STATISTICAL INFERENCE

Olga Vitek

College of Science

College of Computer and Information Science

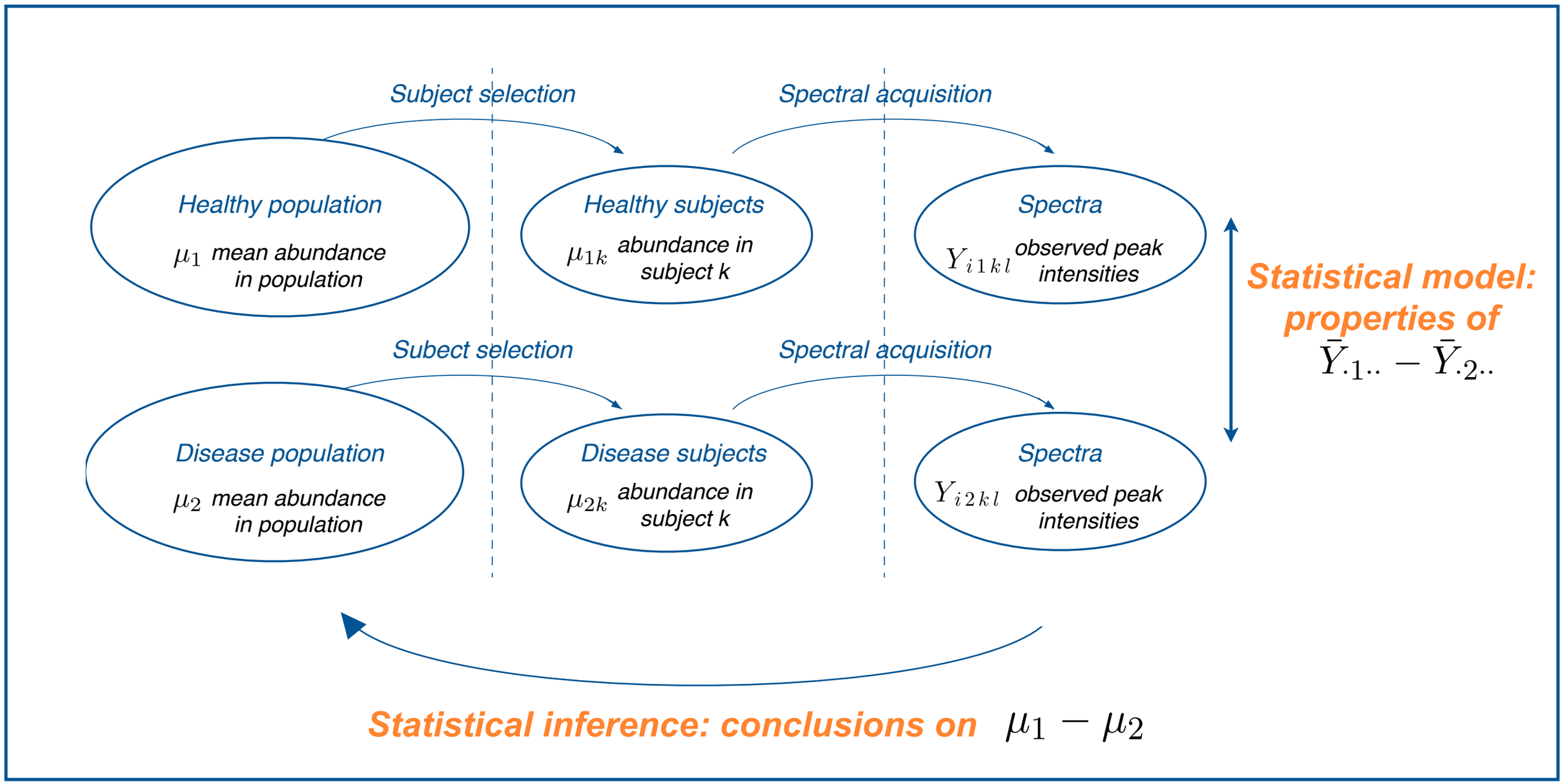
Northeastern University

OUTLINE

- Basic statistical inference
  - T-test and p-values
- P-values: a word of caution
  - Instability, multiplicity, alternative approaches
- So how many replicates do I need?
  - Design of complex experiments

COMPARE DESIGNS

In terms of bias and (in)-efficiency

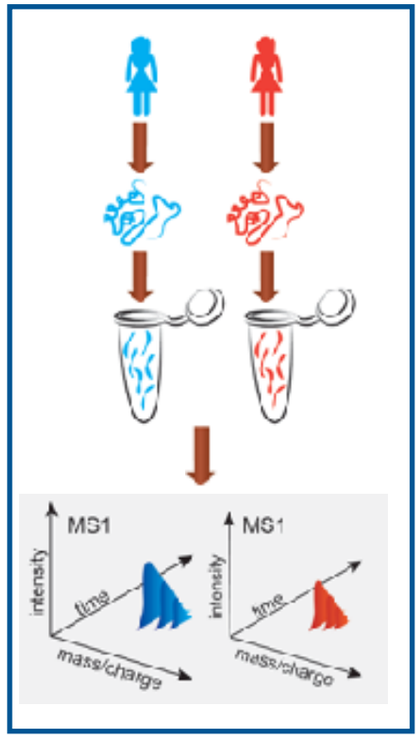


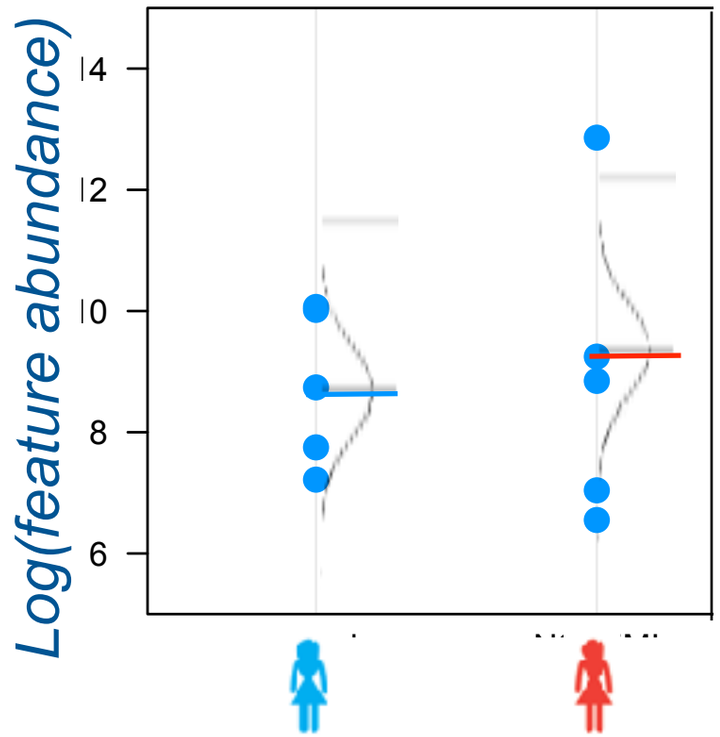
Bias: $\bar{Y}_{.1..} - \bar{Y}_{.2..}$ *systematically different from* $\mu_{1k} - \mu_{2k}$

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

TWO-SAMPLE T-TEST

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





Sample means in each group

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

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

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

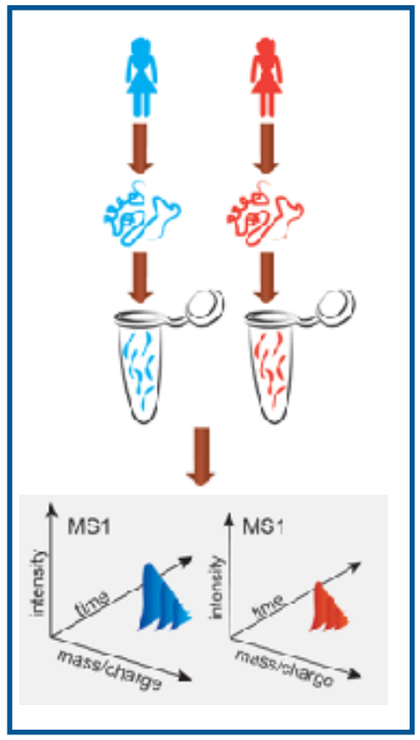
Number of replicates

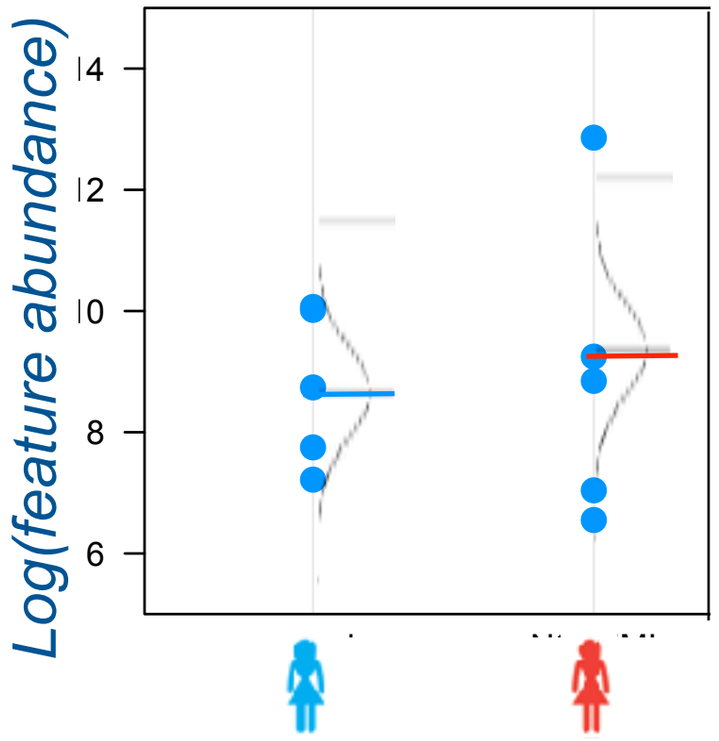
$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_{1\cdot})^2$$

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

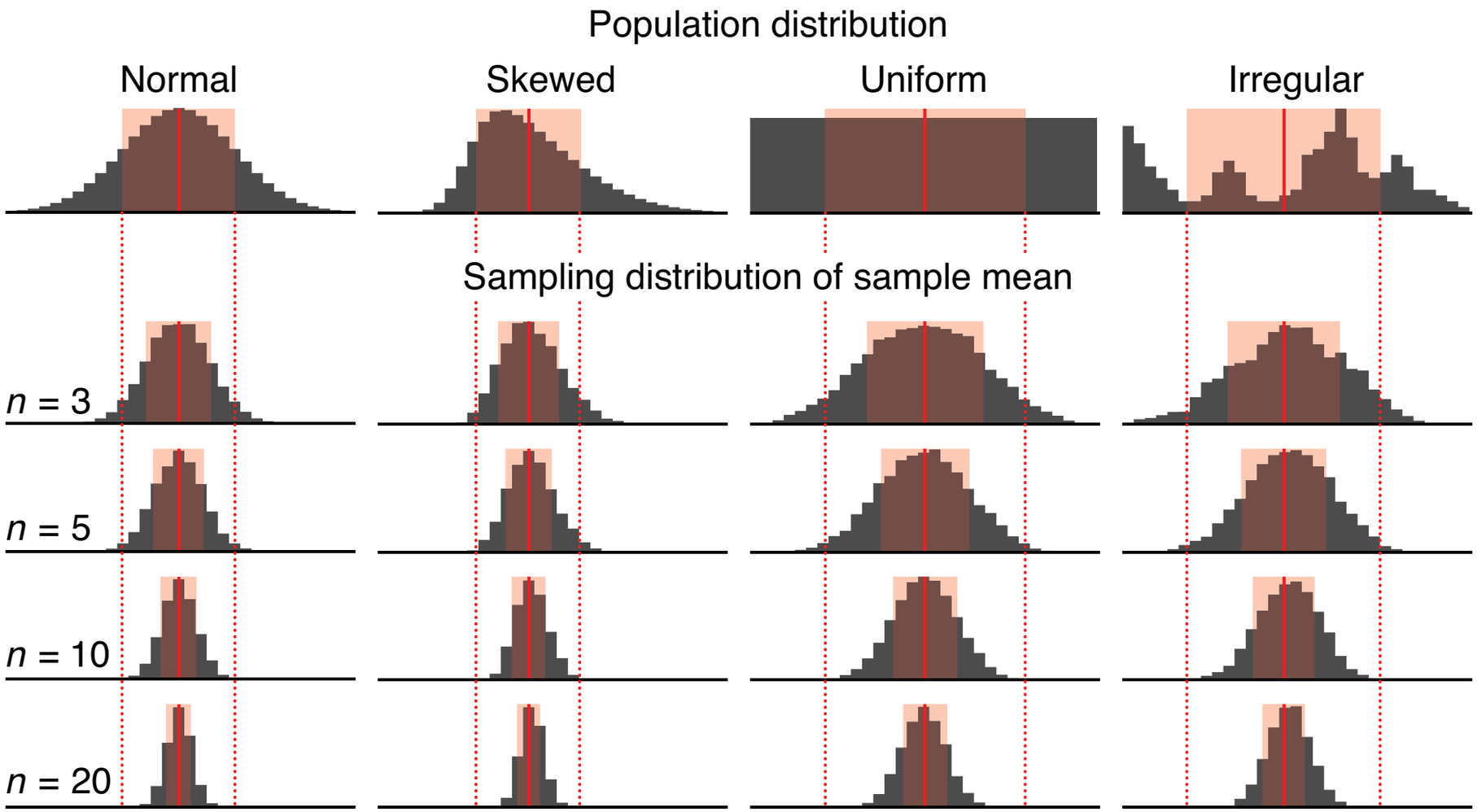
$$\sqrt{\frac{s_1^2}{n_1}}$$

Standard error of the first mean

ASSUMPTION: NORMAL DISTRIBUTION

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

This is the Central Limit Theorem



Probability distribution of the data

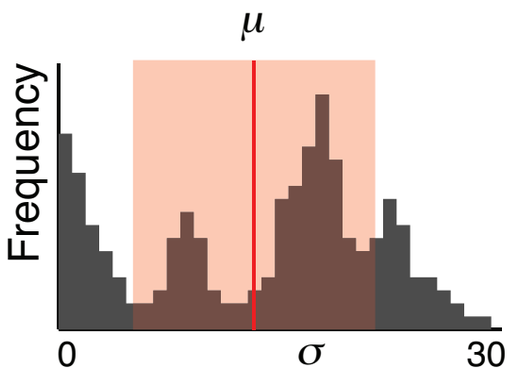
Repeatedly selecting n data points and calculating means

