

STATISTICAL INFERENCE

Olga Vitek

College of Science
College of Computer and Information Science



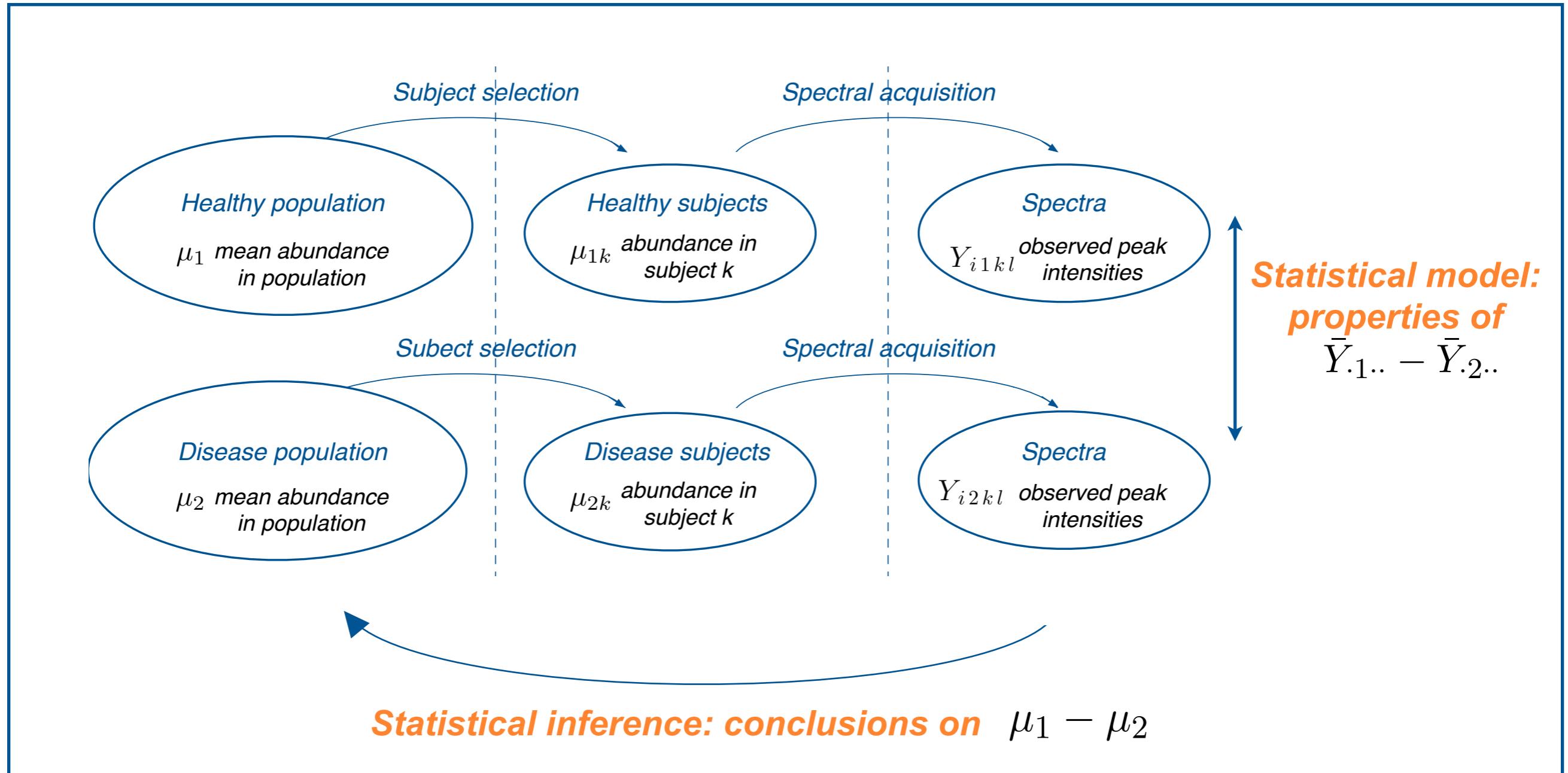
Northeastern University

OUTLINE

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

COMPARE DESIGNS

In terms of bias and (in)-efficiency

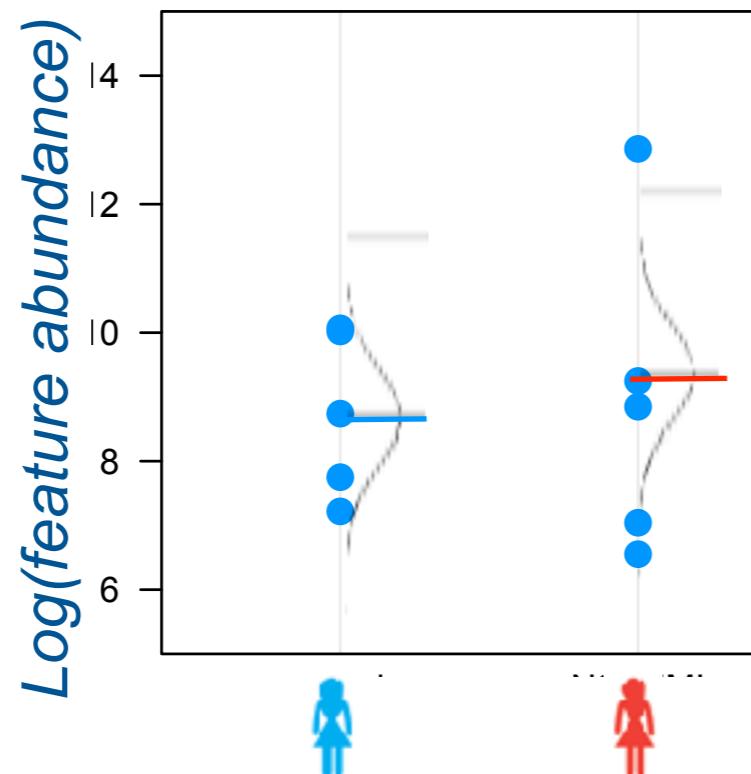
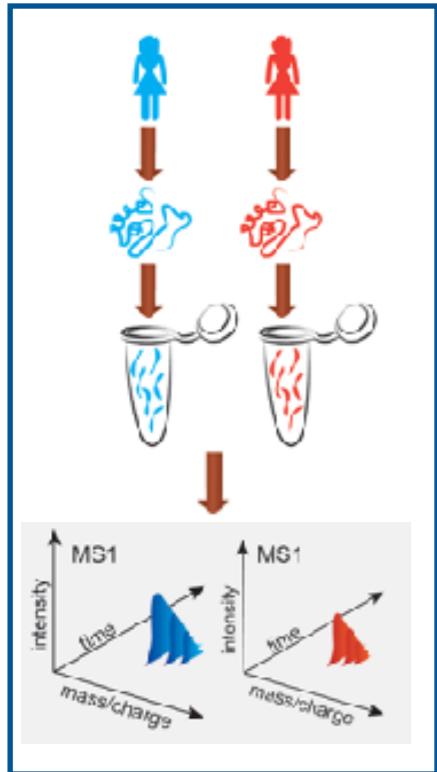


Bias: $\bar{Y}_{1..} - \bar{Y}_{2..}$ systematically different from $\mu_{1k} - \mu_{2k}$

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

TWO-SAMPLE T-TEST

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



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

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

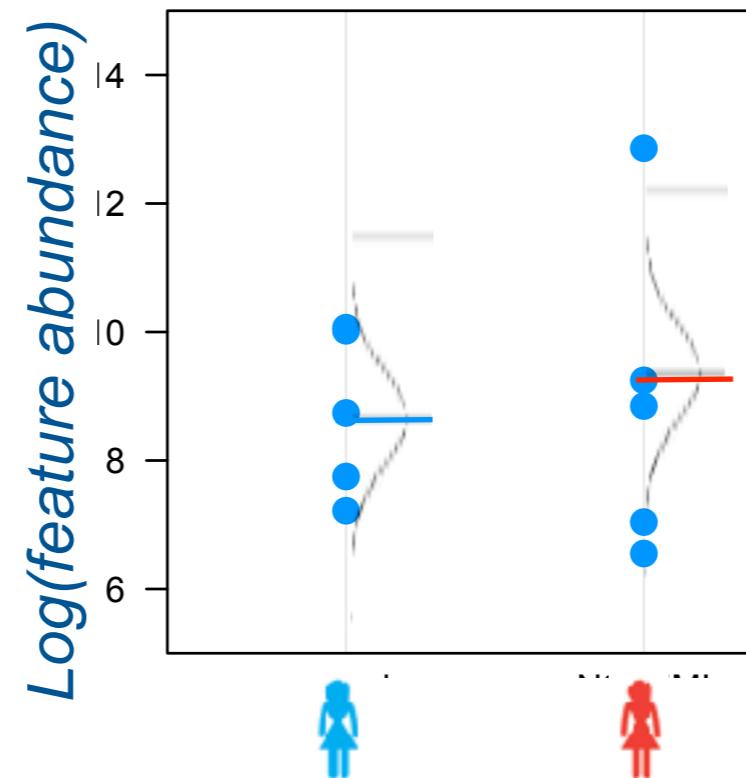
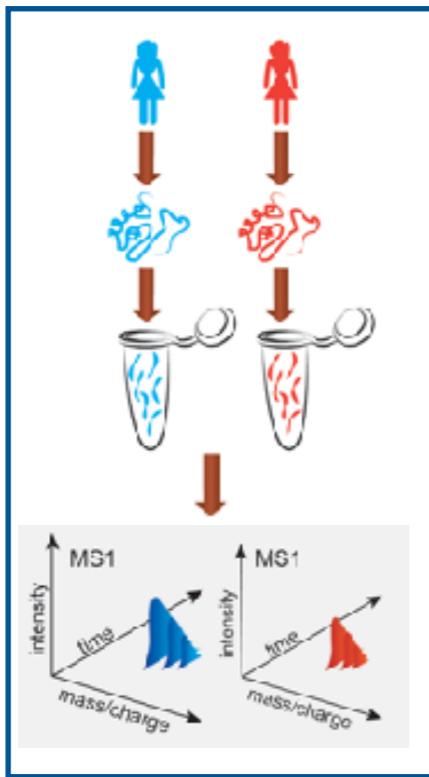
- $\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}$ = estimates log-fold change

