STATISTICAL INFERENCE

Meena Choi, Ting Huang, Olga Vitek

College of Science

College of Computer and Information Science

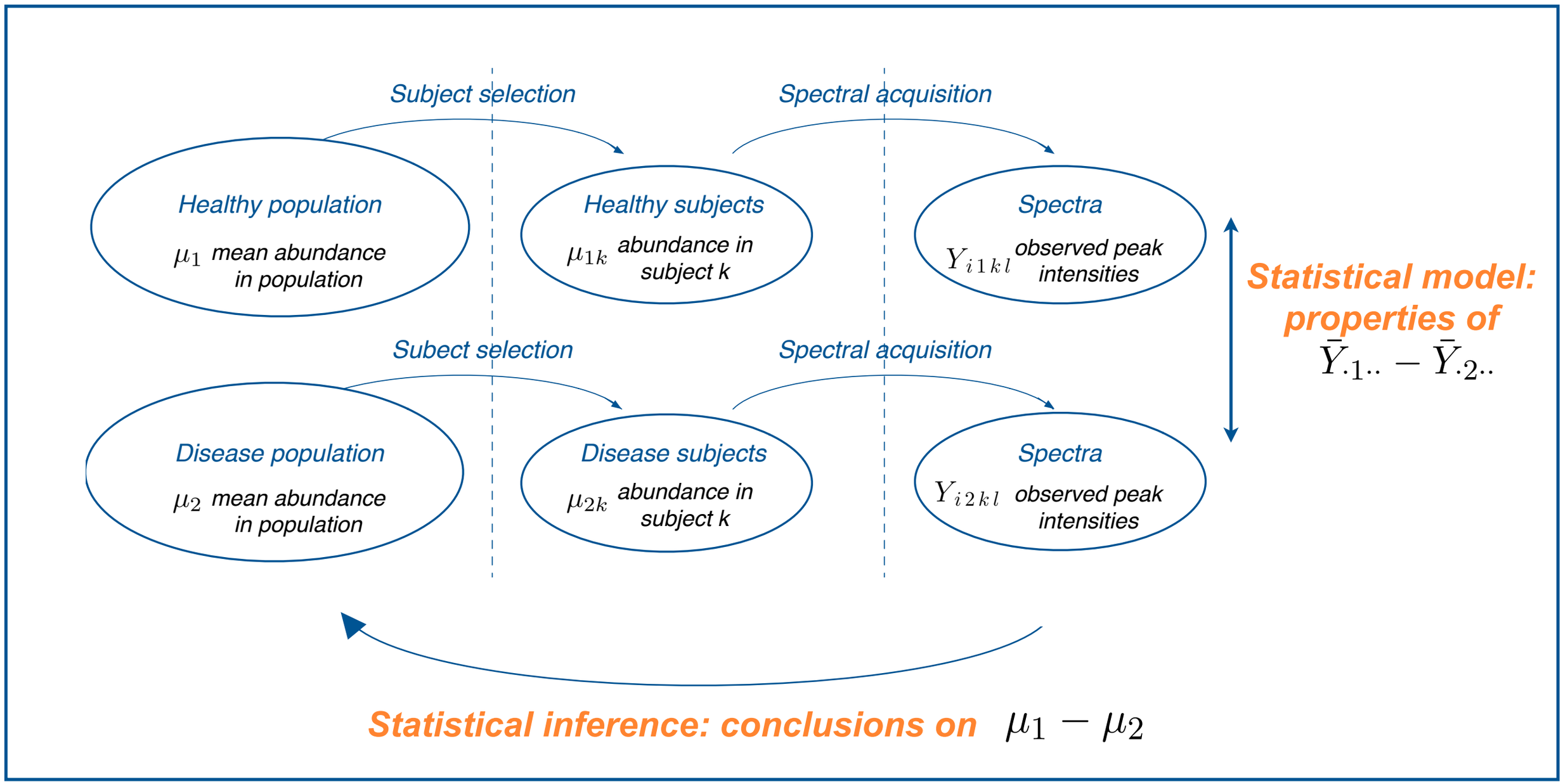
Northeastern University

OUTLINE

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

COMPARE DESIGNS

In terms of bias and (in)-efficiency

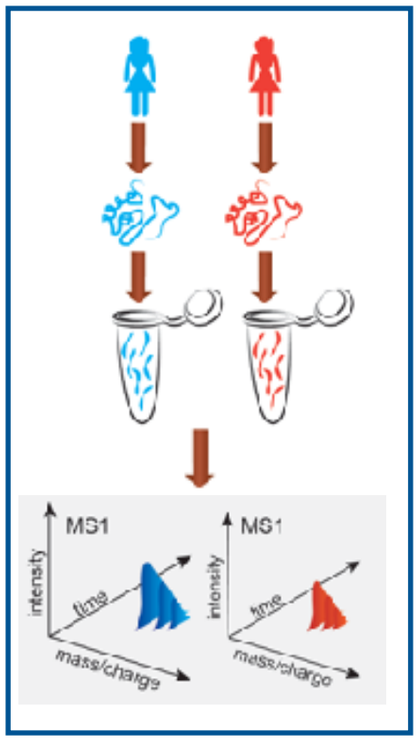


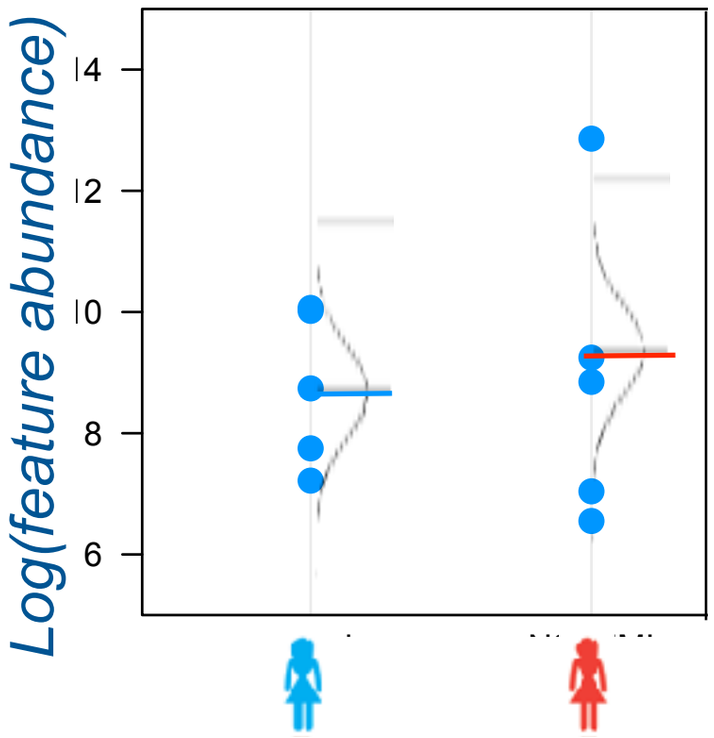
Bias: $\bar{Y}_{.1..} - \bar{Y}_{.2..}$ *systematically different from* $\mu_{1k} - \mu_{2k}$

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

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

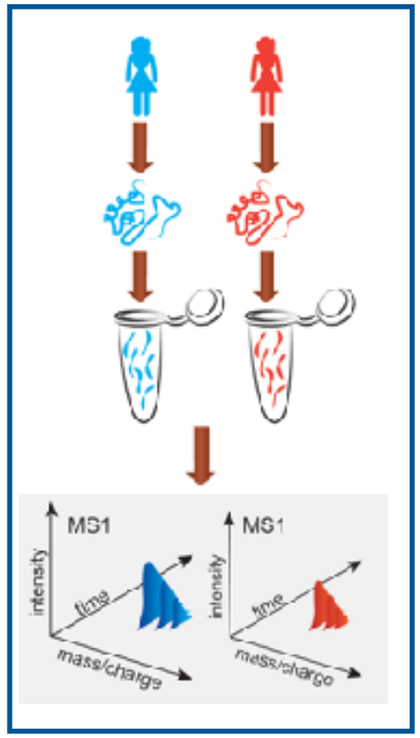
$$log_2\left(\prod_j X_{1j}\right)^{\frac{1}{n_1}} - log_2\left(\prod_j X_{2j}\right)^{\frac{1}{n_2}} = log_2 \frac{\left(\prod_j X_{1j}\right)^{\frac{1}{n_1}}}{\left(\prod_j X_{2j}\right)^{\frac{1}{n_2}}}$$

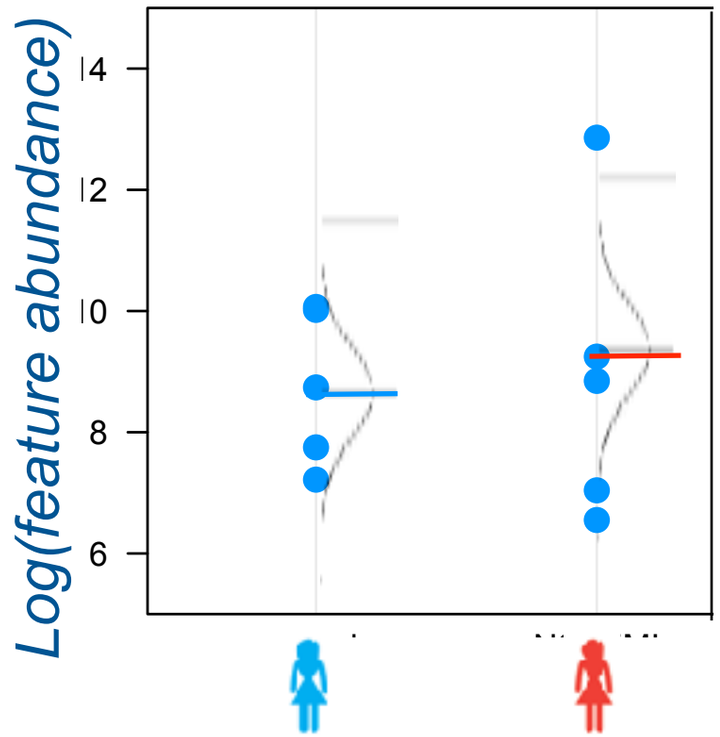
Conclusion:

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, $\mu_1 - \mu_2 \neq 0$

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

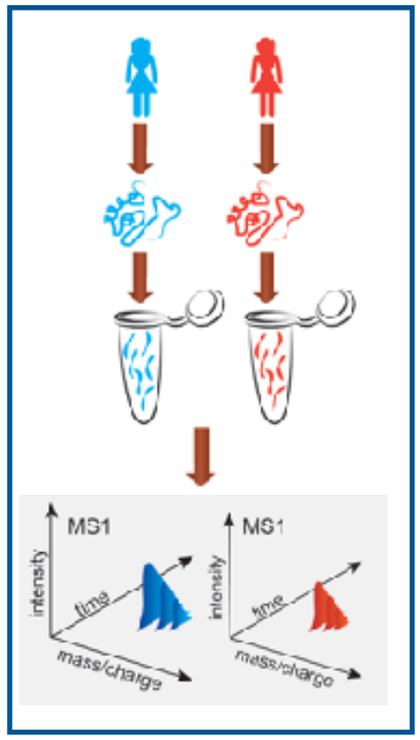
Number of replicates

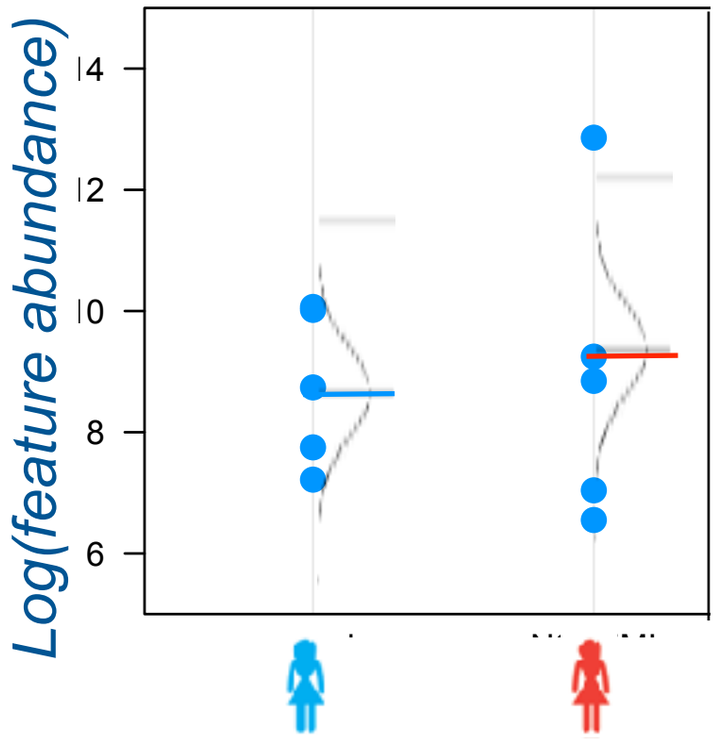
$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_{1\cdot})^2$$

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

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

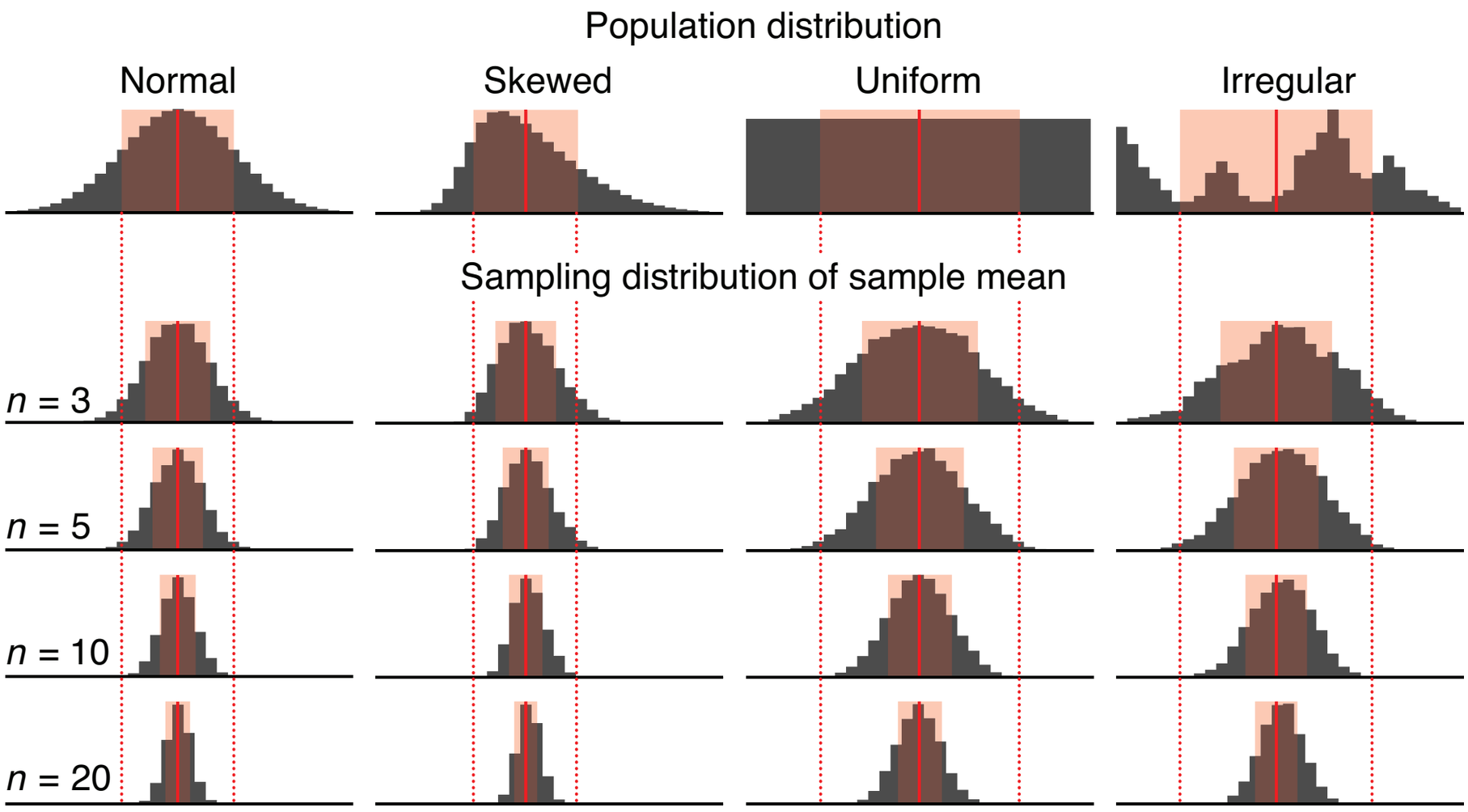
Variance of the sampling distribution of first mean

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

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

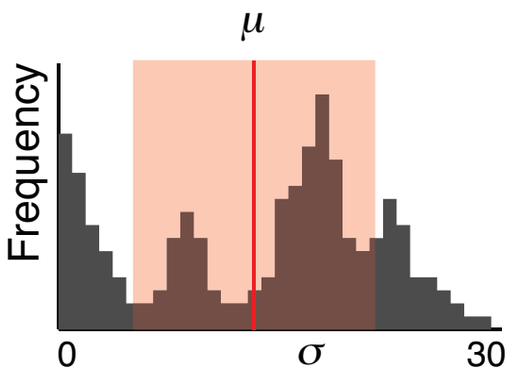
Conclusion:

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

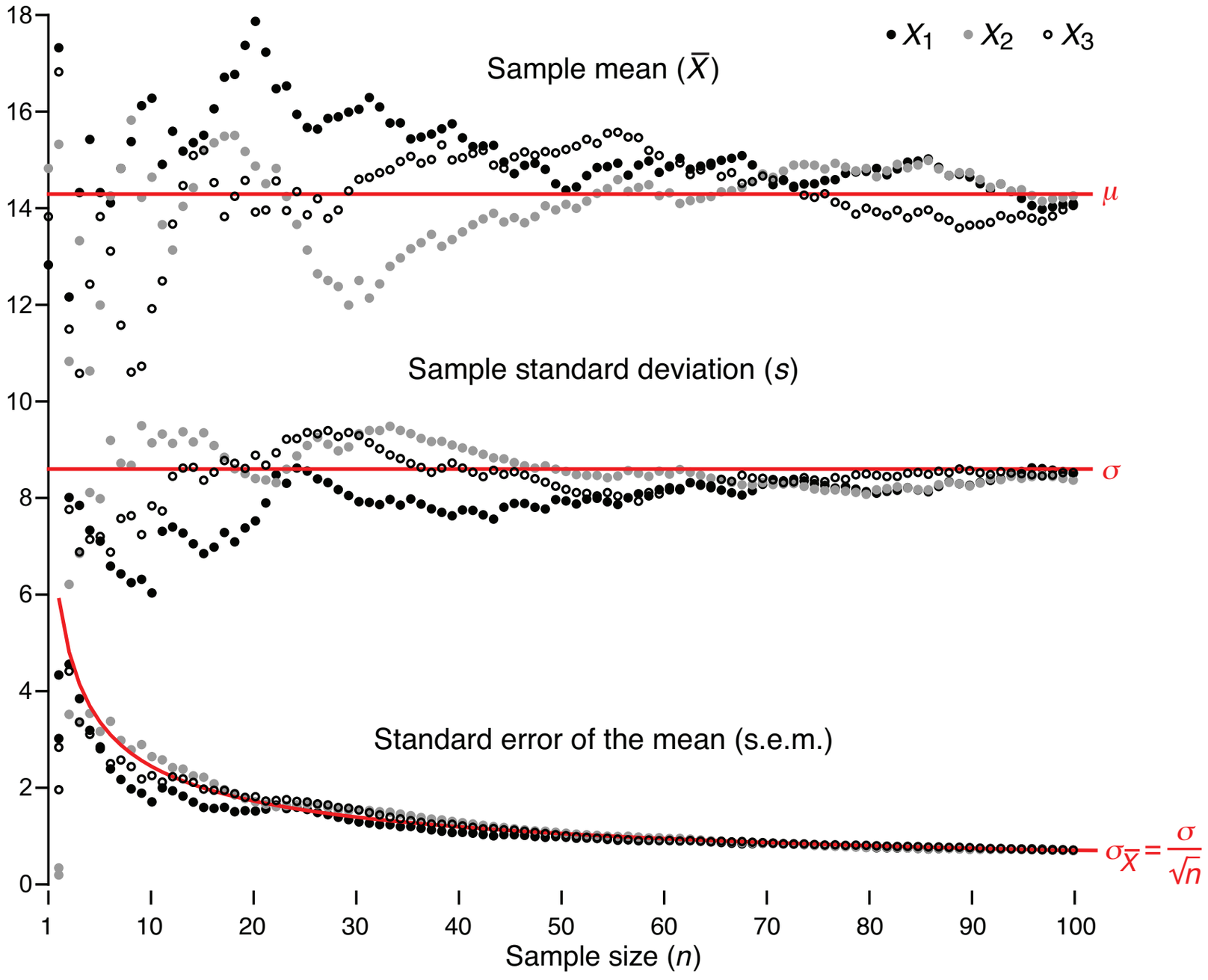
Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

EFFECT OF SAMPLE SIZE

As n increases, the estimates stabilize



Probability distribution of the data

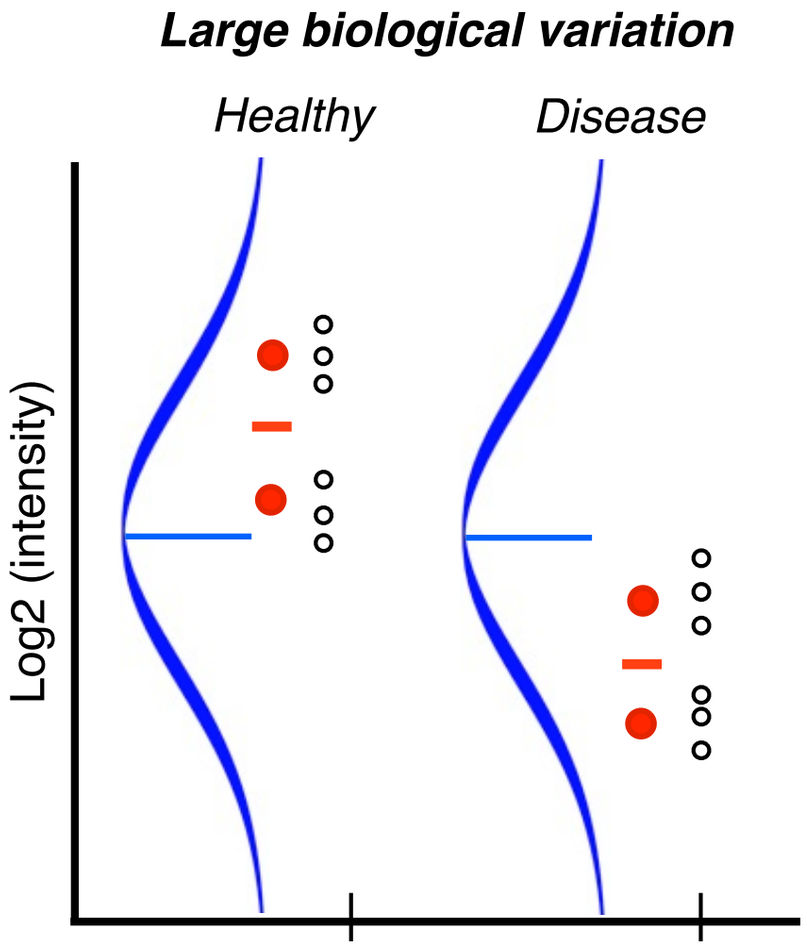


Simulated example

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

EFFECT OF SAMPLE SIZE

As n increases, the estimates stabilize



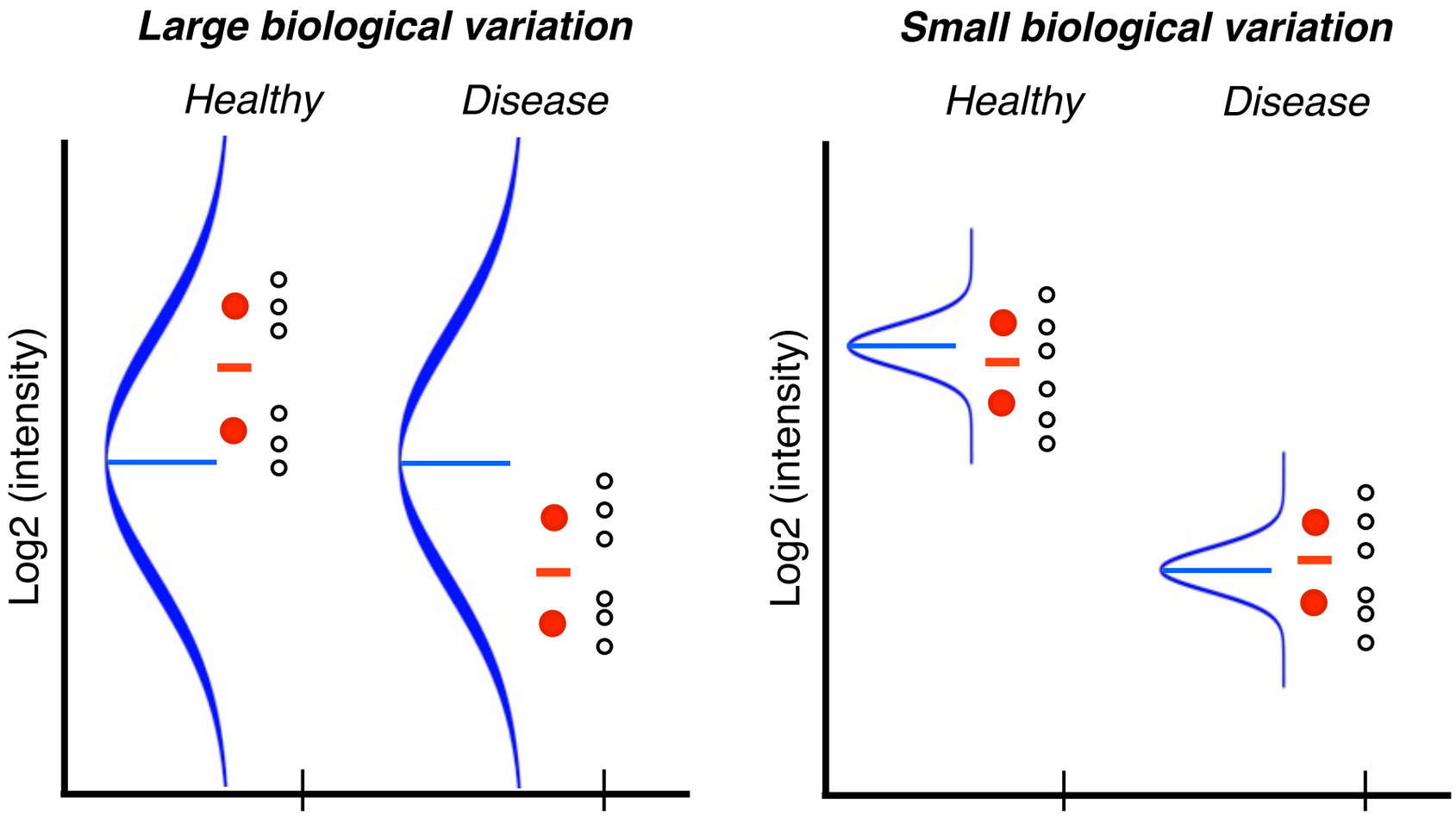
- *Protein abundance in the population*
- *Average abundance in the population*
- *Average abundance in the selected subjects*
- *Abundance in the individual subject*
- *Observed log-intensities of L*