Simulated example

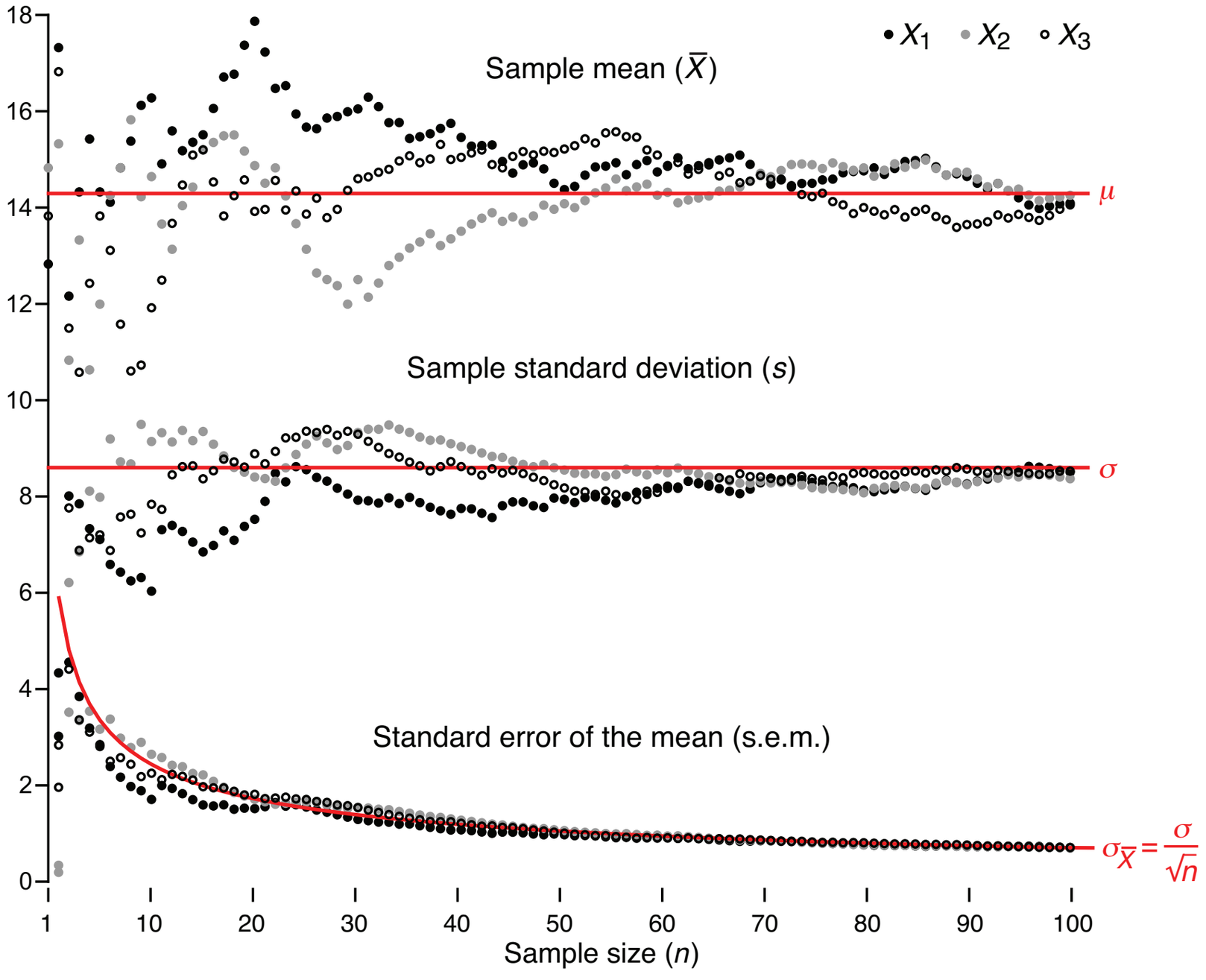
Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

EFFECT OF SAMPLE SIZE

As n increases, the estimates stabilize



Probability distribution of the data

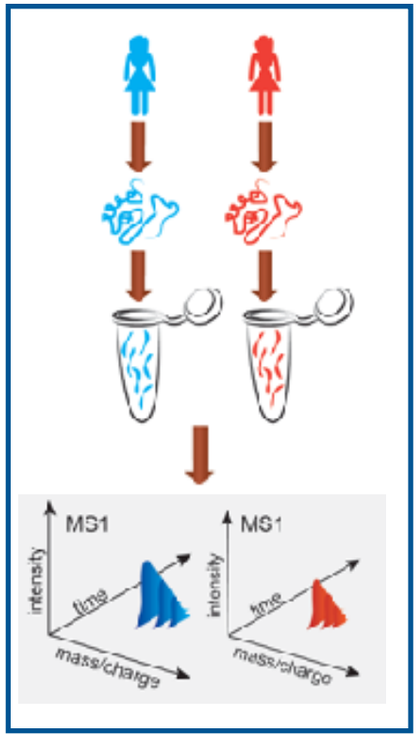


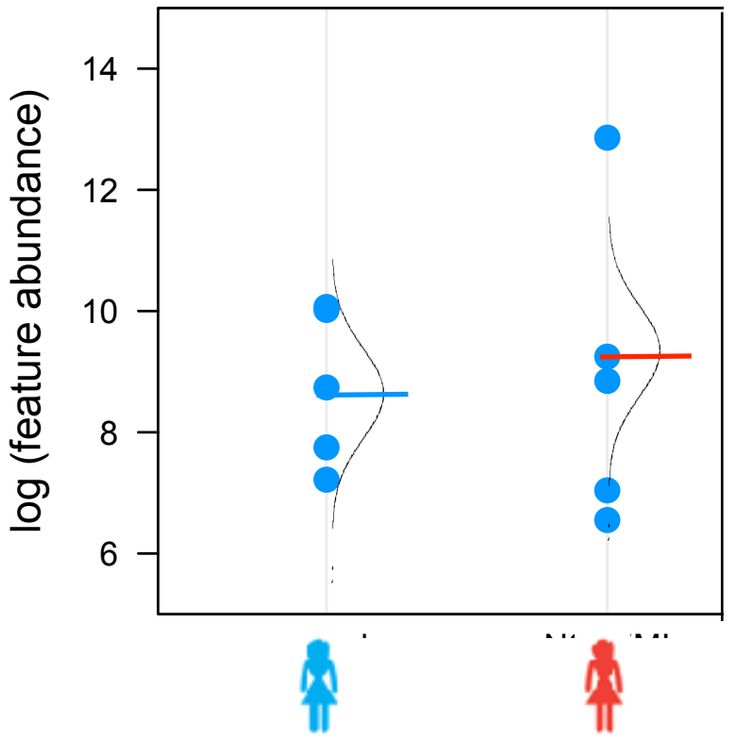
Simulated example

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

FINDING DIFFERENTIALLY ABUNDANT PROTEINS

False positive rate





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

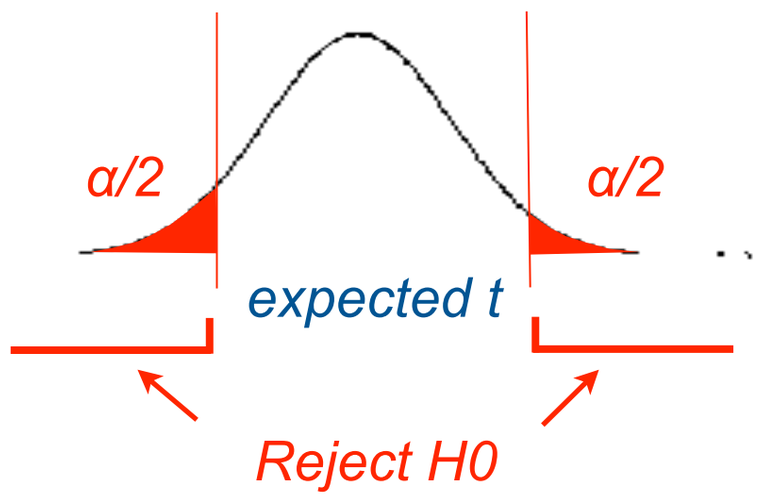
Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

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

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

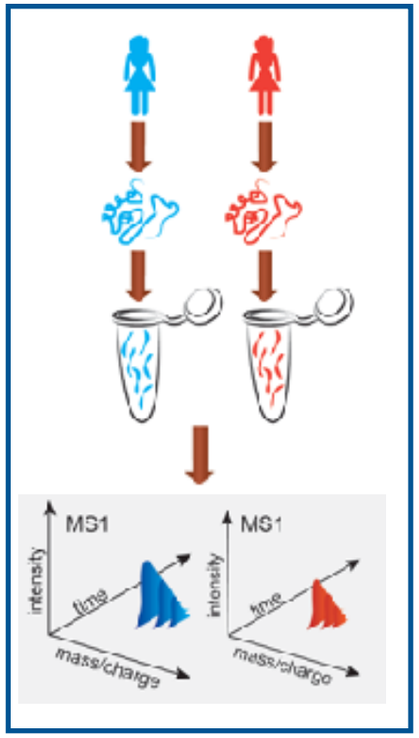
Distribution of the score if H0 is true

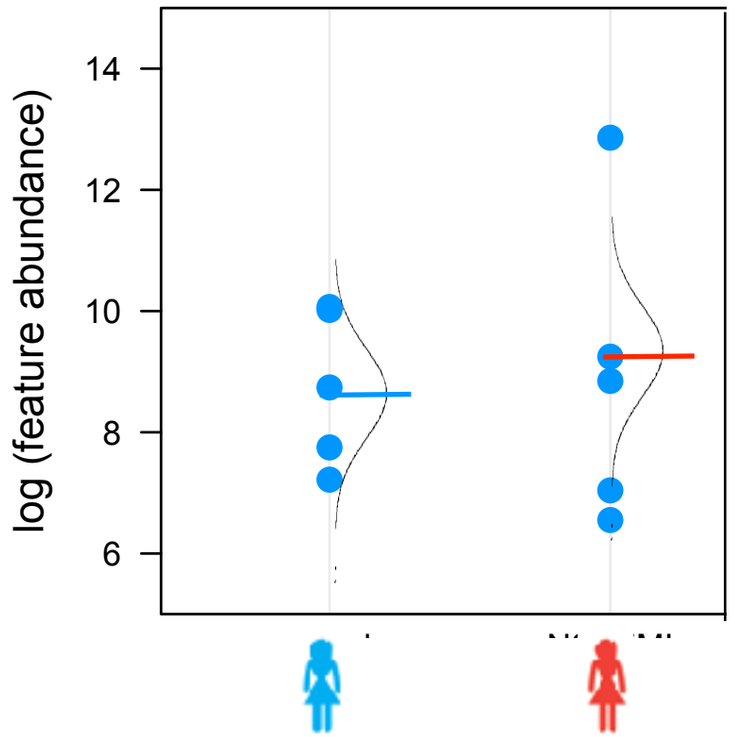
α = False Positive Rate



FINDING DIFFERENTIALLY ABUNDANT PROTEINS

P-value





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

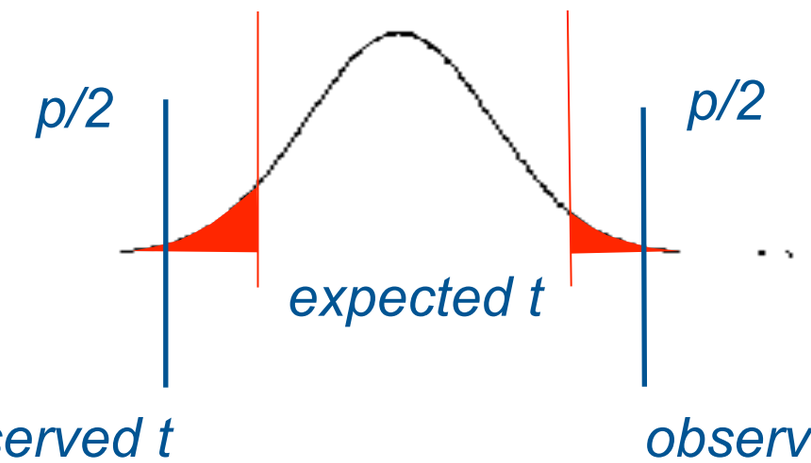
Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

