DESIGN AND ANALYSIS OF QUANTITATIVE PROTEOMIC EXPERIMENTS

Intro to statistical methods and examples using Skyline

Course organizers

Brendan MacLean Senior Scientist, University of Washington
Lead developer of Skyline

Meena Choi PhD Candidate, Purdue University
Lead developer of MSstats

Olga Vitek, Associate Professor, Northeastern University

DESIGN AND ANALYSIS OF QUANTITATIVE PROTEOMIC EXPERIMENTS

Plan for the day

9:00-10:30	Intro to statistical methods. Intro to Skyline.
10:30-10:45	Refreshments.
10:45-12:00	Case study in Skyline.
12:00-1:00	Lunch break on your own.
1:00-2:30	Case study in Skyline. Introduction to MSstats.
2:30-2:45	Refreshments
2:45-4:00	Case study in MSstats, open time for questions

EXPERIMENTAL DESIGN AND BASIC STATISTICS

Olga Vitek

College of Science

College of Computer and Information Science

Northeastern University

WHY STATISTICS?

- Variation and uncertainty are unavoidable
  - *Technical variation:* sampling handling, storage, processing
  - *Instrumental variation:* elution time, ion suppression
  - *Signal processing:* peak boundaries, identity, intensity
  - *Biological variation:* variation in protein abundance

"Statistics: a body of methods for making wise decisions in the face of uncertainty." (W. A. Wallis)

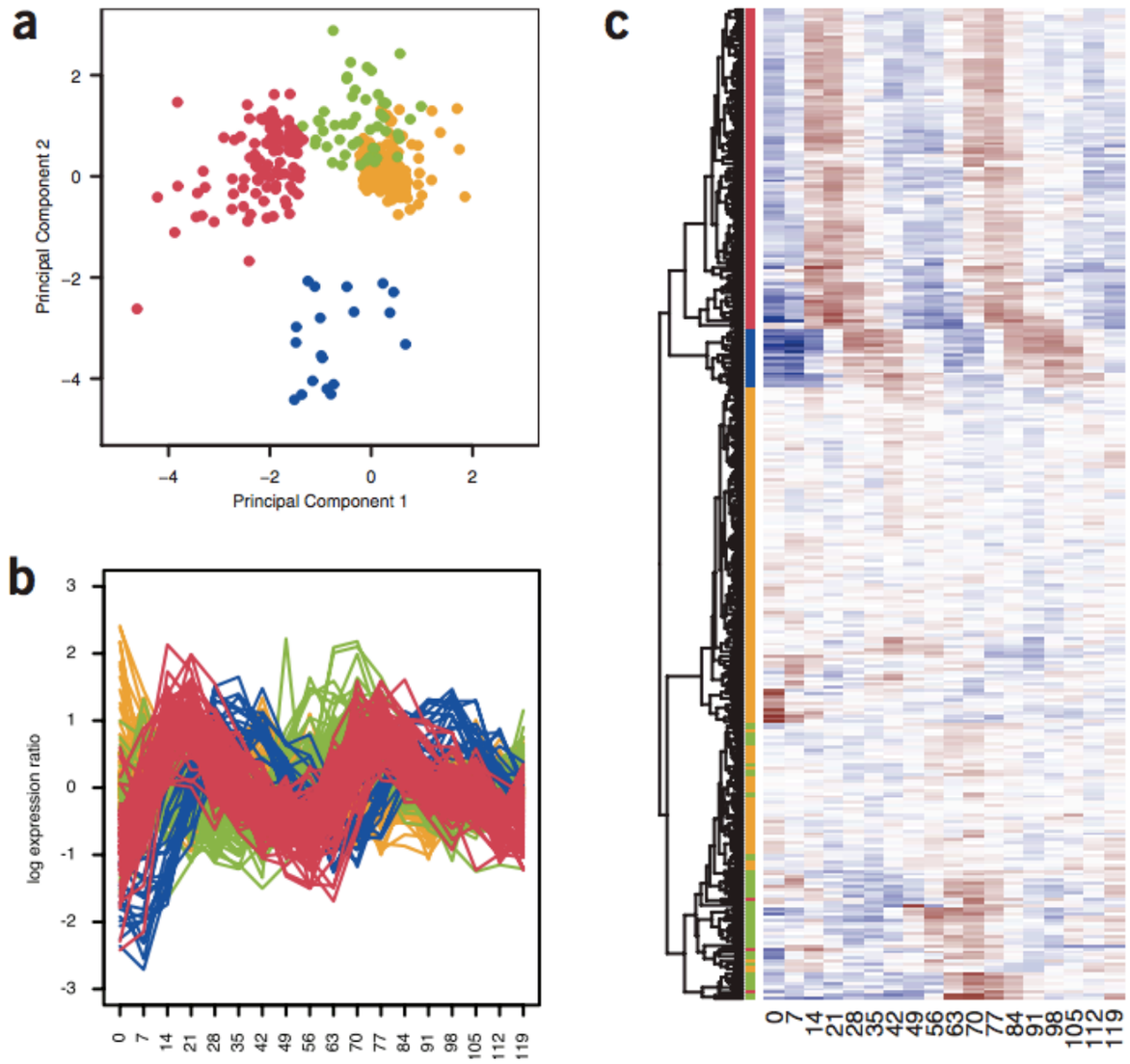
OUTLINE

- Translate scientific question into statistics
  - Statistical terms for 'biomarker' (or 'signature')
- Experimental design
  - Replication, randomization, blocking
- Basic data analysis
  - Simple summaries and models

STATISTICAL GOAL 1: CLASS DISCOVERY

Discover proteins or subjects with similar patterns

- No known class labels
  - E.g., no 'healthy' or 'disease'
  - All variation treated equally
  - No error rates
- Can't find something meaningful if unsure what we look for
  - Best used for visualization

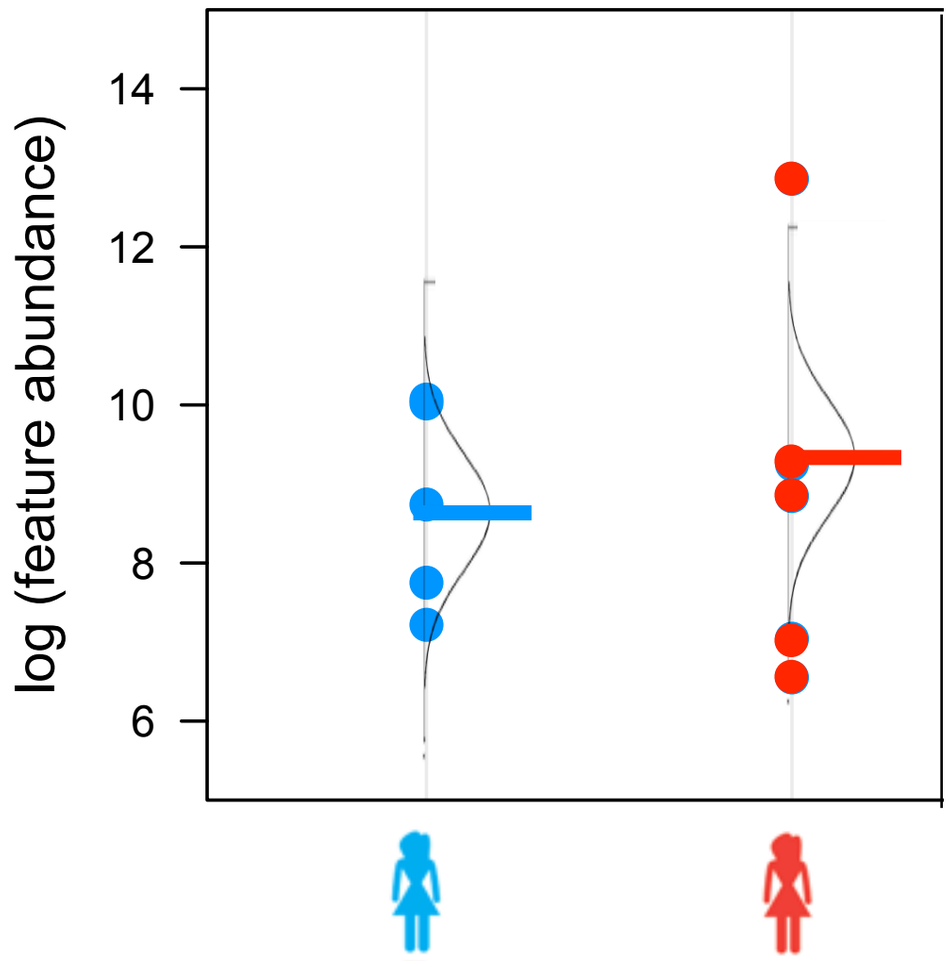


Gehlenborg *et al*, Nature Methods, 2010

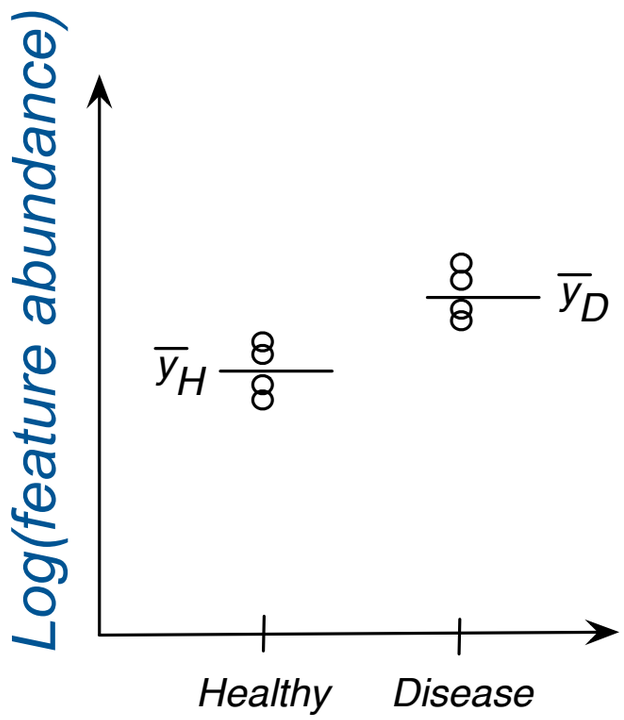
STATISTICAL GOAL 2: CLASS COMPARISON

Compare mean abundances in subject groups

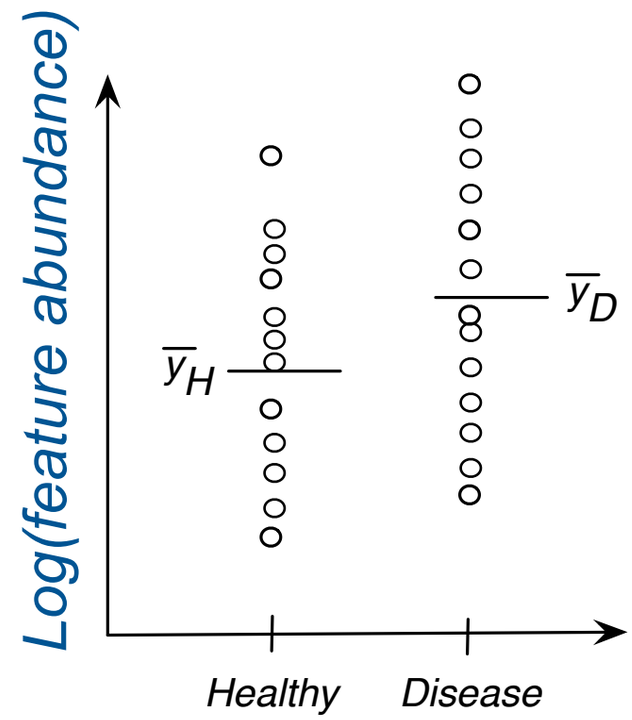
- Known class labels
  - Compare group averages
  - Report p-values, posterior probabilities etc
- Useful when compare groups of subjects
  - Best used for basic biology
  - Initial (Tier III) biomarker discovery screen



DIFFERENTIALLY ABUNDANT PROTEINS ARE NOT ALWAYS BIOMARKERS



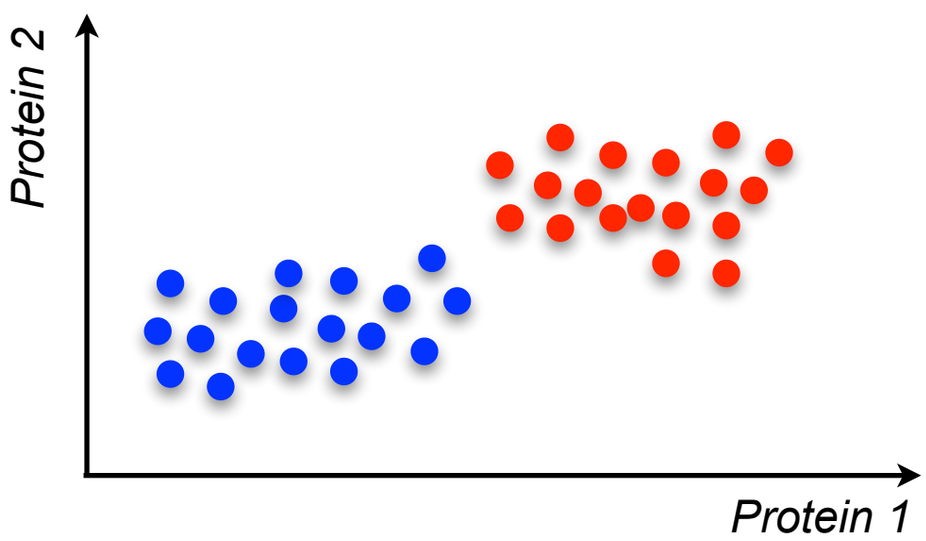
Differentially abundant and predictive



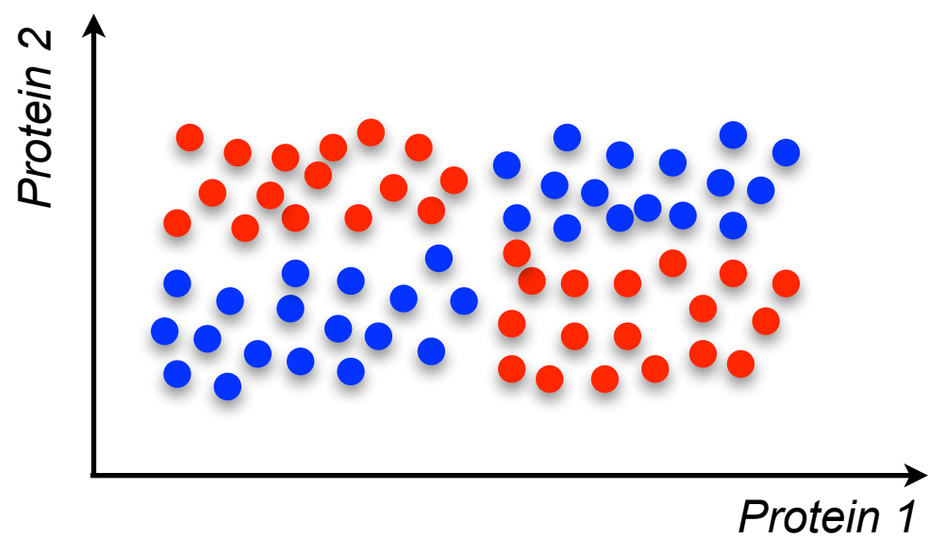
Differentially abundant and not predictive

Single protein: *Differentially abundant* → ✗ *Predictive*

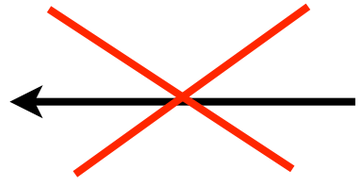
BIOMARKER PROTEINS ARE NOT ALWAYS DIFFERENTIALLY ABUNDANT



Differentially abundant and predictive



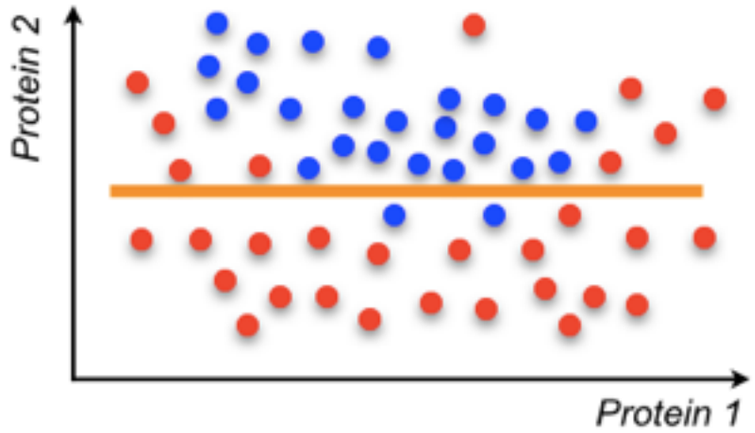
Not differentially abundant but predictive

