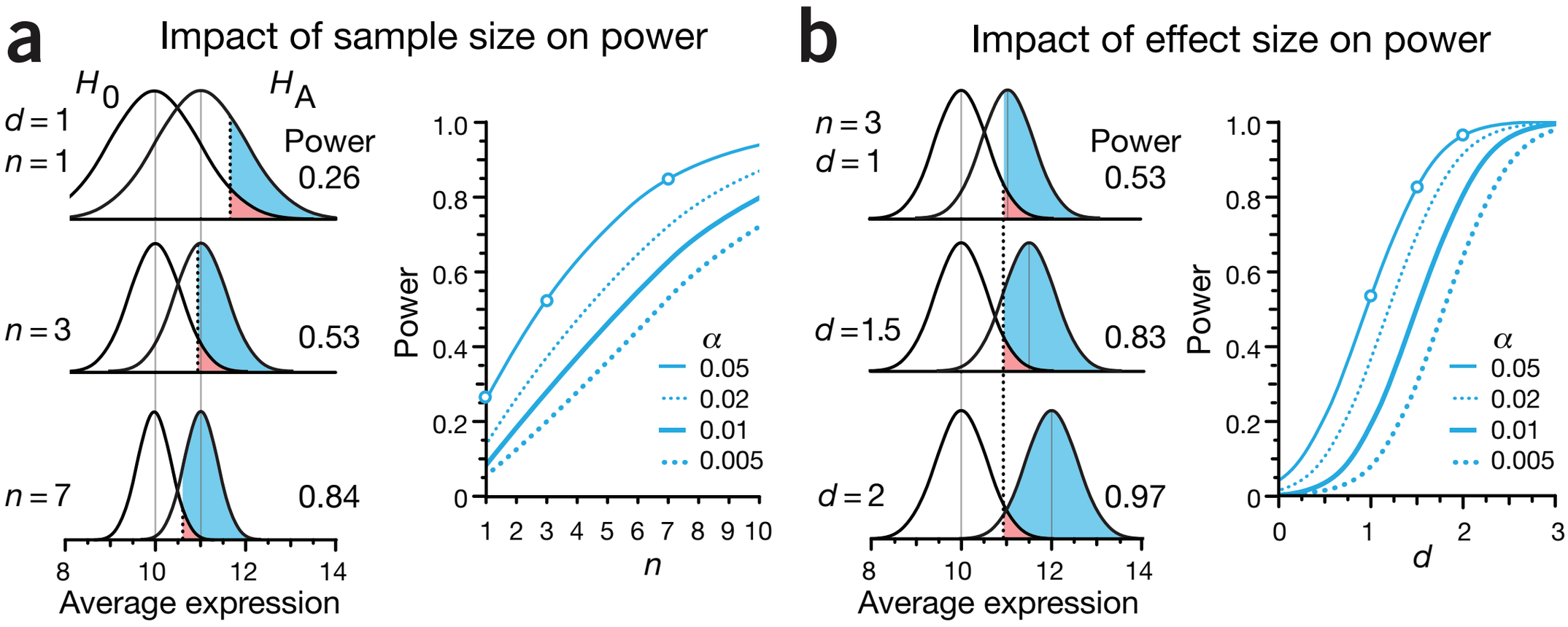


α trades off the sensitivity and the specificity of the test

Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

STATISTICAL POWER

Sample size and true fold change impact the power



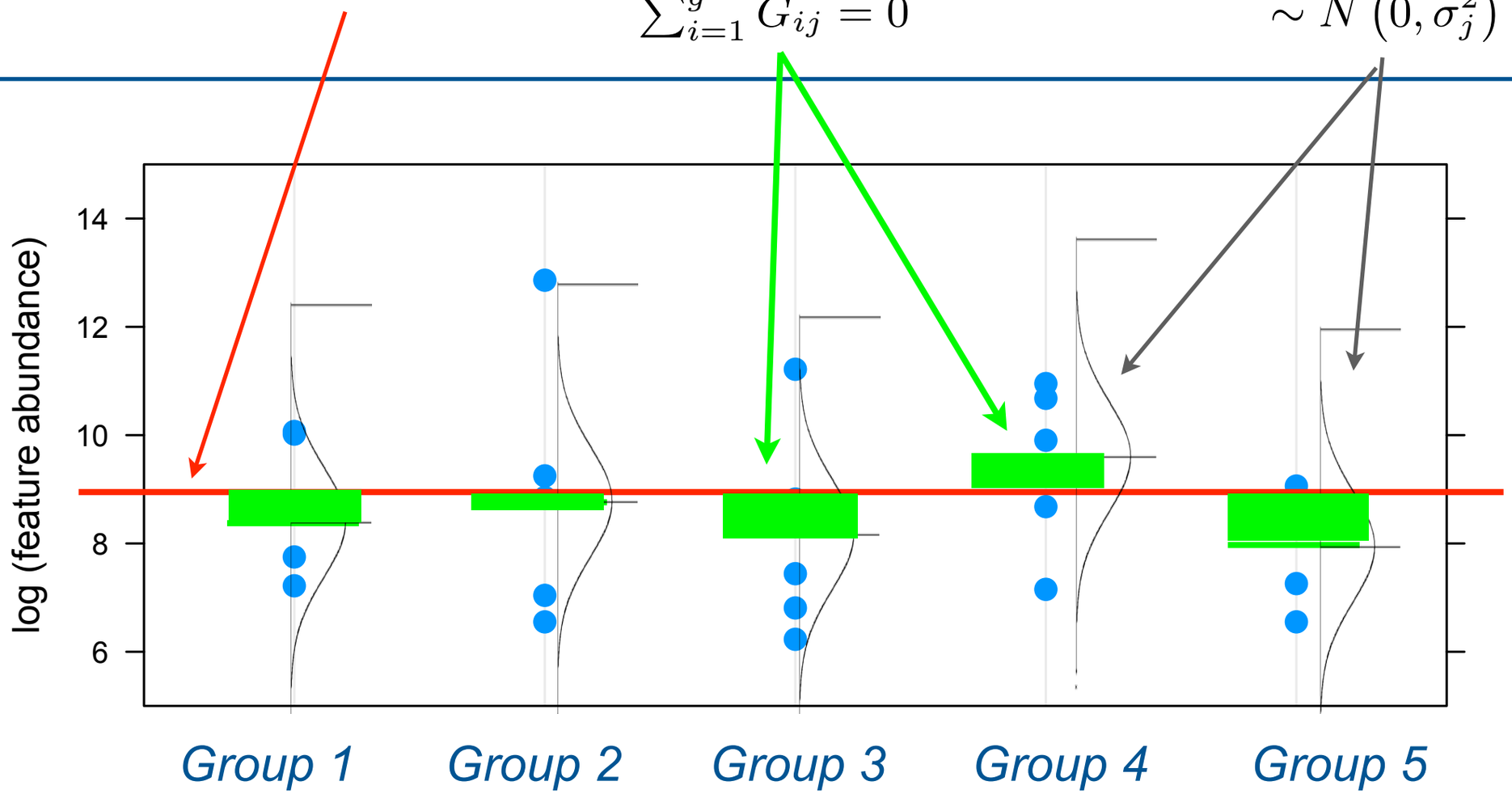
Higher statistical power is achieved when we have

- more replicates
- larger differences between the groups
- smaller biological and technical variation