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

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

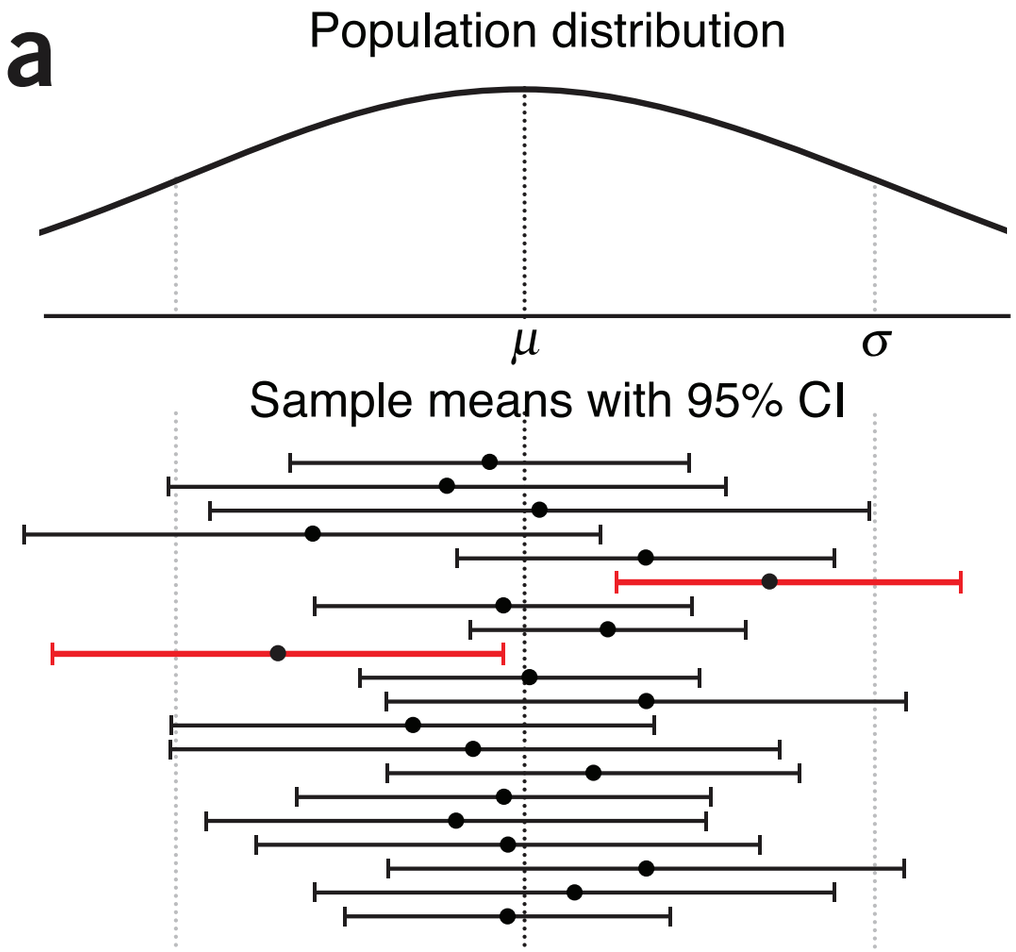
Distribution of the score if H0 is true

p = p-value

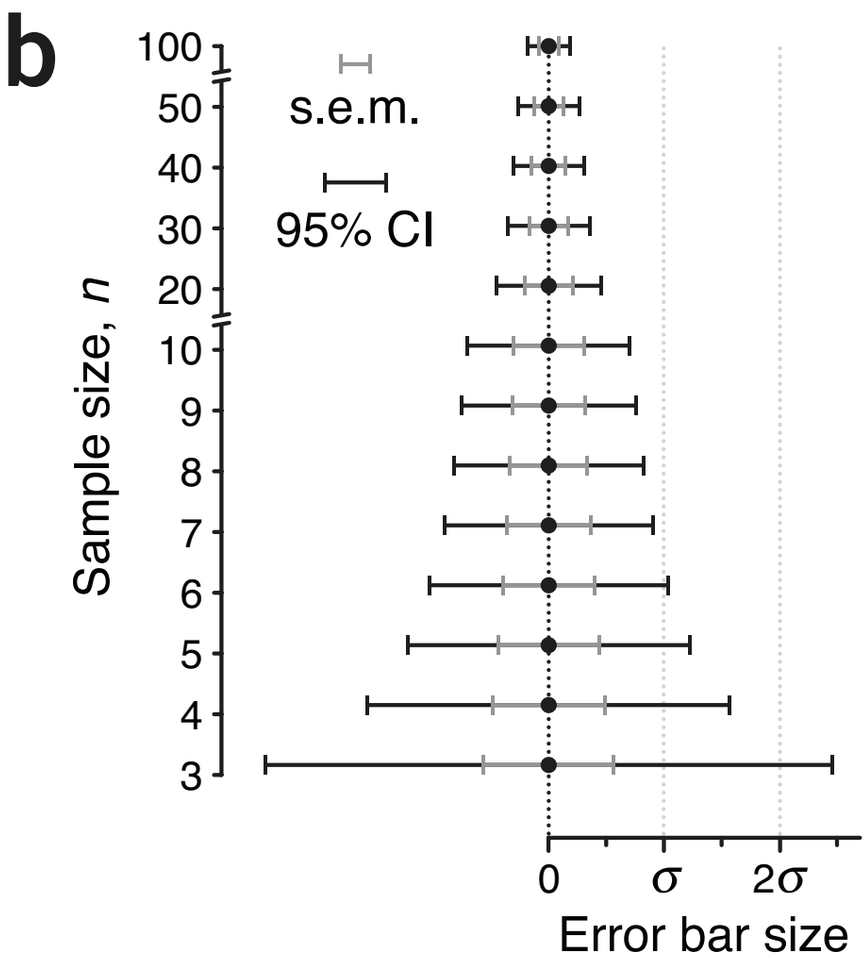


ALTERNATIVE TO TESTING: CONFIDENCE INTERVALS

Not all error bars are made equal



A 95% CI means that 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!

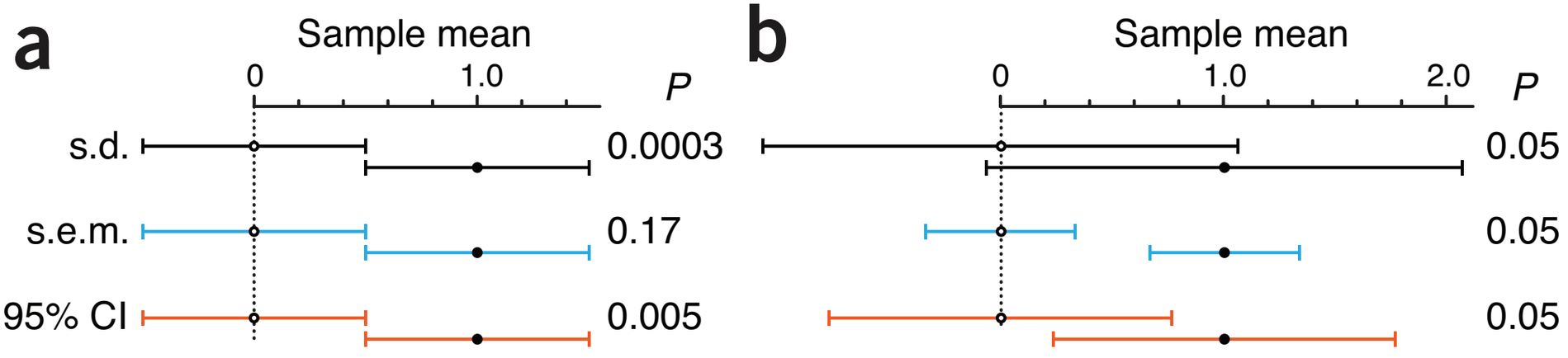
Width of the intervals depends on the sample size

Simulated example

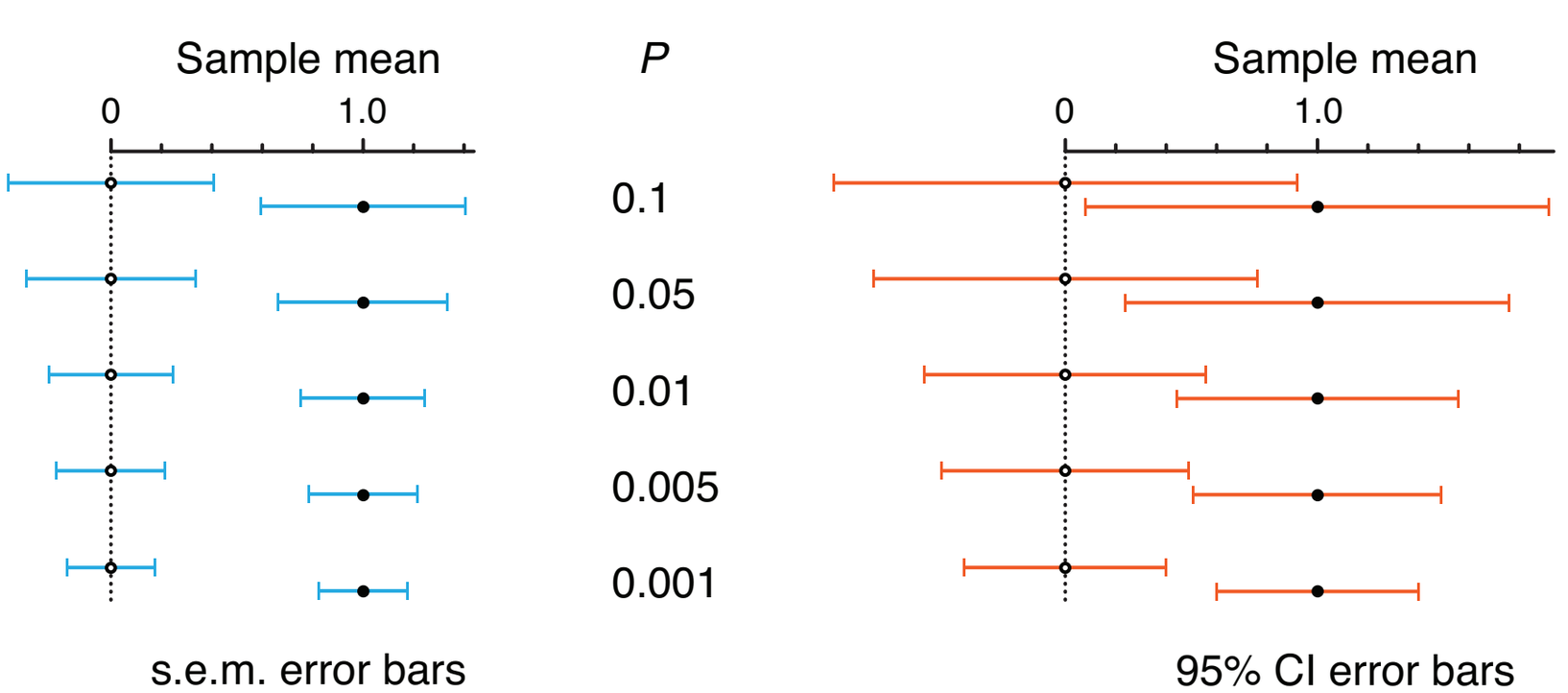
Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

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*

OUTLINE

- Basic statistical inference
  - T-test and p-values
- P-values: a word of caution
  - Instability, multiplicity, alternative approaches
- So how many replicates do I need?
  - Design of complex experiments

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

The American Statistician, February 2016

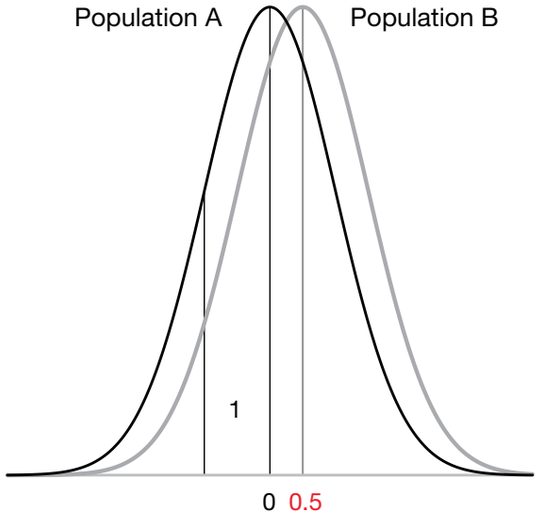
- P-values can indicate how incompatible the data are with a specified statistical model
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance
- Scientific conclusions and business policy decisions should not be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

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

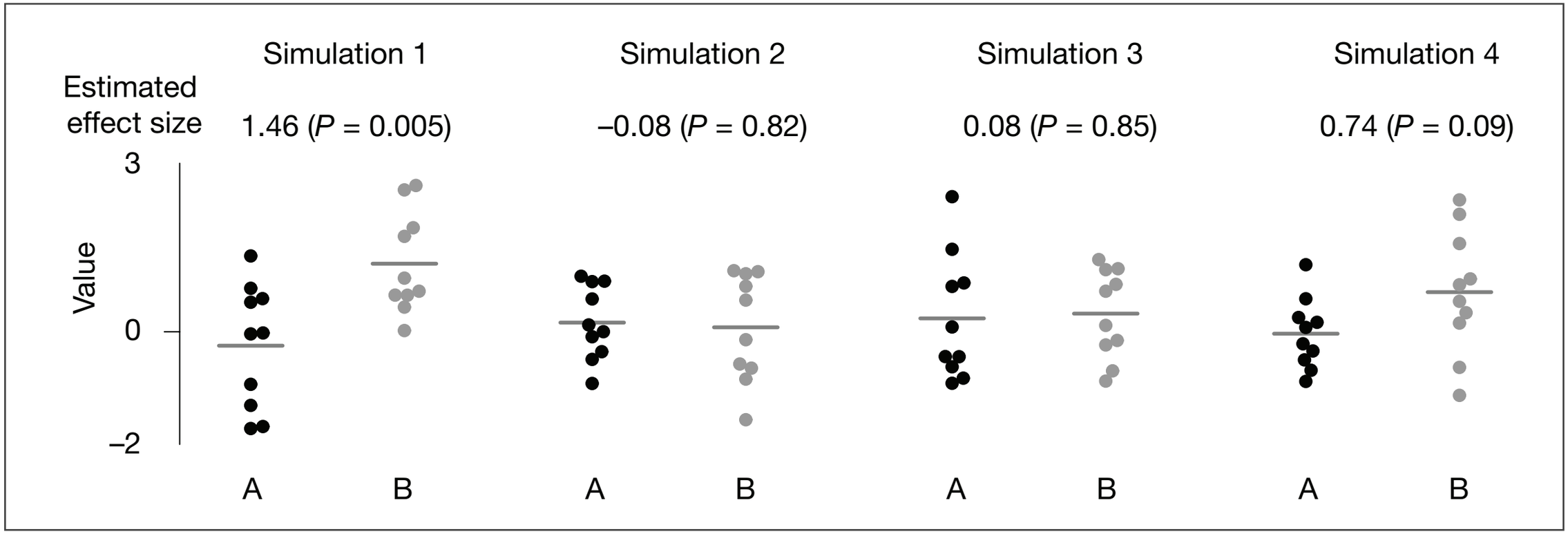
The American Statistician, February 2016