Single protein: *Differentially abundant*  *Predictive*

STATISTICAL GOAL 3: CLASS PREDICTION

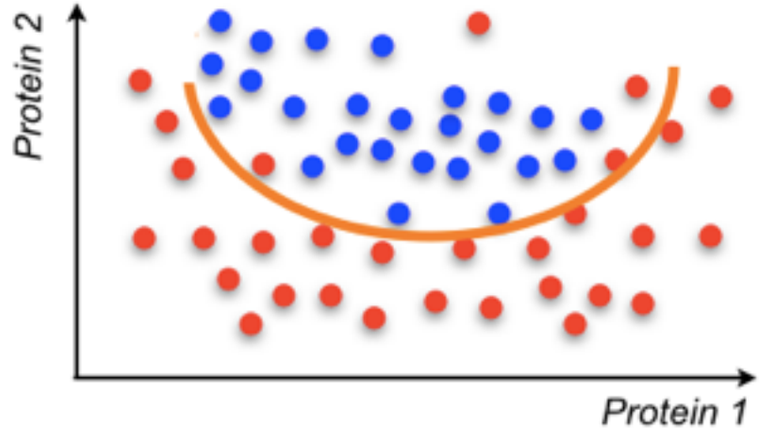
Classify each subject into a known group

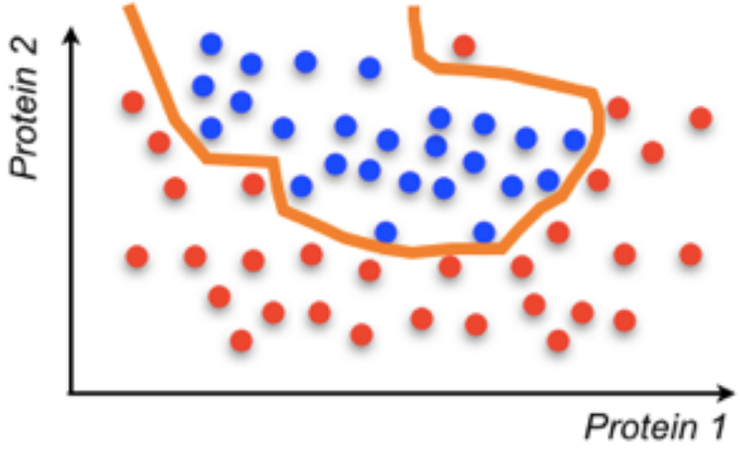
- Known class labels
  - Predict individual subjects
  - Report misclassification error (sensitivity, specificity, predictive value etc)
- Useful when focus on an individual
  - Tier I or Tier II biomarker discovery studies



Linear prediction rule

Quadratic prediction rule



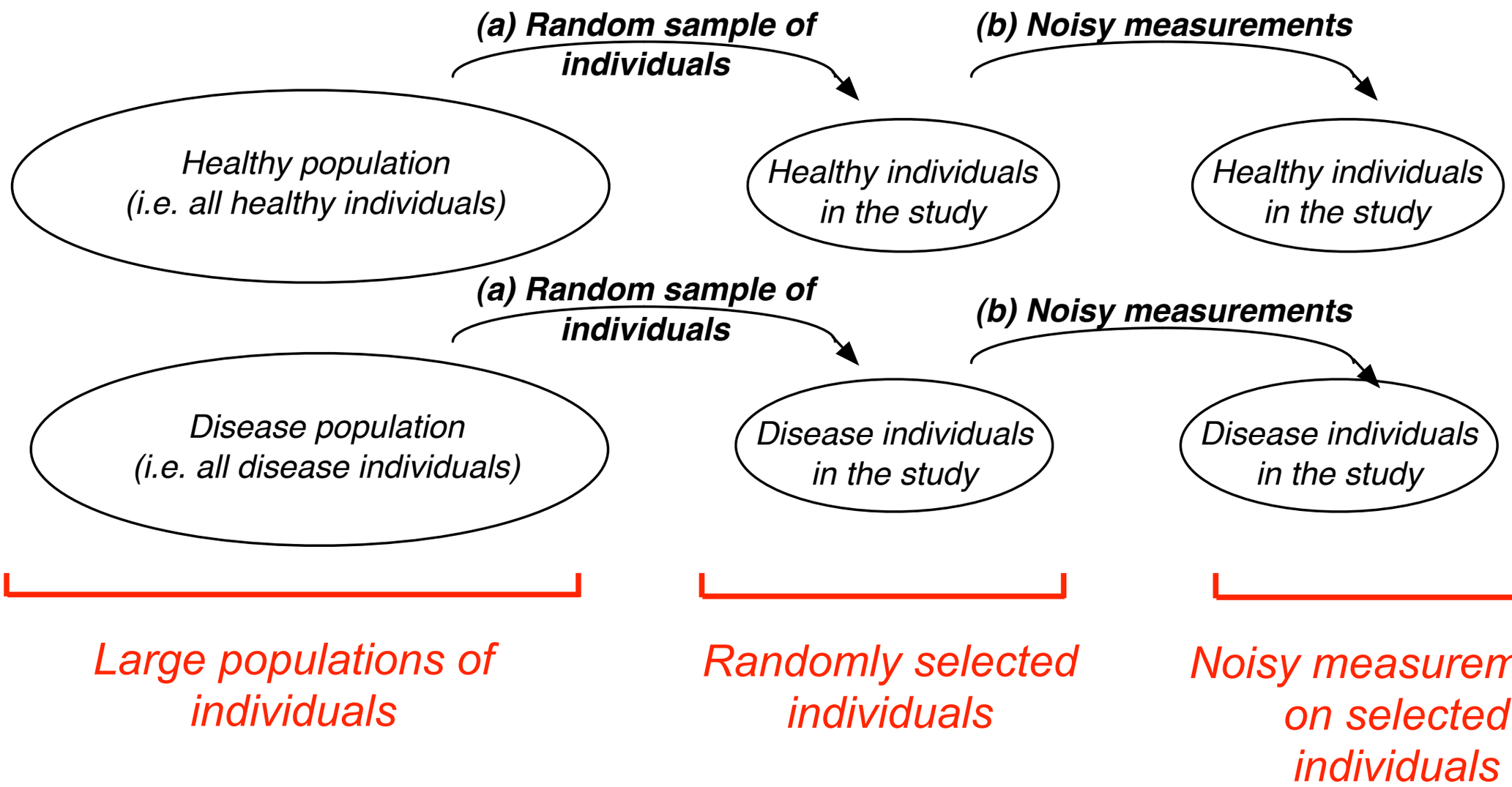


General prediction rule

OUTLINE

- Translate scientific question into statistics
  - Statistical terms for 'biomarker' (or 'signature')
- Experimental design
  - Replication, randomization, blocking
- Basic data analysis
  - Simple summaries and models

A STATISTICIAN'S VIEW OF THE EXPERIMENT

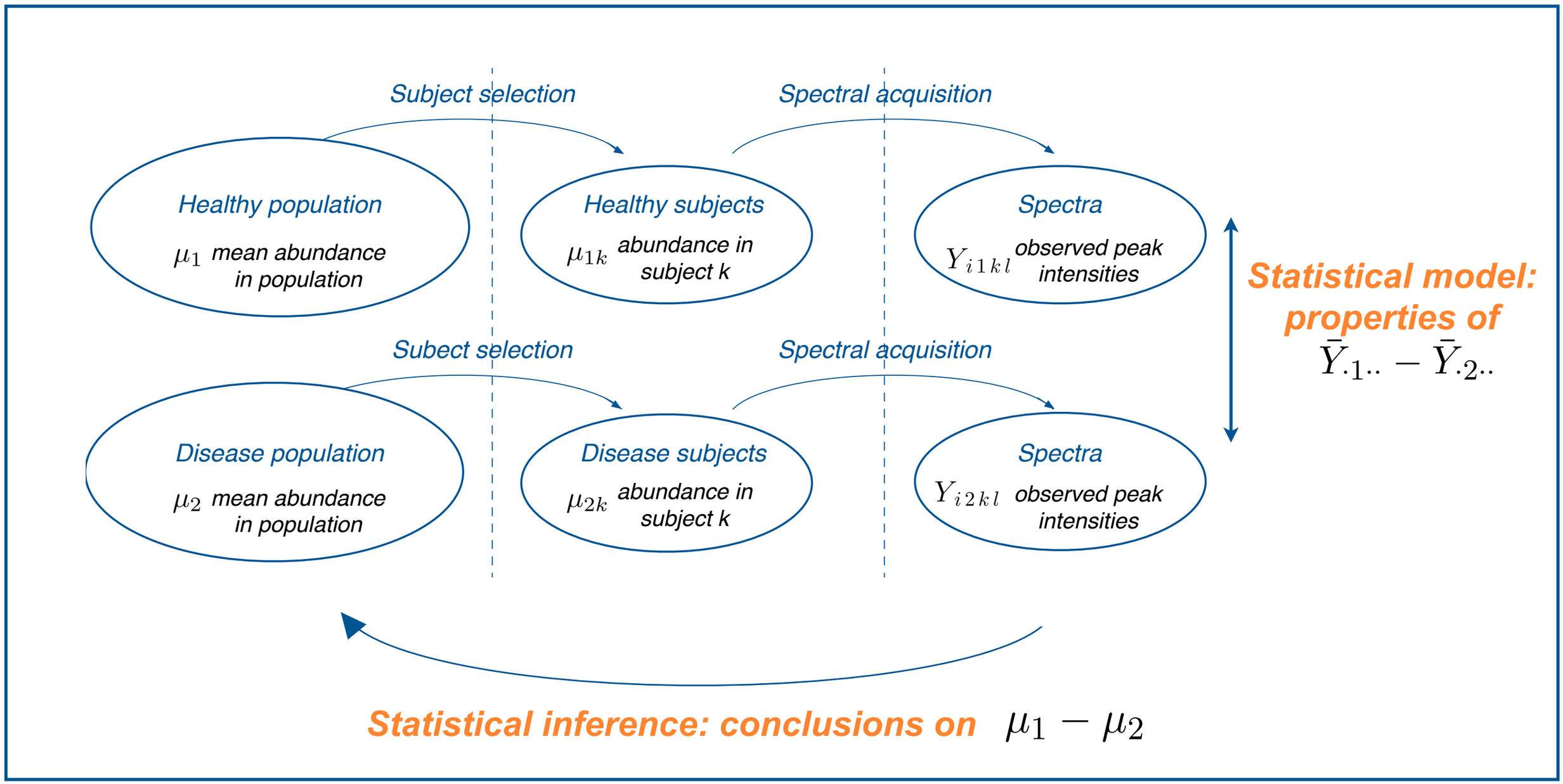


Dangers:

Bias: conclusions systematically differ from truth

Inefficiency: unnecessary variation in the data

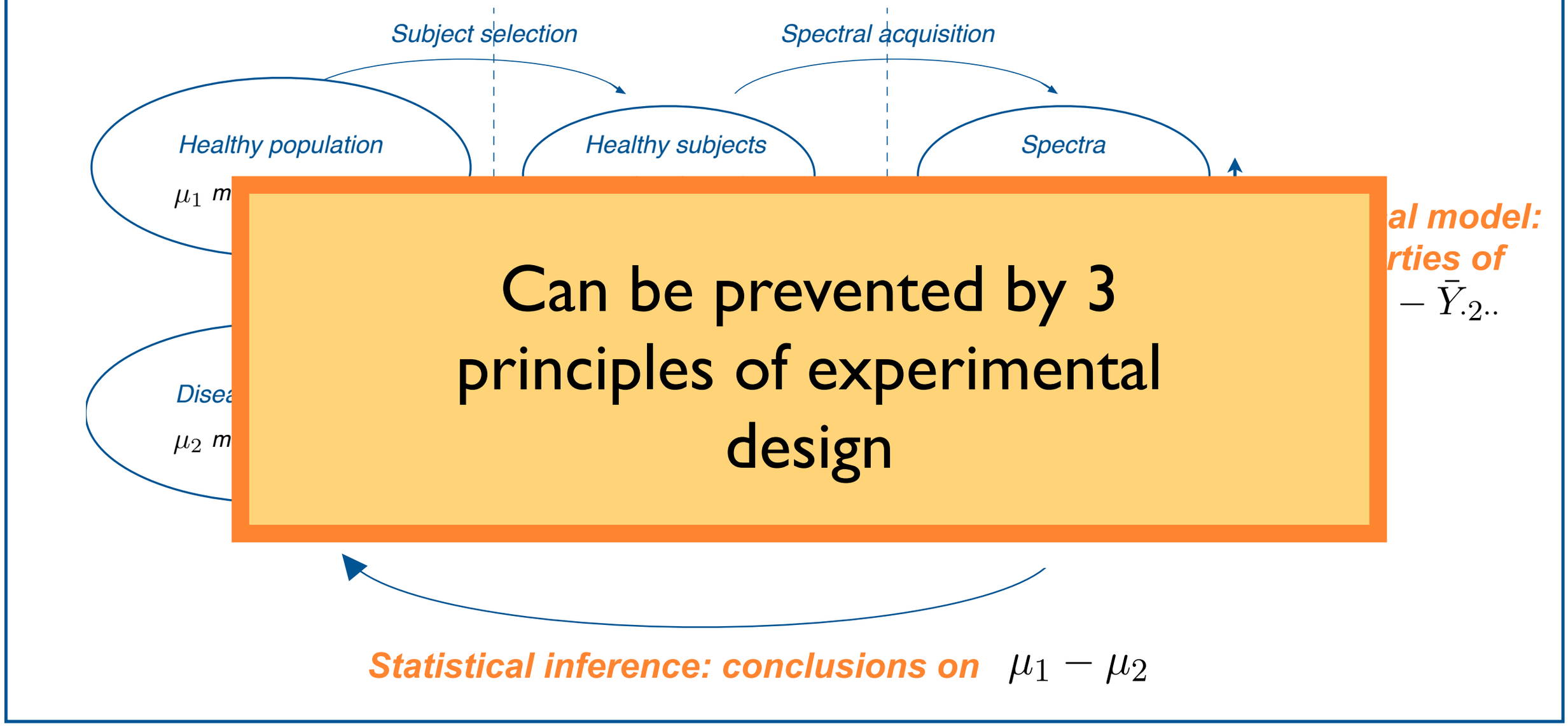
DEFINITION OF BIAS AND INEFFICIENCY



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

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

DEFINITION OF BIAS AND INEFFICIENCY

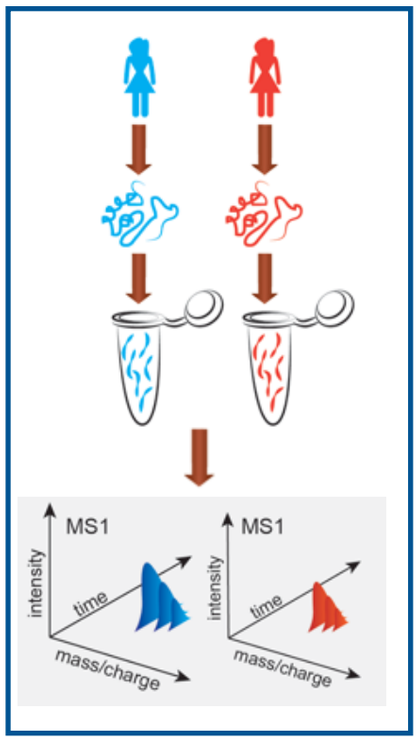


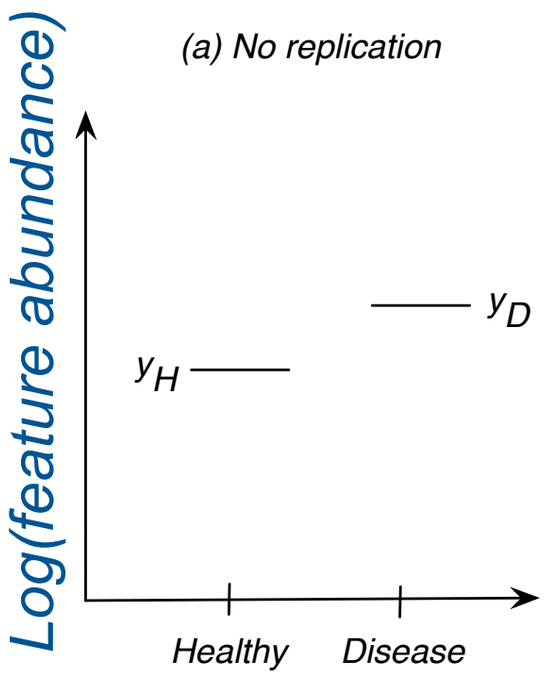
Bias: $\bar{Y}_{.1..} - \bar{Y}_{.2..}$ *systematically different from* $\mu_{1k} - \mu_{2k}$