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

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

H_0 : 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$

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

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

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

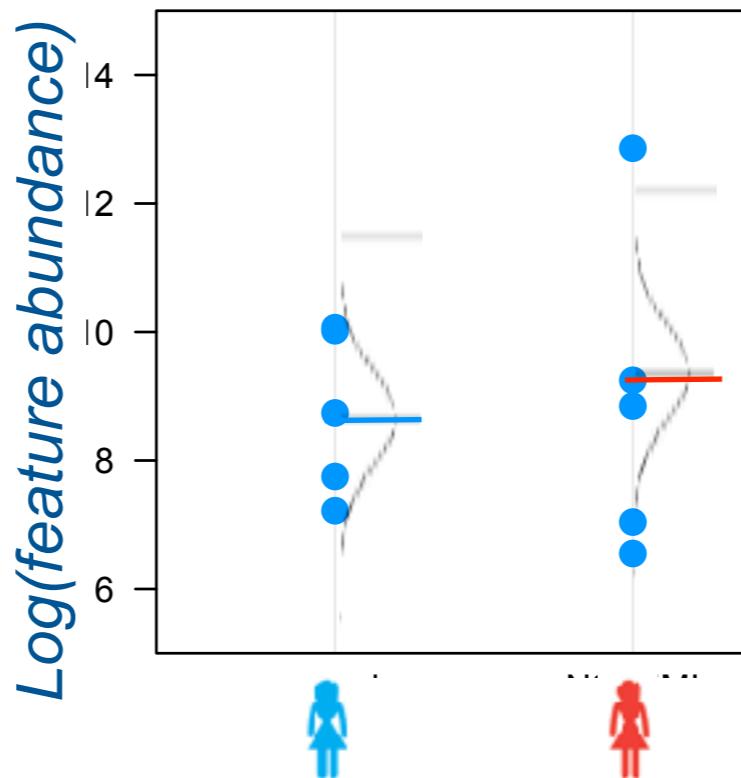
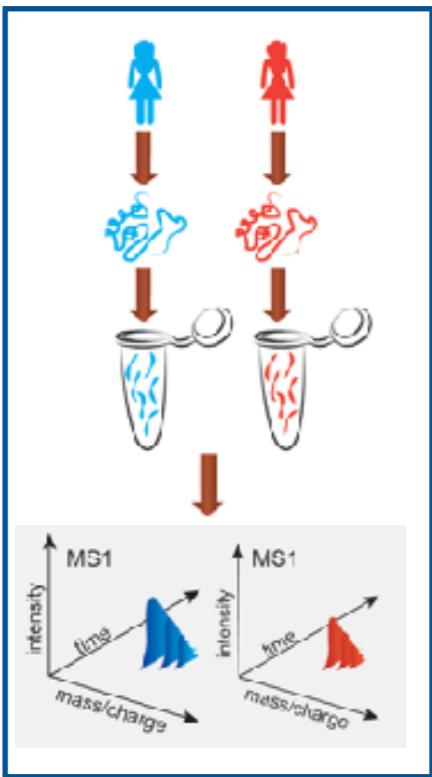
Number of
replicates

$$s_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2$$

Sample variance

TWO-SAMPLE T-TEST

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



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

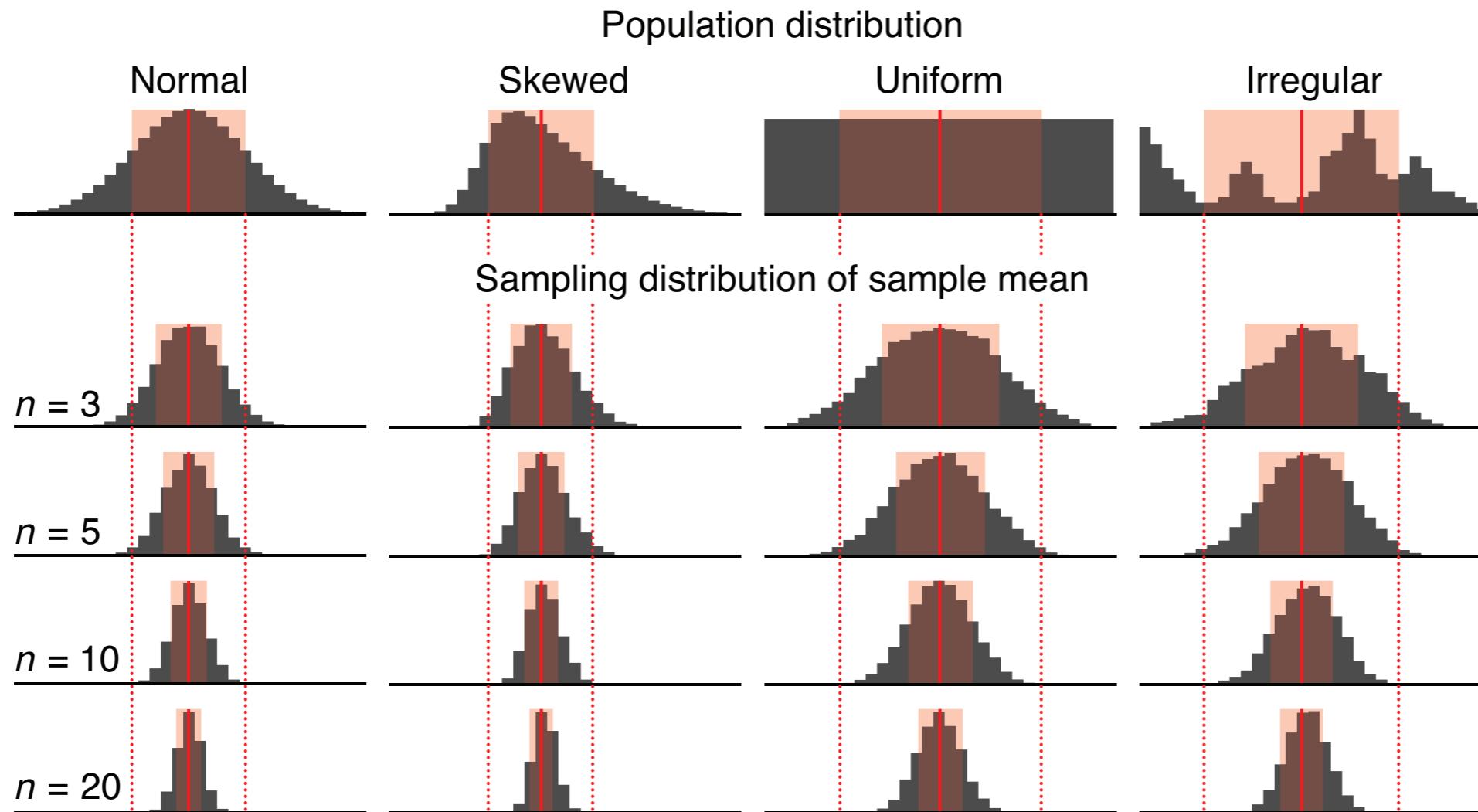
H_0 : 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$
 H_a : change in abundance, $\mu_1 - \mu_2 \neq 0$