- 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



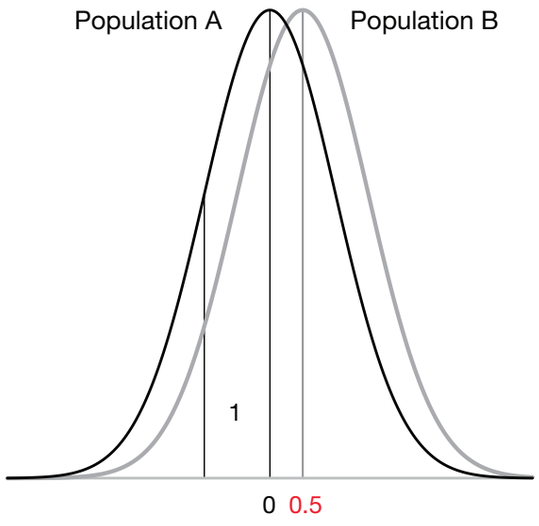
- Repeatedly sampling data leads to different results
- The problem worsens when testing many proteins
- Partial solutions:
  - Larger sample size
  - Adjustment for multiple testing



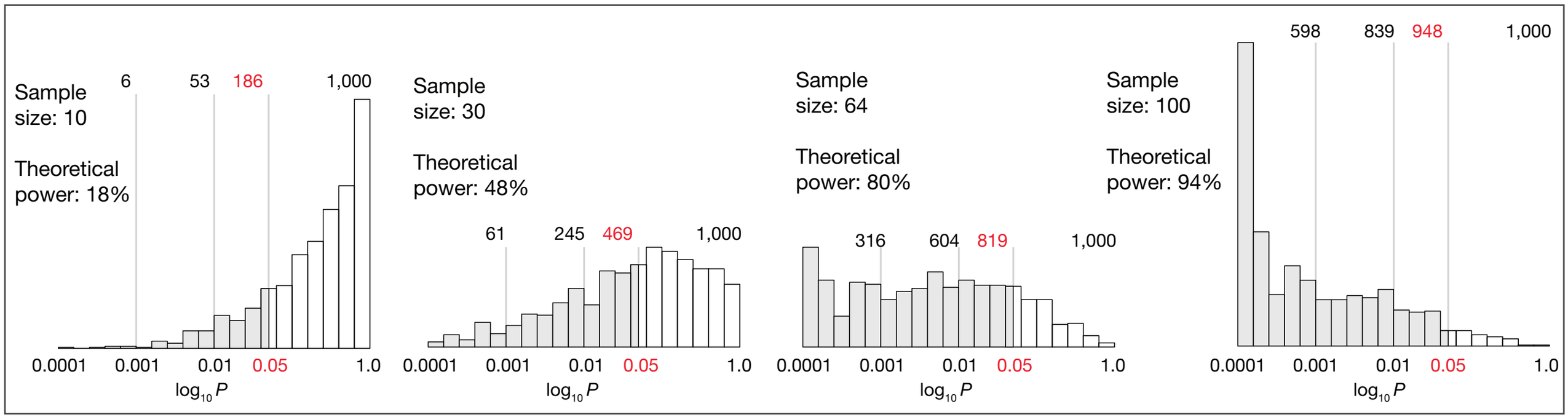
Simulated example

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

WITH SMALL SAMPLE SIZE, P-VALUES ARE UNSTABLE



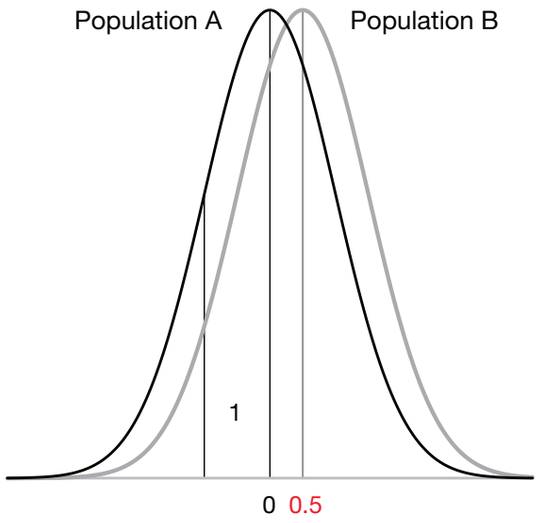
- Repeatedly sampling data leads to different results
- The problem worsens when testing many proteins
- Partial solutions:
  - Larger sample size
  - Adjustment for multiple testing



Simulated example

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

WITH SMALL SAMPLE SIZE, CONCLUSIONS ARE BIASED



Simulated example

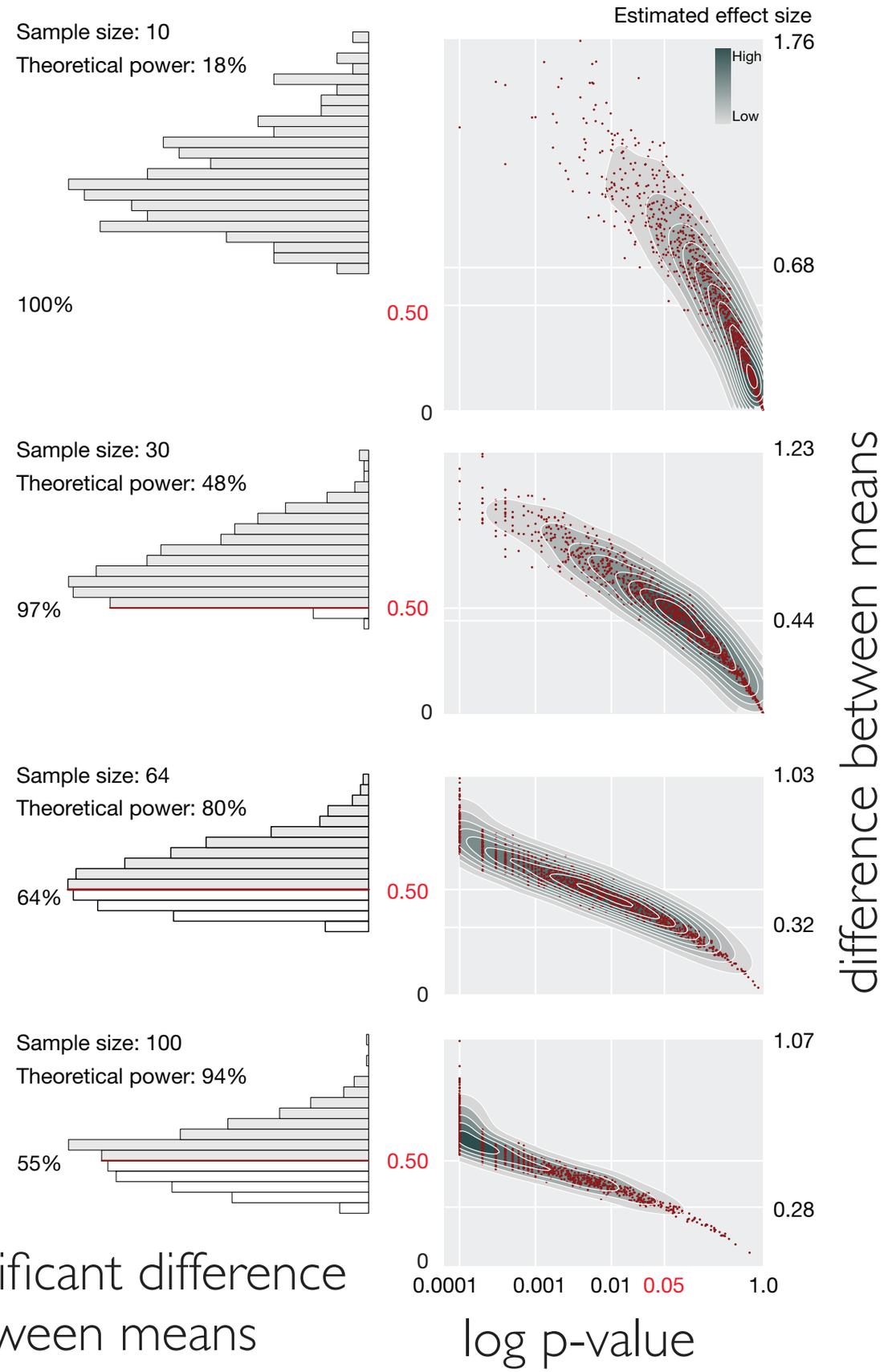
Halsey, Curran-Everett, Volwer and Drummond, *Nature Methods*, 2015