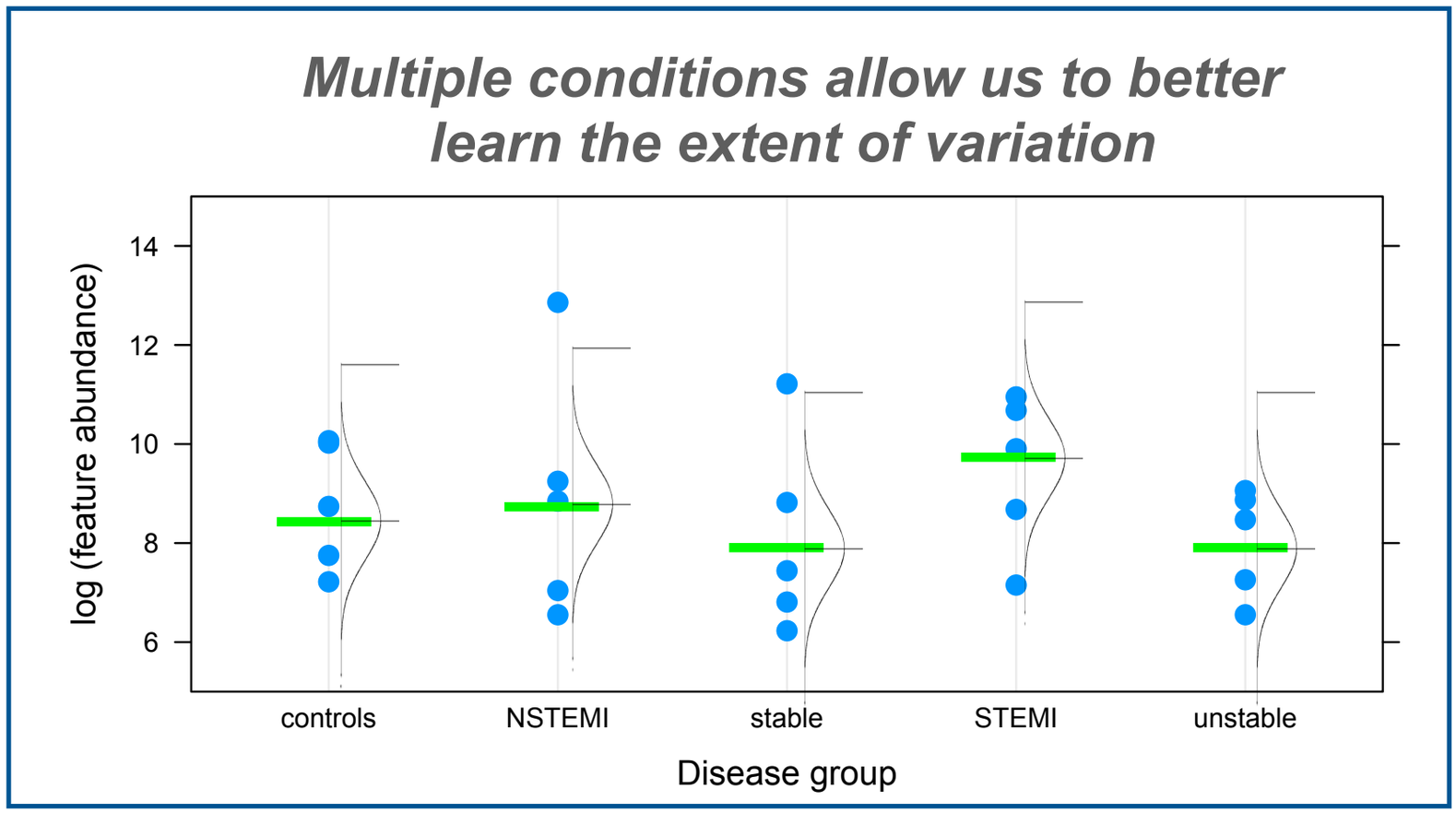
Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

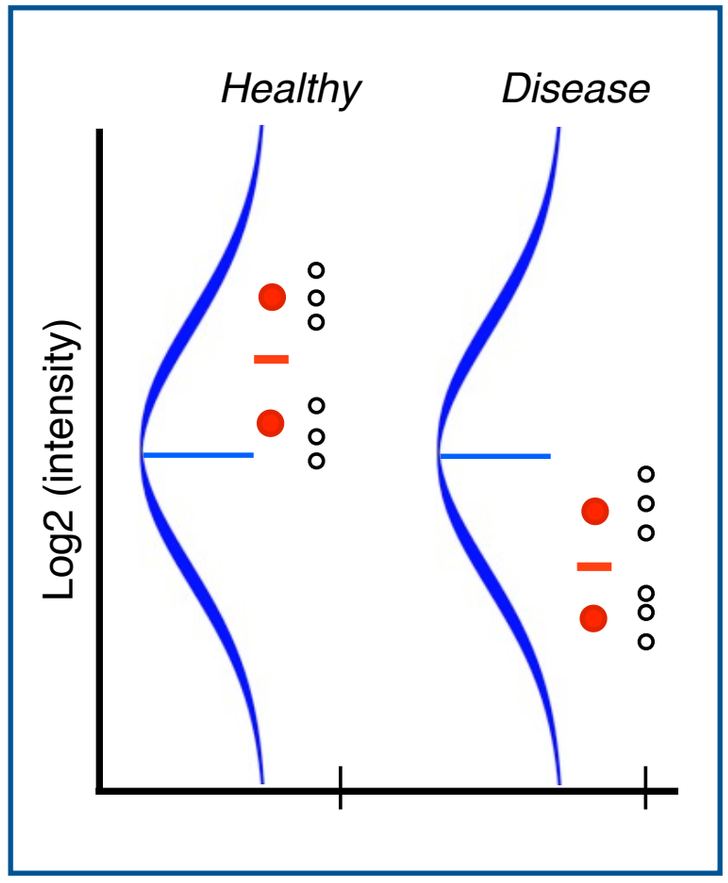
MULTI-GROUP ANALYSIS: A ONE-WAY ANALYSIS OF VARIANCE (ANOVA)

Observed feature intensity	=	Overall feature mean	+	Systematic deviation due to disease group	+	Random deviation due to non-systematic sources of variation
y_{ijk}	=	μ_j	+	G_{ij} $\sum_{i=1}^{g} G_{ij} = 0$	+	ϵ_{ijk} $\sim N\left(0, \sigma_j^2\right)$



LINEAR MIXED MODELS DESCRIBE COMPLEX DESIGNS





Observed feature intensity	=	Systematic mean signal of disease group	+	Systematic/Random deviation of subject	+	Random deviation of measurement error

More complicated models with more terms
More flexibility and accuracy

OUTLINE

- Describing the data
  - Center and variation
- Basic statistical inference
  - T-test and p-values
- P-values: a word of caution
  - Instability, multiplicity, alternative approaches

AMERICAN STATISTICAL ASSOCIATION (ASA) STATEMENT ON STATISTICAL SIGNIFICANCE AND P-VALUES

The American Statistician, February 2016

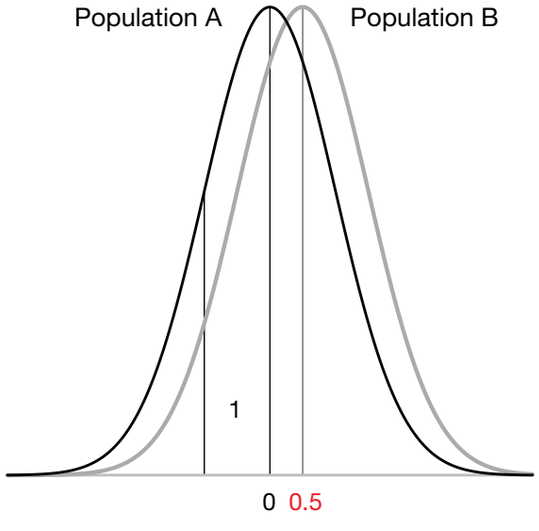
- P-values can indicate how incompatible the data are with a specified statistical model
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance
- Scientific conclusions and business policy decisions should not be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