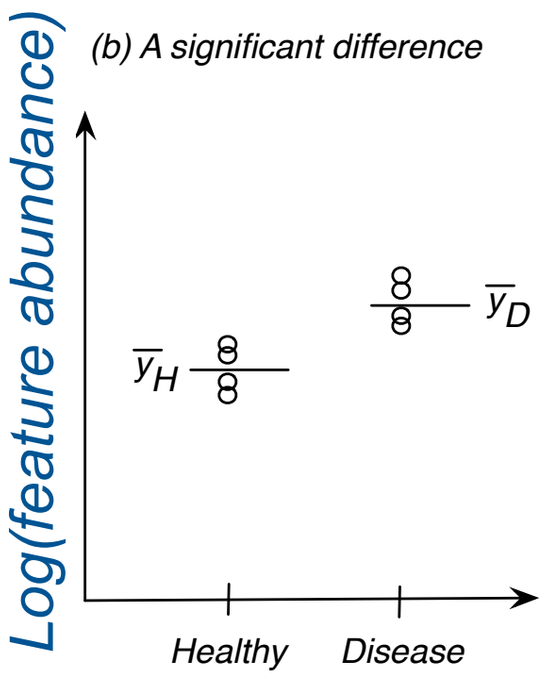
Inefficiency: *Large* $Var(\bar{Y}_{.1..} - \bar{Y}_{.2..})$

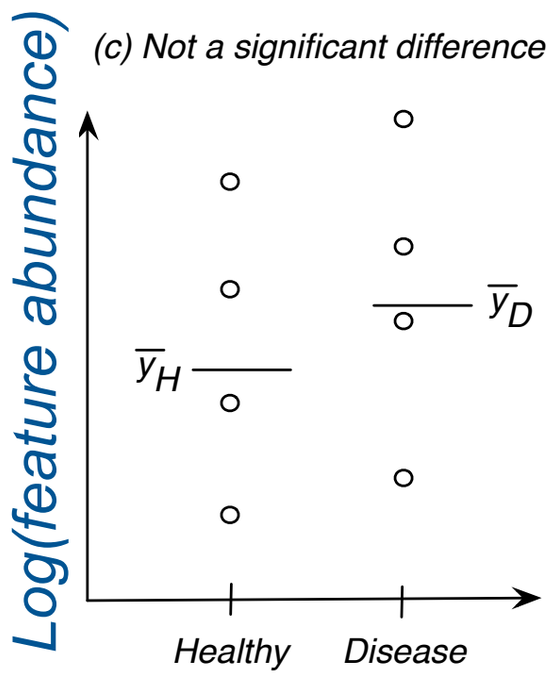
PRINCIPLE 1: REPLICATION

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









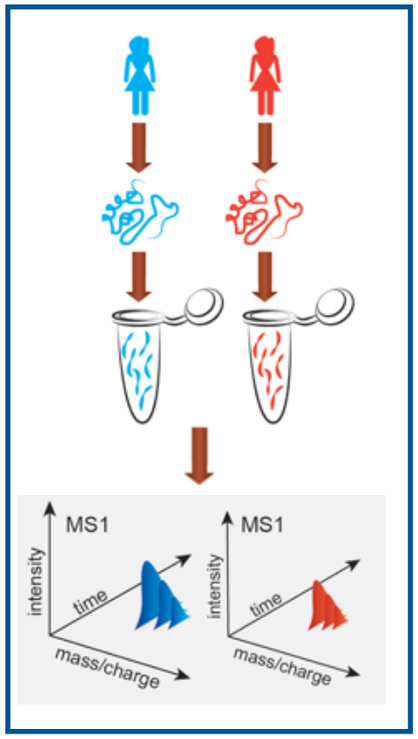
Two levels of randomness imply two types of replication:

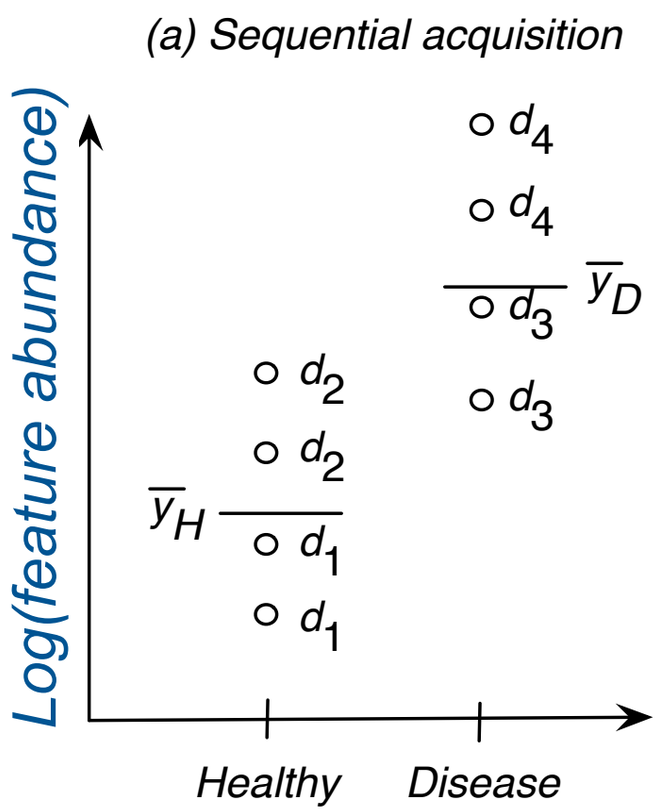
- *Biological replicates:* selecting multiple subjects from the population
- *Technical replicates:* multiple runs per subject

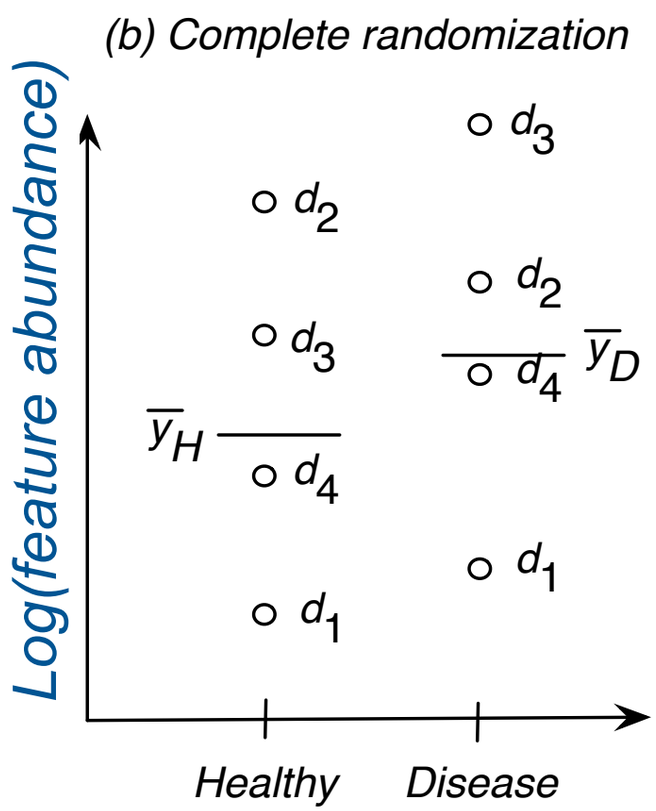
Oberg and Vitek, *J. Proteome Research*, 8, 2009

PRINCIPLE 2: RANDOMIZATION

Prevents bias







No randomization
= confounding
= bias

Complete randomization
= no bias

Two levels of randomness imply two types of randomization:

- *Biological replicates:* random selection of subjects from the population
- *Technical replicates:* random allocation of samples to all processing steps

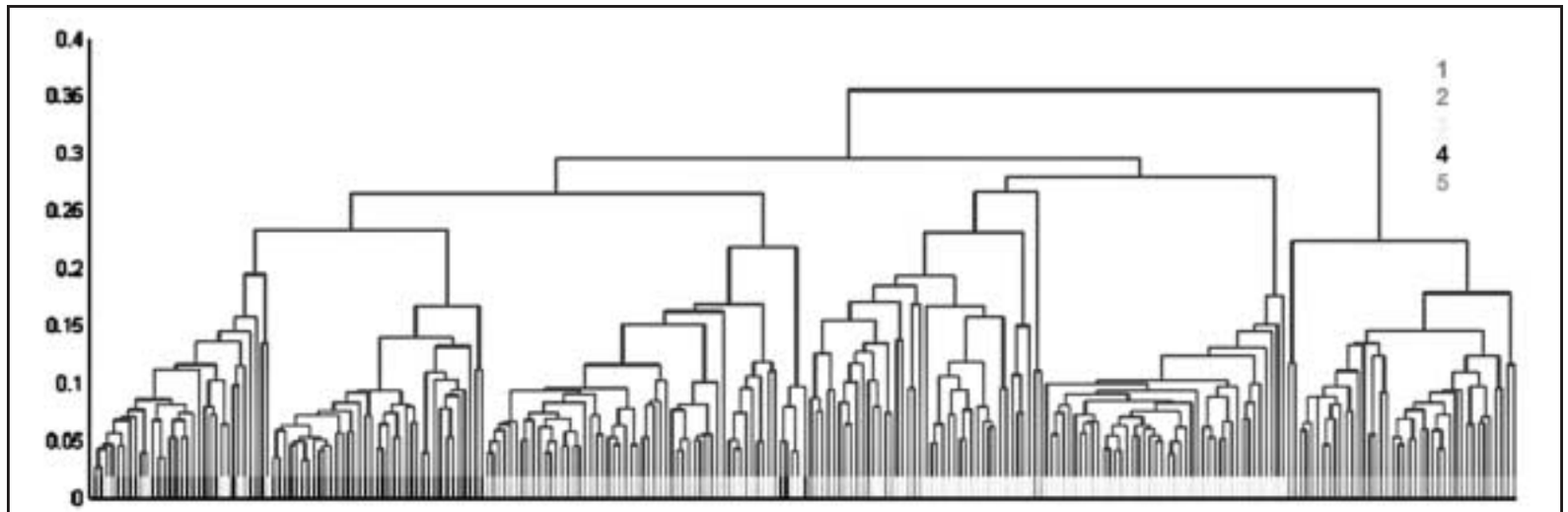
Oberg and Vitek, *J. Proteome Research*, 8, 2009

EXAMPLE: LACK OF RANDOMIZATION

Hu, Coombes, Morris, Baggerly, Briefings in Functional Genomics, 2005

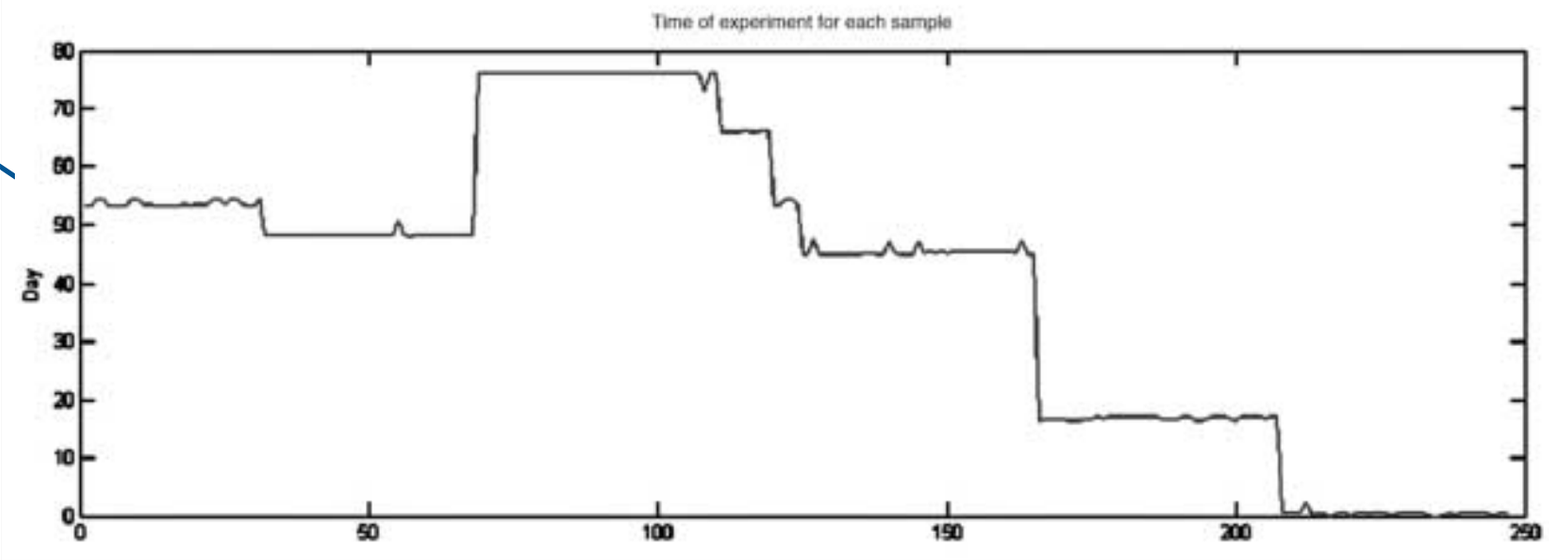
- Serum samples with five types of cancer
- SELDI-TOF MS
  - normalized, peak picked

Hierarchical clustering of samples



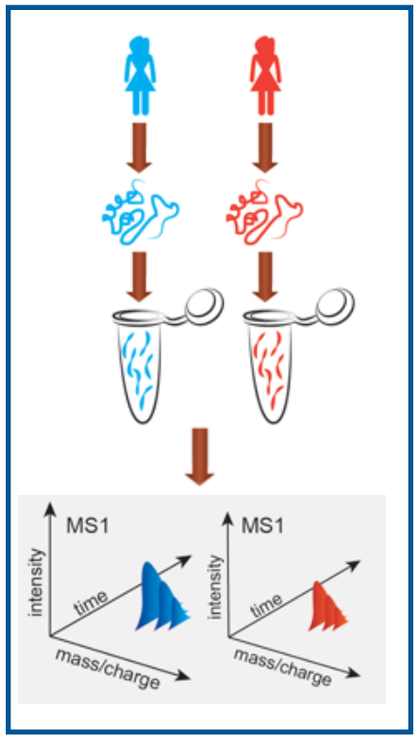
Cancer subtype confounded with time

Same time-based clustering on the QC samples!