10 replicates

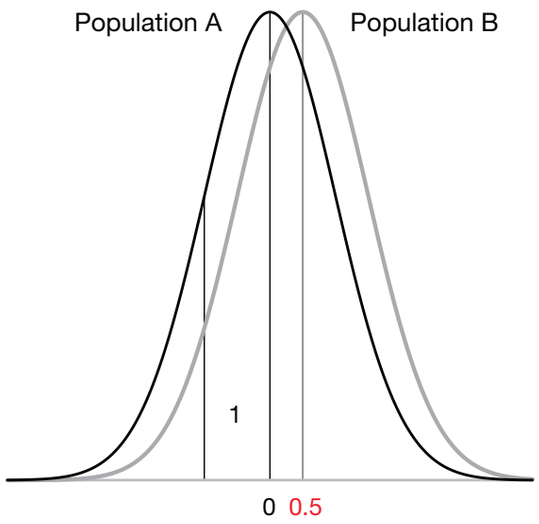
30 replicates

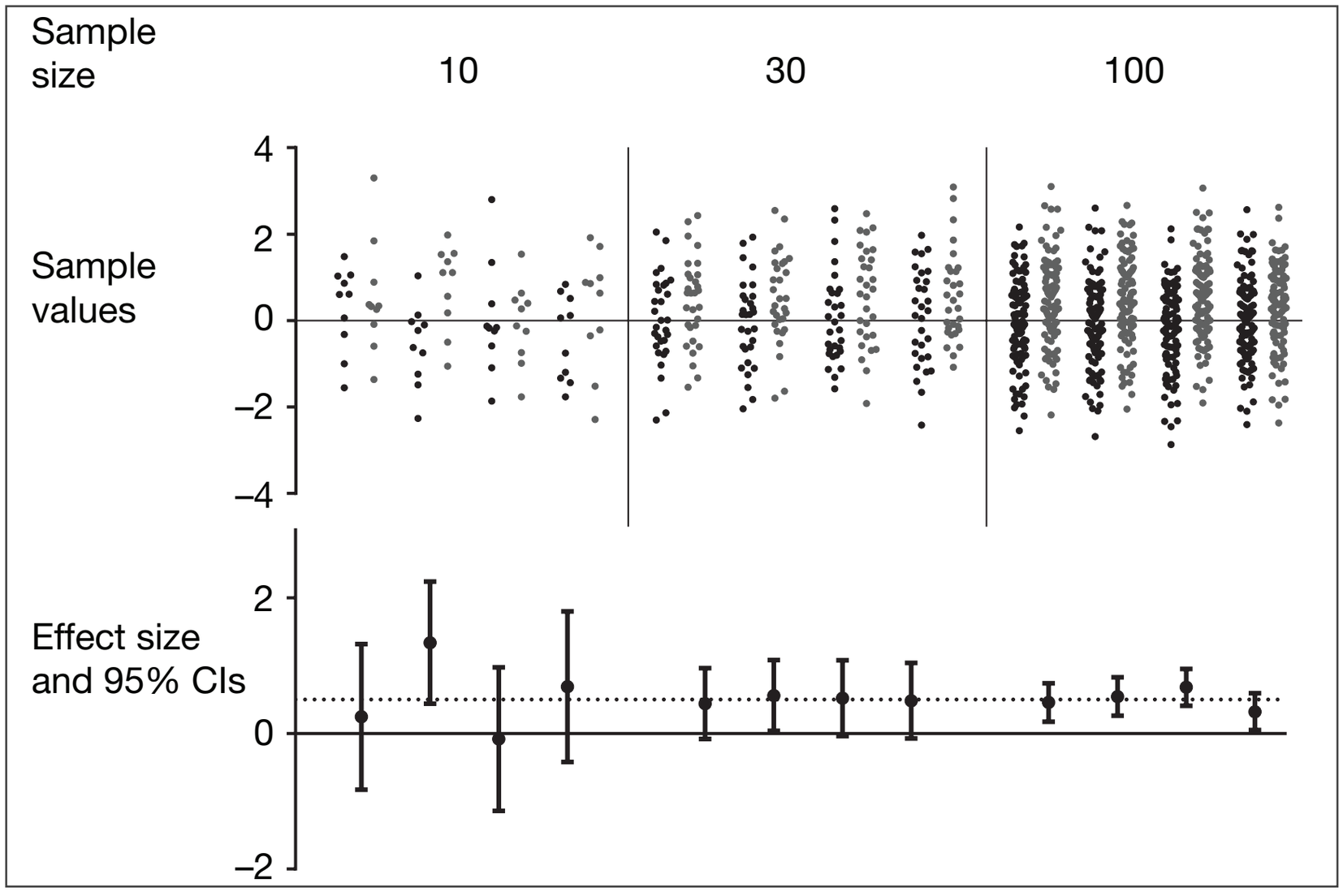
64 replicates

100 replicates



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

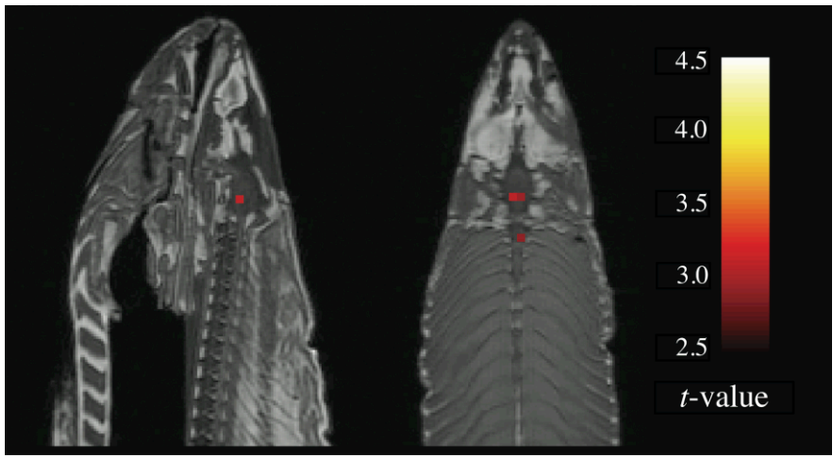
INTRODUCTION

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

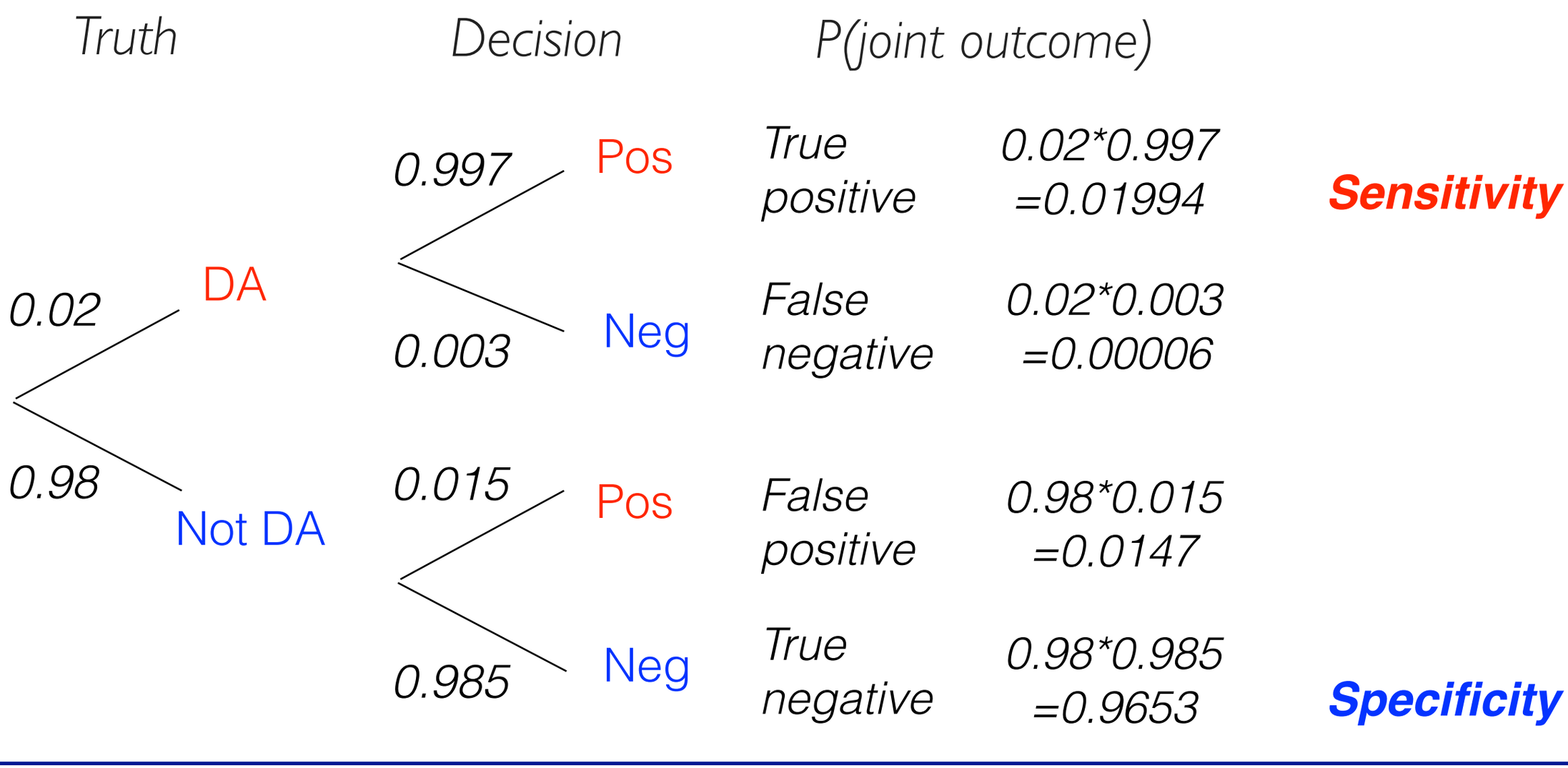
GLM RESULTS



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/

MORE UNCHANGING PROTEINS
=> MORE FALSE DISCOVERIES



$$FDR = P(\text{Not DA} \mid \text{Pos}) = \frac{P(\text{Not DA AND Pos})}{P(\text{Pos})} = \frac{0.0147}{0.01994 + 0.0147} = 0.4243$$

MULTIPLE TESTING

Control False Positive Rate for two proteins

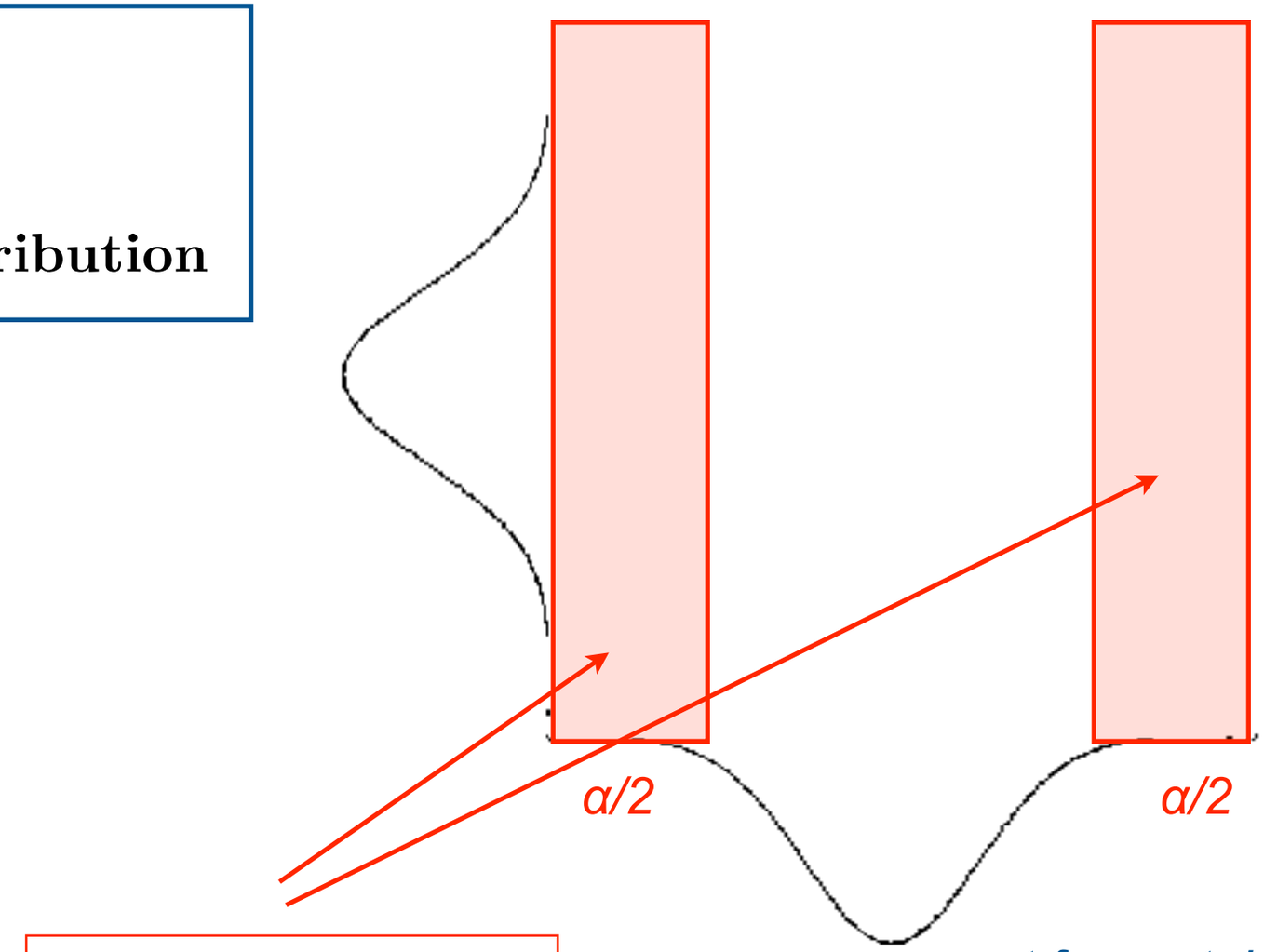
For each protein:

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

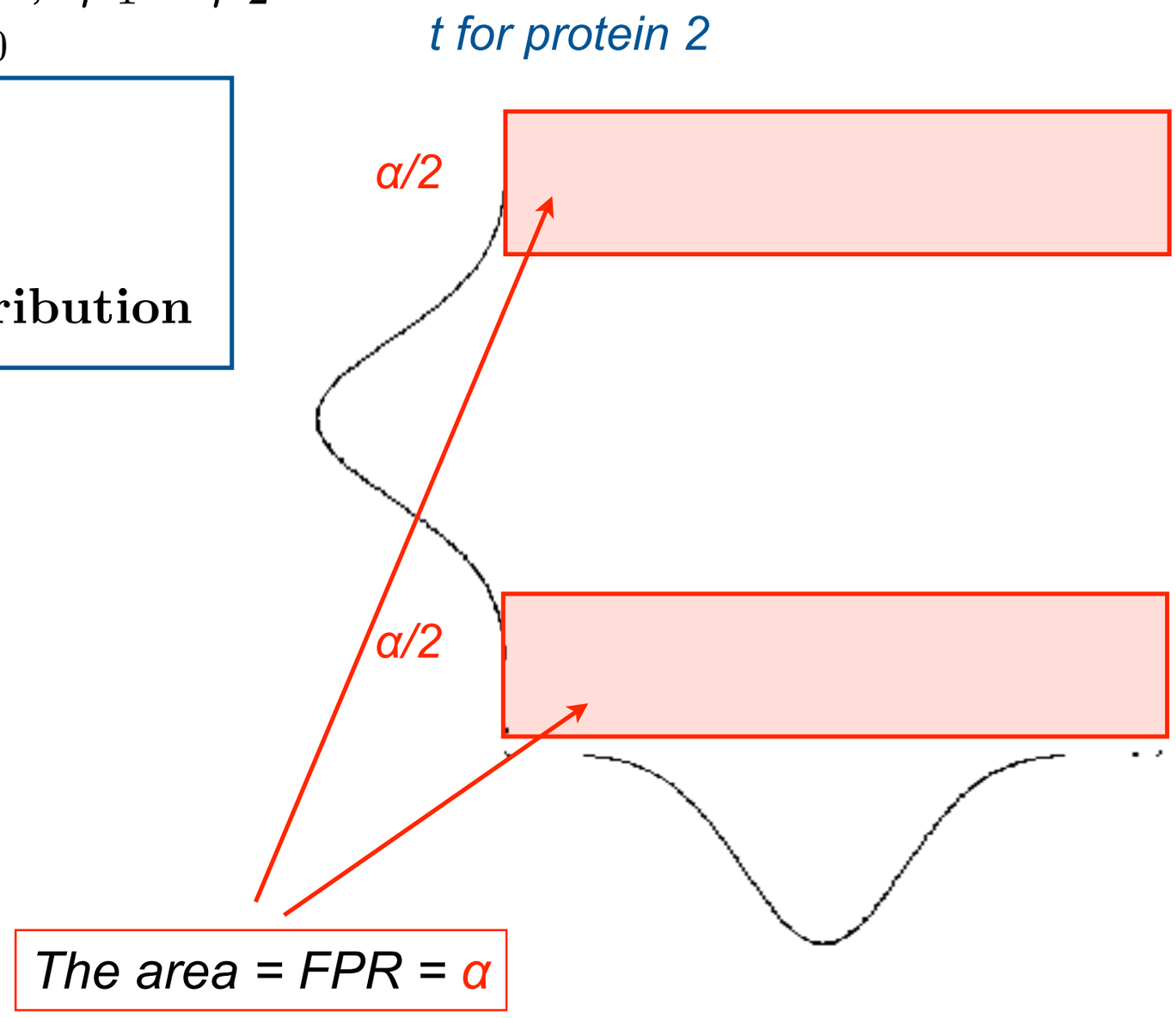
For each protein:

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

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

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

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



MULTIPLE TESTING

Control False Positive Rate for two proteins

For each protein:

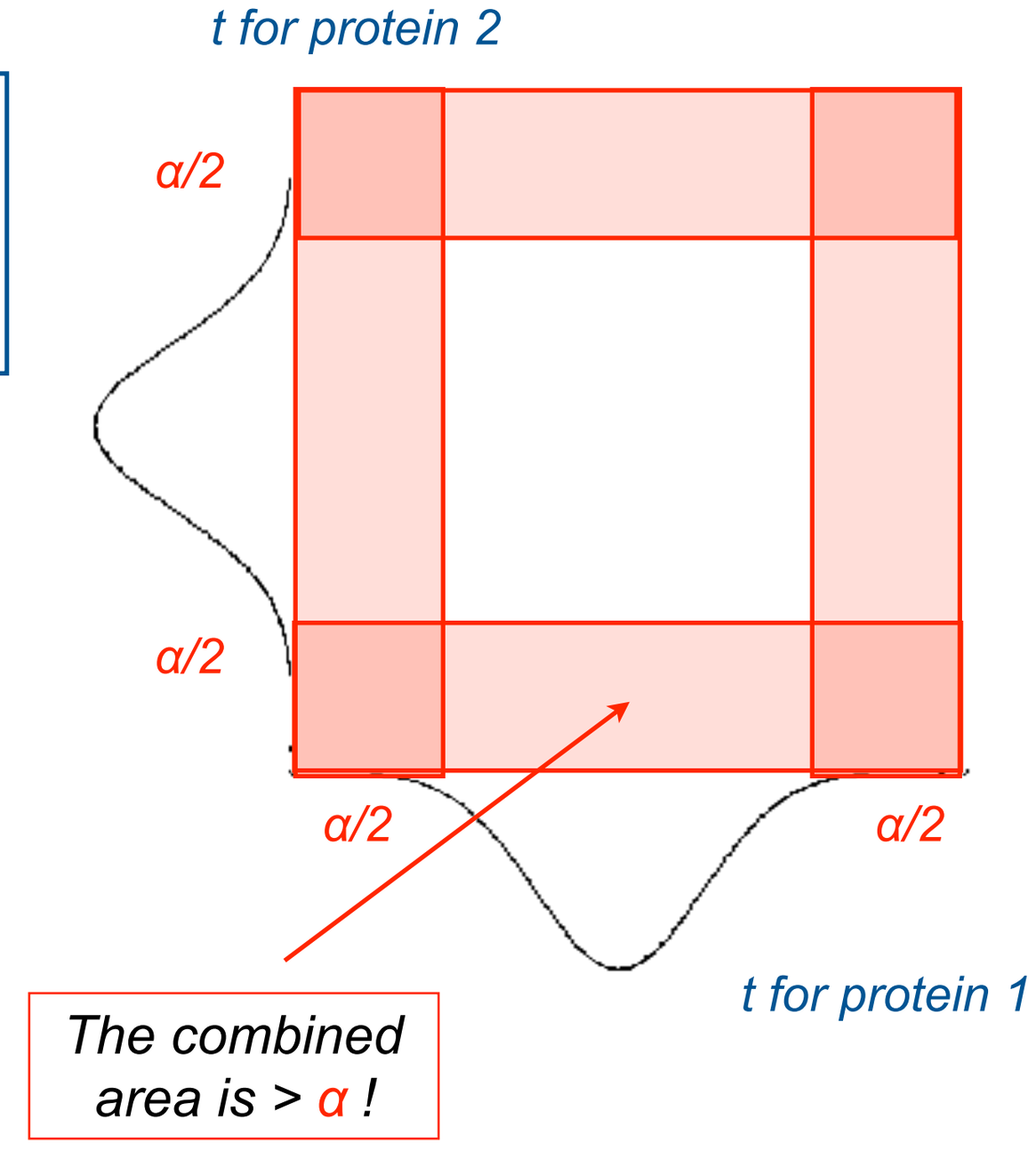
H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

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



MULTIPLE TESTING

Control False Positive Rate for two proteins

For each protein:

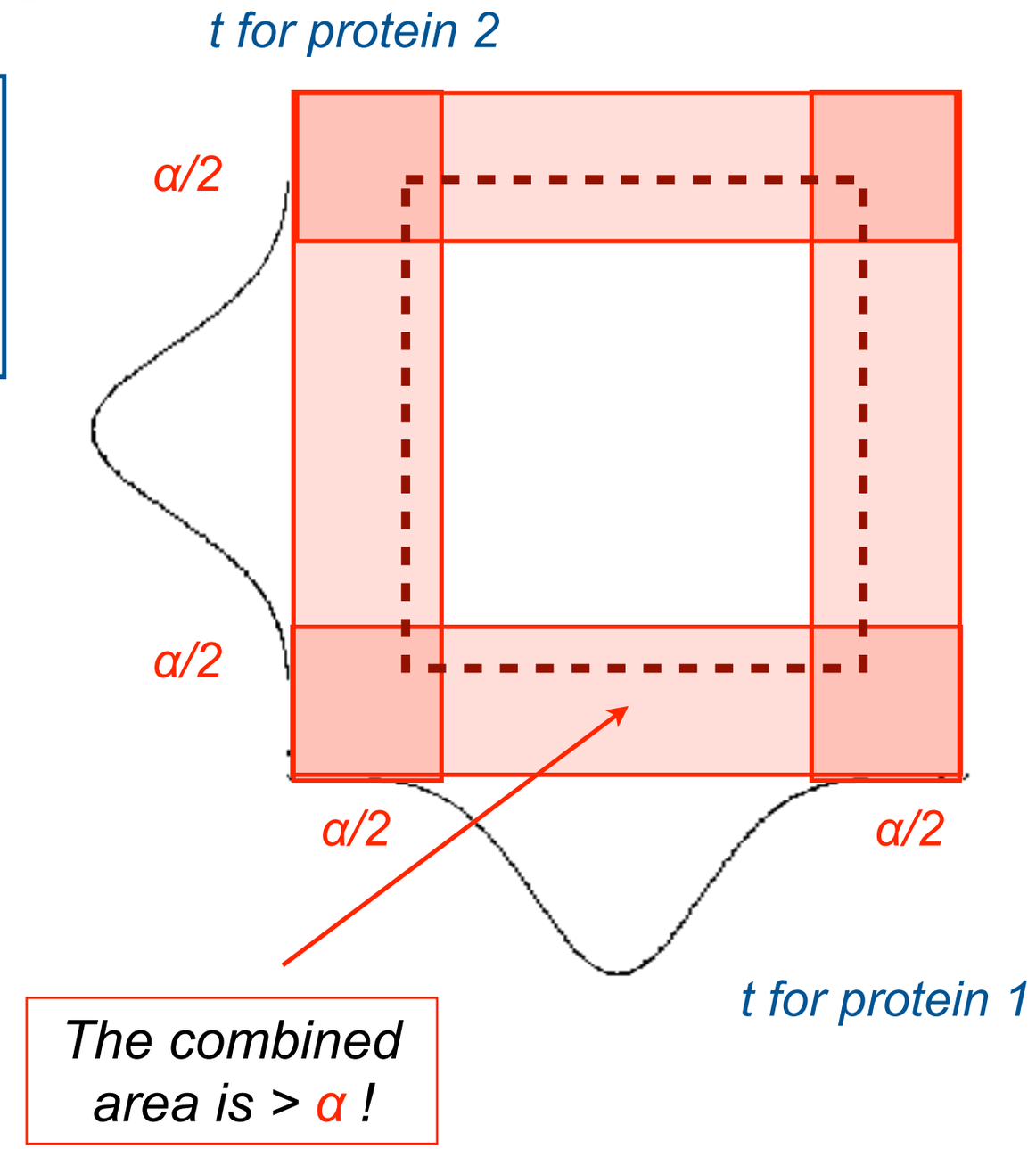
H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

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

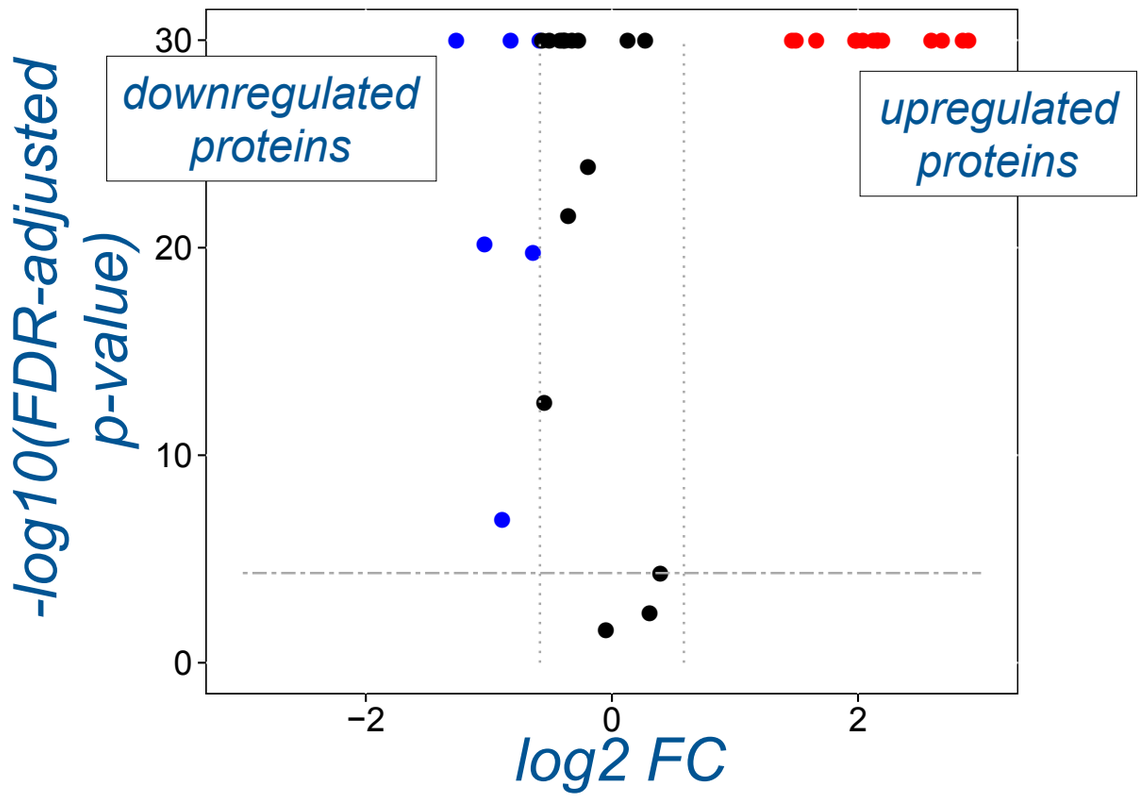
Change criteria from False Positive Rate to False Discovery Rate

	# 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 = $P(V > 0)$



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

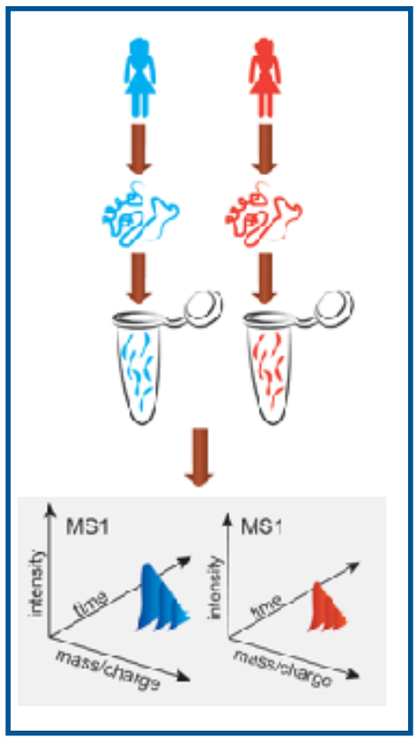
http://simplystatistics.org/2016/02/01/a-menagerie-of-messed-up-data-analyses-and-how-to-avoid-them/

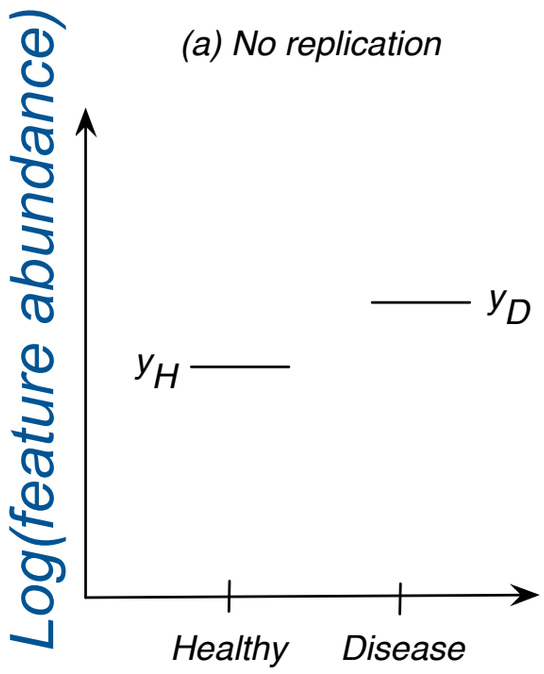
OUTLINE

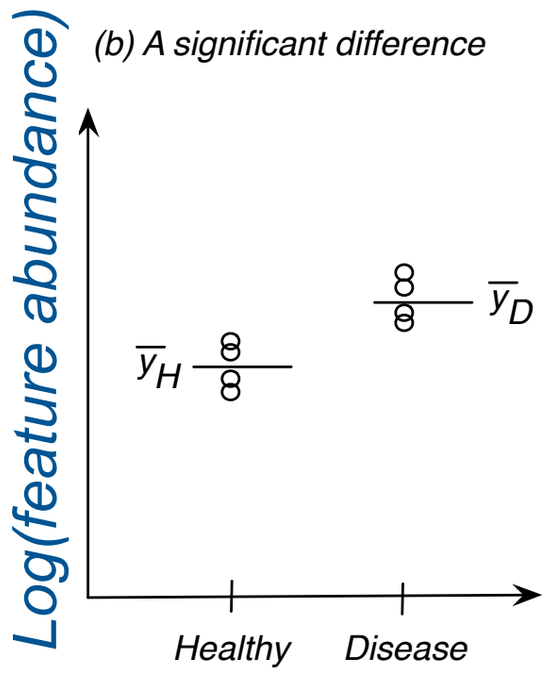
- Basic statistical inference
  - T-test and p-values
- P-values: a word of caution
  - Instability, multiplicity, alternative approaches
- So how many replicates do I need?
  - Design of complex experiments