AMERICAN STATISTICAL ASSOCIATION (ASA) STATEMENT ON STATISTICAL SIGNIFICANCE AND P-VALUES

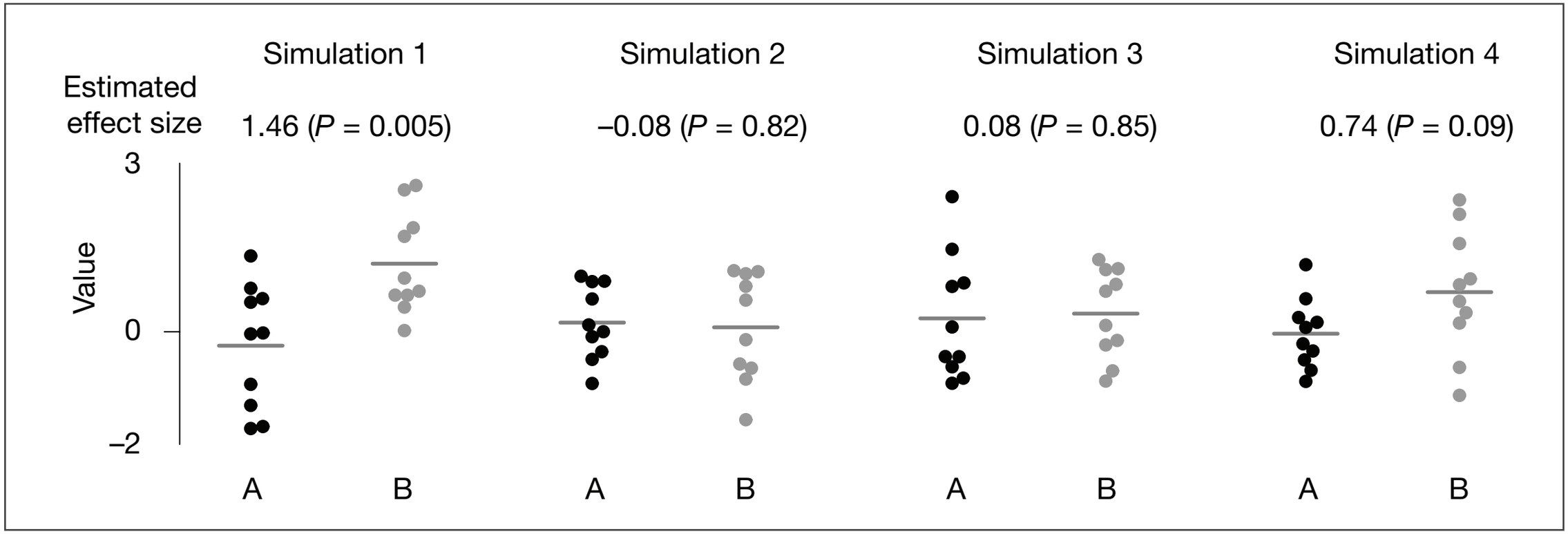
The American Statistician, February 2016

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

WITH SMALL SAMPLE SIZE, P-VALUES ARE UNSTABLE



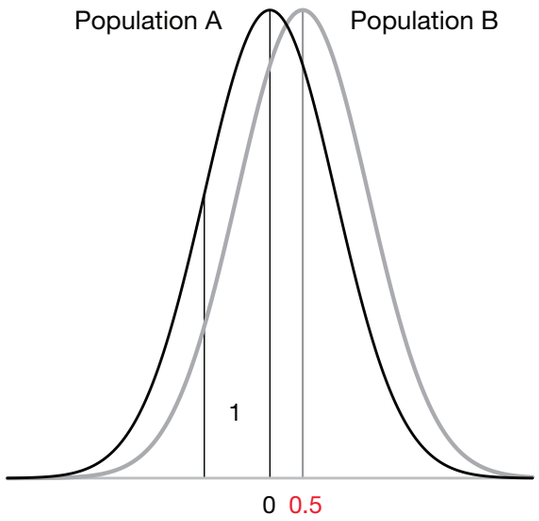
- Repeatedly sampling data leads to different results
- The problem worsens when testing many proteins
- Partial solutions:
  - Larger sample size
  - Adjustment for multiple testing



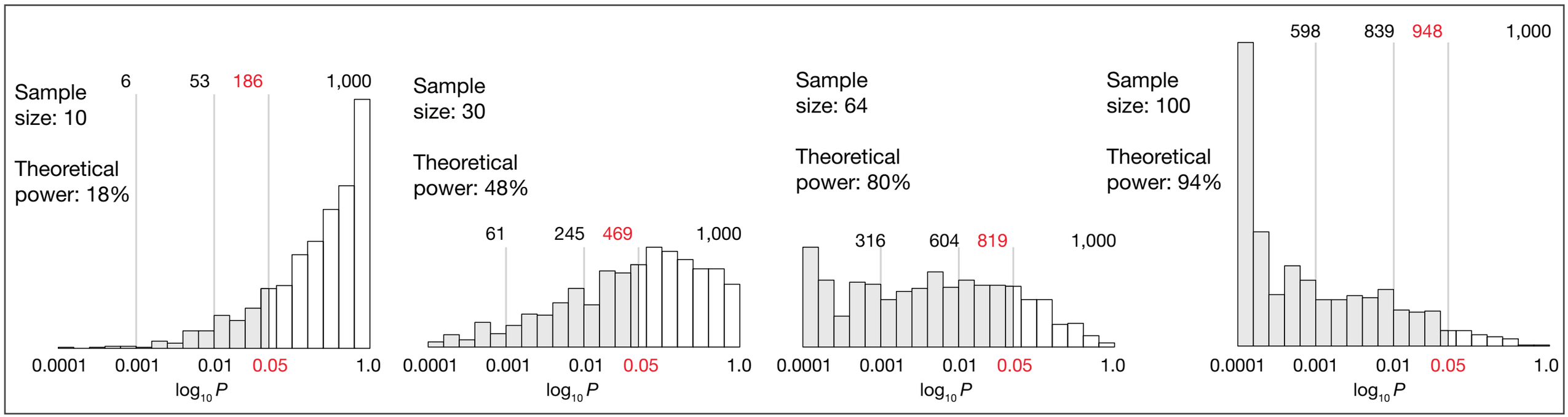
Simulated example

Halsey, Curran-Everett, Vowler and Drummond, *Nature Methods*, 2015

WITH SMALL SAMPLE SIZE, P-VALUES ARE UNSTABLE



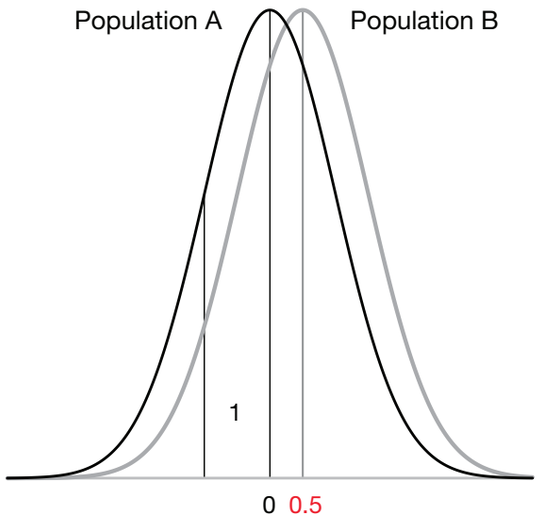
- Repeatedly sampling data leads to different results
- The problem worsens when testing many proteins
- Partial solutions:
  - Larger sample size
  - Adjustment for multiple testing