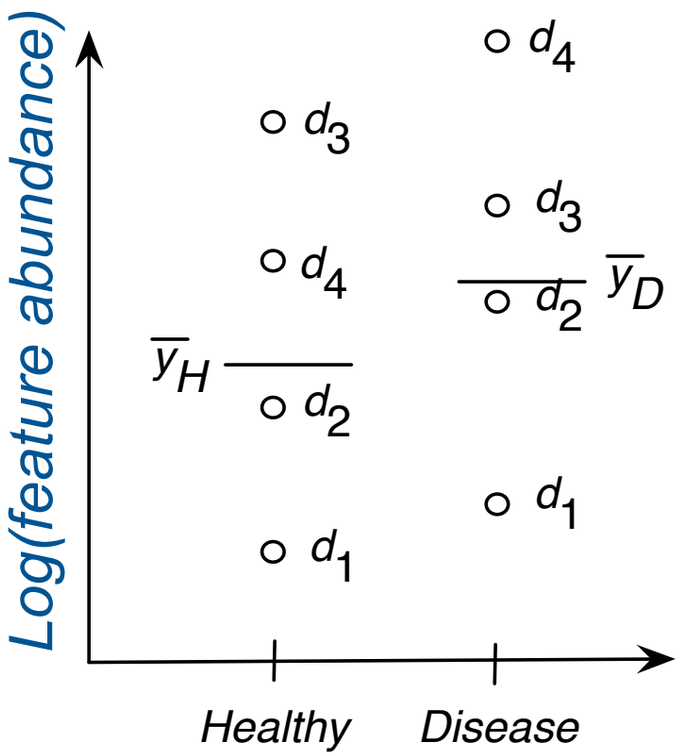


PRINCIPLE 3: BLOCKING

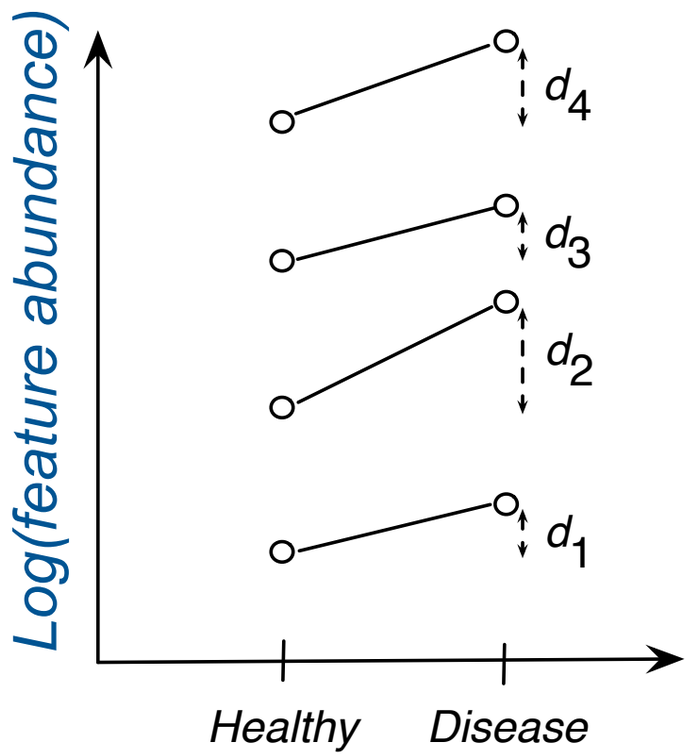
Helps reduce both bias and inefficiency



(b) Complete randomization



(c) Day = block



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

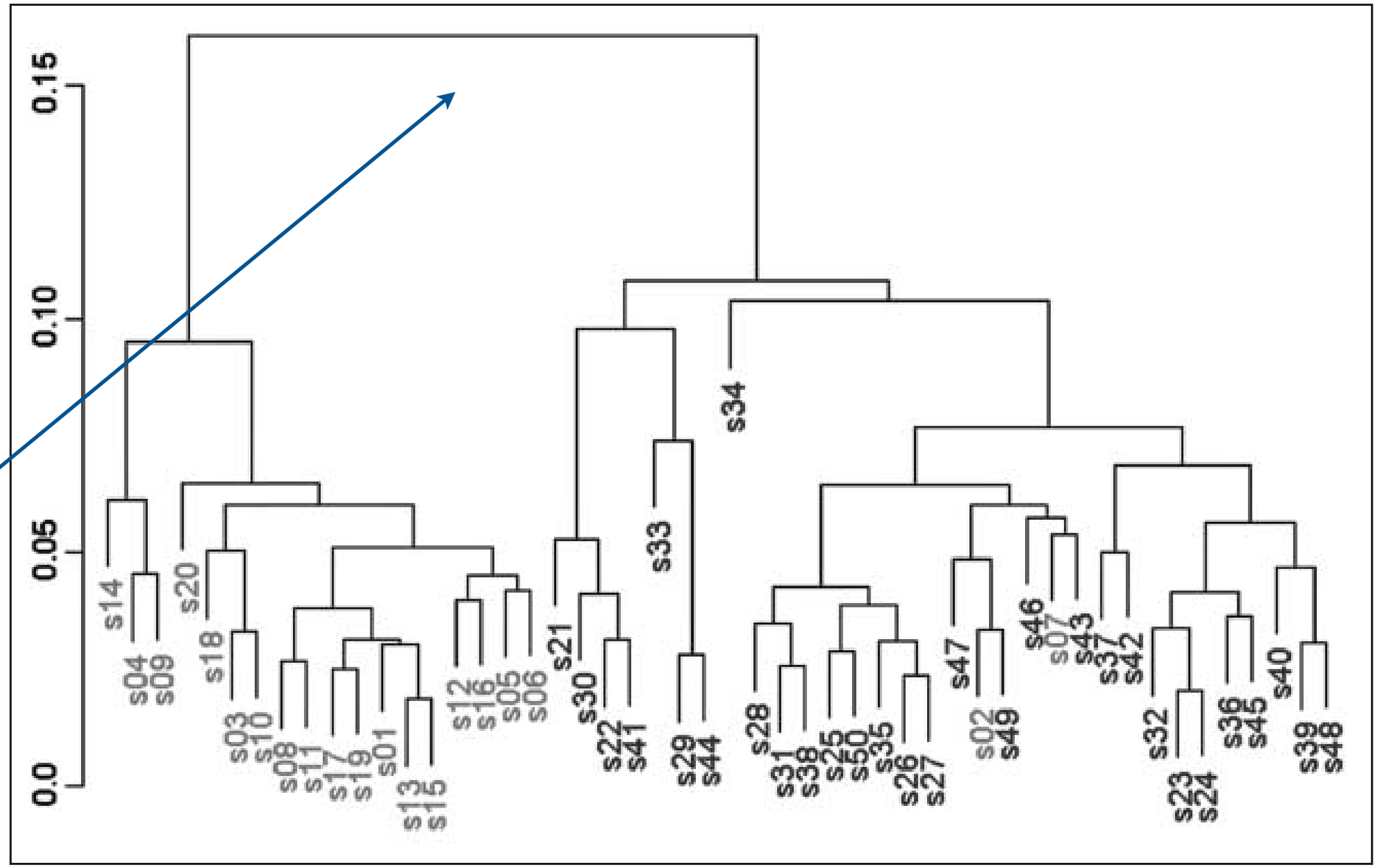
EXAMPLE: LACK OF BLOCKING

Hu, Coombes, Morris, Baggerly, Briefings in Functional Genomics, 2005

- Serum samples with two types of cancer
- SELDI-TOF MS, 3 fractions
  - normalized, peak picked

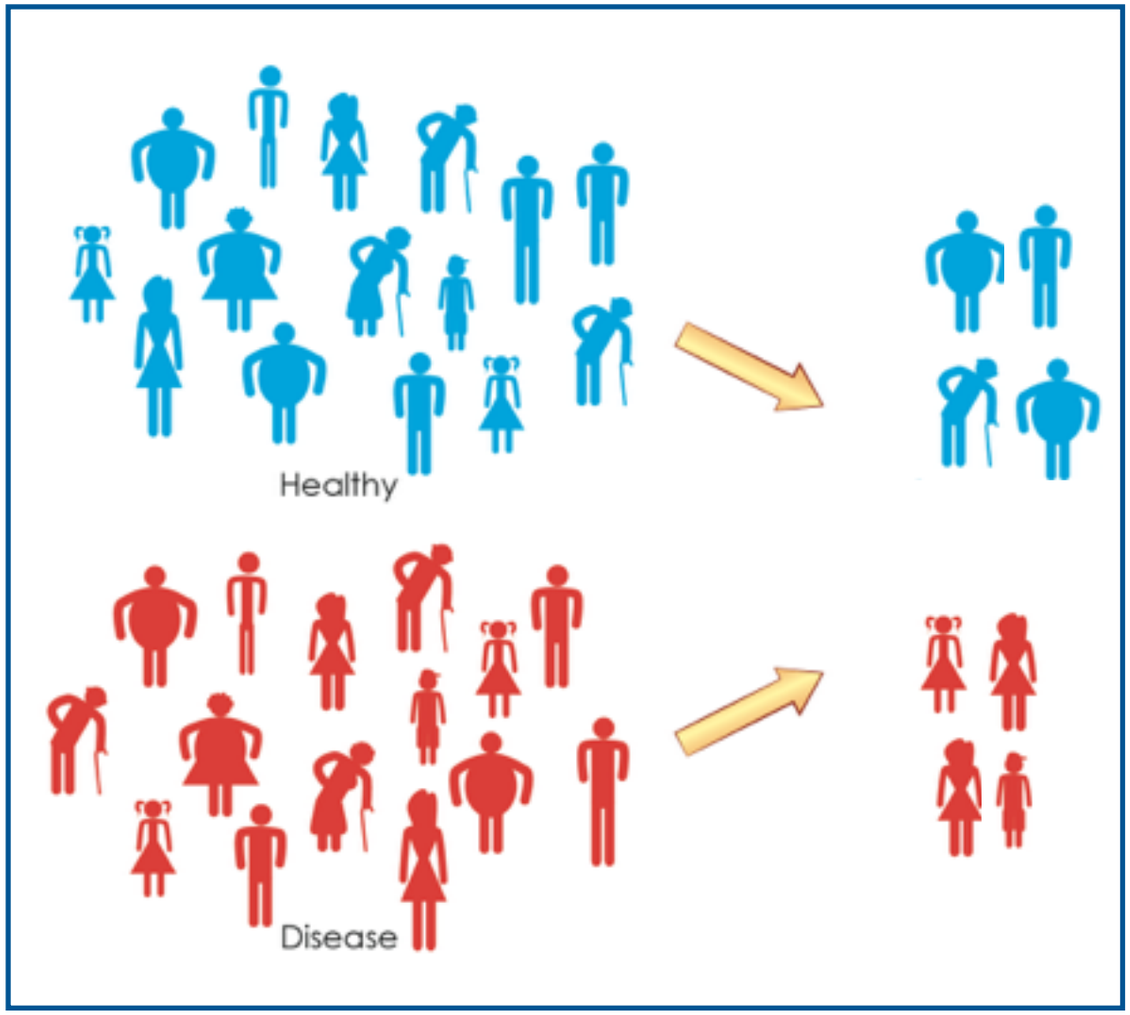
Hierarchical clustering of samples

Protocol change

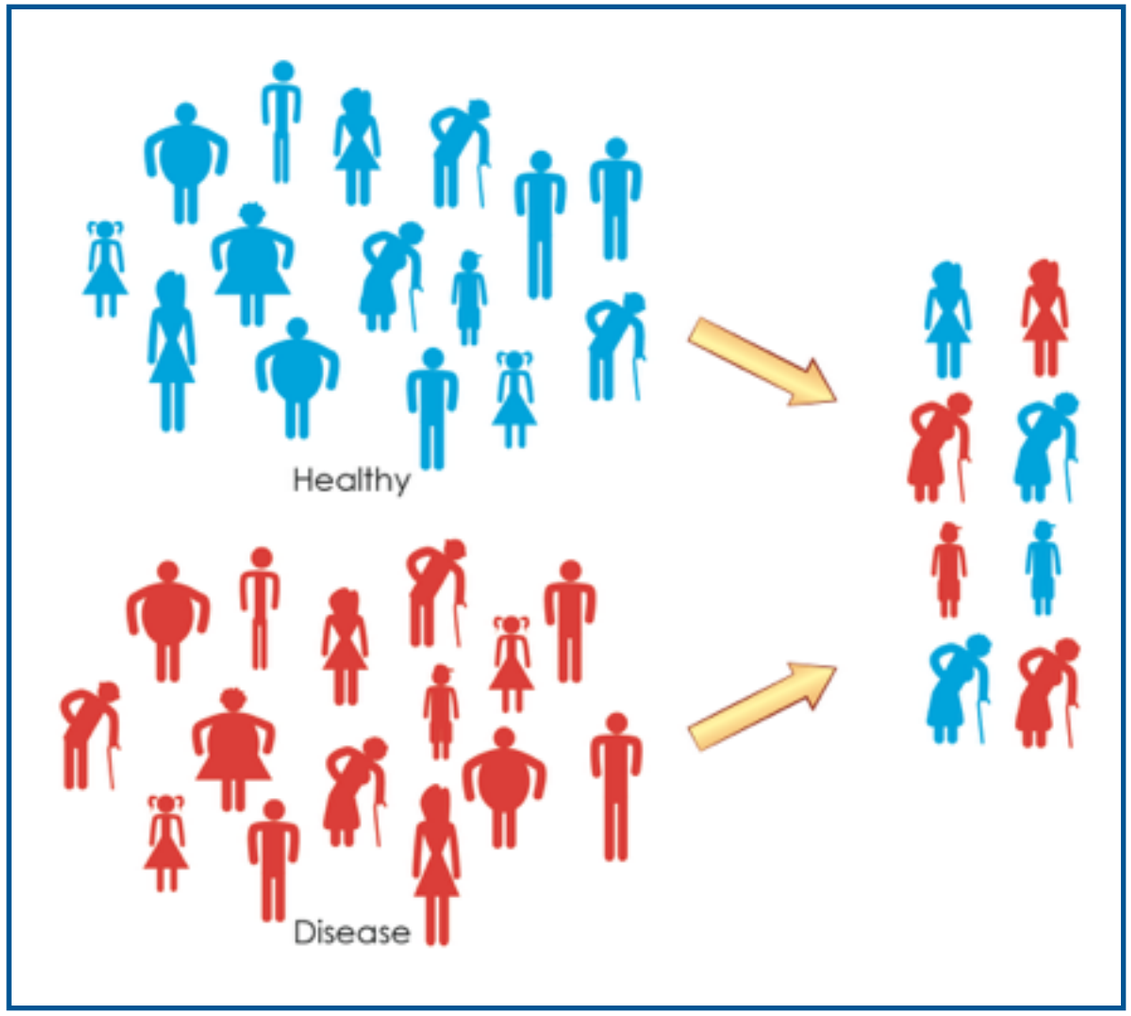


MATCHING

Blocking with respect to biological risk factors



Complete randomization
= inflated variance



Block-randomization
= restriction on randomization
= systematic allocation

Käll and Vitek, *PLoS Computational Biology*, 7, 2011

EXAMPLE

Block-randomized selection of subjects from repository

		Disease group				
		Control	Stable angina	Unstable angina	NSTEMI	STEMI
Stratification	$\geq$ 58 y.o; Female	354	300	49	39	29
	$\geq$ 58 y.o; Male	701	843	143	86	54
	$<$ 58 y.o; Female	80	56	5	5	8
	$<$ 58 y.o; Male	264	190	34	23	27

Counts in the initial repository of samples

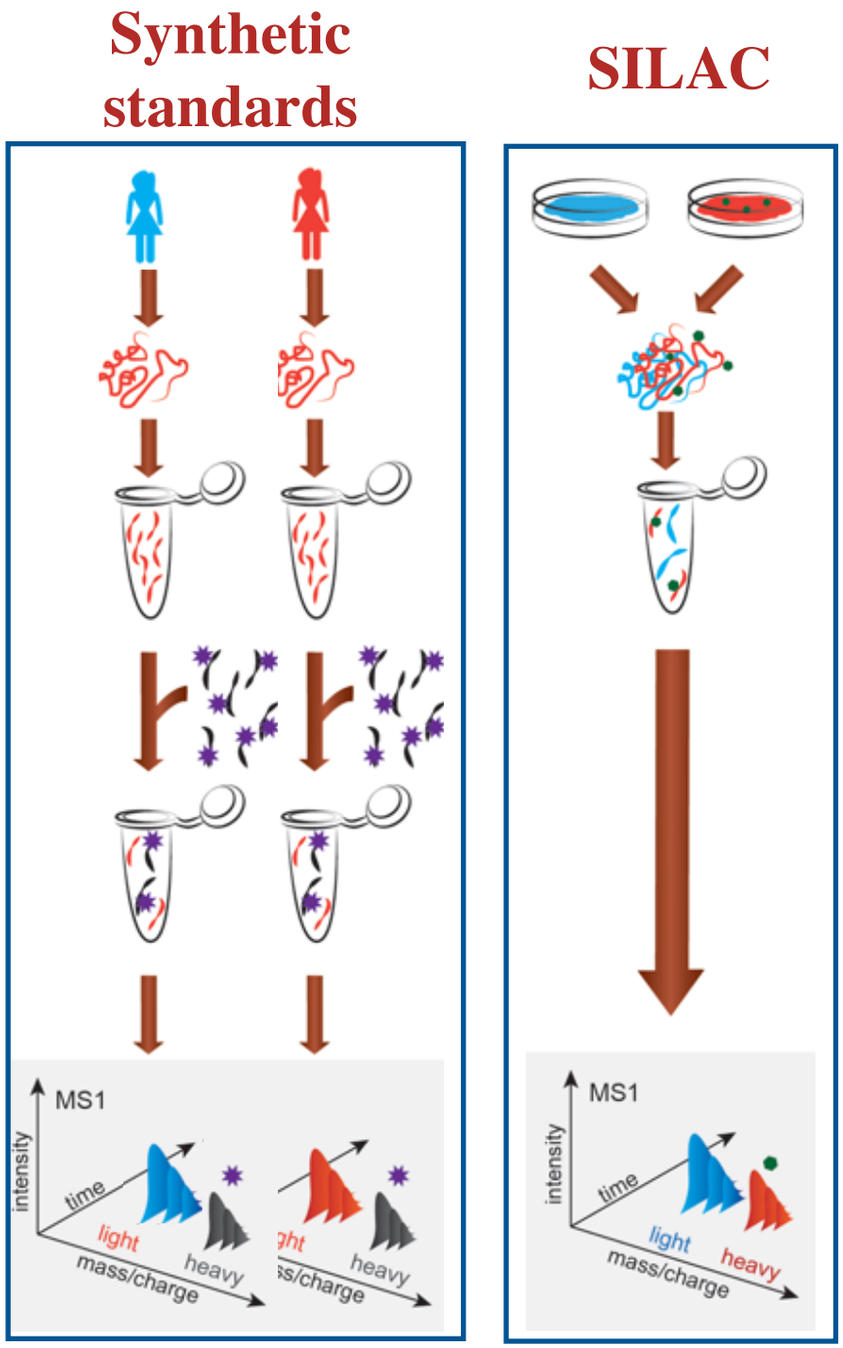
		Disease group				
		Control	Stable angina	Unstable angina	NSTEMI	STEMI
Stratification	$\geq$ 58 y.o; Female	3	3	3	3	3
	$\geq$ 58 y.o; Male	3	3	3	3	3
	$<$ 58 y.o; Female	2	2	2	2	2
	$<$ 58 y.o; Male	2	2	2	2	2