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

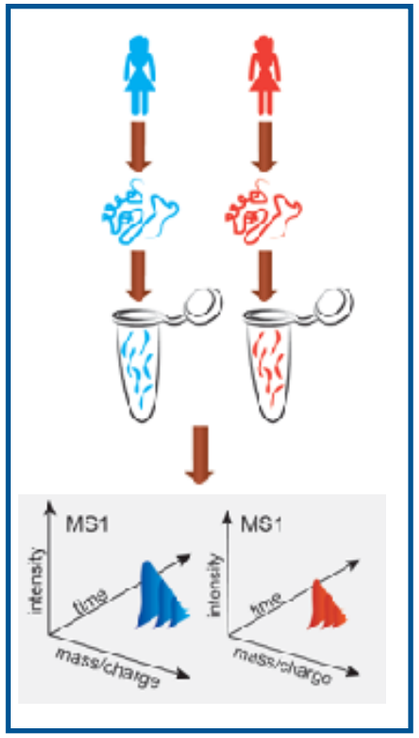


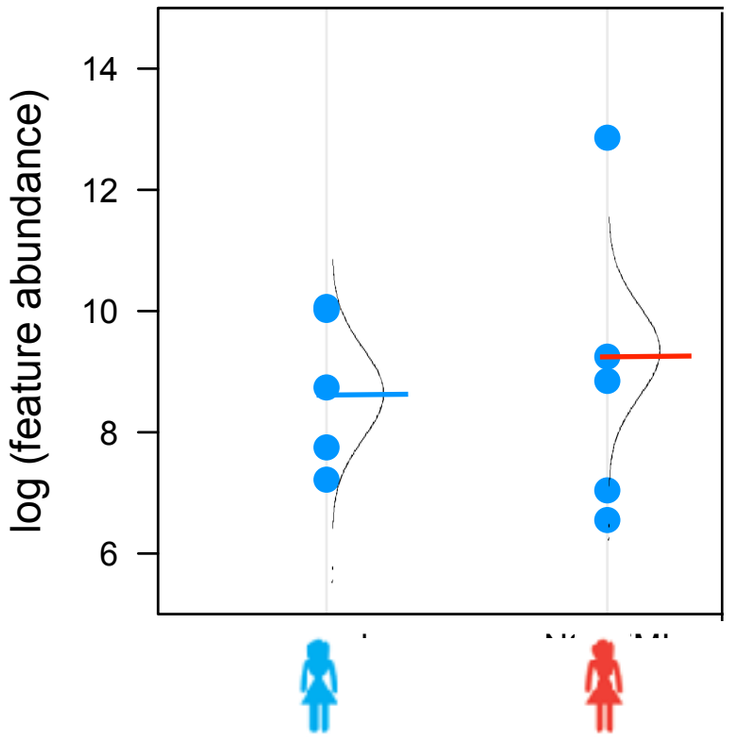
When biological variance is large, more biological replicates are needed to accurately estimate the variance

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, $\mu_1 - \mu_2 = 0$

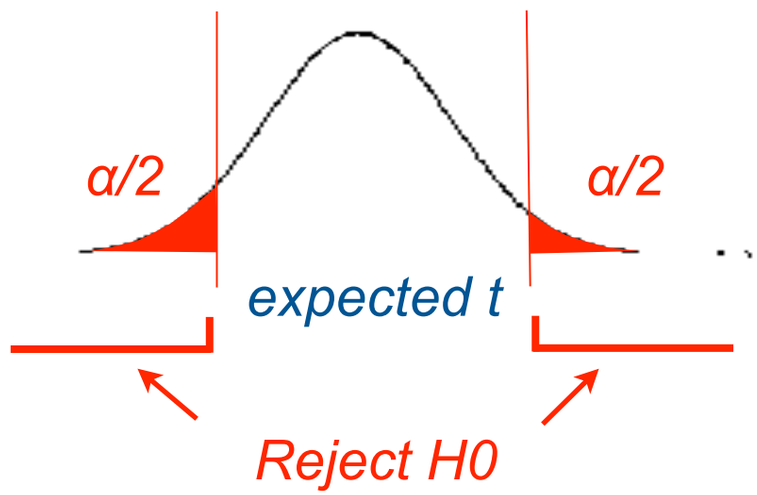
Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

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

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

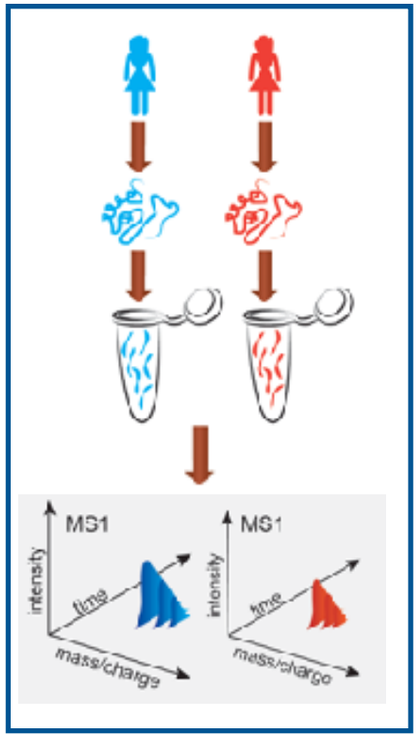
Distribution of the score if H0 is true

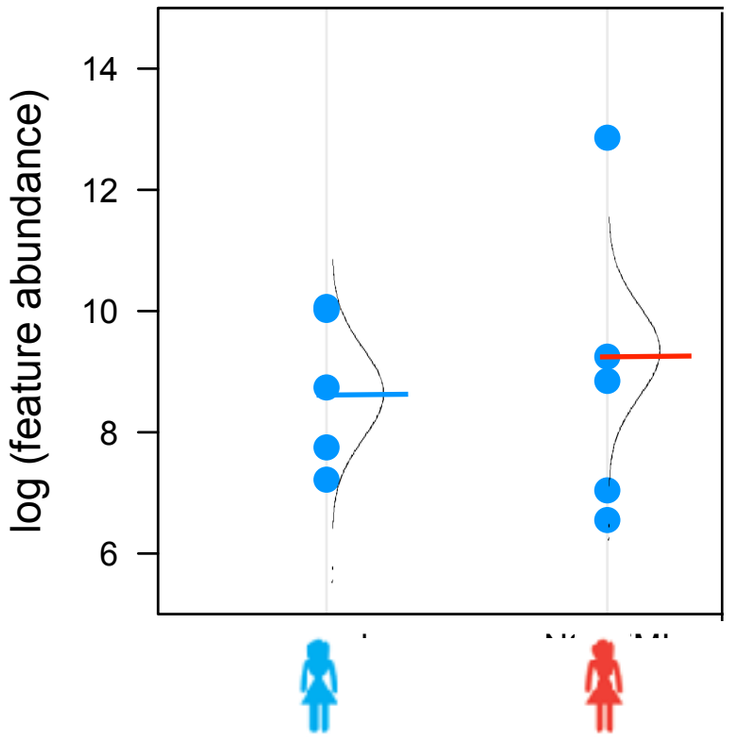
α = False Positive Rate



FINDING DIFFERENTIALLY ABUNDANT PROTEINS

P-value





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

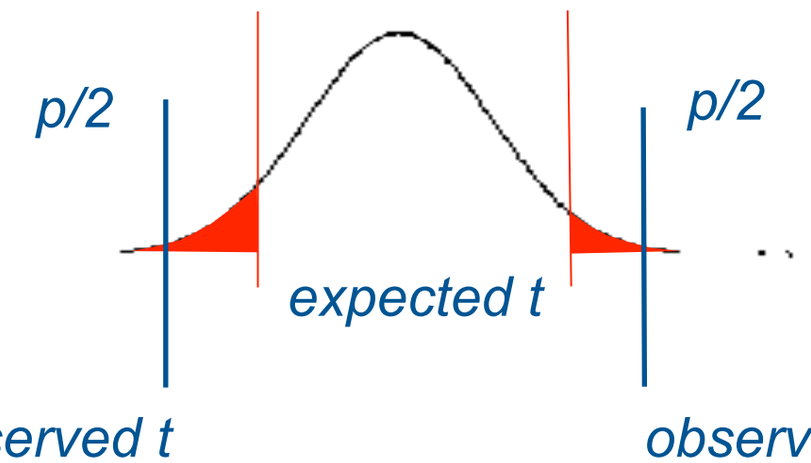
Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