Simulated example

Halsey, Curran-Everett, Volwer and Drummond, *Nature Methods*, 2015

WITH SMALL SAMPLE SIZE, CONCLUSIONS ARE BIASED



Simulated example

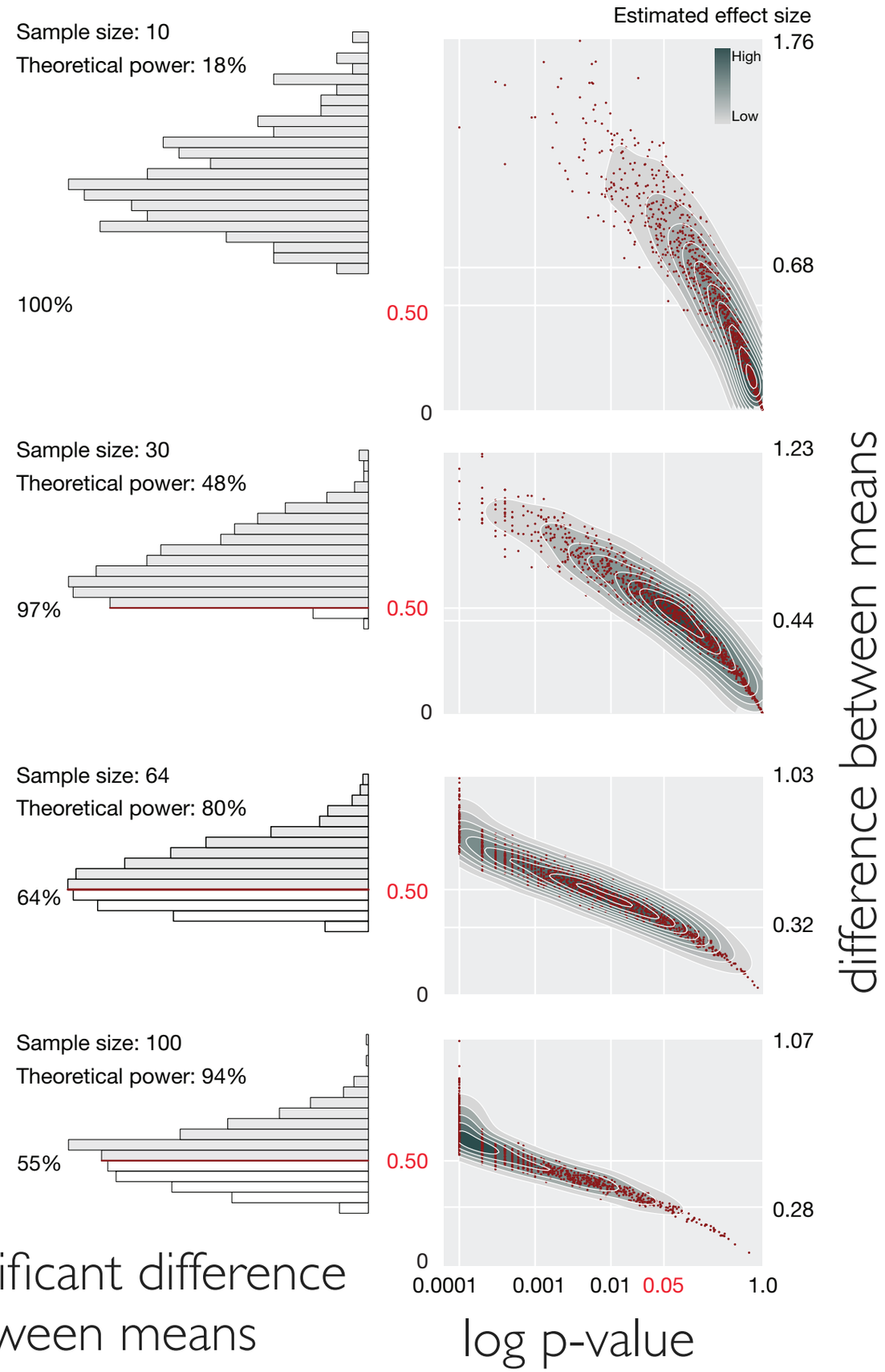
Halsey, Curran-Everett, Volwer and Drummond, *Nature Methods*, 2015