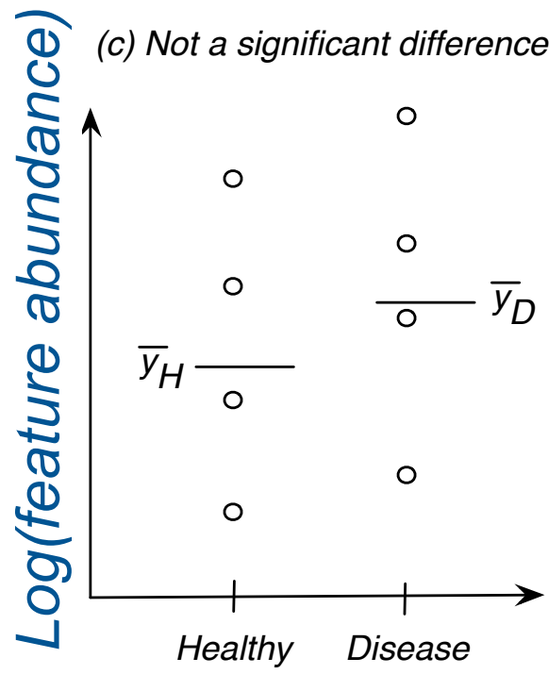
PRINCIPLE 1: REPLICATION

(1) carries out the inference and (2) minimizes inefficiencies









Two levels of randomness imply two types of replication:

- *Biological replicates:* selecting multiple subjects from the population
- *Technical replicates:* multiple runs per subject

Oberg and Vitek, *J. Proteome Research*, 8, 2009

BIOLOGICAL REPLICATION IS MOST IMPORTANT

Observed feature intensity	=	Systematic mean signal of disease group	+	Random deviation due to individual	+	Random deviation due to sample preparation	+	Random deviation due to measurement error
y_{ijkl}	=	Group mean_i	+	$\text{Indiv(Group)}_{j(i)} \sim N\left(0, \sigma^2_{Indiv}\right)$	+	$\text{Prep(Indiv)}_{k(ij)} \sim N\left(0, \sigma^2_{Prep}\right)$	+	$\text{Error}_{l(ijk)} \sim N\left(0, \sigma^2_{Error}\right)$

$$\text{Var}(\bar{y}_H - \bar{y}_D) = 2\left(\frac{\sigma^2_{Indiv}}{I} + \frac{\sigma^2_{Prep}}{IJ} + \frac{\sigma^2_{Error}}{IJK}\right)$$

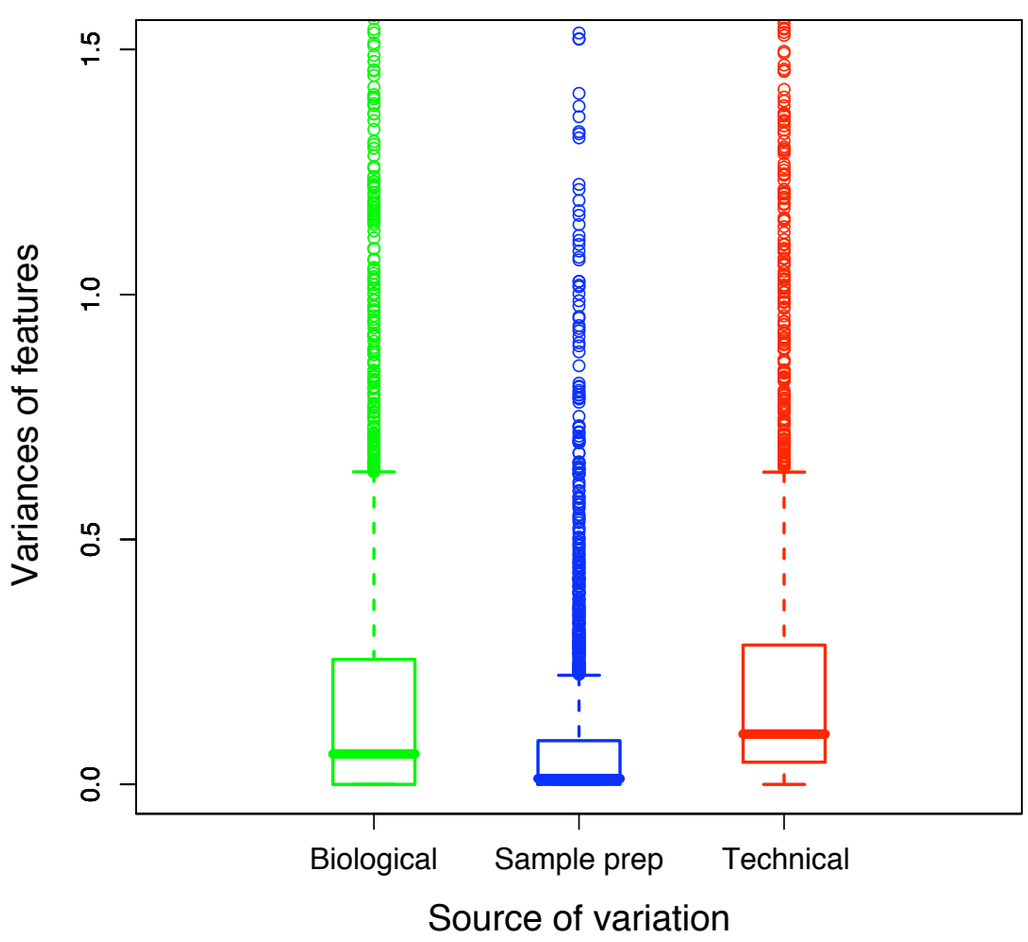
I: # individuals per disease group

J: # sample preps

K: # replicate runs

A pilot experiment

- 2 healthy individuals, 2 with diabetes
- multiple sample preparations
- multiple LC-MS replicates



BIOLOGICAL REPLICATION IS MOST IMPORTANT

Observed feature intensity	=	Systematic mean signal of disease group	+	Random deviation due to individual	+	Random deviation due to sample preparation	+	Random deviation due to measurement error
y_{ijkl}	=	Group mean_i	+	$\text{Indiv(Group)}_{j(i)} \sim N\left(0, \sigma^2_{\text{Indiv}}\right)$	+	$\text{Prep(Indiv)}_{k(ij)} \sim N\left(0, \sigma^2_{\text{Prep}}\right)$	+	$\text{Error}_{l(ijk)} \sim N\left(0, \sigma^2_{\text{Error}}\right)$

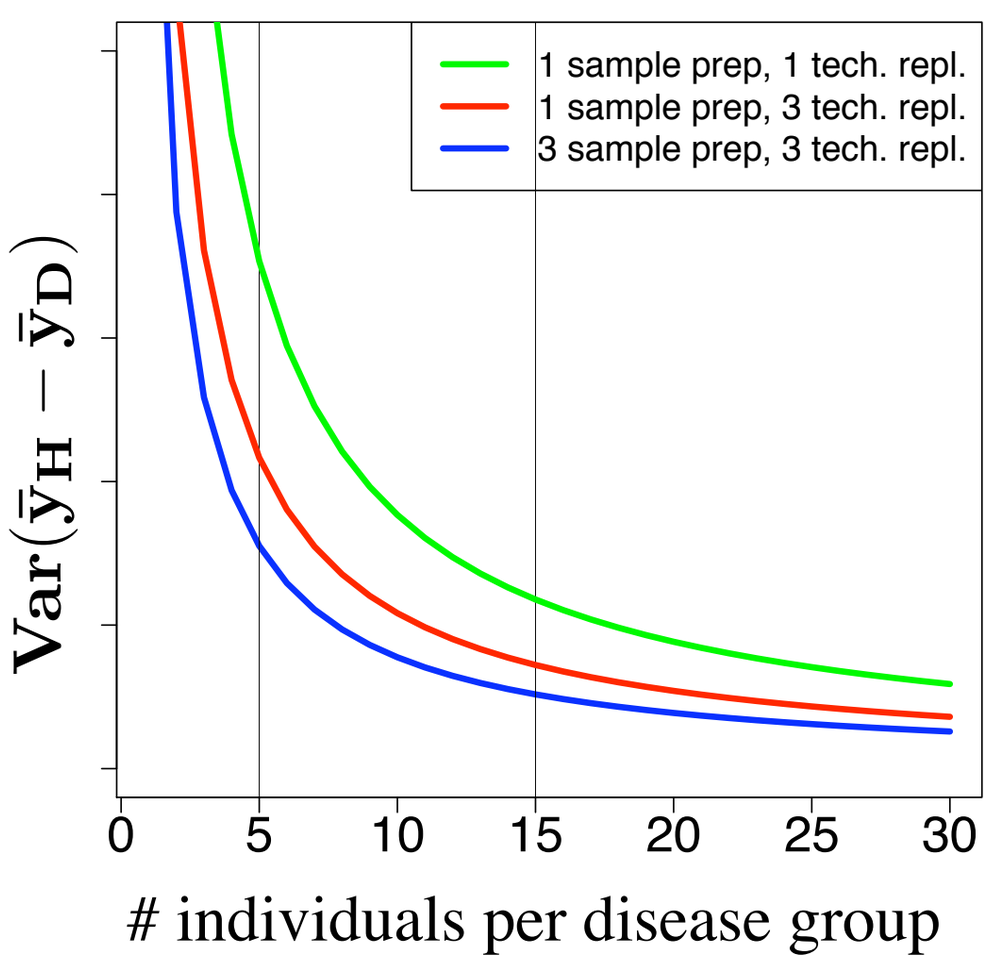
$$\text{Var}(\bar{y}_H - \bar{y}_D) = 2\left(\frac{\sigma^2_{\text{Indiv}}}{I} + \frac{\sigma^2_{\text{Prep}}}{IJ} + \frac{\sigma^2_{\text{Error}}}{IJK}\right)$$

I: # individuals per disease group

J: # sample preps

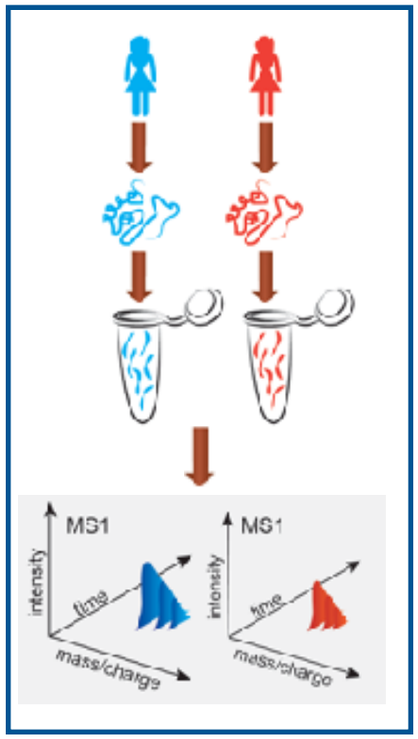
K: # replicate runs

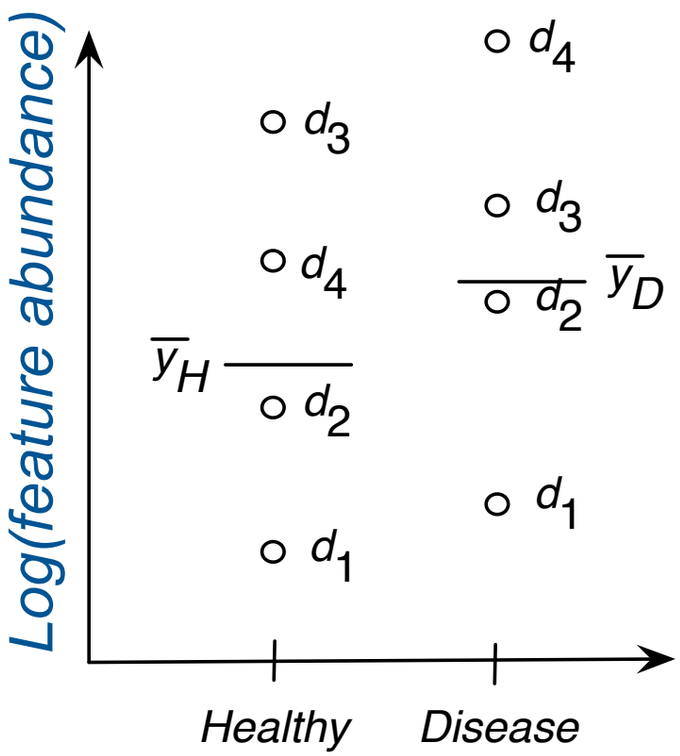
Conclusion 1:
Maximize the number of biological replicates

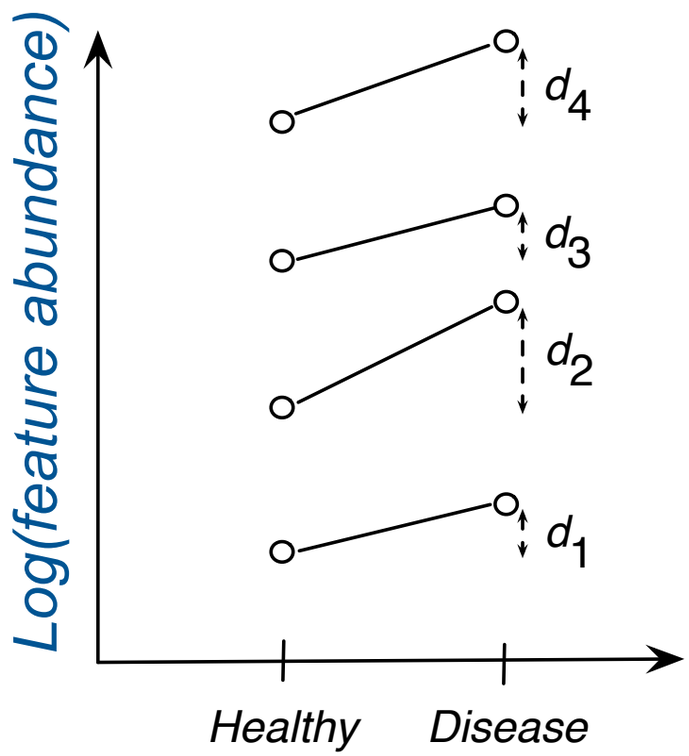


PRINCIPLE 3: BLOCKING

Helps reduce both bias and inefficiency







Complete randomization
= inflated variance

Block-randomization
= restriction on randomization
= systematic allocation

Two levels of randomness imply two types of blocks:

- *Biological replicates:* subjects having similar characteristics (e.g. age)
- *Technical replicates:* samples processed together (e.g. in a same day)

Oberg and Vitek, *J. Proteome Research*, 8, 2009

BLOCKING IS HELPFUL

Especially when between-block variance is large

Observed feature intensity	=	Systematic mean signal of disease group	+	Random deviation due to block (e.g. plate or day)	+	Random deviation due to individual	+	Random deviation due to measurement error
y_{ijkl}	=	Group mean_i	+	$\text{Block}_k \sim N(0, \sigma^2_{Block})$	+	$\text{Indiv(Group)}_{j(i)} \sim N(0, \sigma^2_{Indiv})$	+	$\text{Error}_{l(ijk)} \sim N(0, \sigma^2_{Error})$

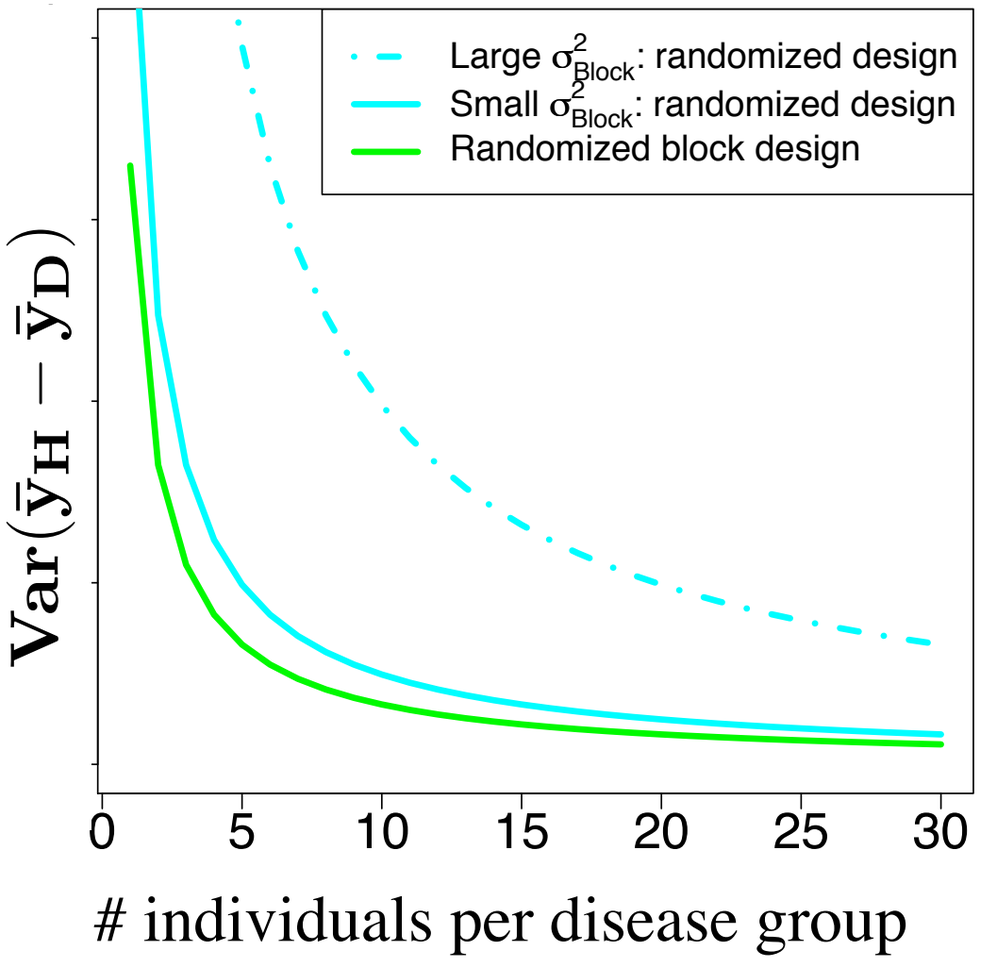
A completely randomized design

I: # individuals per disease group

$$\text{Var}(\bar{y}_H - \bar{y}_D) = 2\left(\frac{\sigma^2_{Block} + \sigma^2_{Indiv} + \sigma^2_{Error}}{I}\right)$$

A block-randomized design

$$\text{Var}(\bar{y}_H - \bar{y}_D) = 2\left(\frac{\sigma^2_{Indiv} + \sigma^2_{Error}}{I}\right)$$



Conclusion 2: Block-randomize
- if can not control a large source of variation
- if moderate sample size

SO HOW MANY REPLICATES DO I NEED?

If we only have one feature:

Fix: α - probability of a false positive discovery

β - probability of a true positive discovery

$\boldsymbol{\Delta}$ - anticipated fold change

$\sigma^2_{\mathbf{Indiv}}$ and $\sigma^2_{\mathbf{Error}}$ - anticipated variability

Write:

$$\mathbf{Var}(\bar{\mathbf{y}}_{\mathbf{H}} - \bar{\mathbf{y}}_{\mathbf{D}}) \leq \left(\frac{\boldsymbol{\Delta}}{\mathbf{z}_{\mathbf{1}-\beta} + \mathbf{z}_{\mathbf{1}-\alpha/\mathbf{2}}} \right)^{\mathbf{2}}$$

where $\mathbf{z}_{\mathbf{1}-\beta}$ and $\mathbf{z}_{\mathbf{1}-\alpha/\mathbf{2}}$ are quantiles of the Normal distribution

A completely randomized design

I: # individuals per disease group

$$\mathbf{Var}(\bar{\mathbf{y}}_{\mathbf{H}} - \bar{\mathbf{y}}_{\mathbf{D}}) = \mathbf{2}\left(\frac{\sigma^2_{\mathbf{Block}} + \sigma^2_{\mathbf{Indiv}} + \sigma^2_{\mathbf{Error}}}{\mathbf{I}} \right)$$

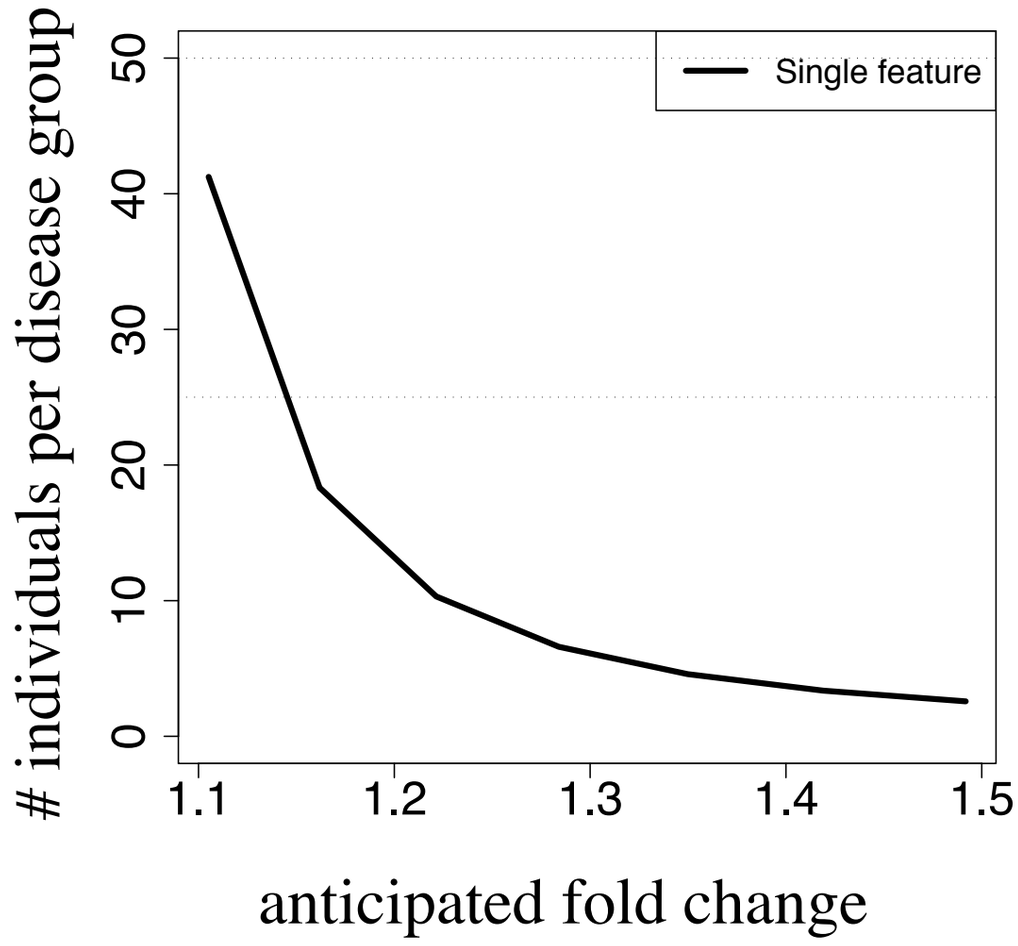
A block-randomized design

$$\mathbf{Var}(\bar{\mathbf{y}}_{\mathbf{H}} - \bar{\mathbf{y}}_{\mathbf{D}}) = \mathbf{2}\left(\frac{\sigma^2_{\mathbf{Indiv}} + \sigma^2_{\mathbf{Error}}}{\mathbf{I}} \right)$$

→ solve for the number of individuals I

SO HOW MANY REPLICATES DO I NEED?

Example: pilot study with diabetes patients.
A block-randomized design
If we only had one feature:



Conclusion 3:
The smaller the anticipated difference, the larger the sample size

DIFFICULTY: MANY FEATURES ARE OF INTEREST

Would like to control the False Discovery Rate:

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

$$\mathbf{q} = \mathbf{E}\left[\frac{\mathbf{V}}{\mathbf{max(R, 1)}}\right]$$ = the "average" proportion of false positives

This changes the sample size calculation:

Fix: $\mathbf{q}$ - the False Discovery Rate

$\mathbf{m_0/m_1}$ - anticipated ratio of unchanging features

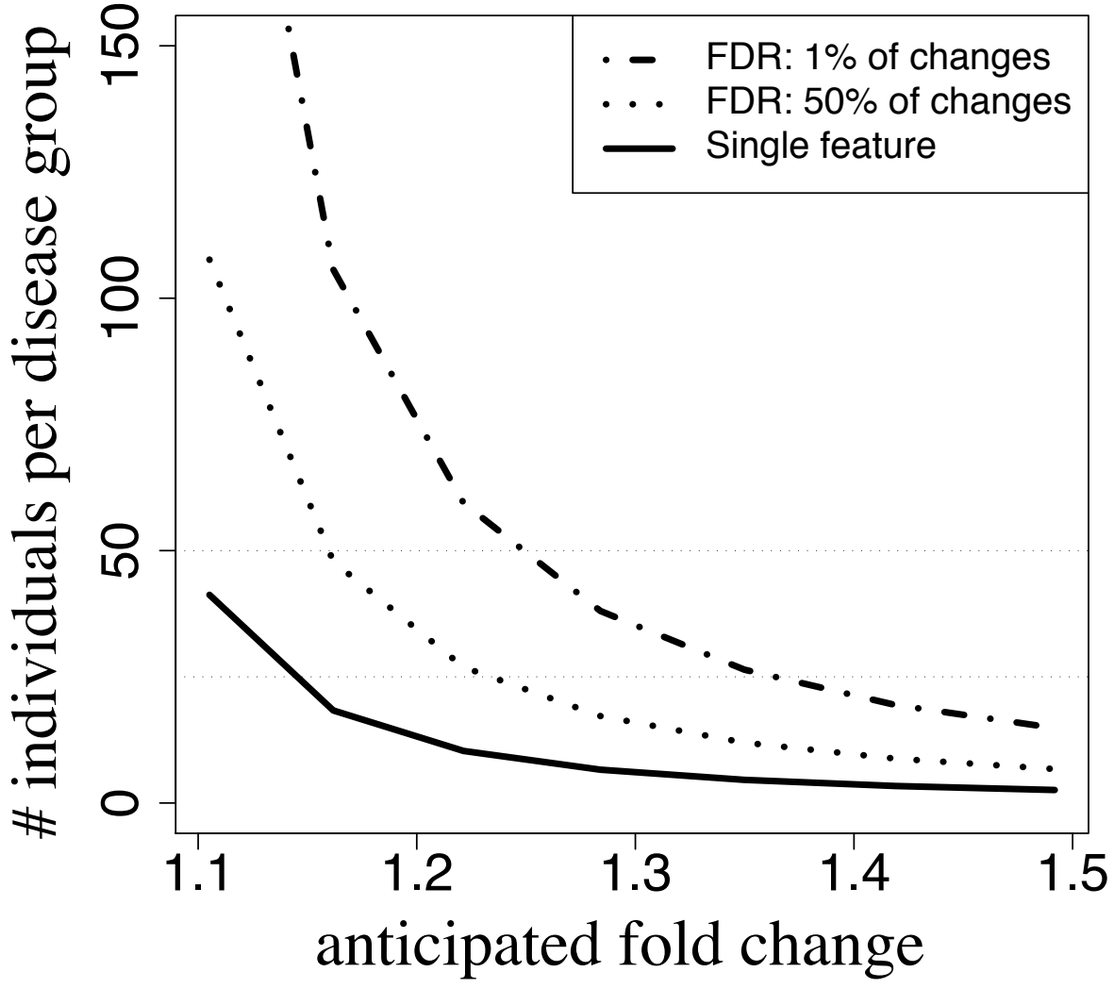
This defines

$$\alpha_{\text{ave}} \leq (\mathbf{1} - \beta)_{\text{ave}} \cdot \mathbf{q}\frac{\mathbf{1}}{\mathbf{1} + (\mathbf{1} - \mathbf{q}) \cdot \mathbf{m_0/m_1}},$$

- average probability of a false positive discovery

SO HOW MANY REPLICATES DO I NEED?

Example: pilot study with diabetes patients.
A block-randomized design



Conclusion 4:
The fewer changes we expect, the larger the sample size

Oberg and Vitek, *JPR*, 2009