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

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

ASSUMPTION: NORMAL DISTRIBUTION

The Central Limit Theorem



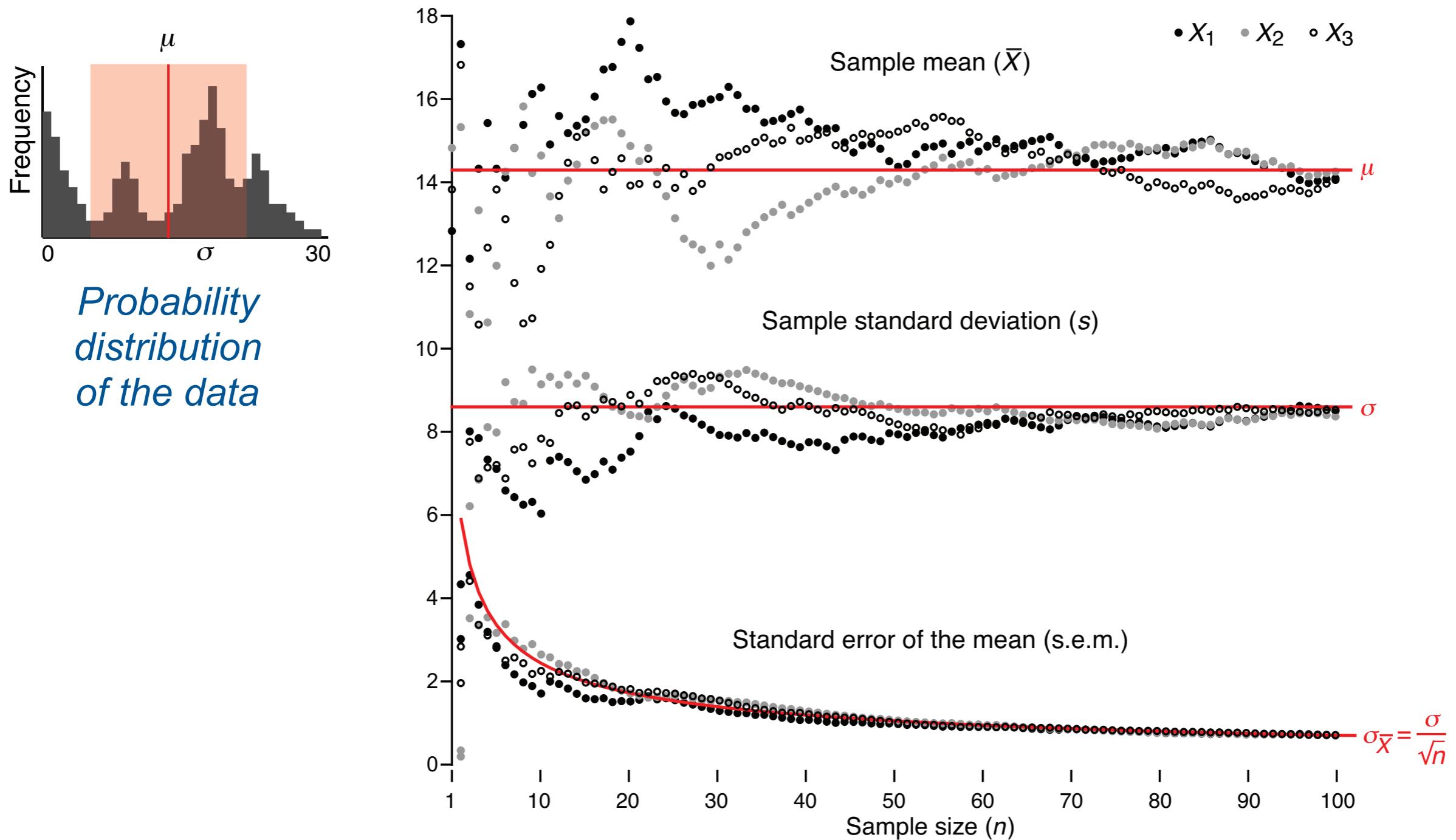
Probability distribution of the data

Repeatedly selecting n data points and calculating means

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

EFFECT OF SAMPLE SIZE

As n increases, the estimates stabilize

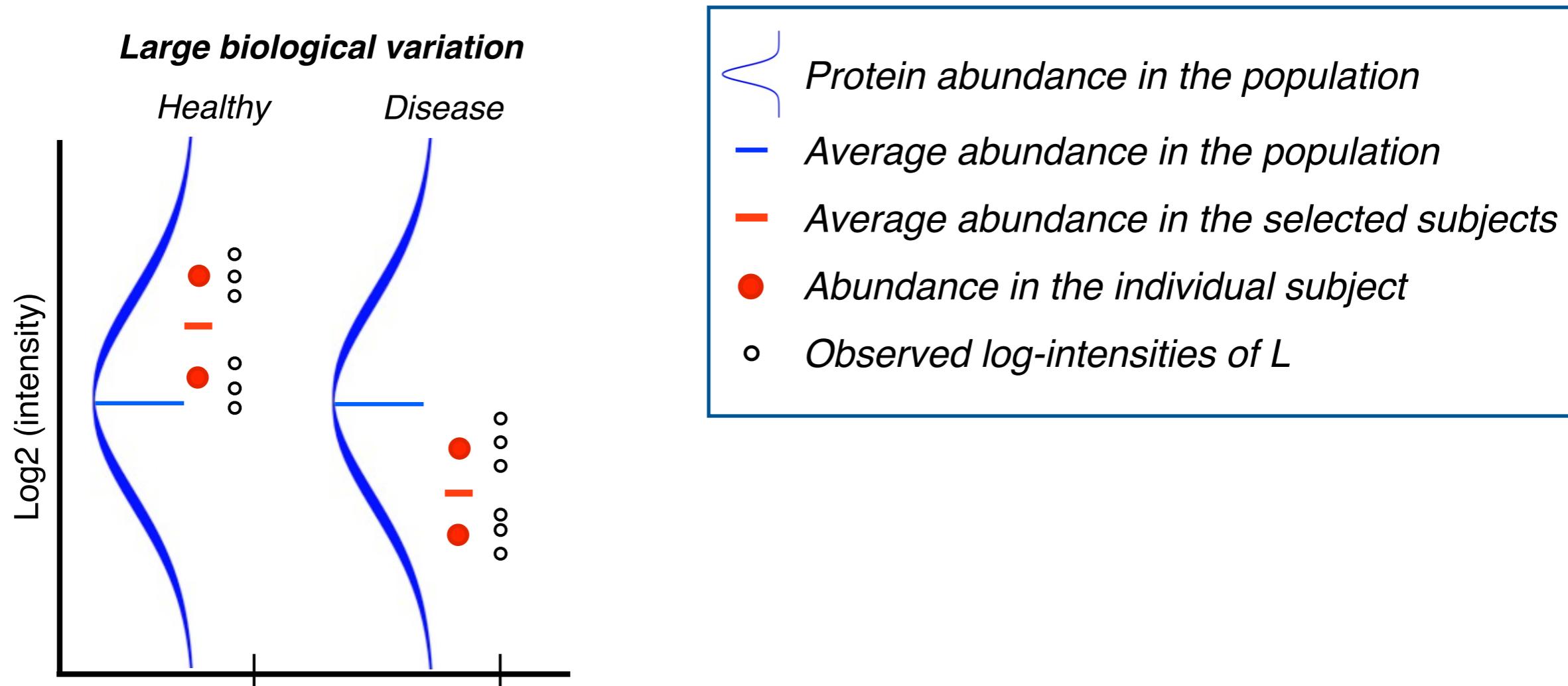


Simulated example

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

EFFECT OF SAMPLE SIZE

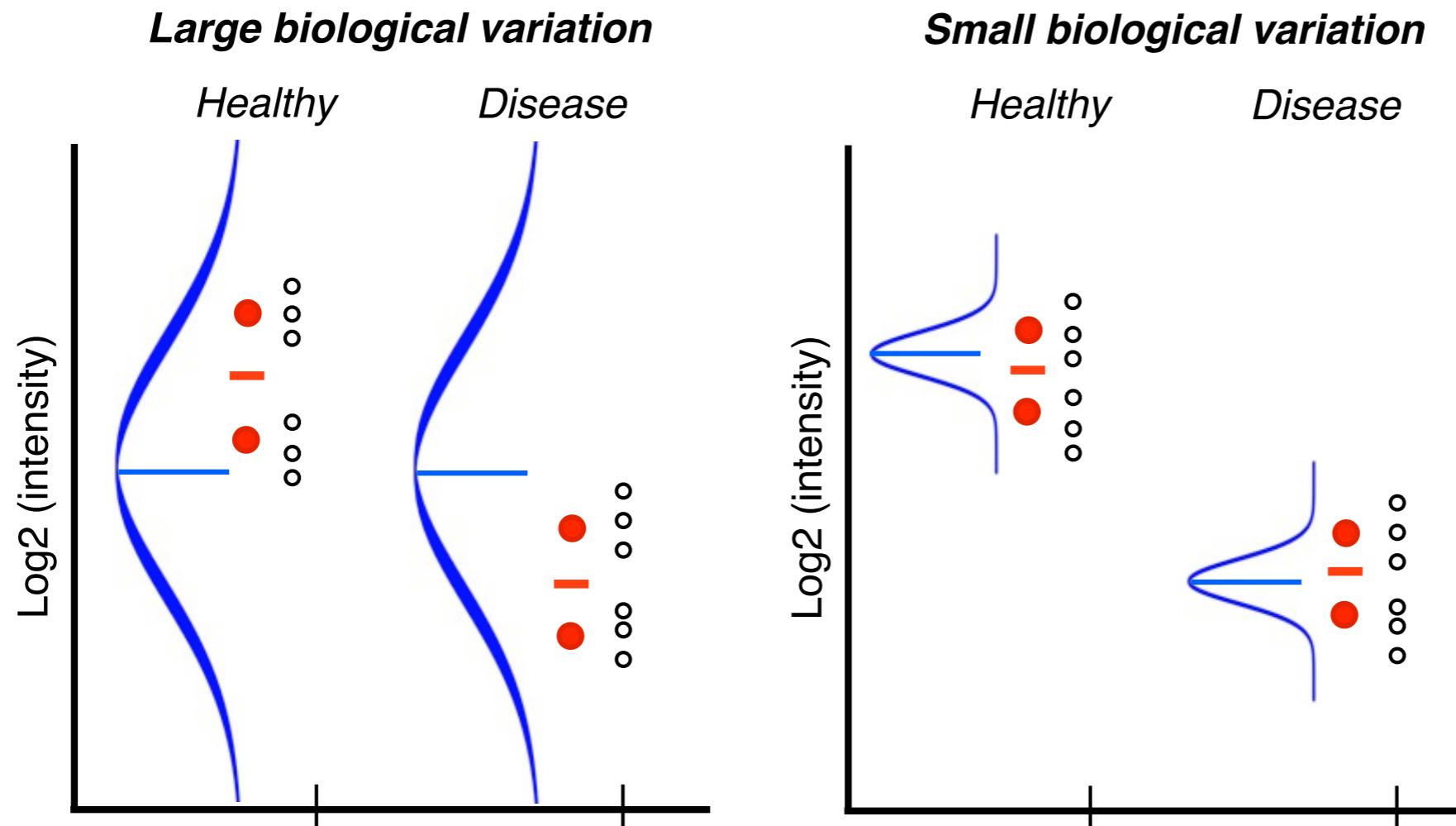
As n increases, the estimates stabilize