10 replicates

30 replicates

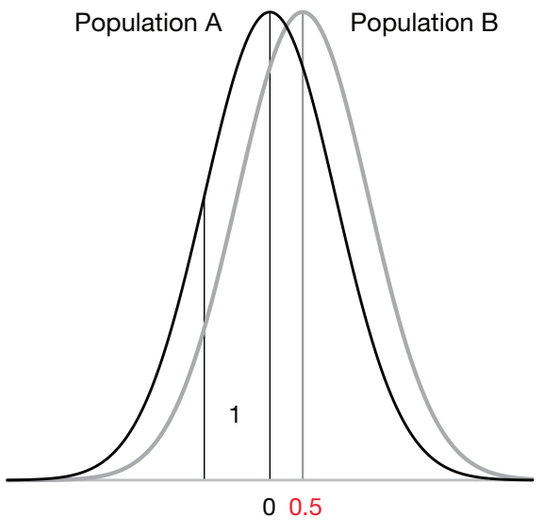
64 replicates

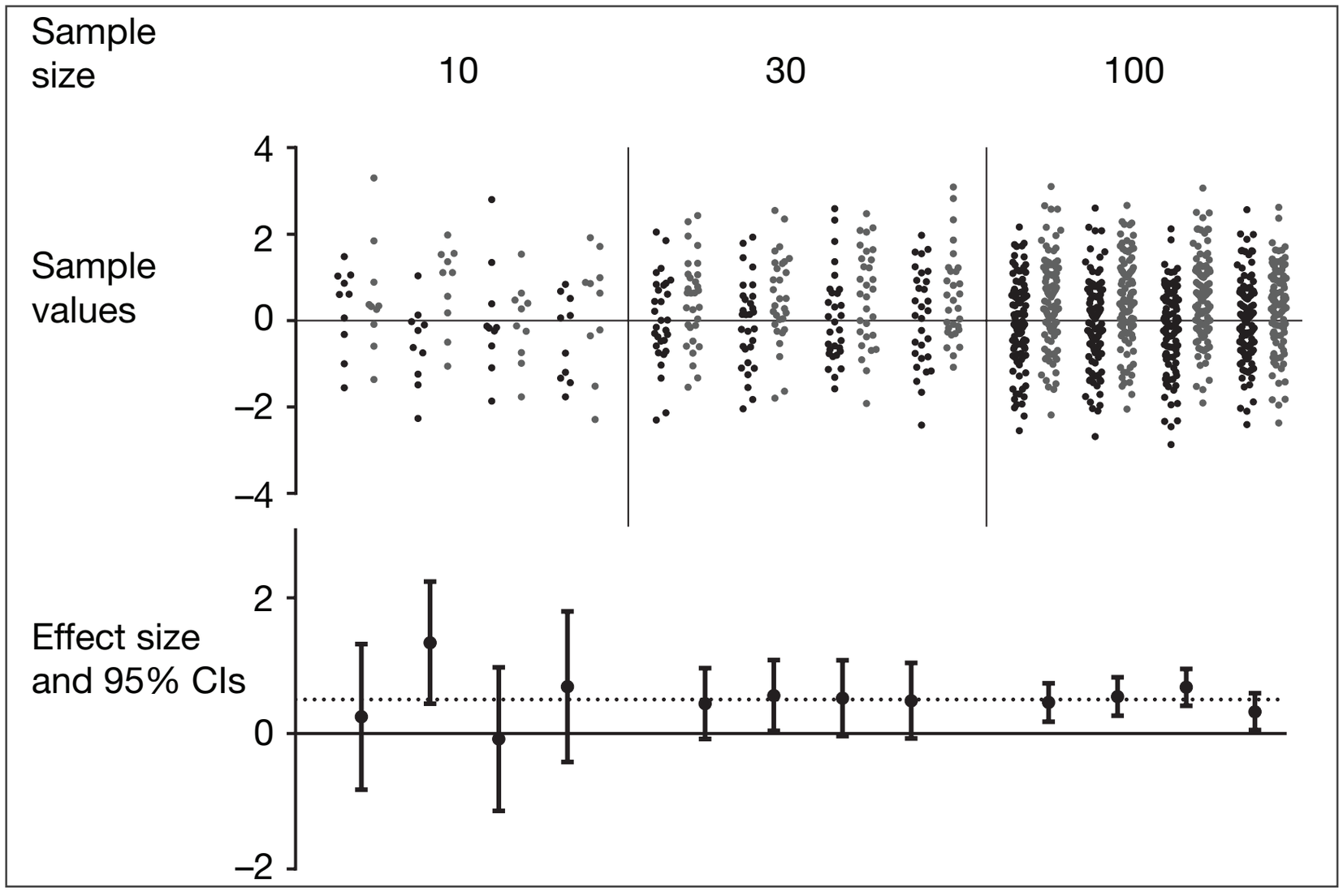
100 replicates



significant difference between means

CONFIDENCE INTERVALS PROVIDE COMPLEMENTARY INSIGHT





Simulated example

Halsey, Curran-Everett, Volwer and Drummond, *Nature Methods*, 2015

PITFALL: MULTIPLE TESTING

- An fMRI on dead fish
- Found many active brain regions
  - All background noise and random variation

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY;
[3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

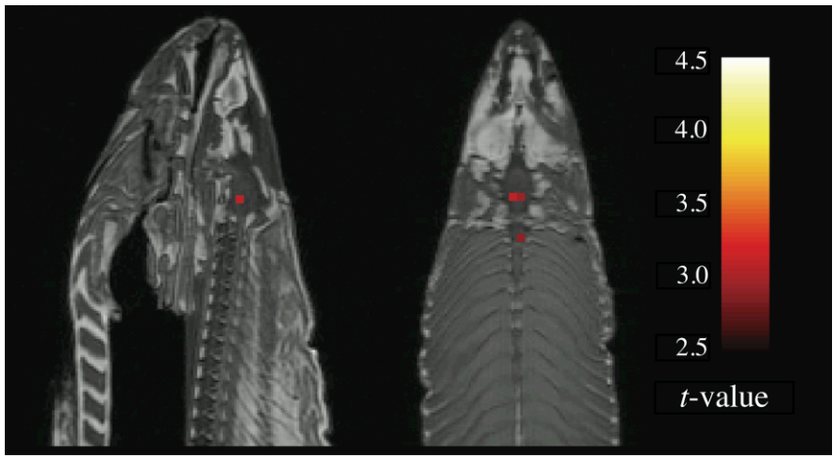
INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at

GLM RESULTS



Source: a blog by Jeff Leek, Biostatistics, John Hopkins University

http://simplystatistics.org/2016/02/01/a-menagerie-of-messed-up-data-analyses-and-how-to-avoid-them/

MULTIPLE TESTING

Control False Positive Rate for two proteins

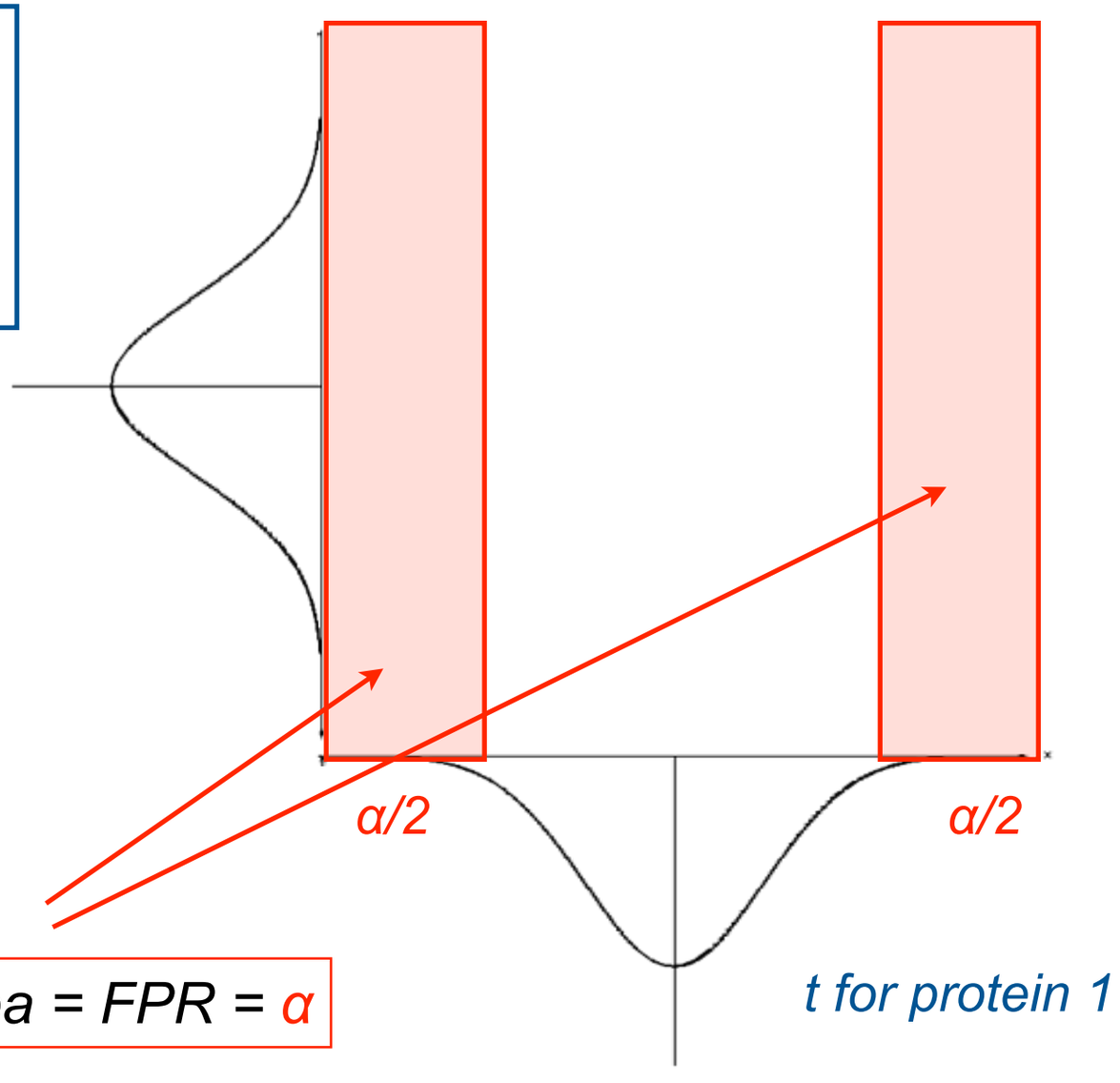
For each protein:

H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

$$\textbf{observed t} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

$$\overset{\textbf{no difference}}{\sim} \textbf{Student distribution}$$



The area = FPR = α

MULTIPLE TESTING

Control False Positive Rate for two proteins

For each protein:

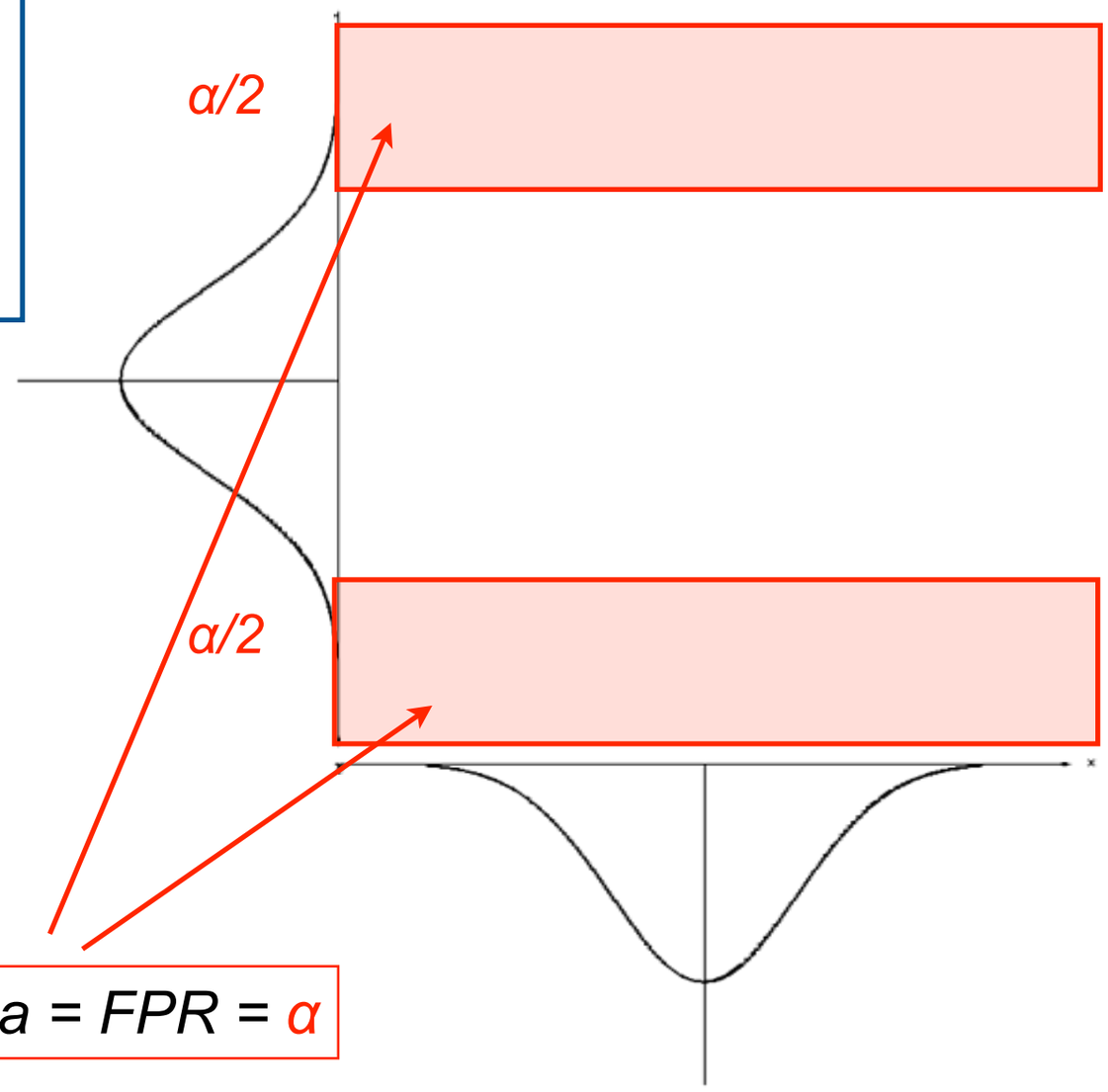
H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

$$\textbf{observed t} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

$$\overset{\textbf{no difference}}{\sim} \textbf{Student distribution}$$

t for protein 2



The area = FPR = α

MULTIPLE TESTING

Control False Positive Rate for two proteins

For each protein:

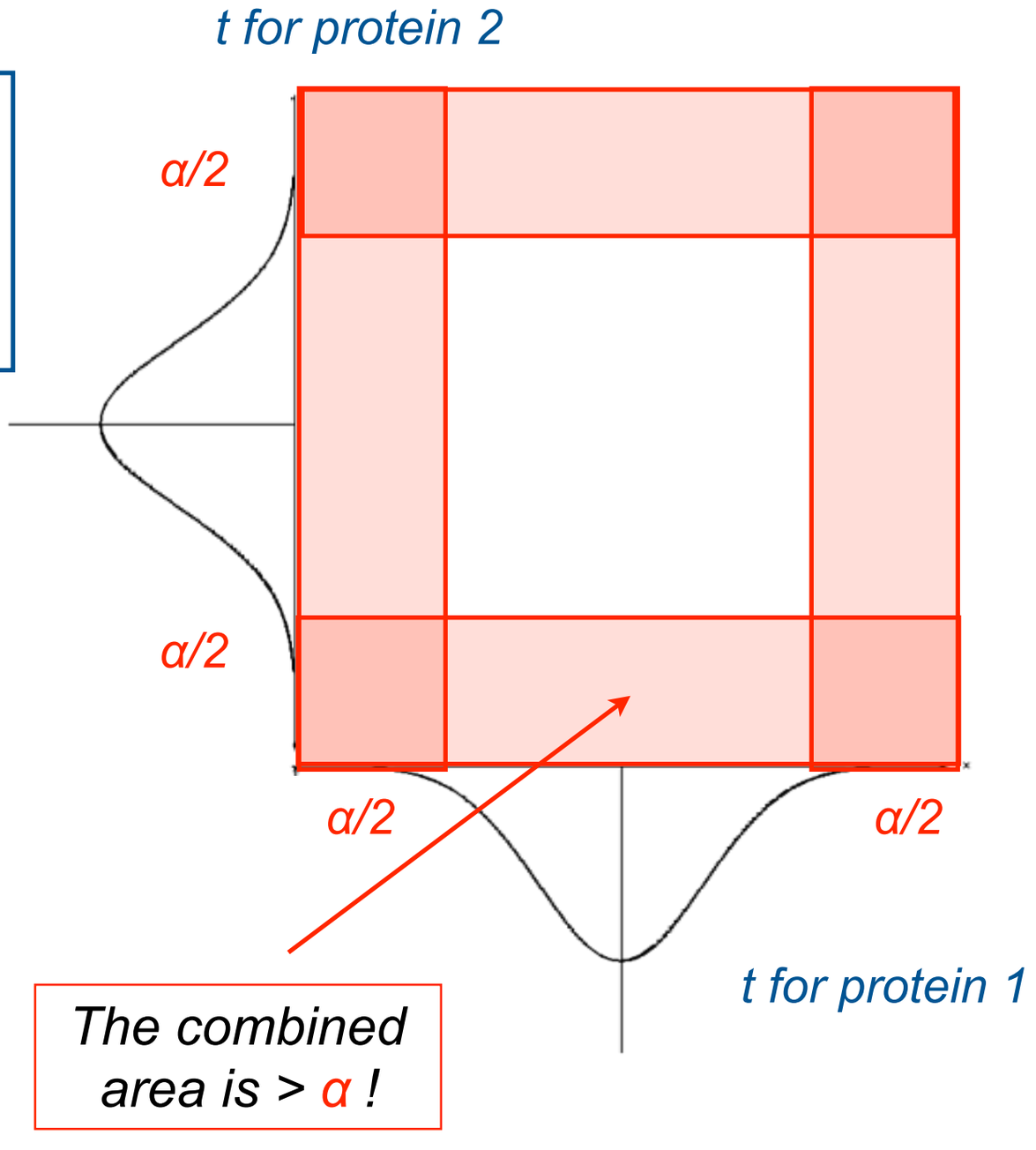
H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

$$\text{observed t} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

$$\overset{\text{no difference}}{\sim} \textbf{Student distribution}$$

- P(at least one incorrect decision) > α
- The univariate FPR does not hold for the list
- Need to define a *multivariate* error rate



