SAMPLE SIZE, ASSOCIATIONS AND CATEGORICAL DATA

Olga Vitek

College of Science

College of Computer and Information Science

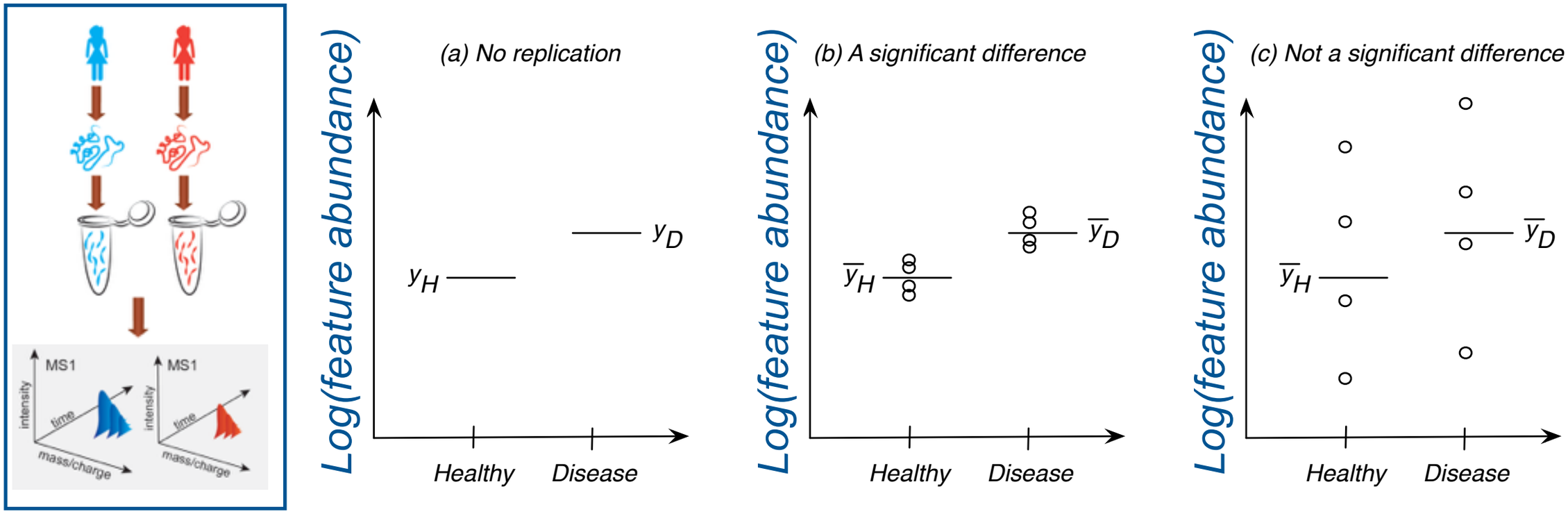
Northeastern University

OUTLINE

- So how many replicates do I need?
  - Design of complex experiments
- Associations between two measurements
  - Correlation and linear regression
- Statistical analysis of count data
  - Comparing proportions and spectral counts