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

EFFECT OF SAMPLE SIZE

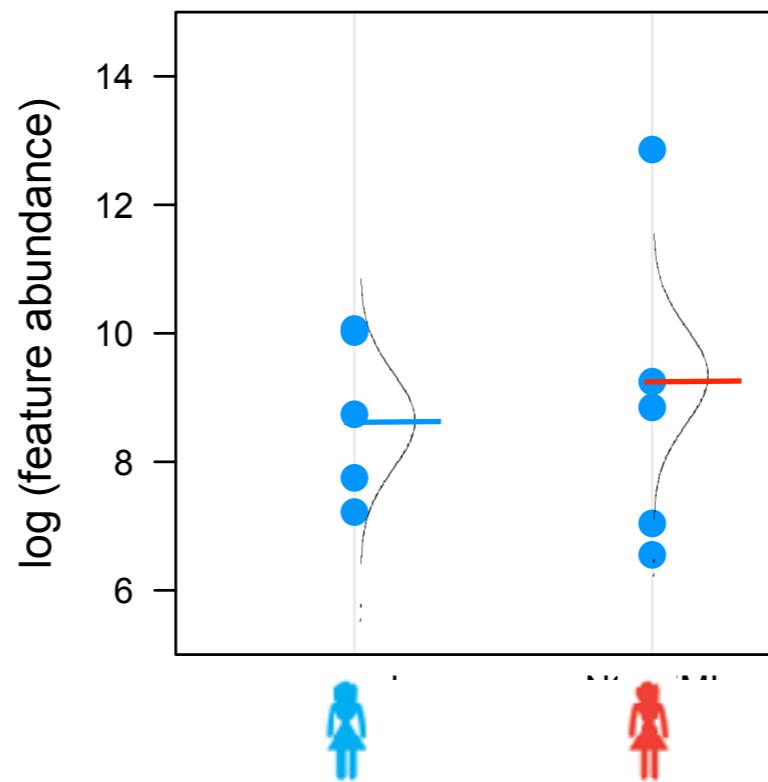
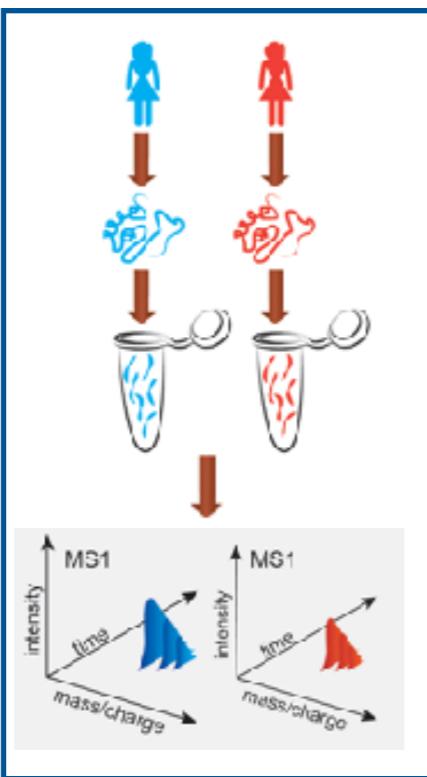
As n increases, the estimates stabilize



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

FINDING DIFFERENTIALLY ABUNDANT PROTEINS

False positive rate

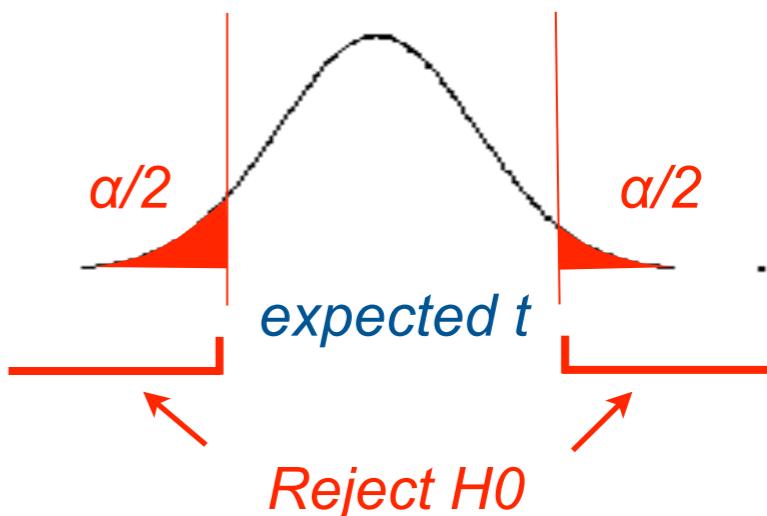


H_0 : 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$
 H_a : change in abundance, $\mu_1 - \mu_2 \neq 0$

Distribution of the score if H_0 is true

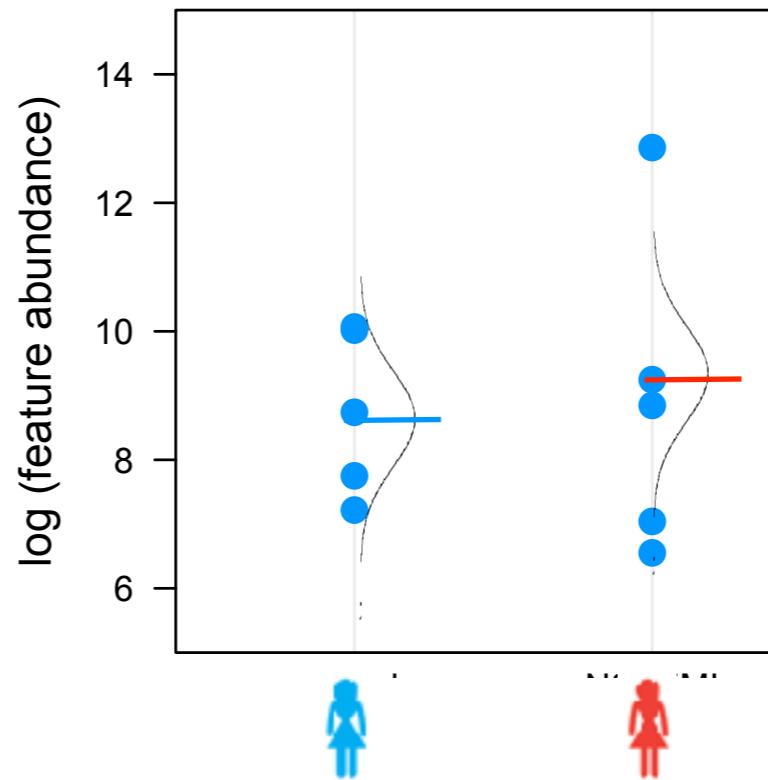
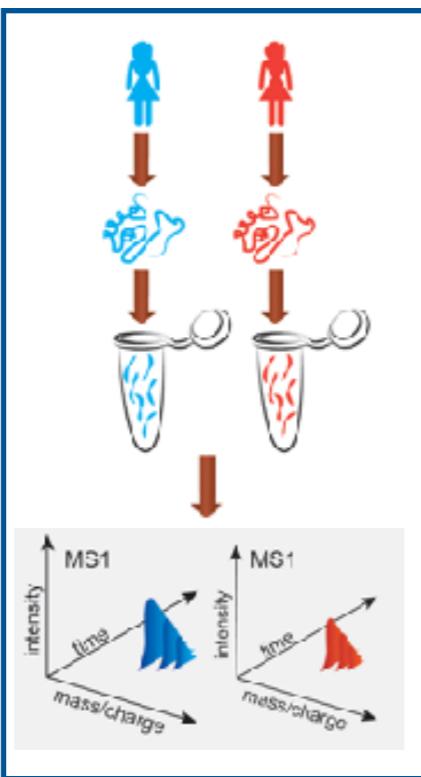
α = False Positive Rate