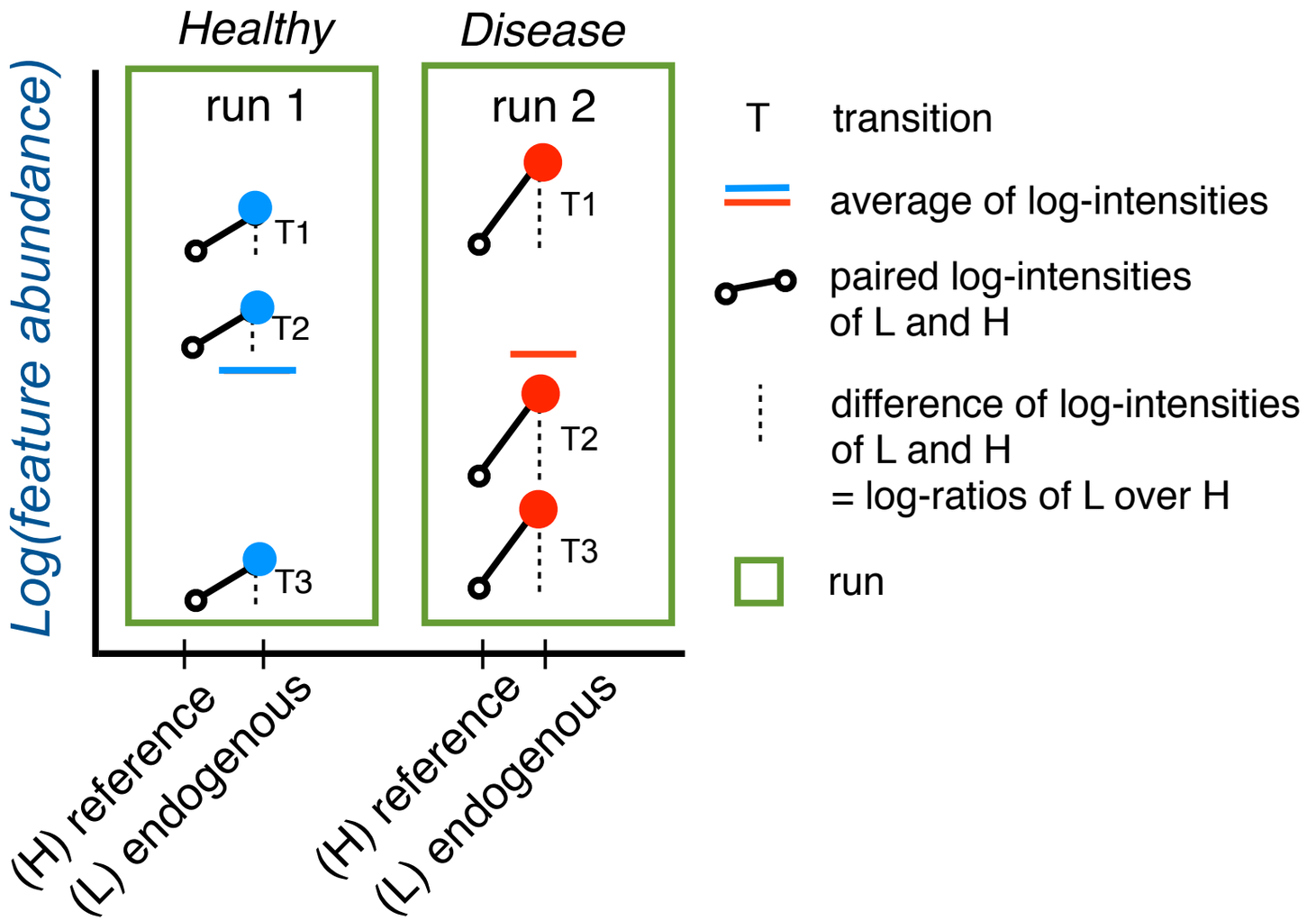
Counts of subjects included in the study

Mass spectra acquired without technical replication

MULTIPLEXING

Blocking with respect to mass spectrometry run





Multiplexing reduces both bias and variance

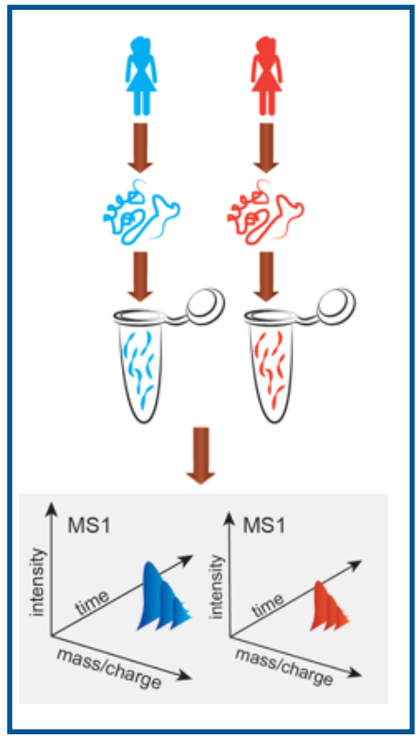
(assuming that extra sample handling does not introduce extra variation)

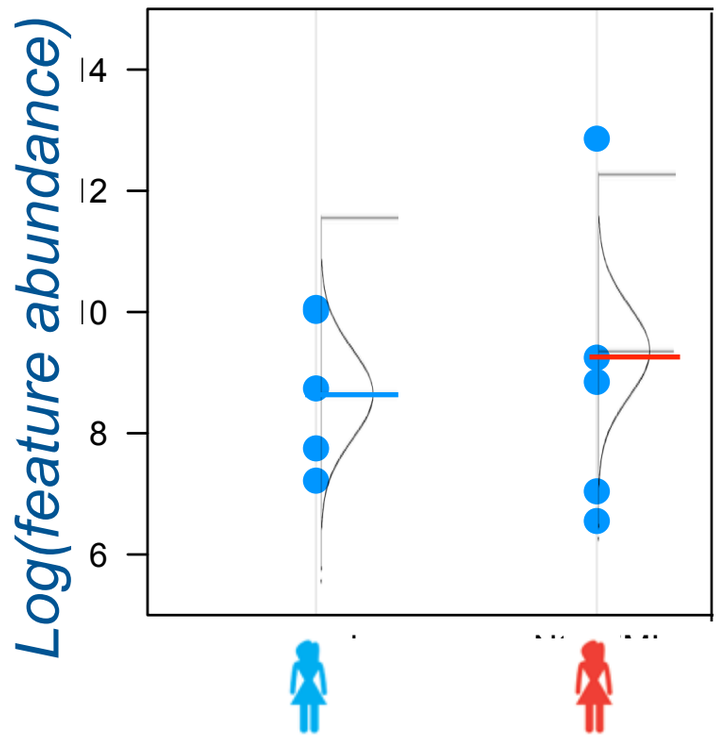
OUTLINE

- Translate scientific question into statistics
  - Statistical terms for 'biomarker' (or 'signature')
- Experimental design
  - Replication, randomization, blocking
- Basic data analysis
  - Simple summaries and models

TWO-SAMPLE T-TEST

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





Sample means in each group

H0: 'status quo', no change in abundance, $\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0} = \mathbf{0}$

Ha: change in abundance, $\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0} \neq \mathbf{0}$

$$\textbf{observed t} = \frac{\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0}}{\sqrt{\textbf{Estimate of variation}}} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

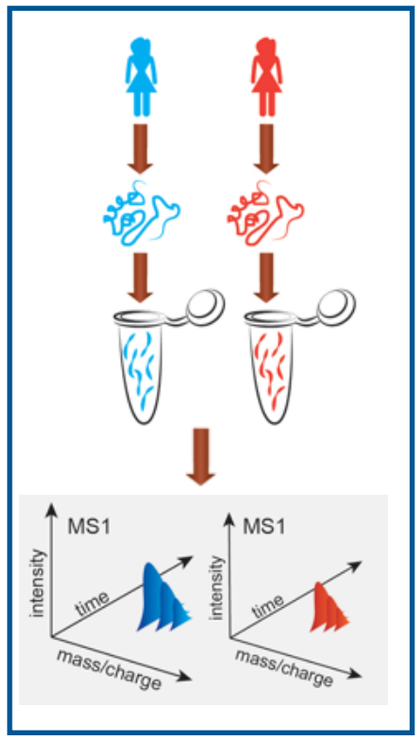
Number of replicates

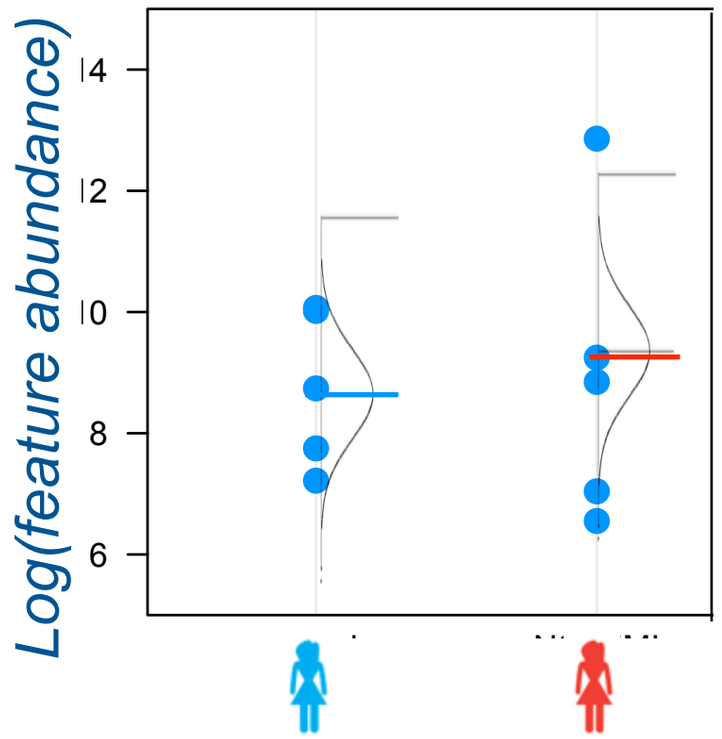
$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_{1\cdot})^2$$

Sample variance

TWO-SAMPLE T-TEST

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





H0: 'status quo', no change in abundance, $\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0} = \mathbf{0}$

Ha: change in abundance, $\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0} \neq \mathbf{0}$

$$\textbf{observed t} = \frac{\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0}}{\sqrt{\textbf{Estimate of variation}}} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

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

Properties of the means

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

Variance of the sampling distribution of first mean

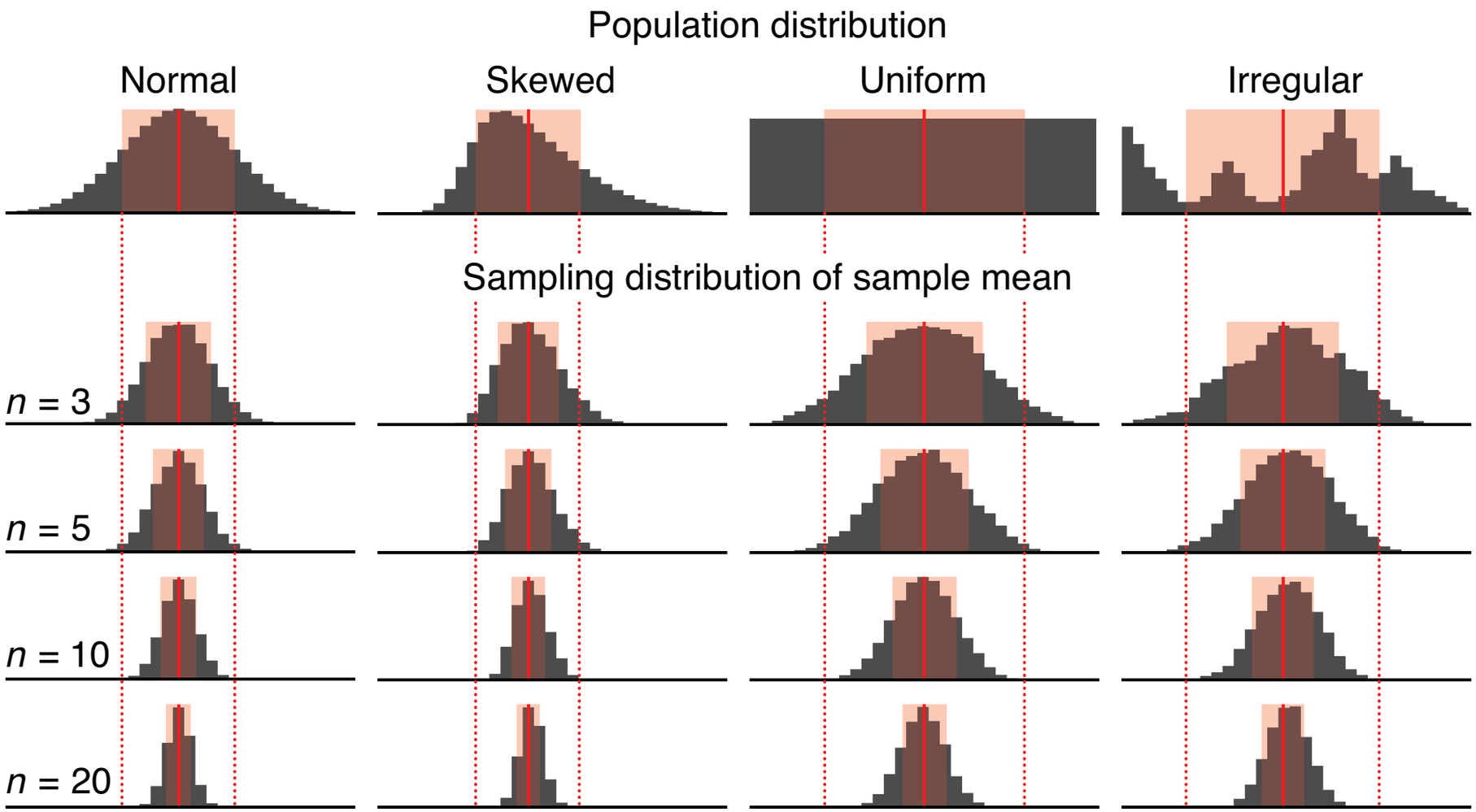
$$\sqrt{\frac{s_1^2}{n_1}}$$

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

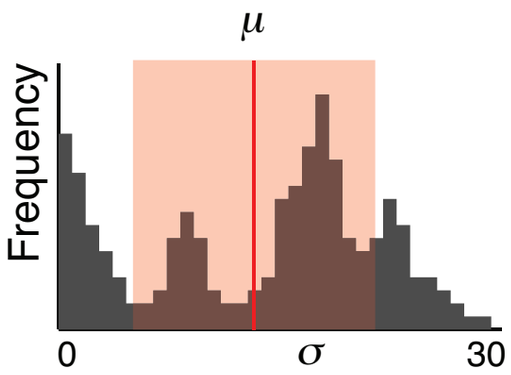
Repeatedly selecting n data points and calculating means