MULTIPLE TESTING

Control False Positive Rate for two proteins

For each protein:

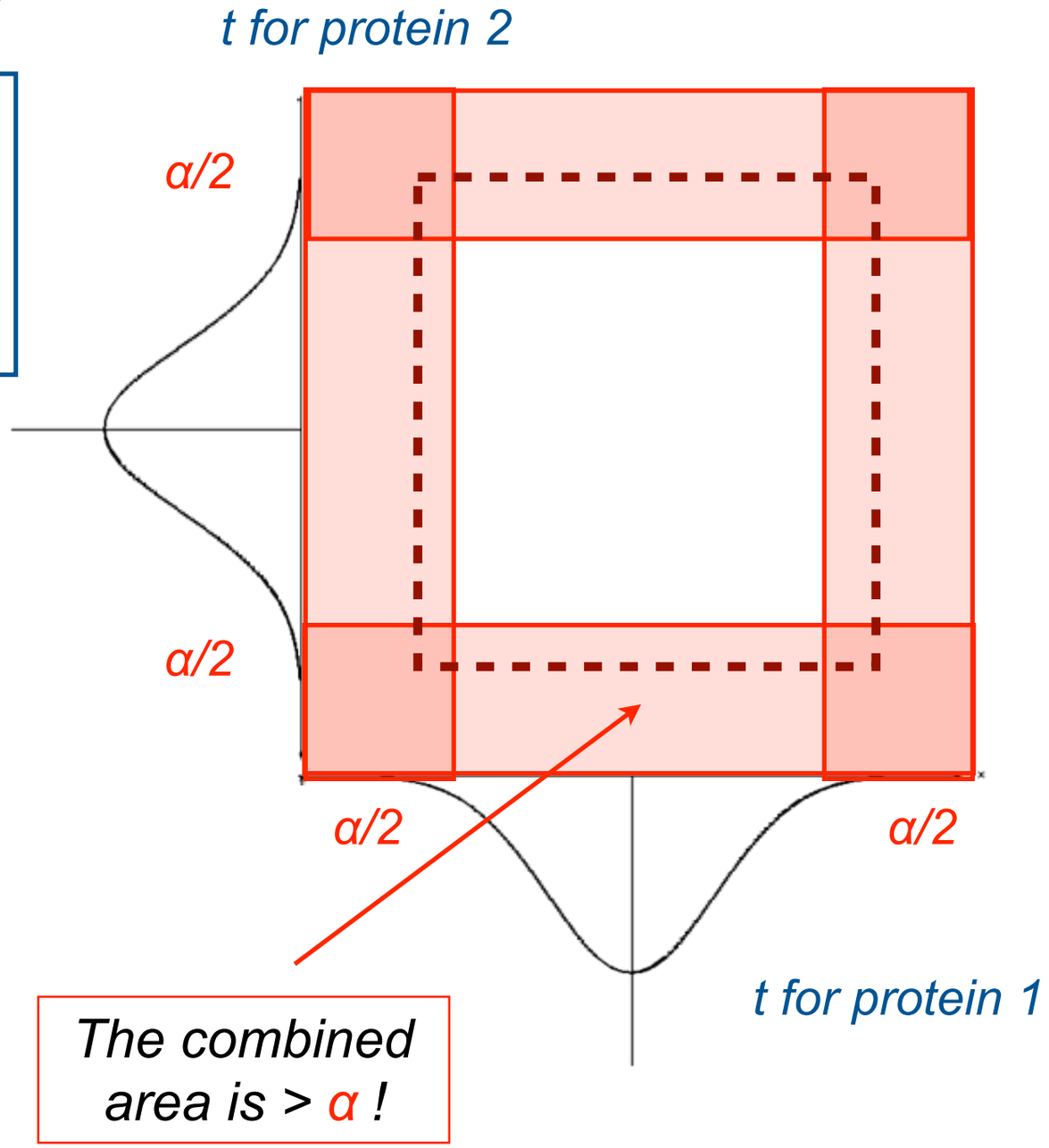
H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

$$\textbf{observed t} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

$$\overset{\textbf{no difference}}{\sim} \textbf{Student distribution}$$

- P(at least one incorrect decision) > α
- The univariate FPR does not hold for the list
- Need to define a *multivariate* error rate



TESTING M PROTEINS

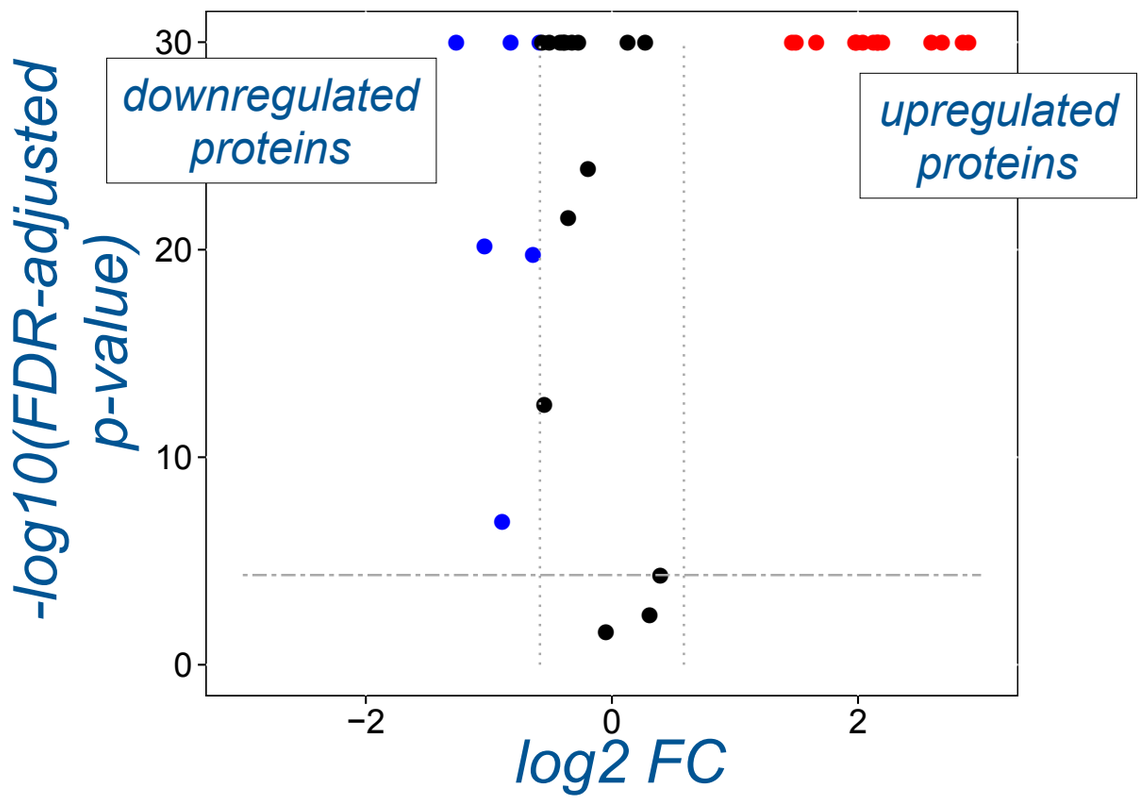
Change criteria from False Positive Rate to False Discovery Rate