observed $t = \frac{\text{difference of group means}}{\text{estimate of variation}}$
no difference \sim Student distribution



FINDING DIFFERENTIALLY ABUNDANT PROTEINS

P-value

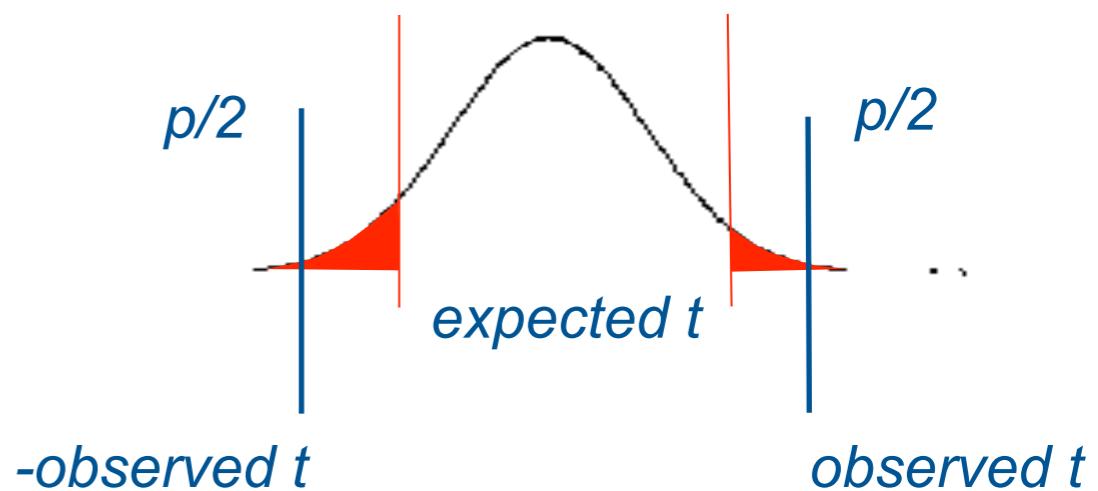


H_0 : 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$
 H_a : change in abundance, $\mu_1 - \mu_2 \neq 0$

observed $t = \frac{\text{difference of group means}}{\text{estimate of variation}}$
no difference \sim Student distribution

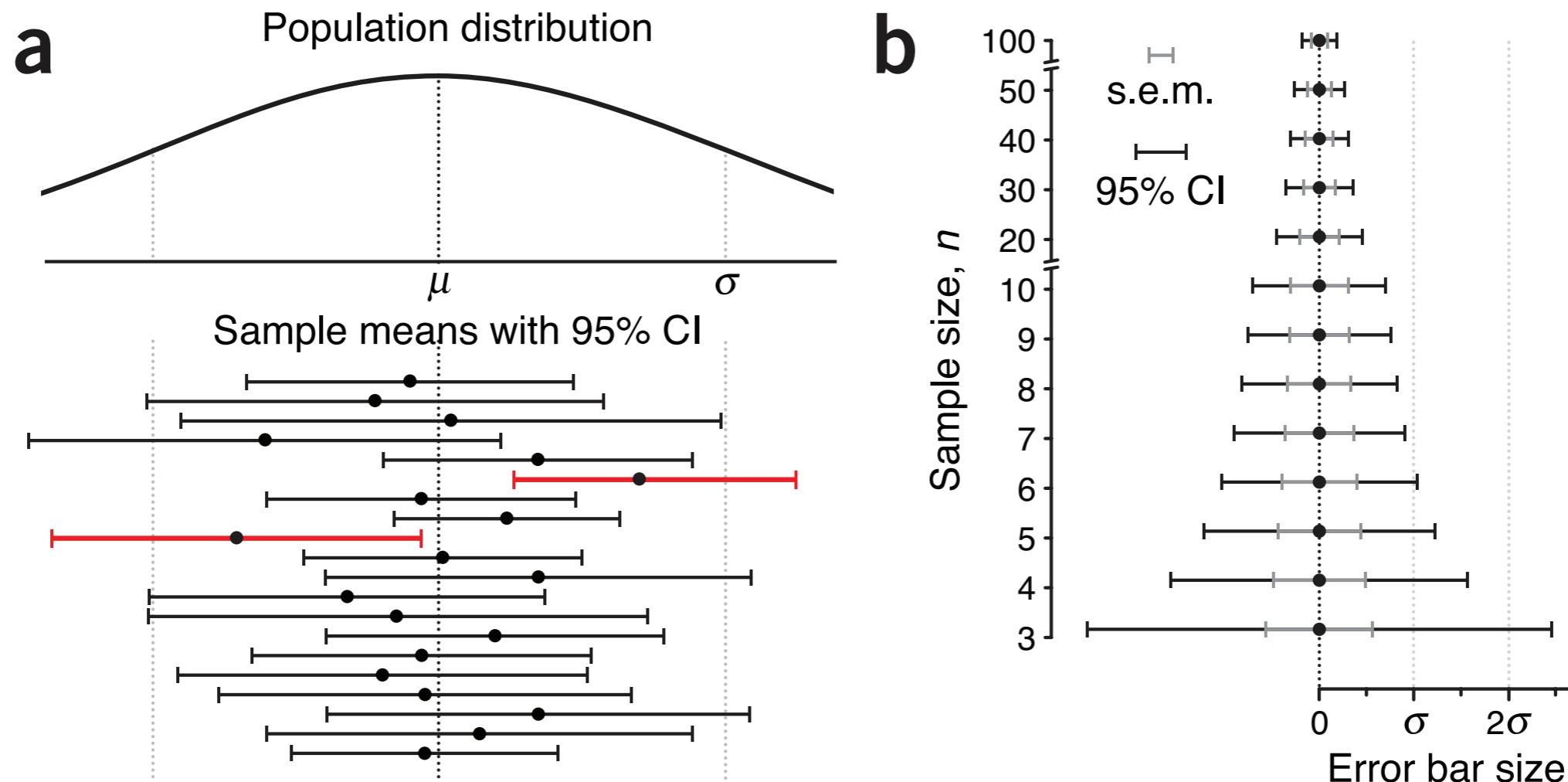
Distribution of the score if H_0 is true