Simulated example

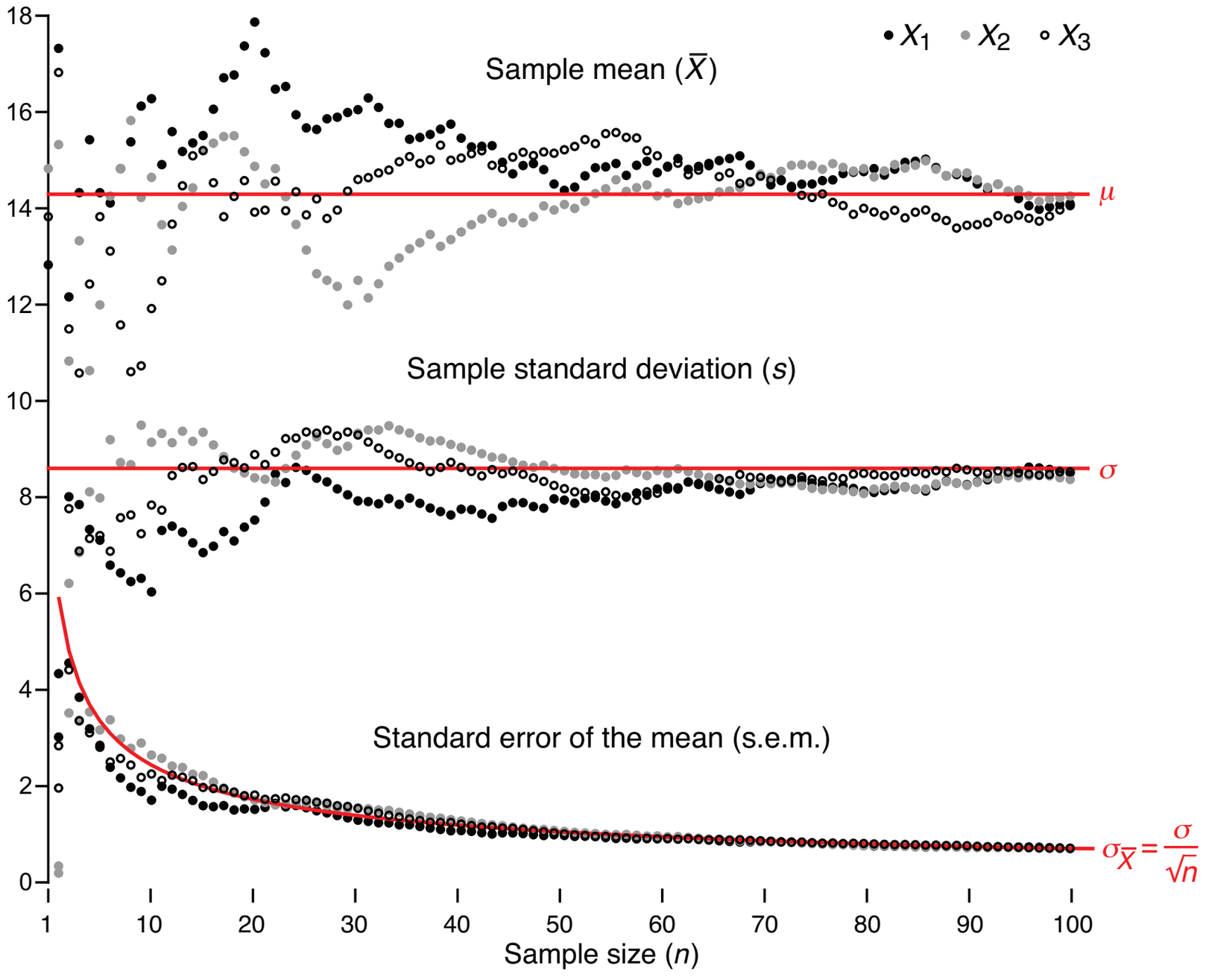
Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

EFFECT OF SAMPLE SIZE

As n increases, the estimates stabilize



Probability distribution of the data

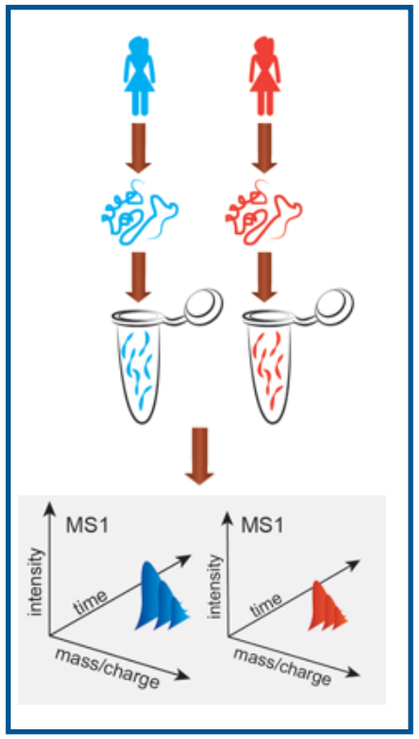


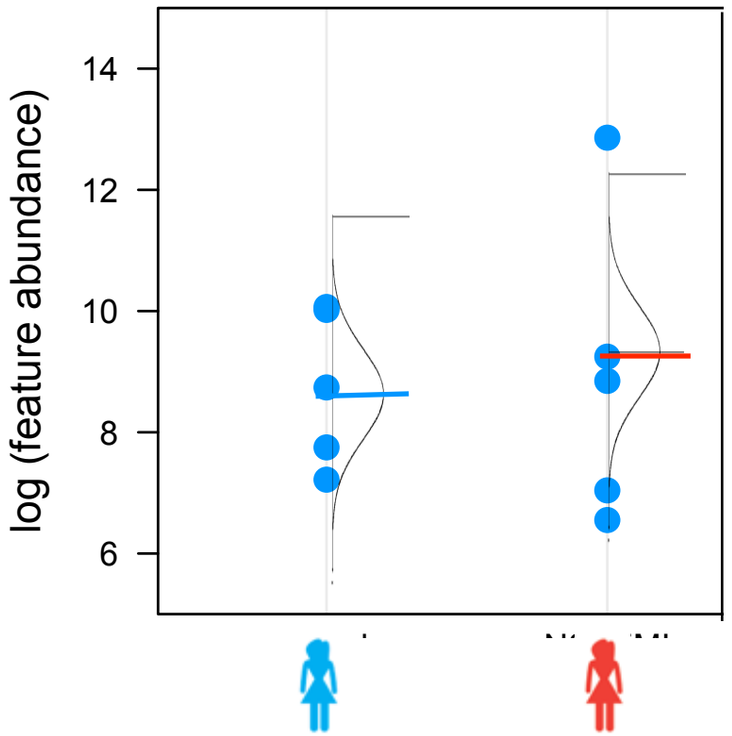
Simulated example

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

FINDING DIFFERENTIALLY ABUNDANT PROTEINS

False positive rate





H0: 'status quo', no change in abundance, $\hat{G}_1 - \hat{G}_0 = 0$

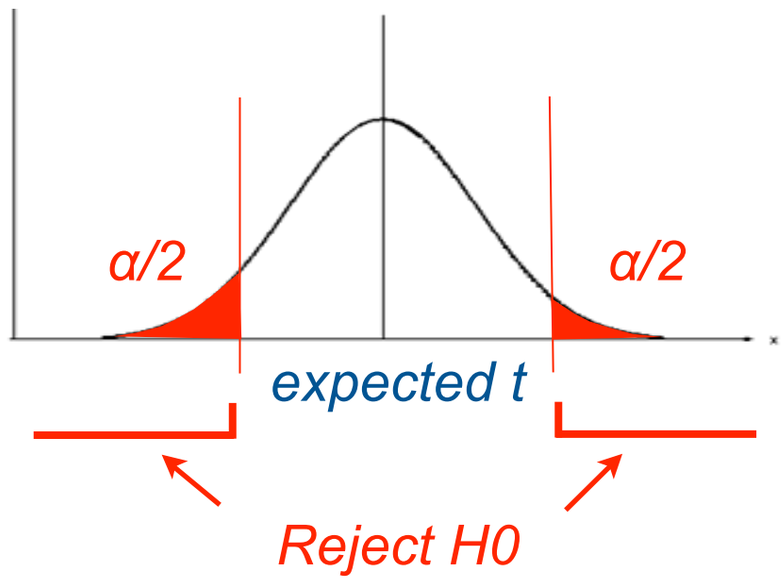
Ha: change in abundance, $\hat{G}_1 - \hat{G}_0 \neq 0$

$$\text{observed t} = \frac{\hat{G}_1 - \hat{G}_0}{\sqrt{\text{Estimate of variation}}}$$

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

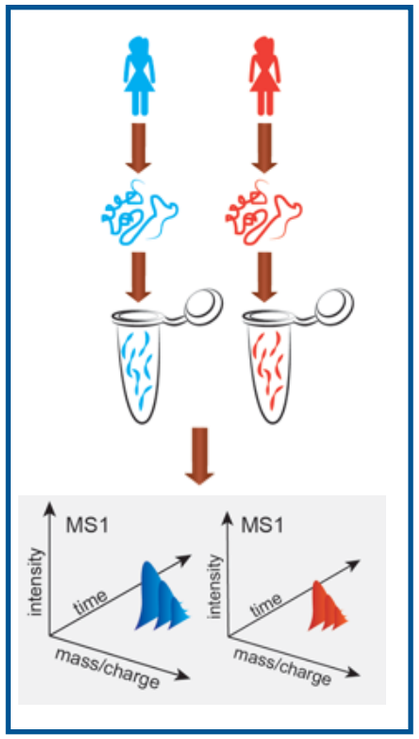
Distribution of the score if H0 is true

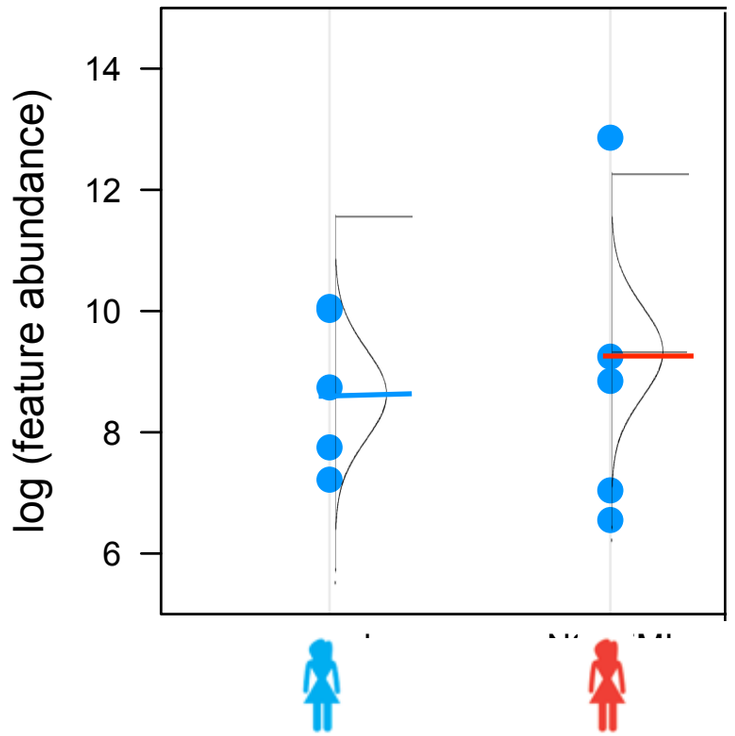
α = False Positive Rate



FINDING DIFFERENTIALLY ABUNDANT PROTEINS

P-value





H0: 'status quo', no change in abundance, $\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0} = \mathbf{0}$

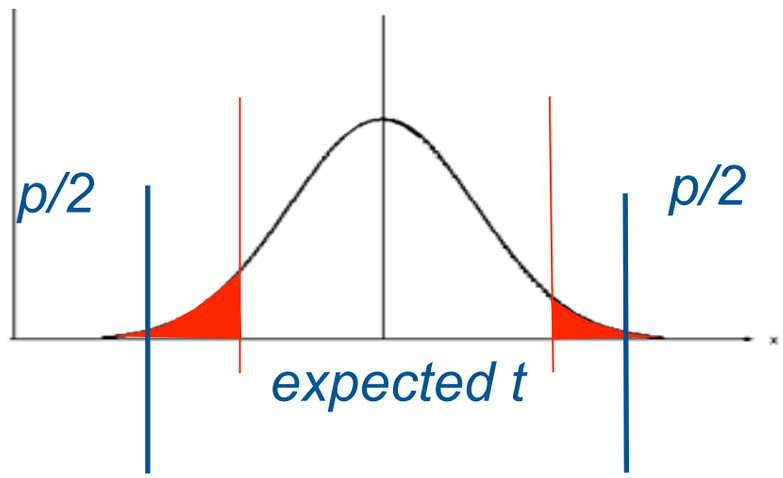
Ha: change in abundance, $\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0} \neq \mathbf{0}$

$$\textbf{observed t} = \frac{\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0}}{\sqrt{\textbf{Estimate of variation}}}$$

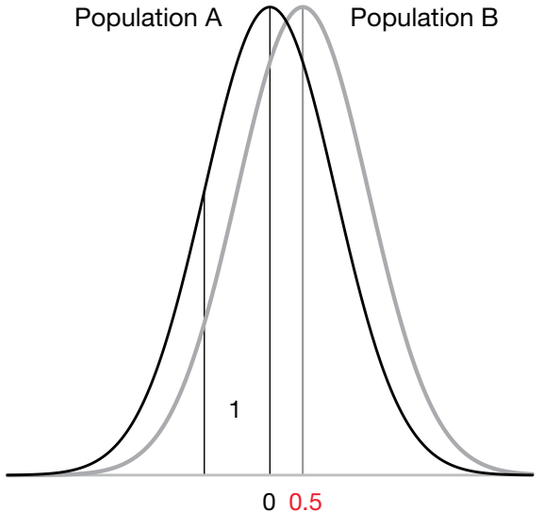
$$\overset{\textbf{no difference}}{\sim} \textbf{Student distribution}$$