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

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

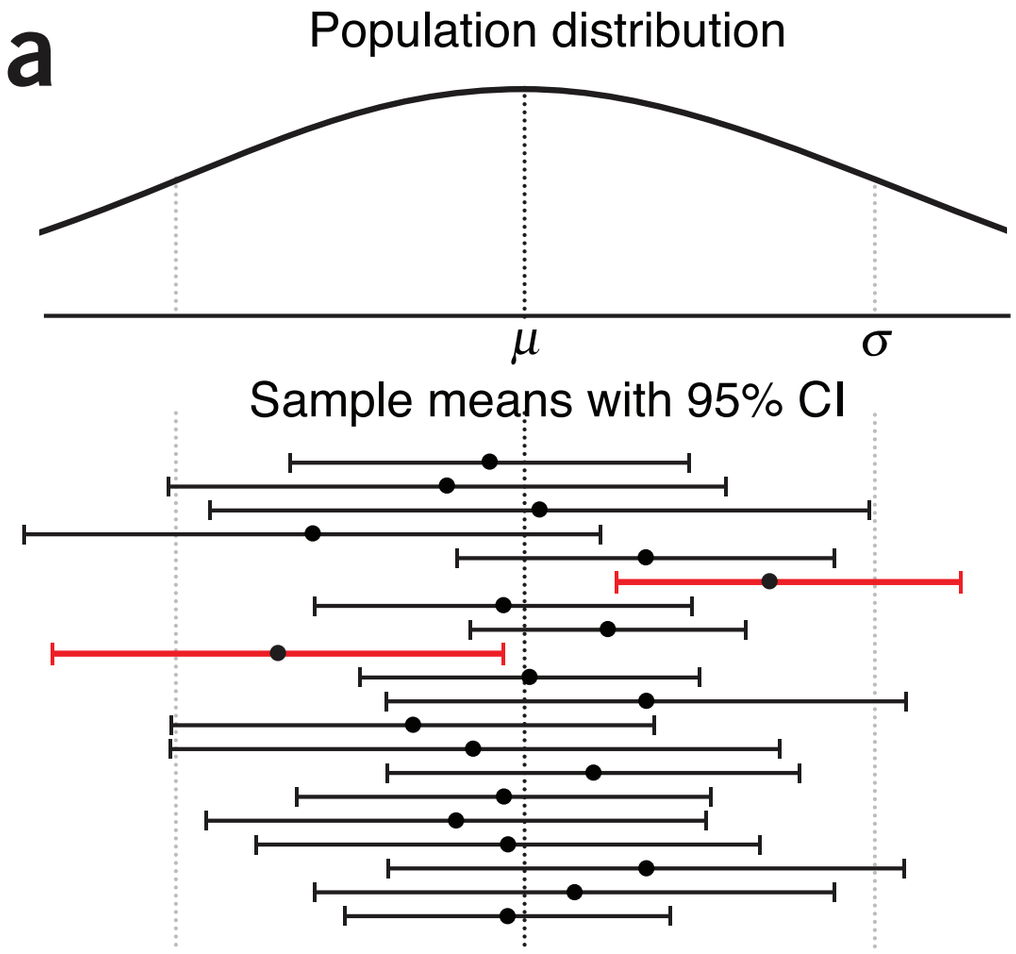
Distribution of the score if H0 is true

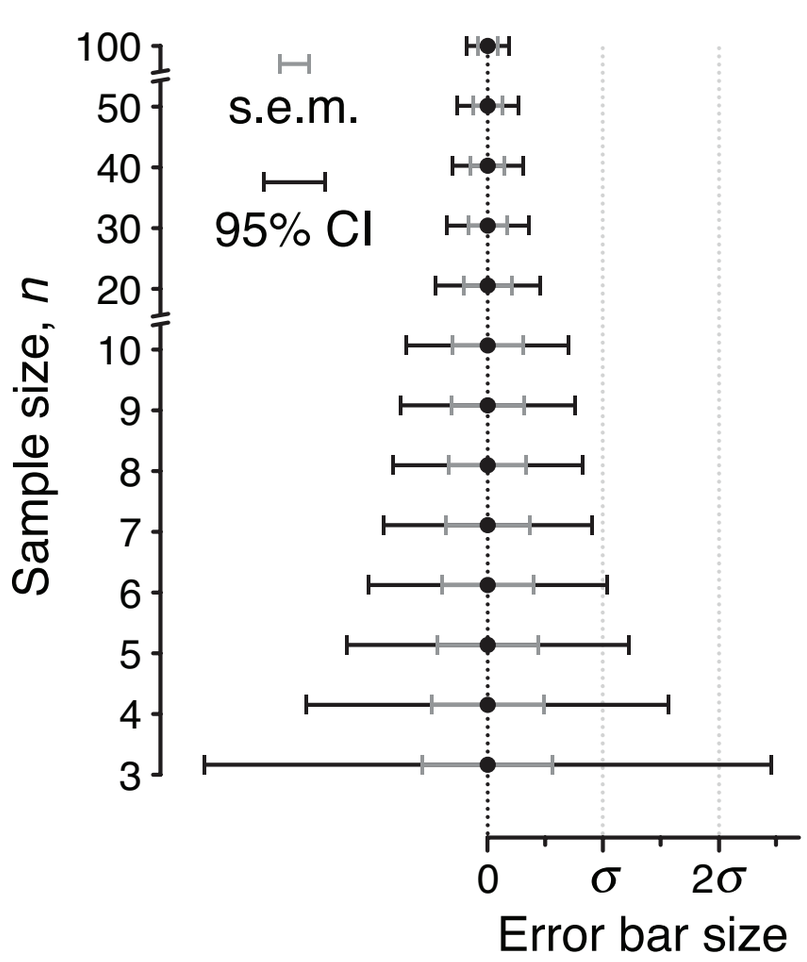
p = p-value



ALTERNATIVE TO TESTING: CONFIDENCE INTERVALS

Not all error bars are made equal





A 95% CI: 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!

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

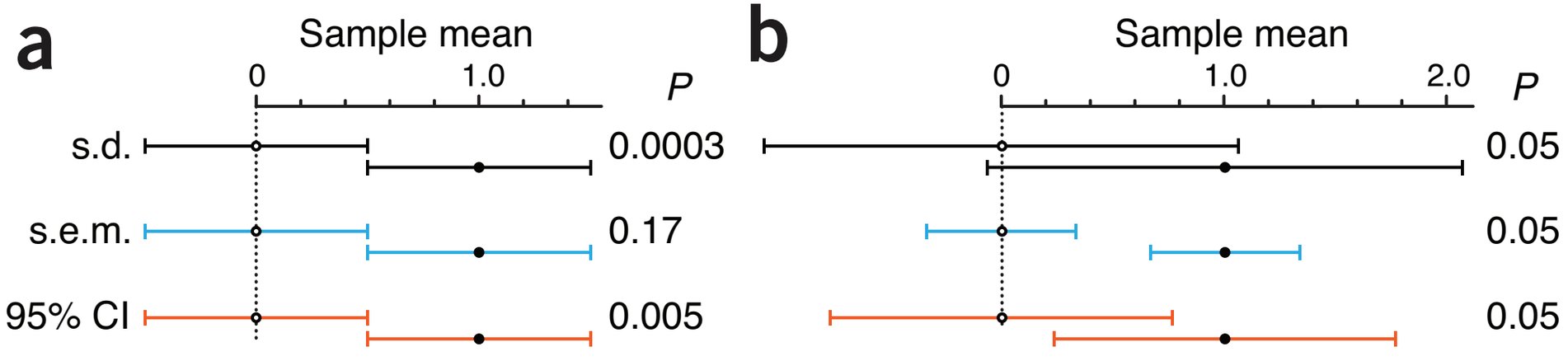
Width of the intervals depends on the sample size

Simulated example

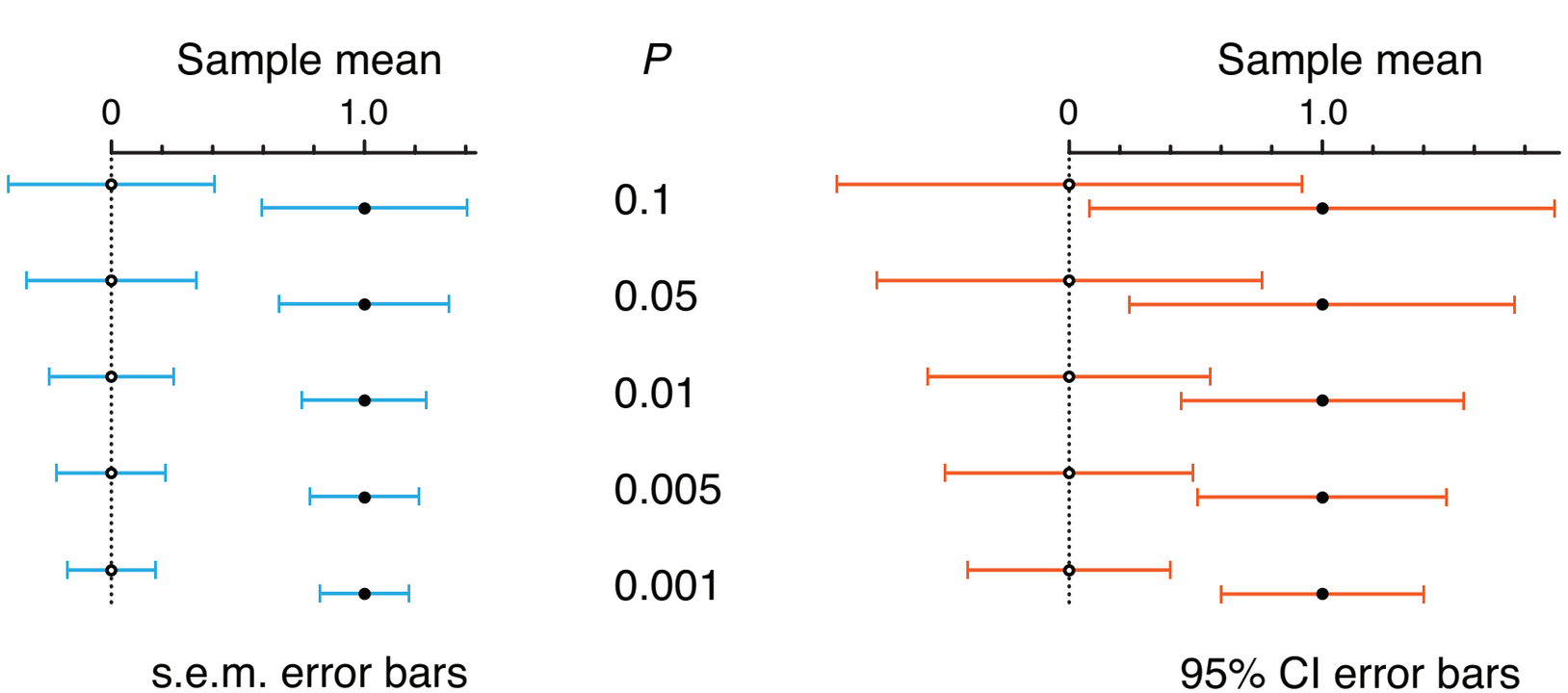
Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