$p = p\text{-value}$



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

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

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

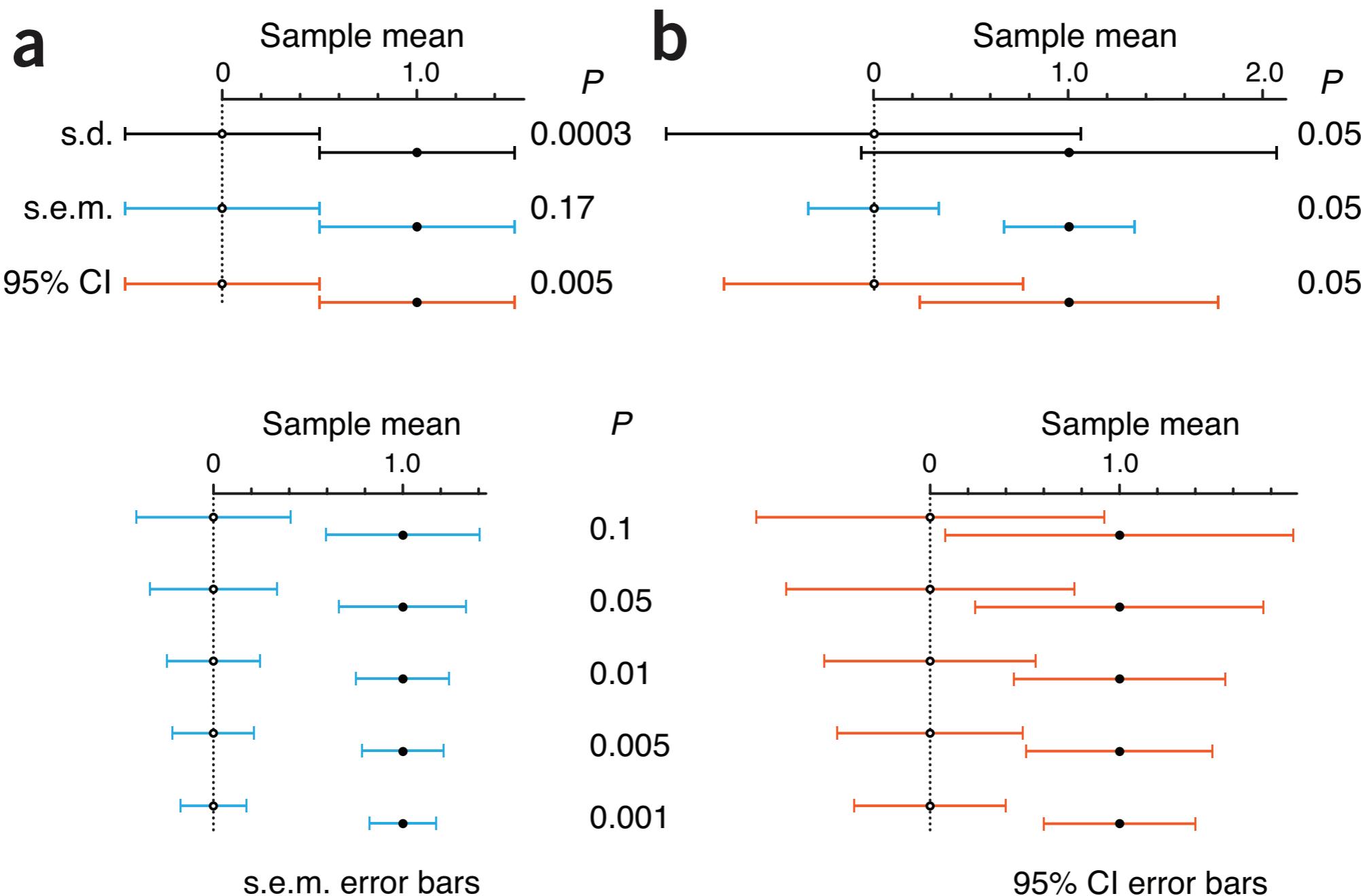
Width of the intervals depends on the sample size

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

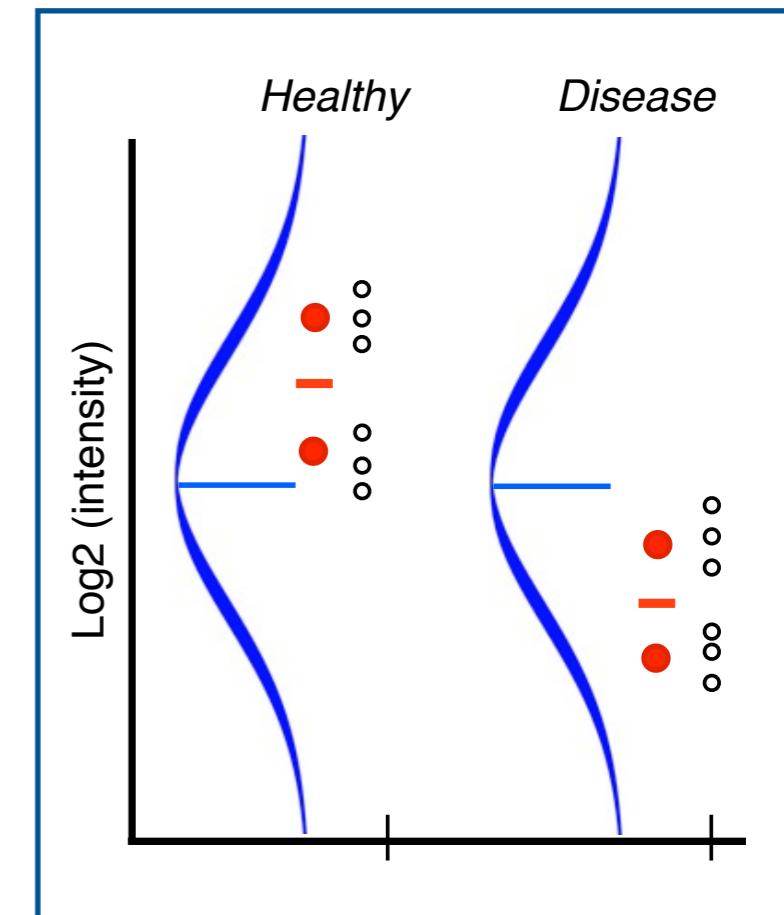
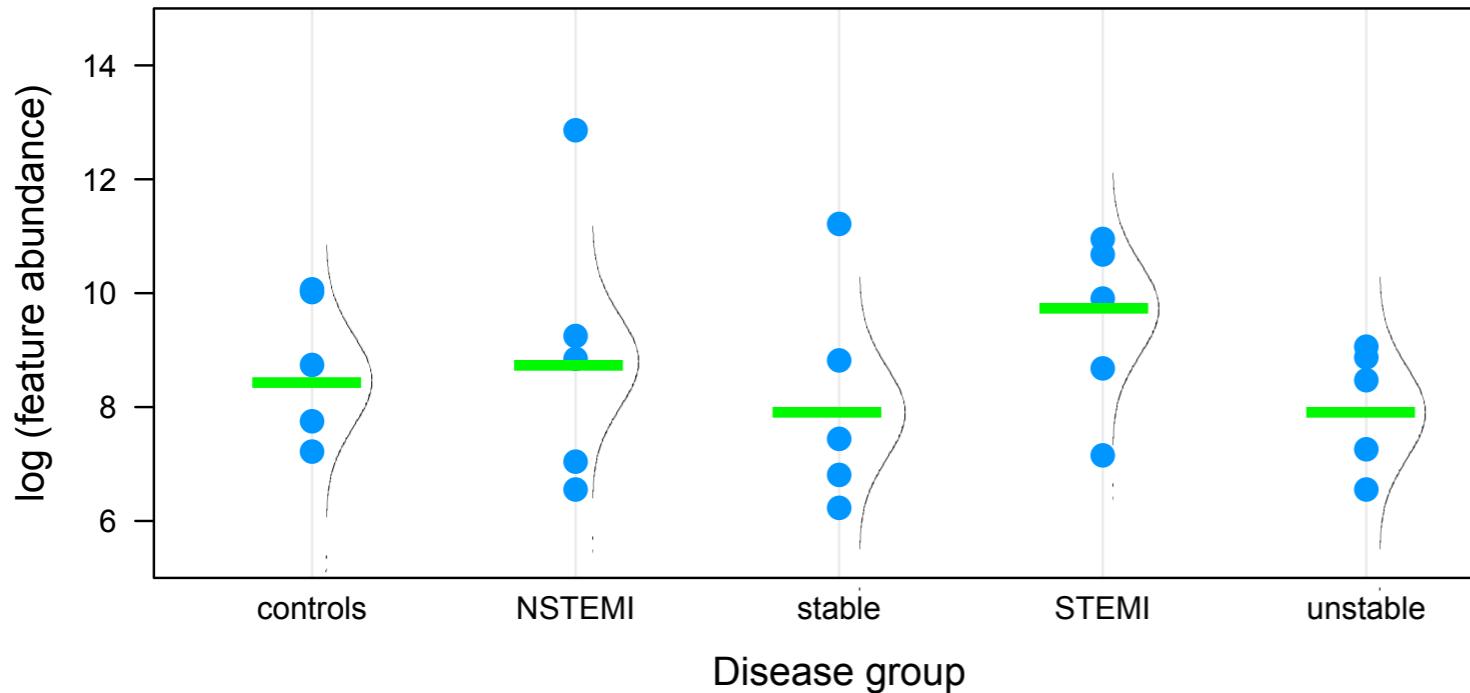


Simulated example

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

LINEAR MIXED MODELS DESCRIBE COMPLEX DESIGNS

Multiple conditions allow us to better learn the extent of variation



$$\text{Observed feature intensity} = \text{Systematic mean signal of disease group} + \text{Systematic/Random deviation of subject} + \text{Random deviation of measurement error}$$