Distribution of the score if H0 is true

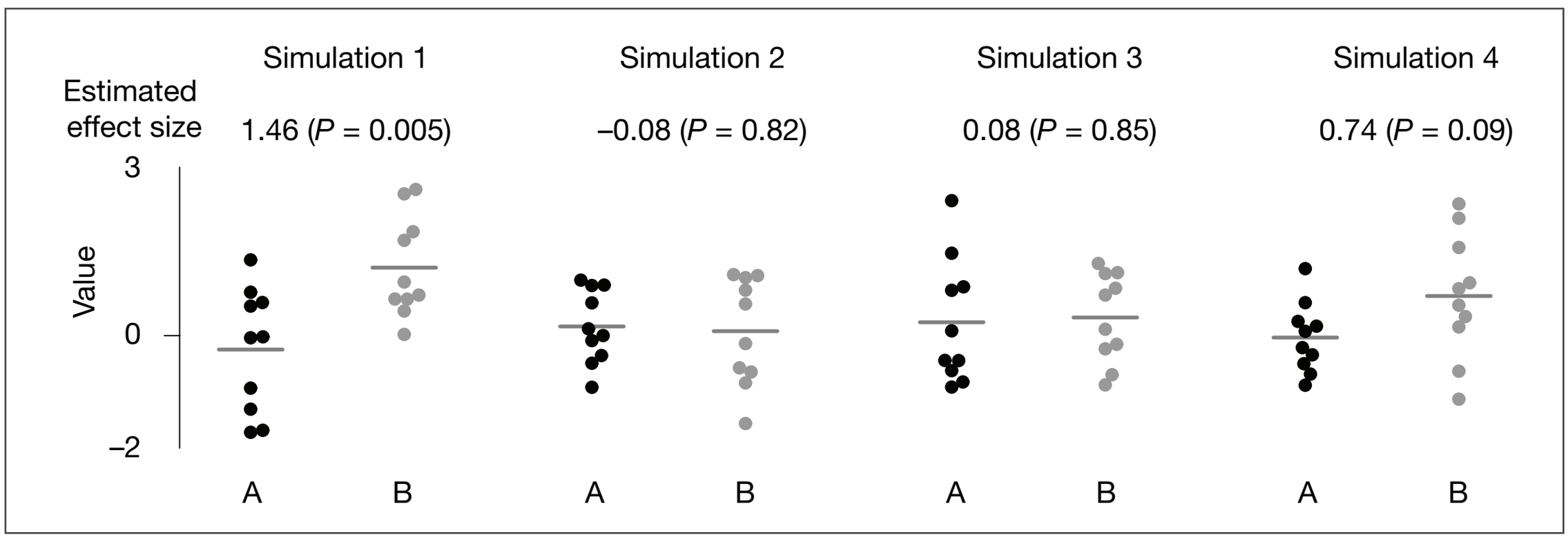
p = p-value



WITH SMALL SAMPLE SIZE, P-VALUES ARE UNSTABLE



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



Simulated example

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

MULTIPLE TESTING

Control False Positive Rate for two proteins

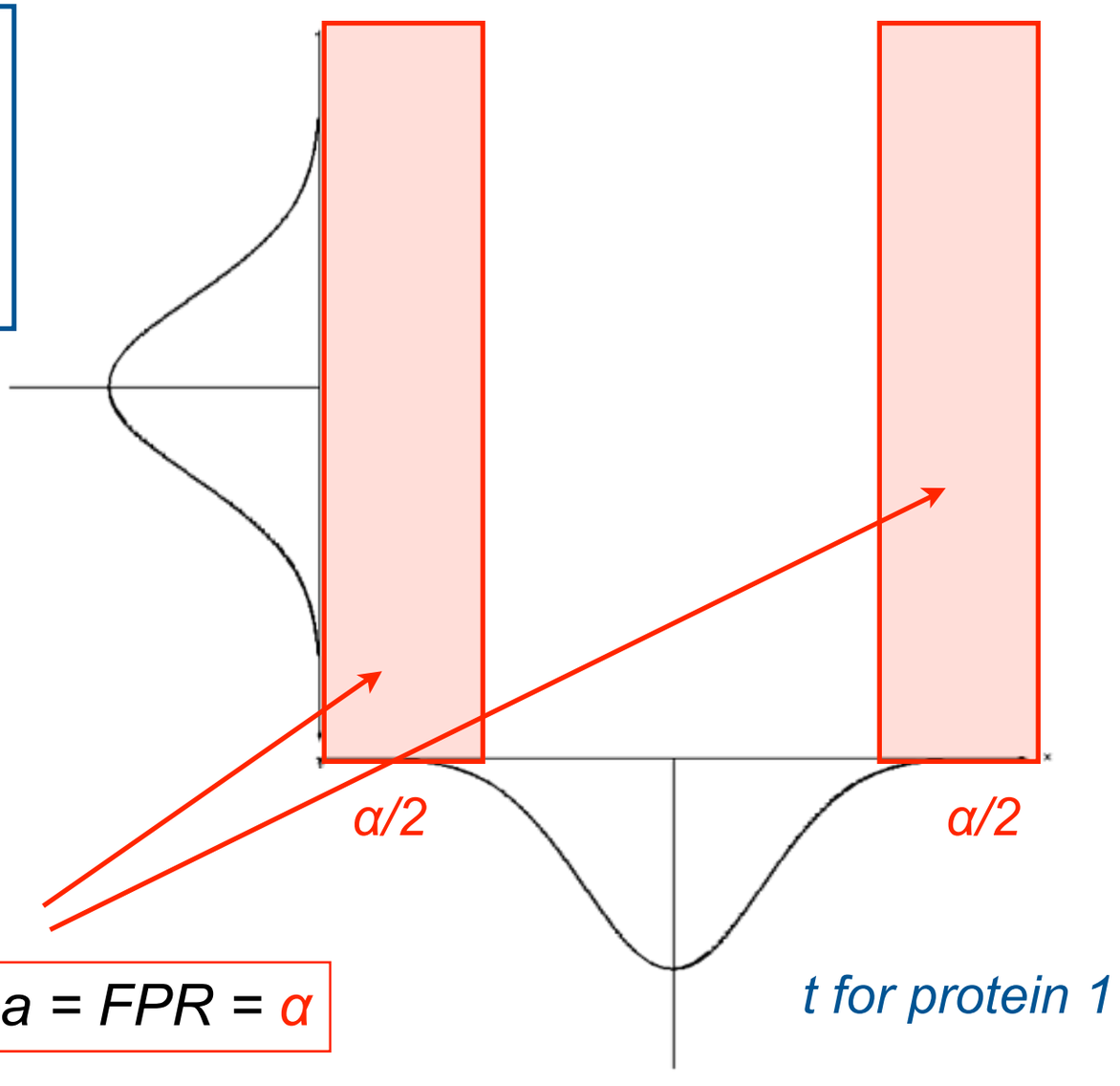
For each protein:

H0: 'status quo', no change in abundance, $\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0} = \mathbf{0}$

Ha: change in abundance, $\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0} \neq \mathbf{0}$

$$\textbf{observed t} = \frac{\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0}}{\sqrt{\textbf{Estimate of variation}}}$$

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



The area = FPR = α

MULTIPLE TESTING

Control False Positive Rate for two proteins

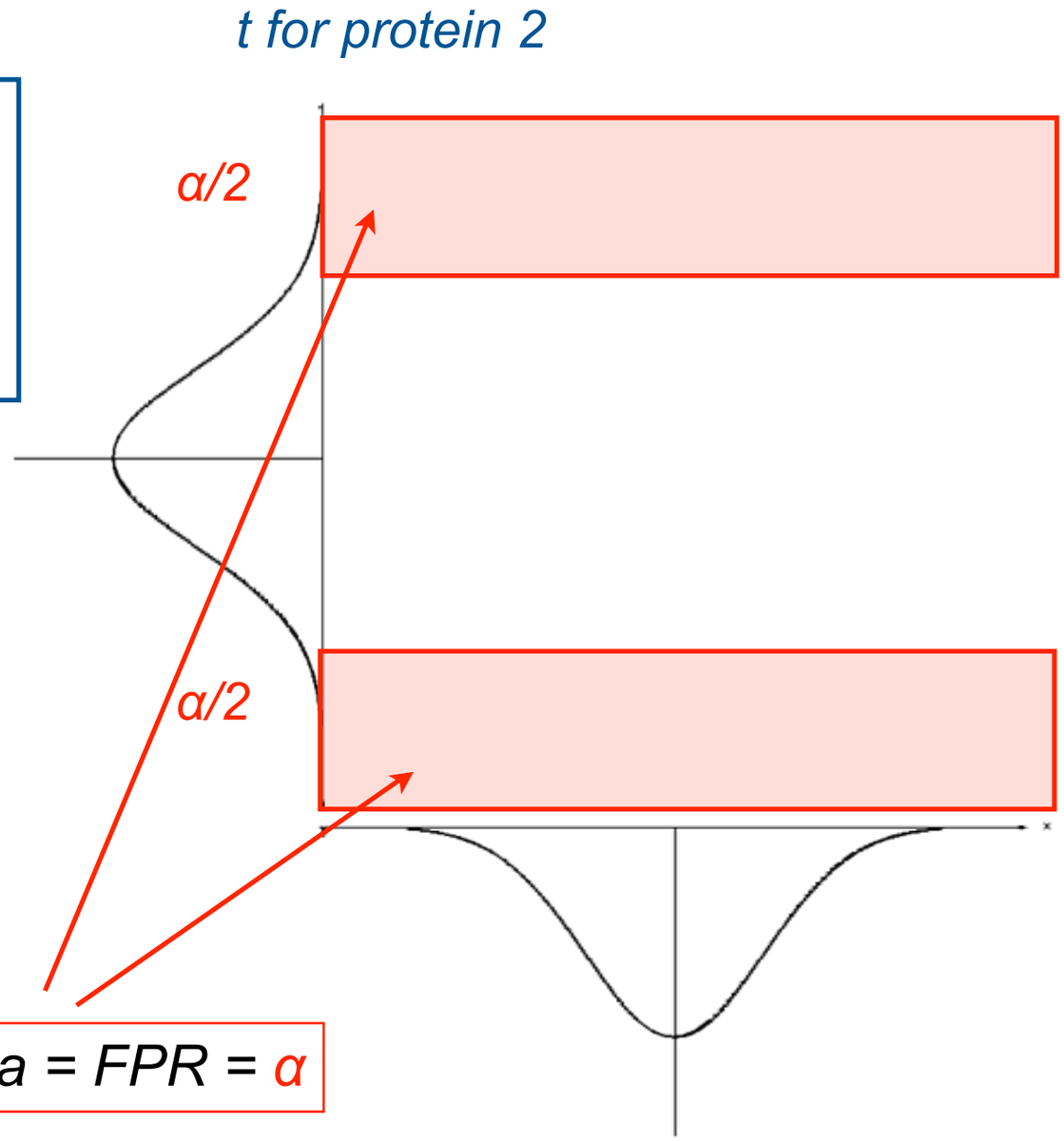
For each protein:

H0: 'status quo', no change in abundance, $\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0} = \mathbf{0}$

Ha: change in abundance, $\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0} \neq \mathbf{0}$

$$\textbf{observed t} = \frac{\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0}}{\sqrt{\textbf{Estimate of variation}}}$$

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



MULTIPLE TESTING

Control False Positive Rate for two proteins

For each protein:

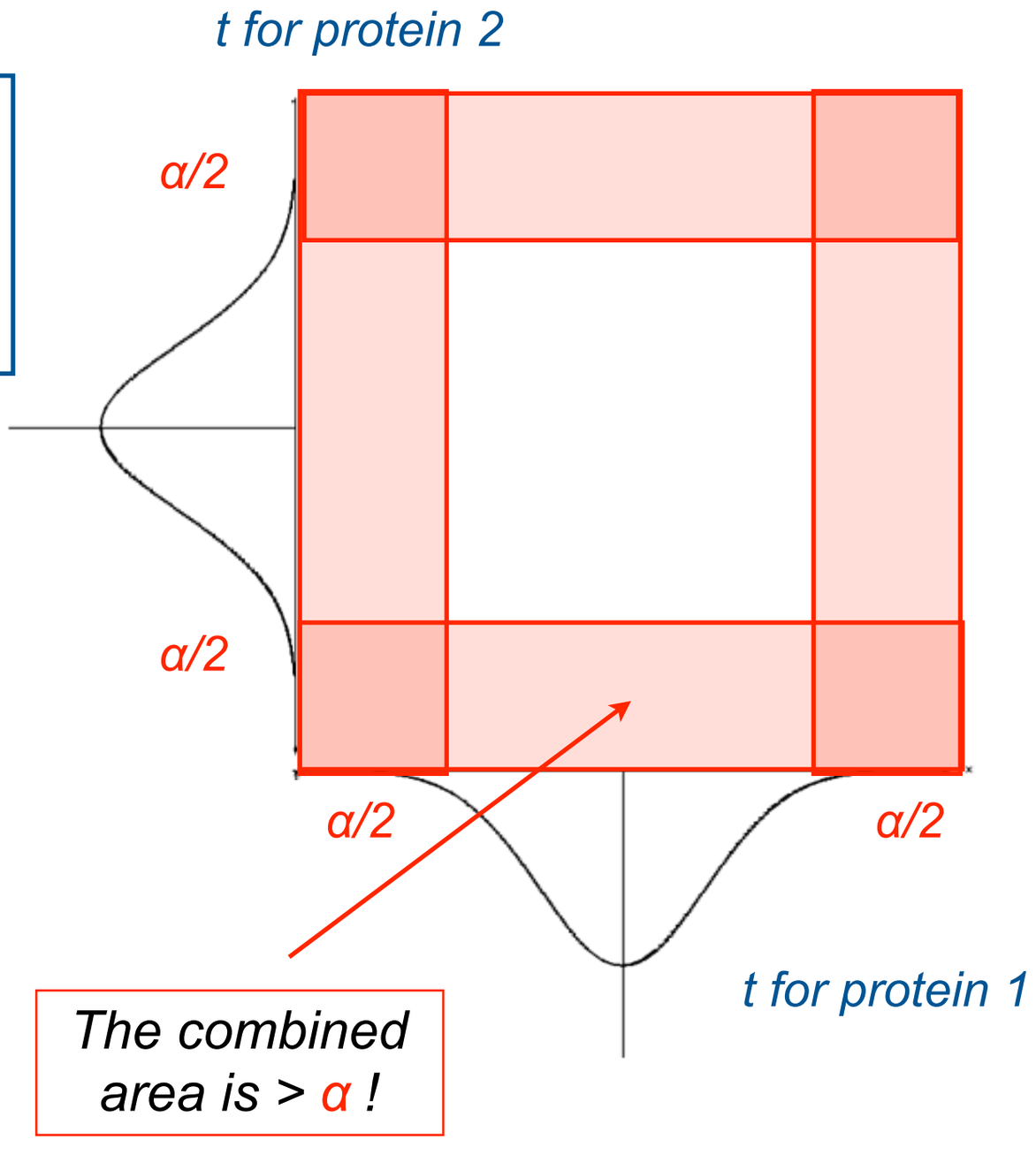
H0: 'status quo', no change in abundance, $\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0} = \mathbf{0}$

Ha: change in abundance, $\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0} \neq \mathbf{0}$

$$\textbf{observed t} = \frac{\hat{\mathbf{G}}_\mathbf{1} - \hat{\mathbf{G}}_\mathbf{0}}{\sqrt{\textbf{Estimate of variation}}}$$

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

- 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:

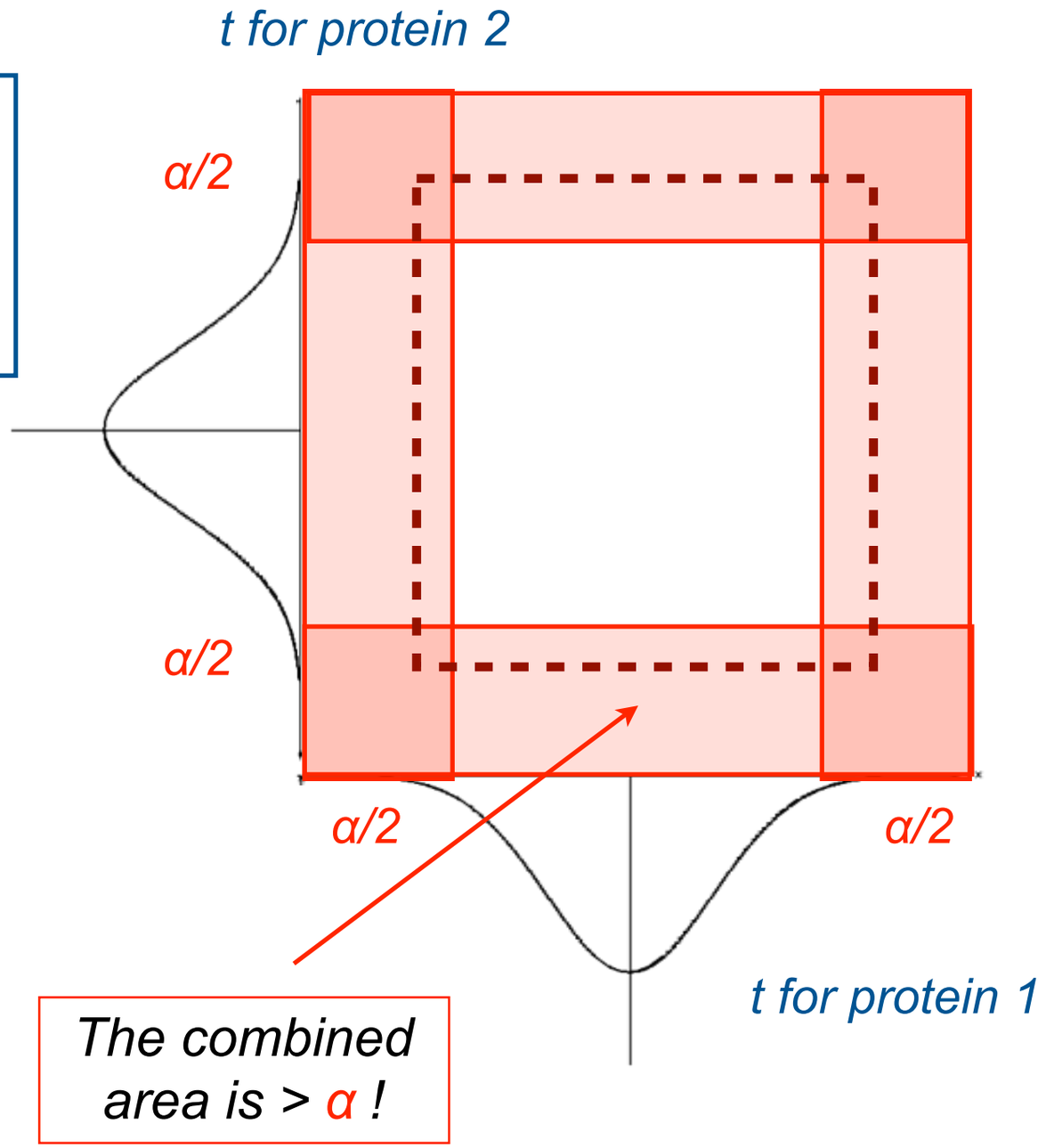
H0: 'status quo', no change in abundance, $\hat{G}_1 - \hat{G}_0 = 0$