ERROR BARS PROVIDE DIFFERENT INSIGHT

Absence of overlap does not always mean stat. significance



n=10



s.e.m. error bars

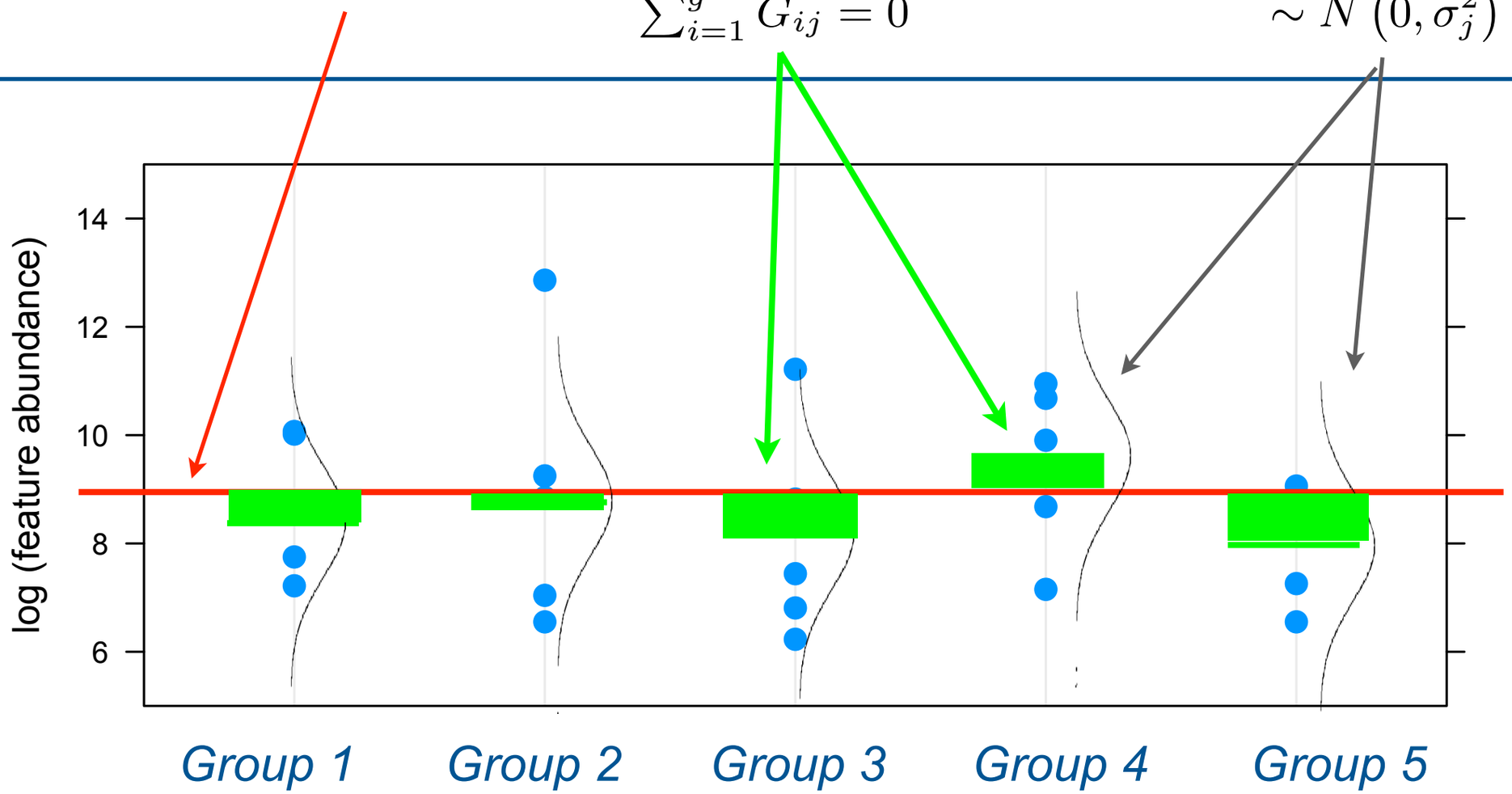
95% CI error bars

Simulated example

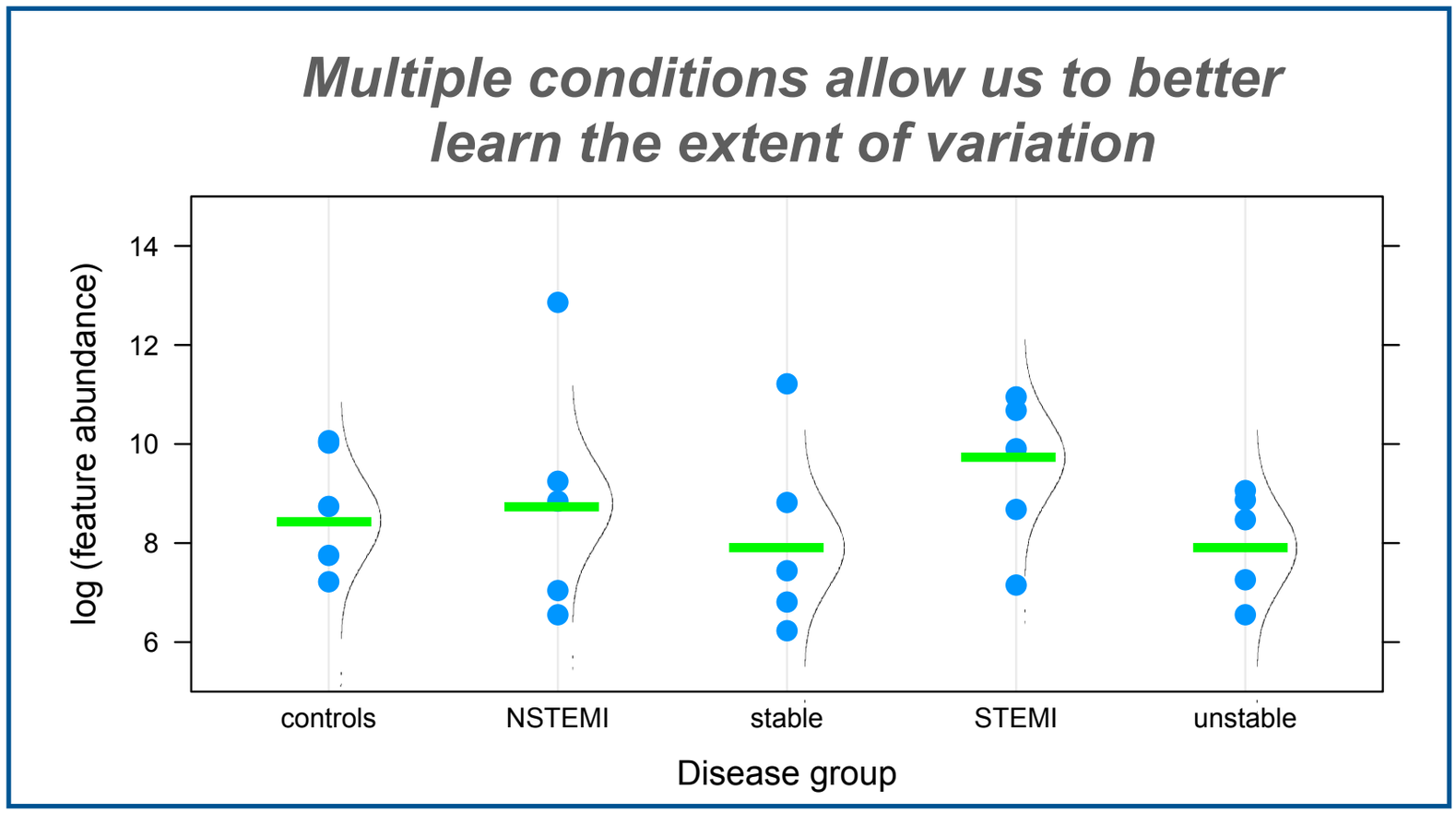
Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