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

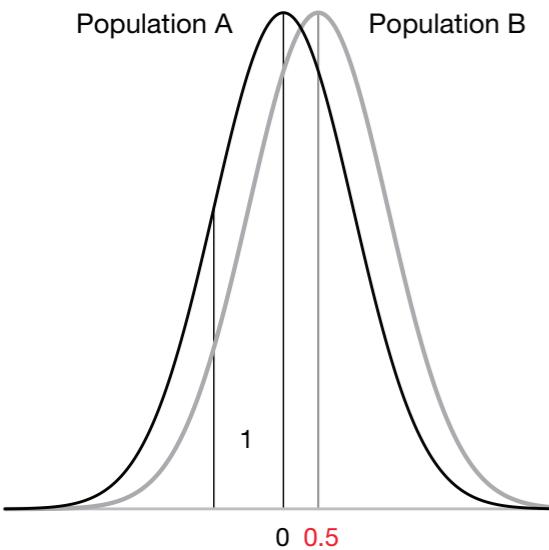
- P-values can indicate how incompatible the data are with a specified statistical model
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance
- Scientific conclusions and business policy decisions should not be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

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

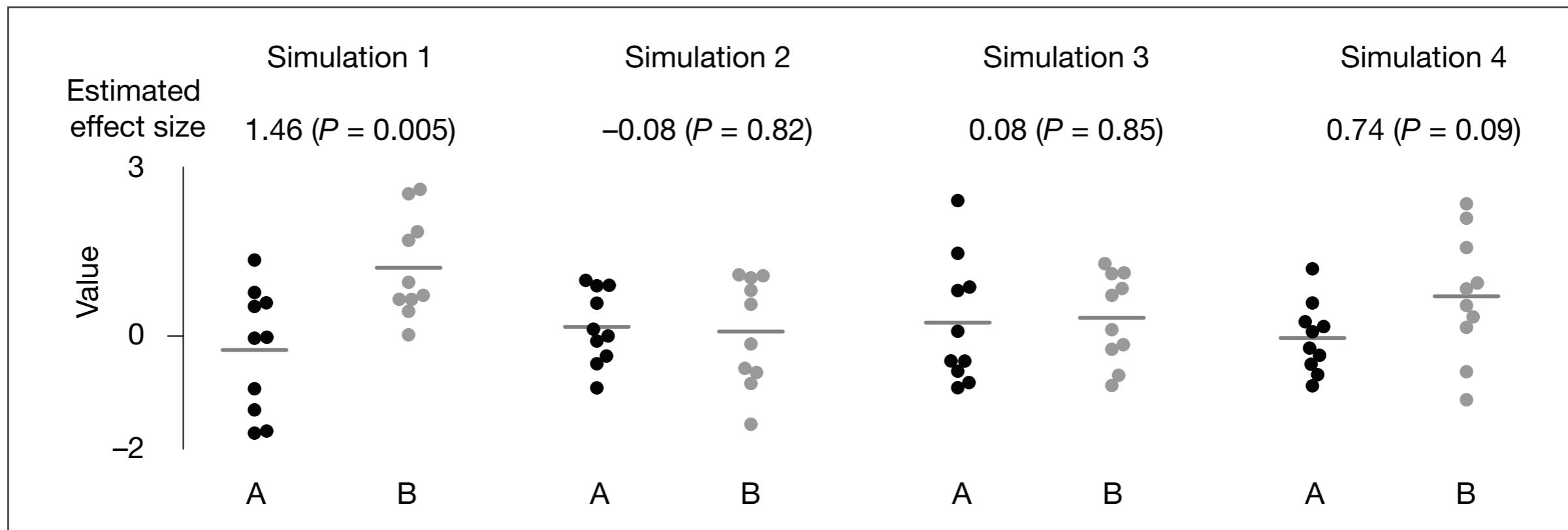
The American Statistician, February 2016

- 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



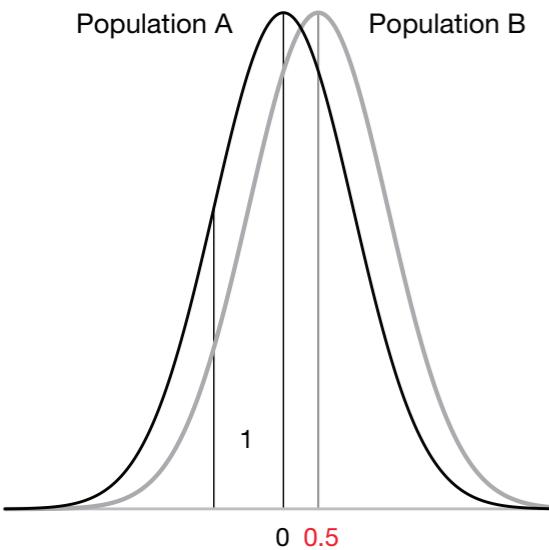
- Repeatedly sampling data leads to different results
- The problem worsens when testing many proteins
- Partial solutions:
 - Larger sample size
 - Adjustment for multiple testing



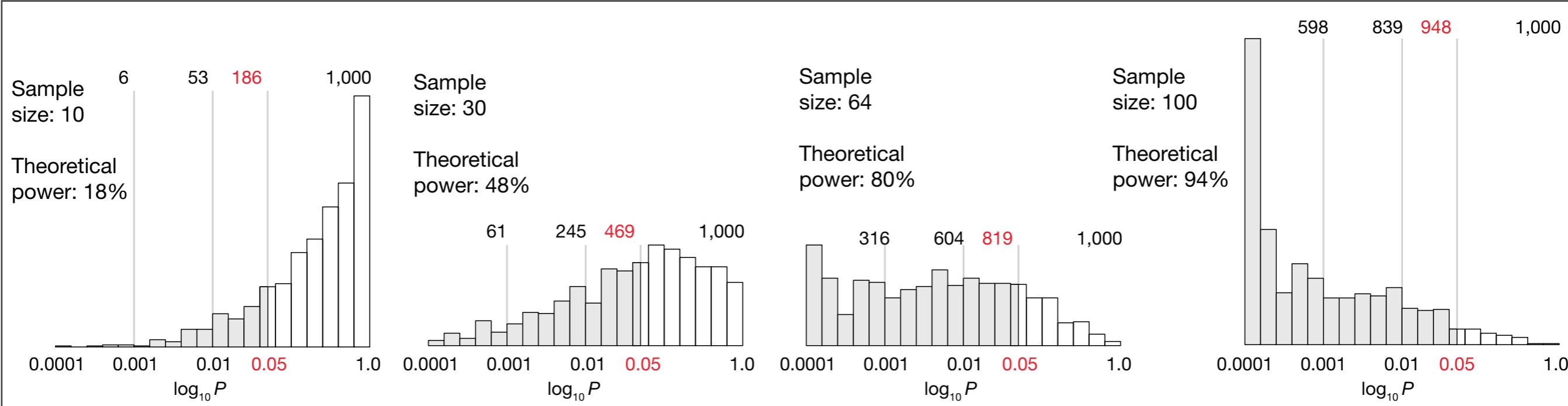
Simulated example

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

WITH SMALL SAMPLE SIZE, P-VALUES ARE UNSTABLE



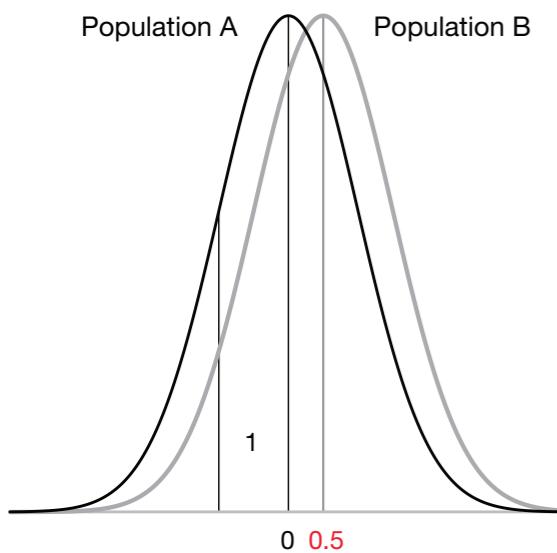
- Repeatedly sampling data leads to different results
- The problem worsens when testing many proteins
- Partial solutions:
 - Larger sample size
 - Adjustment for multiple testing