Ha: change in abundance, $\hat{G}_1 - \hat{G}_0 \neq 0$

$$\text{observed t} = \frac{\hat{G}_1 - \hat{G}_0}{\sqrt{\text{Estimate of variation}}}$$

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

- 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

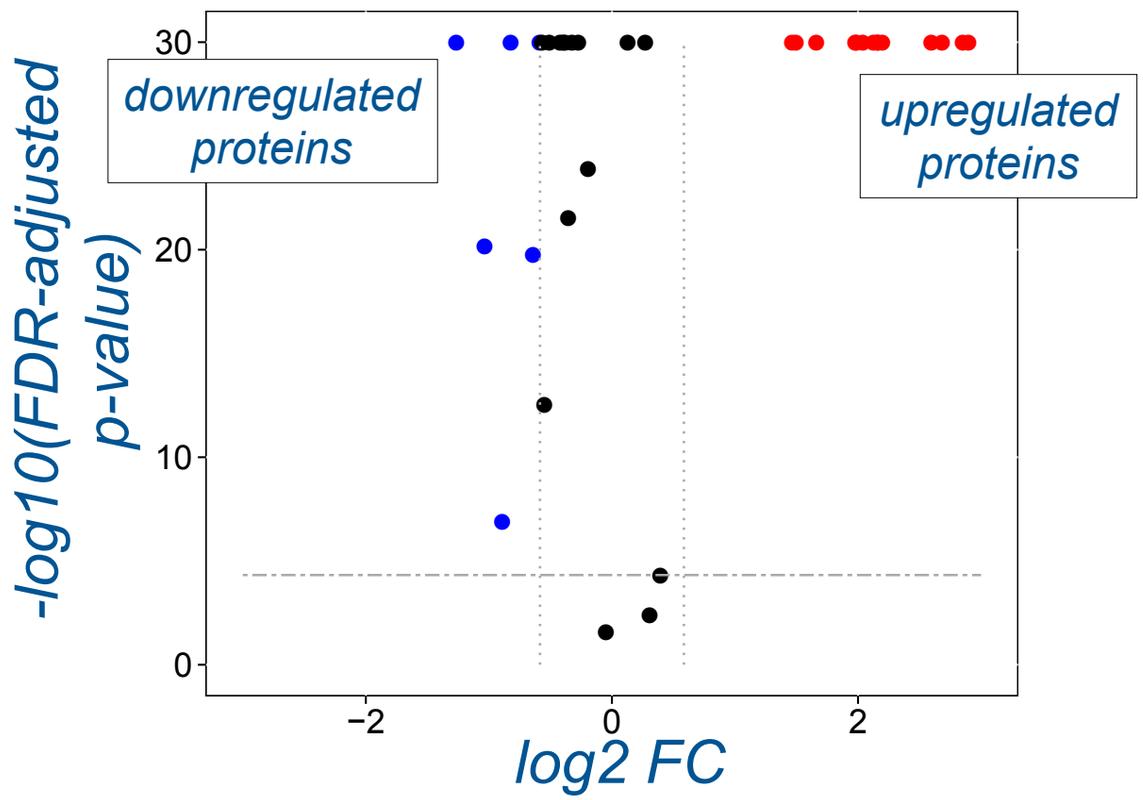
Change criteria from False Positive Rate to False Discovery Rate

	# 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

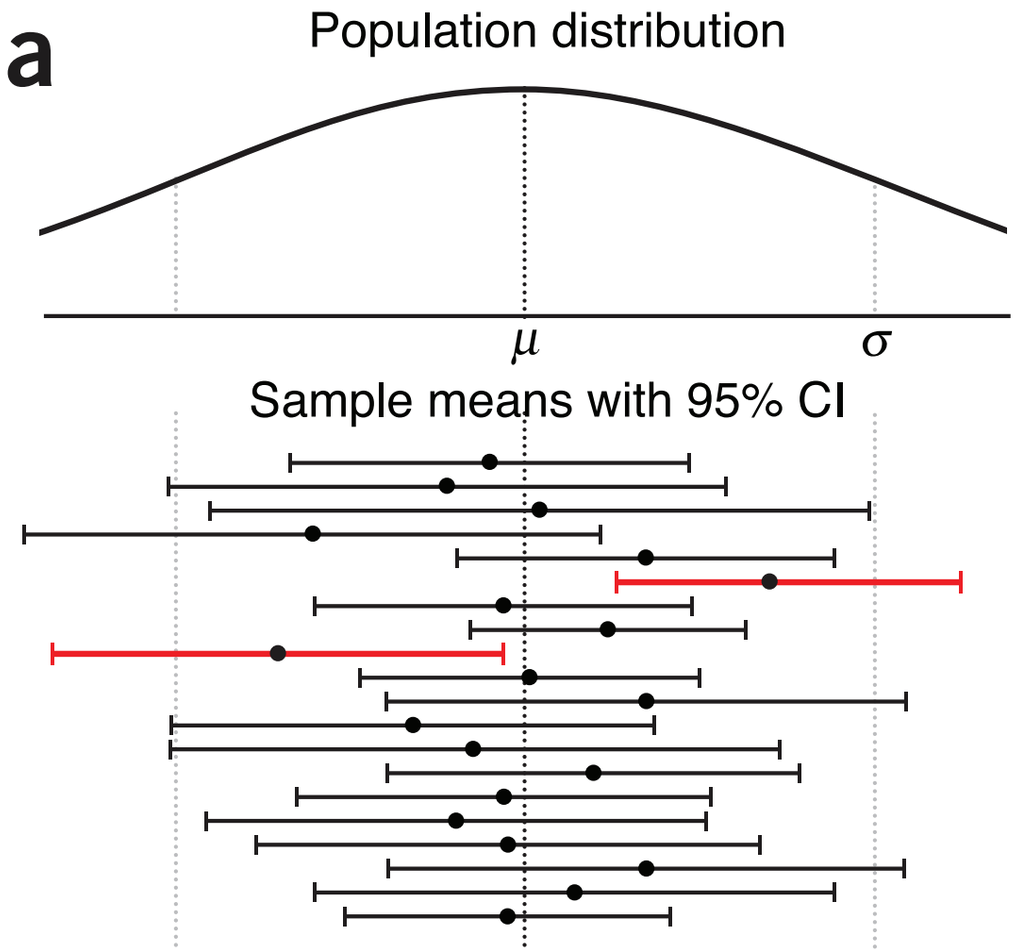
$$\mathbf{FDR} = \mathbf{E}\left[\frac{\mathbf{V}}{\max(\mathbf{R}, 1)}\right]$$

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

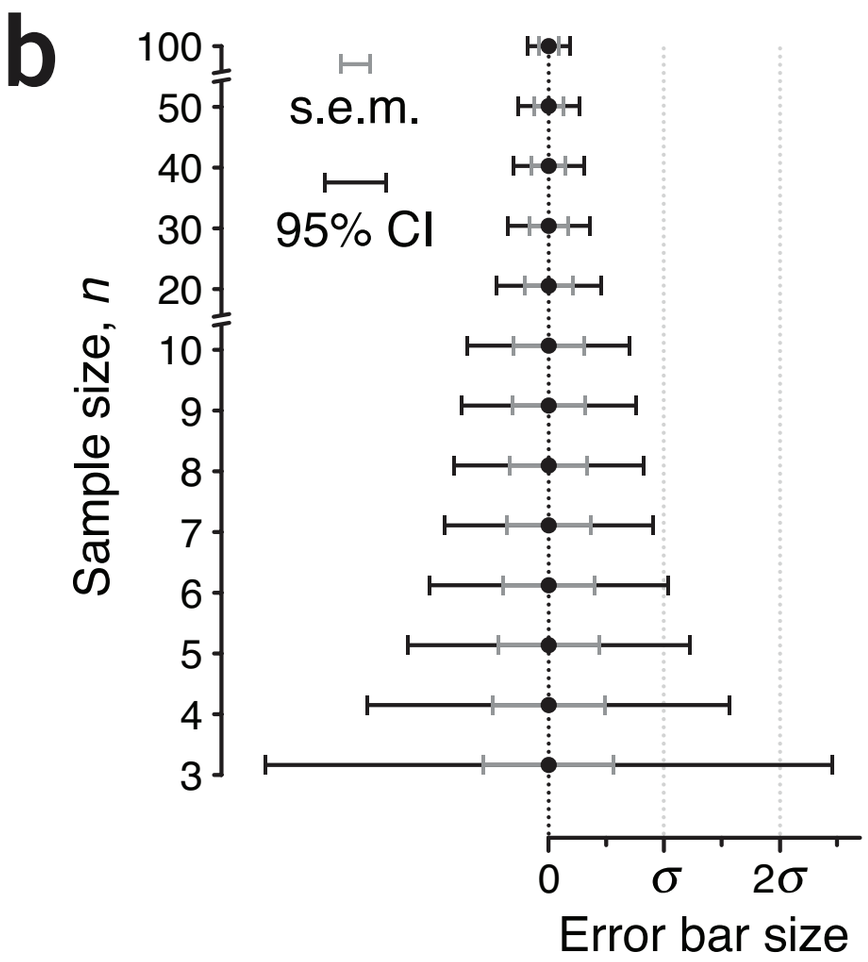


ALTERNATIVE TO TESTING: CONFIDENCE INTERVALS

Not all error bars are made the same



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!

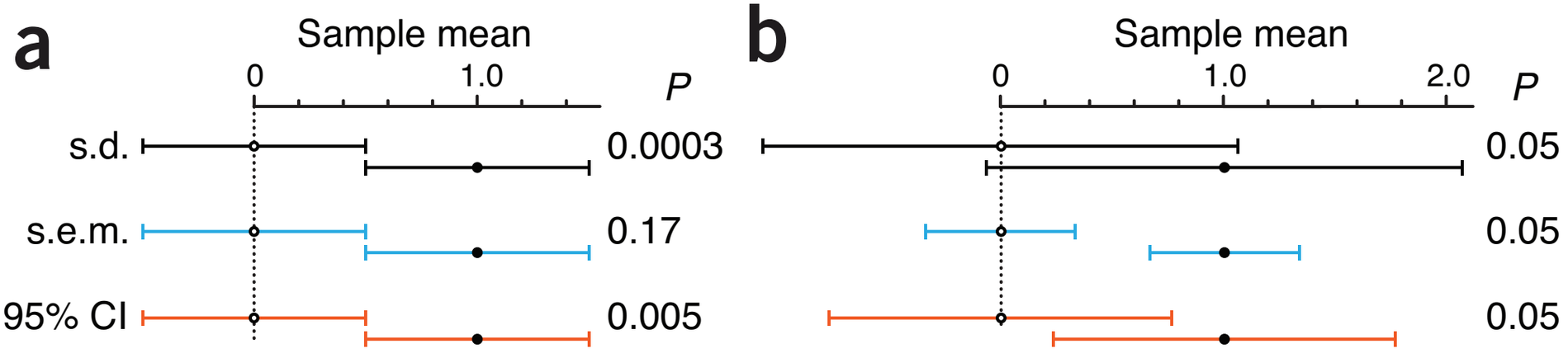
Width of the intervals depends on the sample size

Simulated example

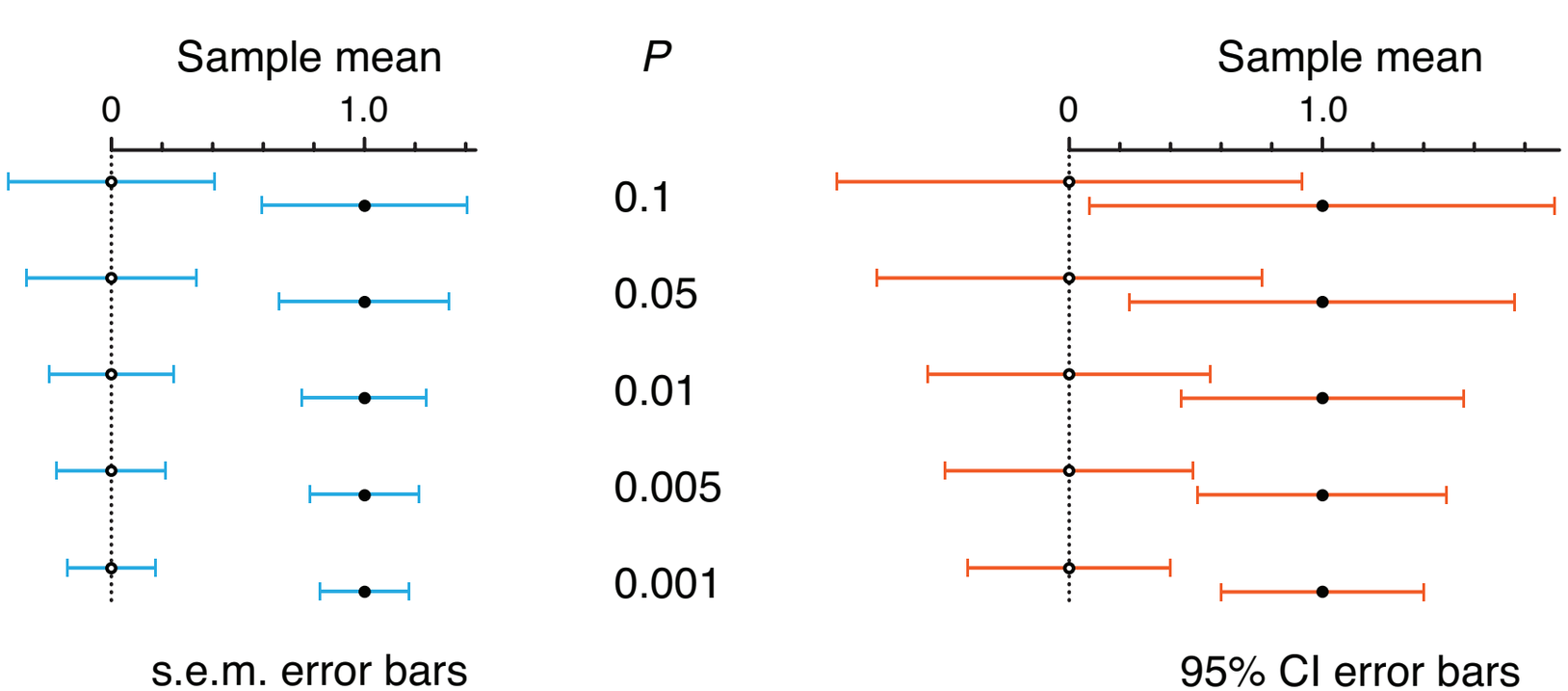
Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

ERROR BARS PROVIDE DIFFERENT INSIGHT

Absence of overlap does not always mean stat. significance



n=10



Simulated example

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

Martin Krzywinski & Naomi Altman

nature methods
Techniques for life scientists and chemists

NATURE METHODS | THIS MONTH

Points of significance: Importance of being uncertain

Martin Krzywinski & Naomi Altman

Affiliations

Nature Methods **10**, 809–810 (2013) | doi:10.1038/nmeth.2613

Points of significance: Comparing samples—part I

Martin Krzywinski & Naomi Altman

Points of Significance: Error bars

Martin Krzywinski & Naomi Altman

Affiliations

Nature Methods **10**, 921–922 (2013) | doi:10.1038/nmeth.2659

Points of significance: Power and sample size

Martin Krzywinski & Naomi Altman

Affiliations

Nature Methods **10**, 1139–1140 (2013) | doi:10.

Points of significance: Significance, *P* values and *t*-tests

Martin Krzywinski & Naomi Altman