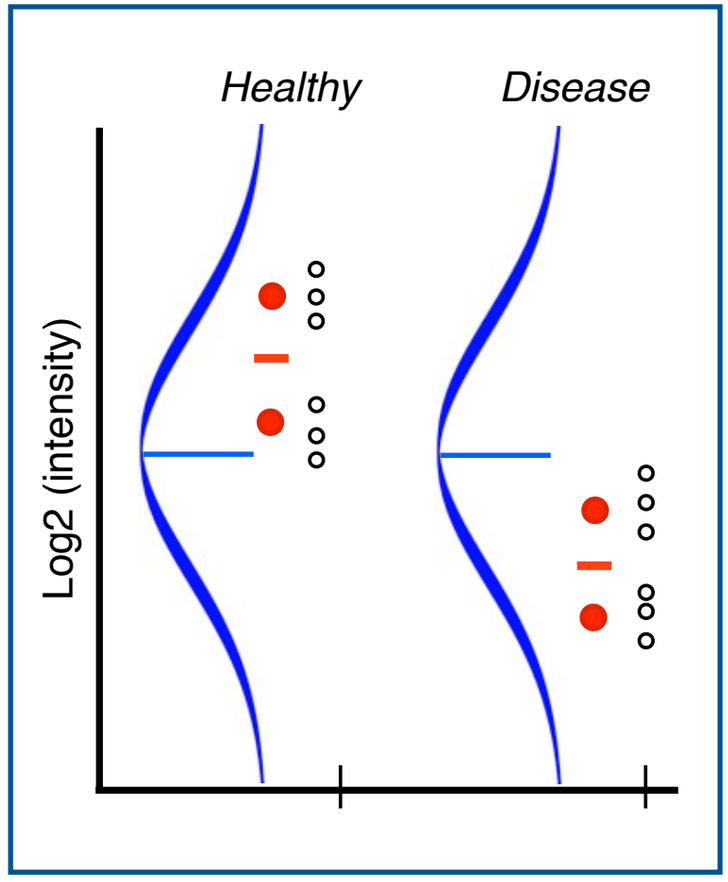
MULTI-GROUP ANALYSIS: A ONE-WAY ANALYSIS OF VARIANCE (ANOVA)

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

- 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

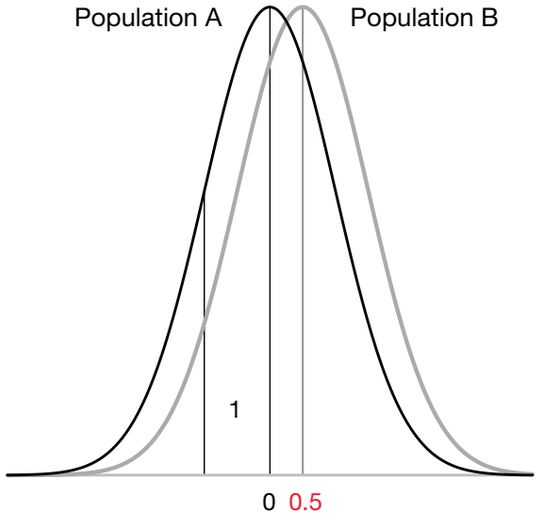
- P-values can indicate how incompatible the data are with a specified statistical model
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance
- Scientific conclusions and business policy decisions should not be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

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

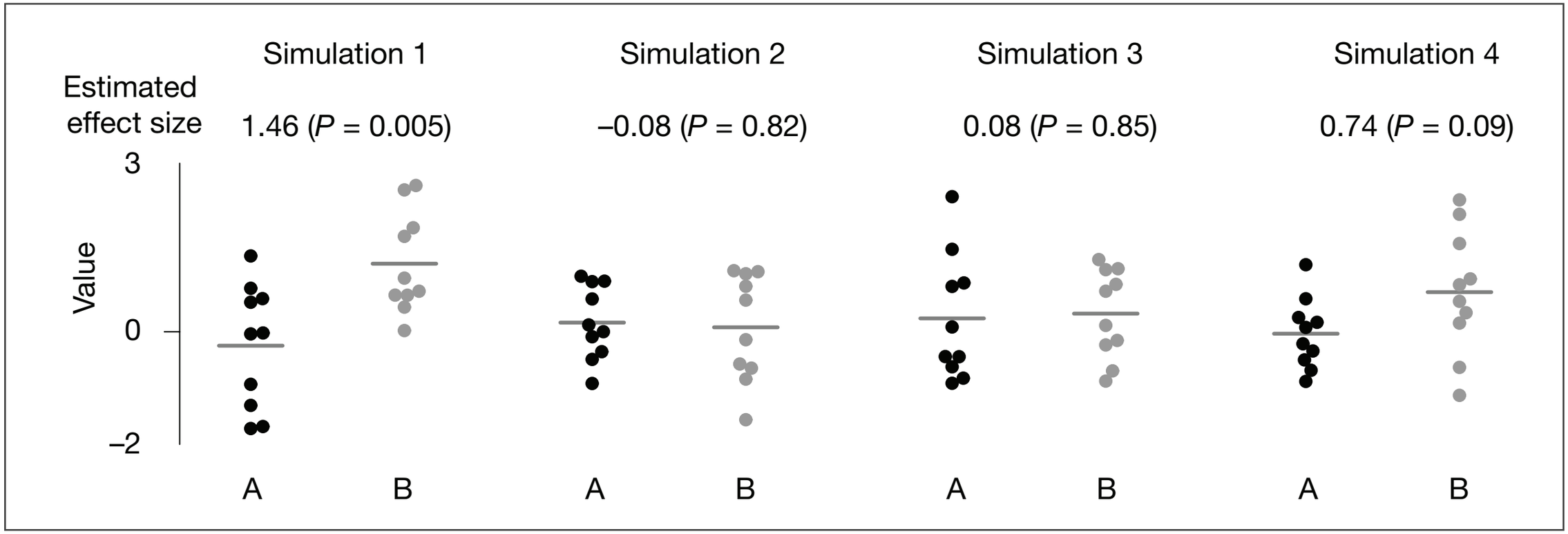
The American Statistician, February 2016

- 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



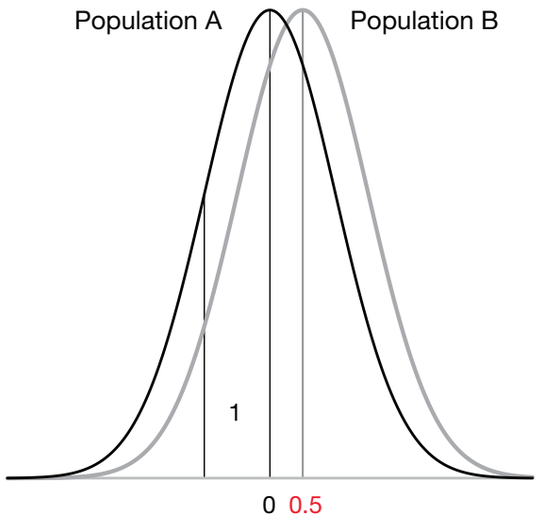
- Repeatedly sampling data leads to different results
- The problem worsens when testing many proteins
- Partial solutions:
  - Larger sample size
  - Adjustment for multiple testing



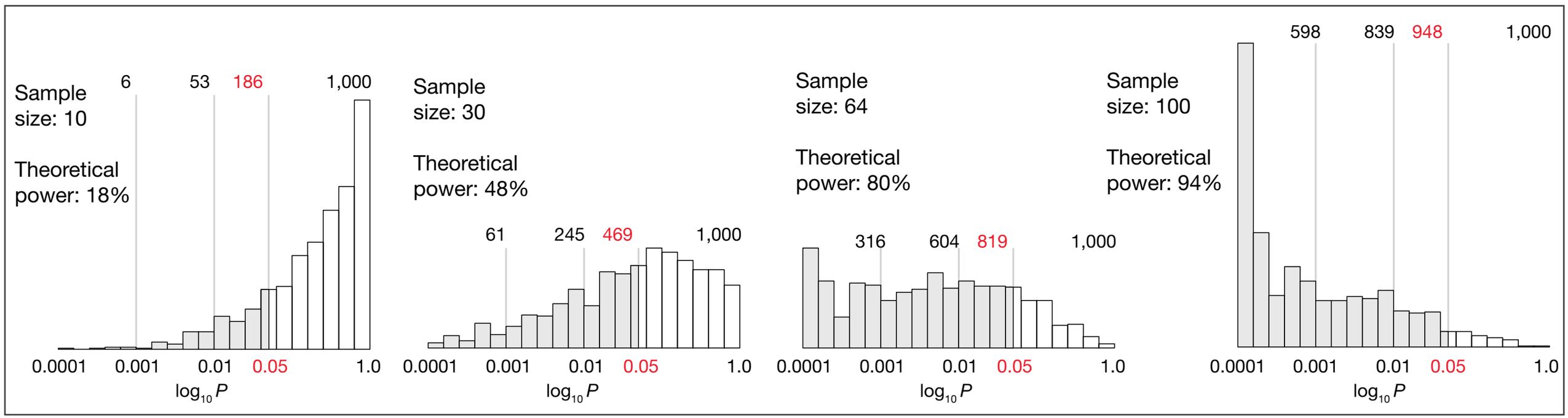
Simulated example

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

WITH SMALL SAMPLE SIZE, P-VALUES ARE UNSTABLE



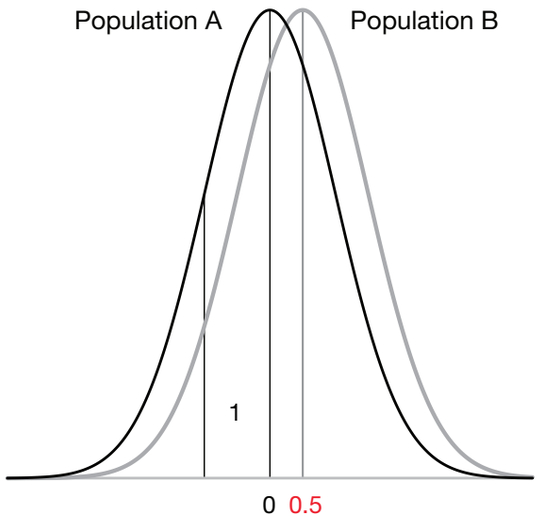
- Repeatedly sampling data leads to different results
- The problem worsens when testing many proteins
- Partial solutions:
  - Larger sample size
  - Adjustment for multiple testing