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

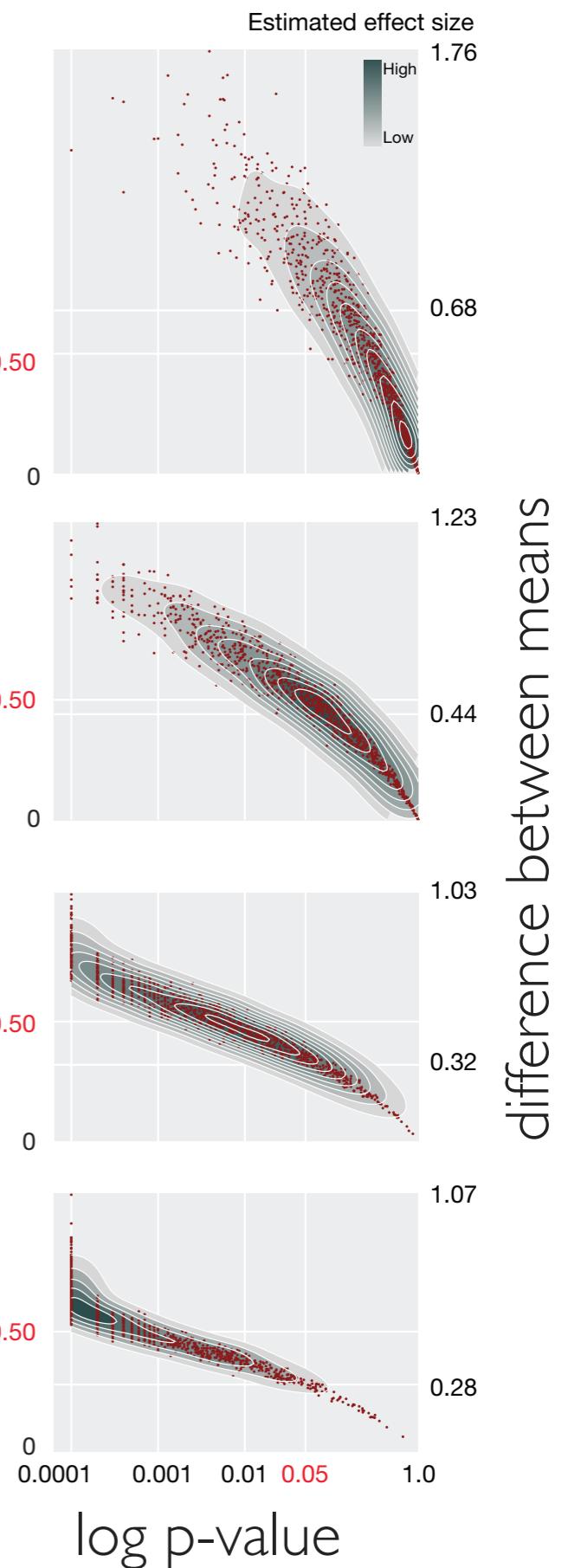
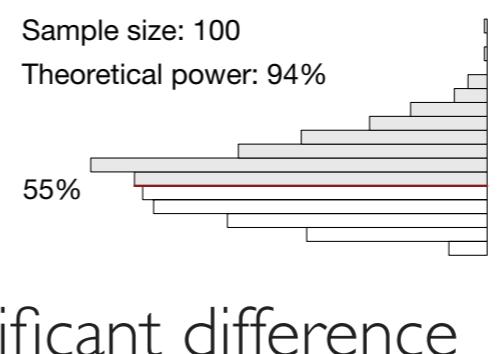
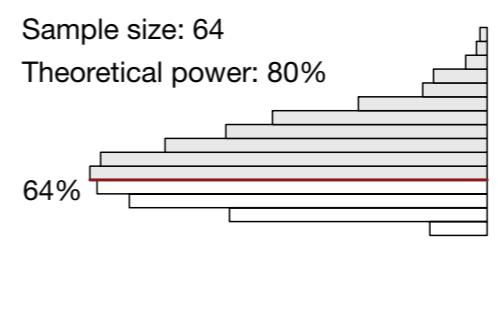
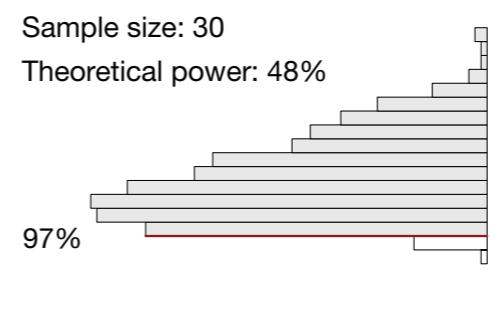
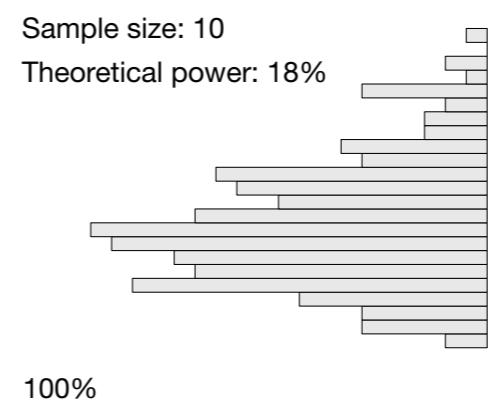
10 replicates

30 replicates

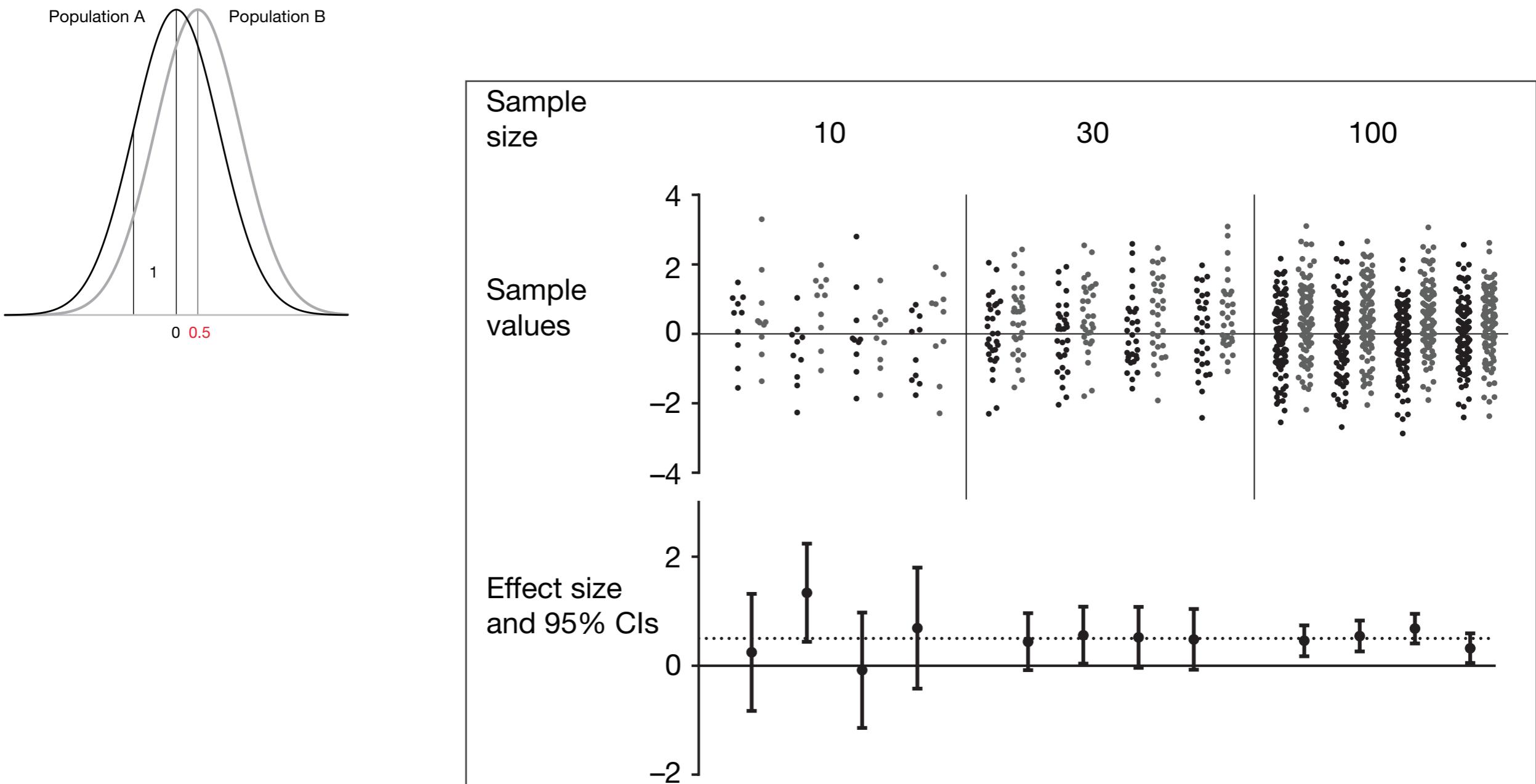
64 replicates

100 replicates

significant difference
between means



CONFIDENCE INTERVALS PROVIDE COMPLEMENTARY INSIGHT



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

PITFALL: OUTCOME SWITCHING

- Anti-depressant Paxil was studied for several main outcomes
 - None showed an effect
 - Some secondary outcomes did
- 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/>