	# of proteins with no detected difference	# of proteins with detected difference	Total
# true non-diff. proteins	U	V	m_0
# true diff. proteins	T	S	$m_1 = m - m_0$
Total	$m - R$	R	m

- False discovery rate (FDR)
  - An infinite number of measurements on same proteins
  - FDR: the *average* proportion of false discoveries

$$\mathbf{FDR} = \mathbf{E}\left[\frac{\mathbf{V}}{\max(\mathbf{R}, 1)}\right]$$

Bonferroni approach controls family-wise error rate = $P(V > 0)$



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Vide

Archive > Volume 520 > Issue 7549 > Comment > Article

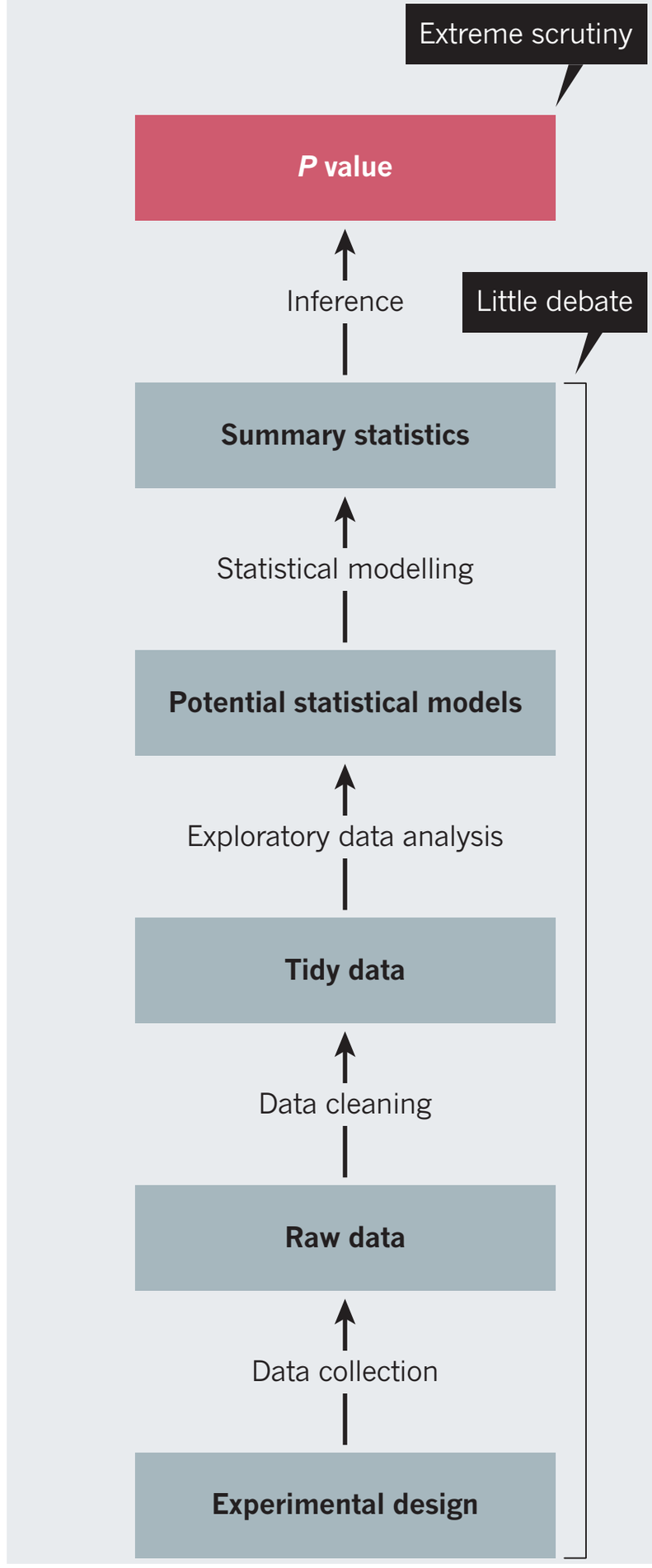
NATURE | COMMENT

Statistics: *P* values are just the tip of the iceberg

Jeffrey T. Leek & Roger D. Peng

28 April 2015

Statistical considerations are key at every step



PITFALL: OUTCOME SWITCHING

- Anti-depressant Paxil was studied for several main outcomes
  - None showed an effect
  - Some secondary outcomes dis
- Switched the outcome of the trial and used to market the drug

Vox SCIENCE & HEALTH

How researchers dupe the public with a sneaky practice called "outcome switching"

Updated by Julia Belluz on December 29, 2015, 8:10 a.m. ET

julia.belluz@voxmedia.com

Source: a blog by Jeff Leek, Biostatistics, John Hopkins University

http://simplystatistics.org/2016/02/01/a-menagerie-of-messed-up-data-analyses-and-how-to-avoid-them/

PITFALL: NOT PRE-SPECIFIED DATA SELECTION AND ANALYSIS

- Compare 2 groups: women at peak and off peak fertility cycle
  - A series of choices of which women to include in which comparison group
- Conclude that at peak fertility women are more likely to wear red or pink shirts

The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*

Andrew Gelman† and Eric Loken‡

14 Nov 2013

Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of *potential* comparisons when the details of data analysis are highly contingent on data, without the researcher having to perform any conscious procedure of fishing or examining multiple p-values. We discuss in the context of several examples of published papers where data-analysis decisions were theoretically-motivated based on previous literature, but where the details of data selection and analysis were not pre-specified and, as a result, were contingent on data.

Source: a blog by Jeff Leek, Biostatistics, John Hopkins University

http://simplystatistics.org/2016/02/01/a-menagerie-of-messed-up-data-analyses-and-how-to-avoid-them/

RESEARCHER DEGREE OF FREEDOM

Psychological SCIENCE
A Journal of the Association for Psychological Science

False-Positive Psychology

Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1],

Leif D. Nelson[2] **and**

Uri Simonsohn[1]

+ Author Affiliations

Joseph P. Simmons, The Wharton School, University of Pennsylvania, 551 Jon M. Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104 E-mail: jsimmo@wharton.upenn.edu
Leif D. Nelson, Haas School of Business, University of California, Berkeley, Berkeley, CA 94720-1900 E-mail: leif_nelson@haas.berkeley.edu
Uri Simonsohn, The Wharton School, University of Pennsylvania, 548 Jon M. Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104 E-mail: uws@wharton.upenn.edu

This Article

Published online before print October 17, 2011, doi: 10.1177/0956797611417632

Psychological Science **November 2011** vol. 22 no. 11 **1359-1366**

» **Abstract** ***Free***
Full Text ***Free***
Full Text (PDF) ***Free***

All Versions of this Article:
» Version of Record - Nov 7, 2011
0956797611417632v1 - Oct 17, 2011
What's this?