Simulated example

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

WITH SMALL SAMPLE SIZE, CONCLUSIONS ARE BIASED



Simulated example

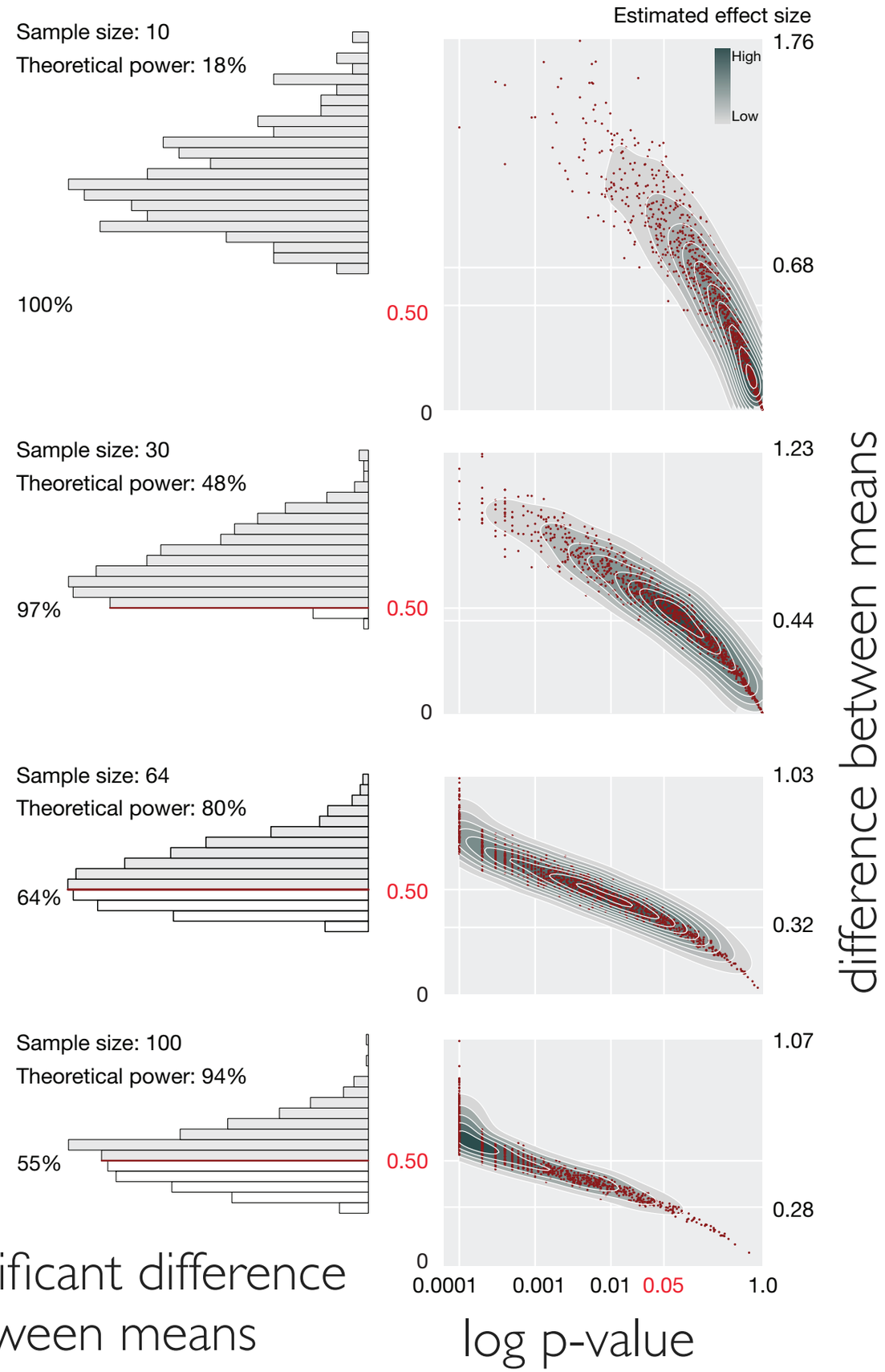
Halsey, Curran-Everett, Volwer and Drummond, *Nature Methods*, 2015

10 replicates

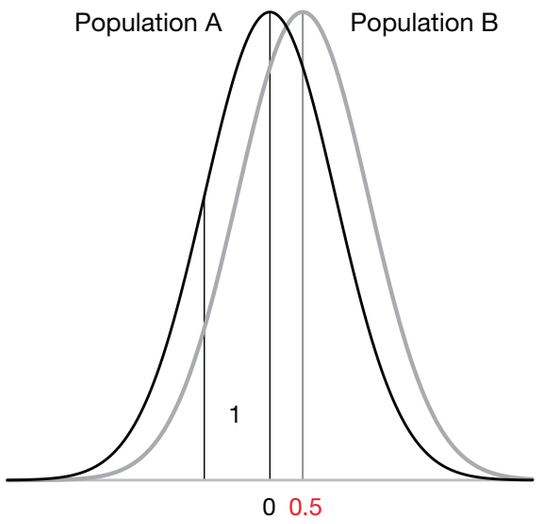
30 replicates

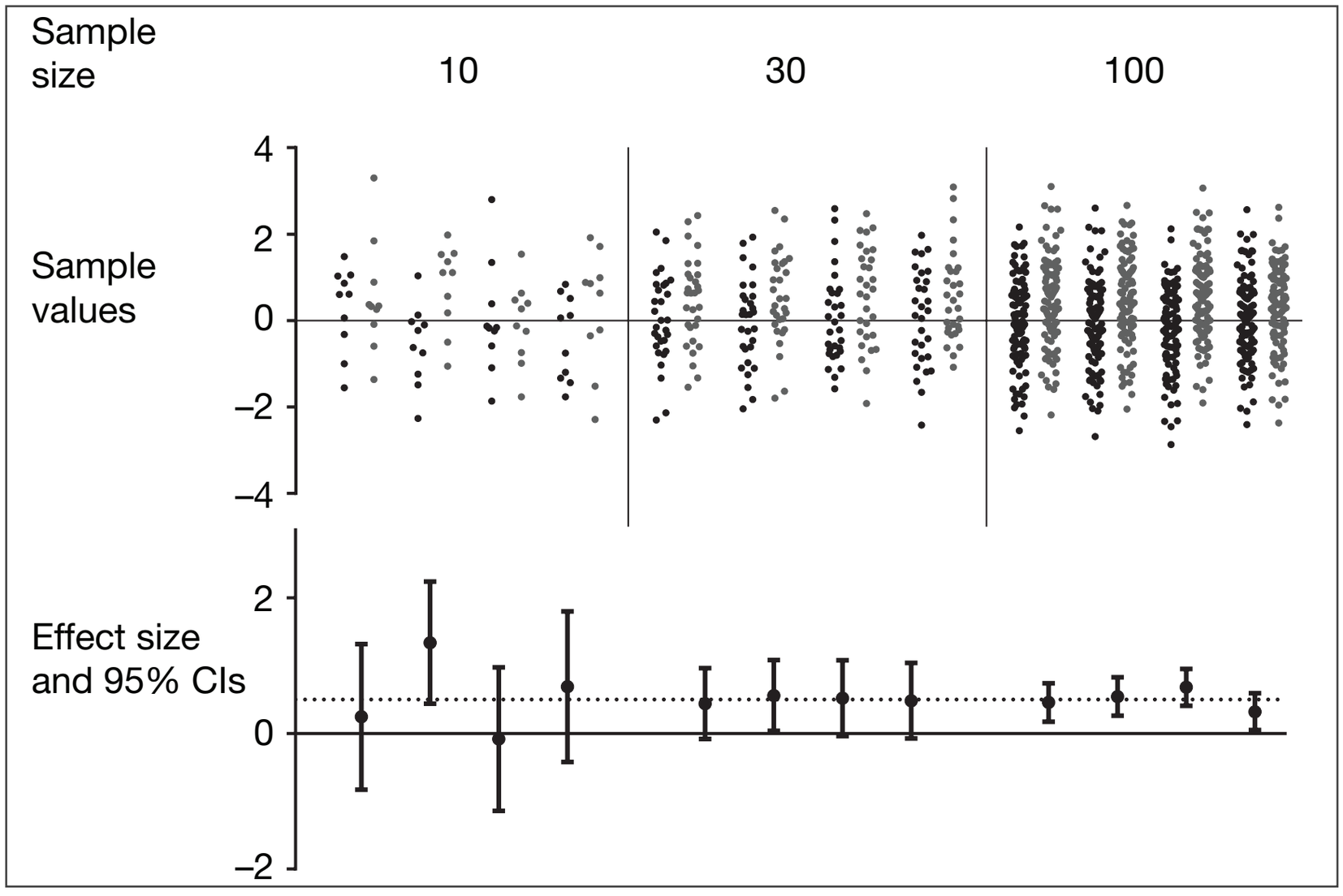
64 replicates

100 replicates



CONFIDENCE INTERVALS PROVIDE COMPLEMENTARY INSIGHT





Simulated example

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

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/