PRINCIPLE 1: REPLICATION

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



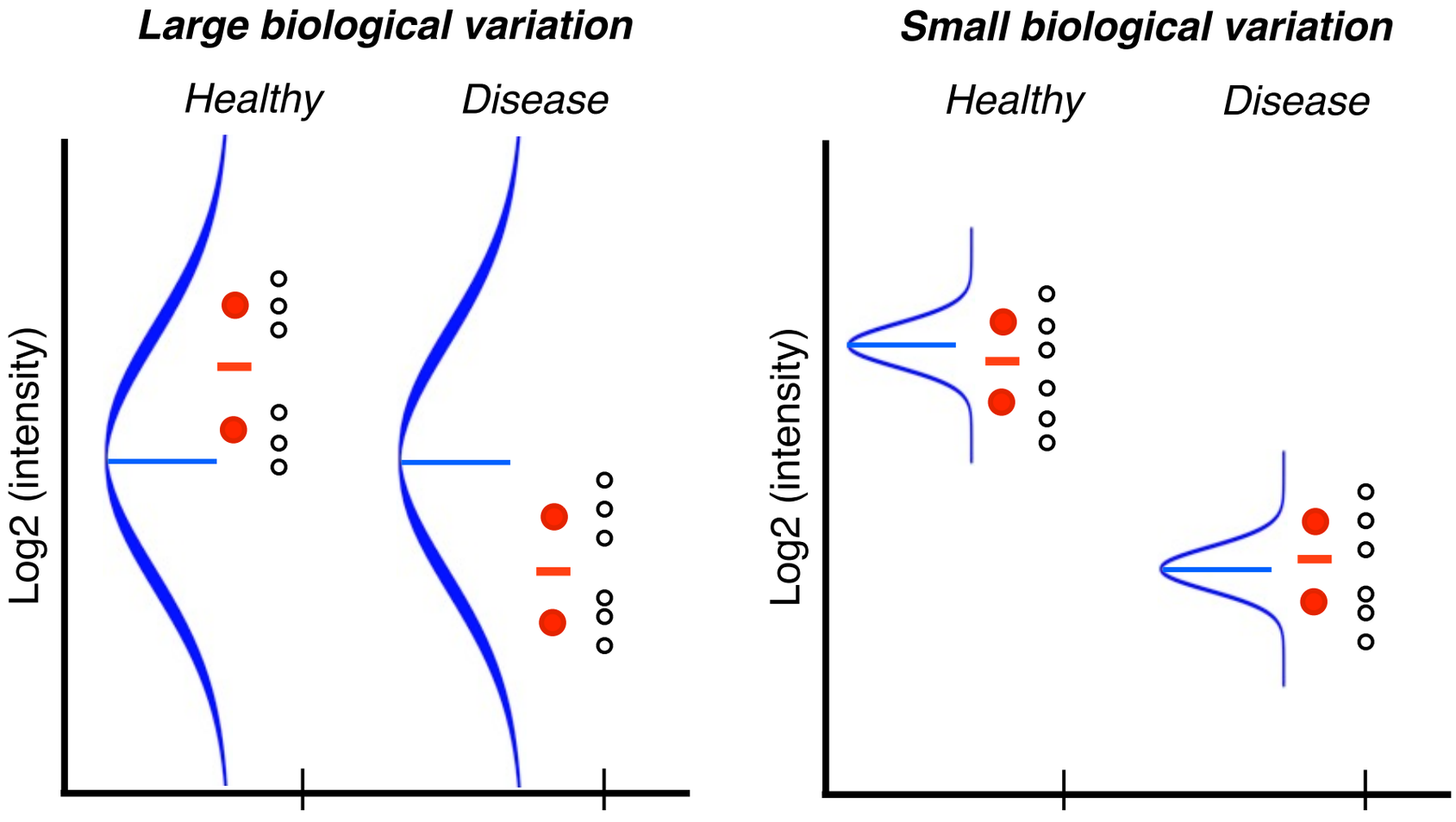
Two levels of randomness imply two types of replication:

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

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

MULTI-LAYER DESIGN AND ANALYSIS

Multiple sources are responsible for variation in measurements

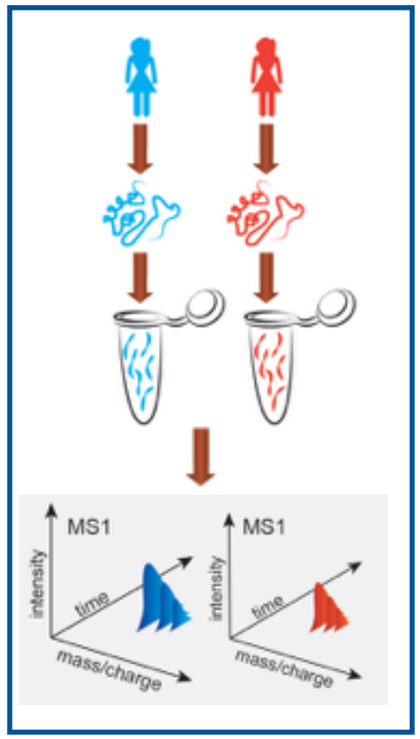


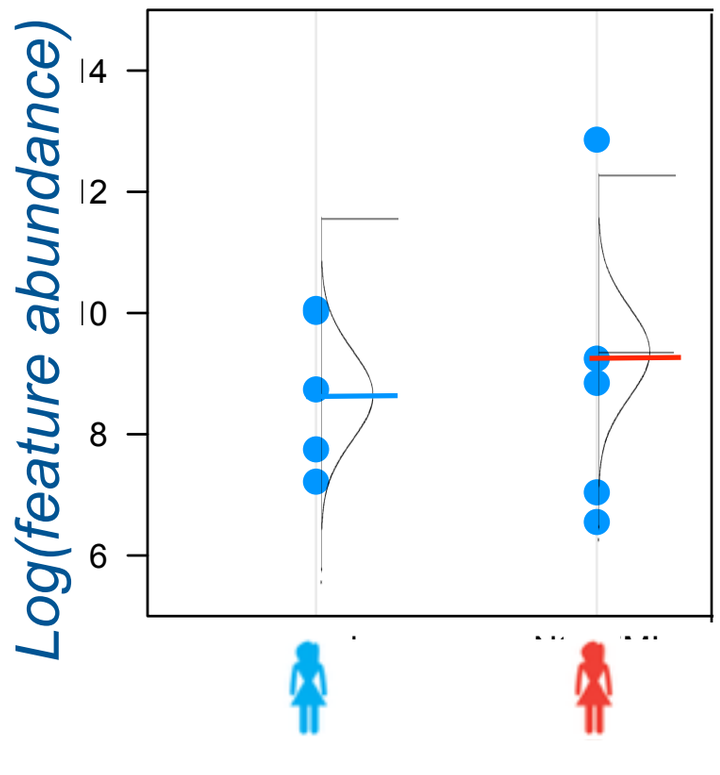
When biological variance is large, more biological replicates are needed to accurately estimate the variance

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

TWO-SAMPLE T-TEST

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



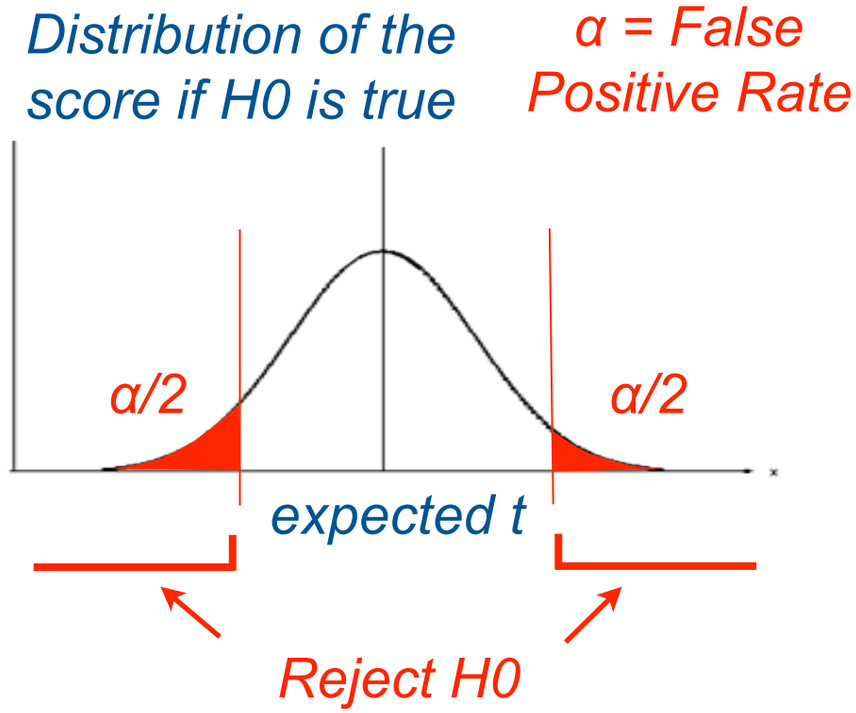


H0: no change in abundance, $\mu_1 - \mu_2 = 0$

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

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

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



SO HOW MANY REPLICATES DO I NEED?

One protein, two-group comparison

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:

$$\mathbf{Var}(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) \leq \left(\frac{\boldsymbol{\Delta}}{\mathbf{z_{1-\beta}} + \mathbf{z_{1-\alpha/2}}}\right)^{\mathbf{2}}$$

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

$$\mathbf{Var}(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) = s_1^2/n_1 + s_2^2/n_2 \leq \left(\frac{\boldsymbol{\Delta}}{\mathbf{z_{1-\beta}} + \mathbf{z_{1-\alpha/2}}}\right)^{\mathbf{2}}$$

→ solve for the number of individuals n_1 and n_2

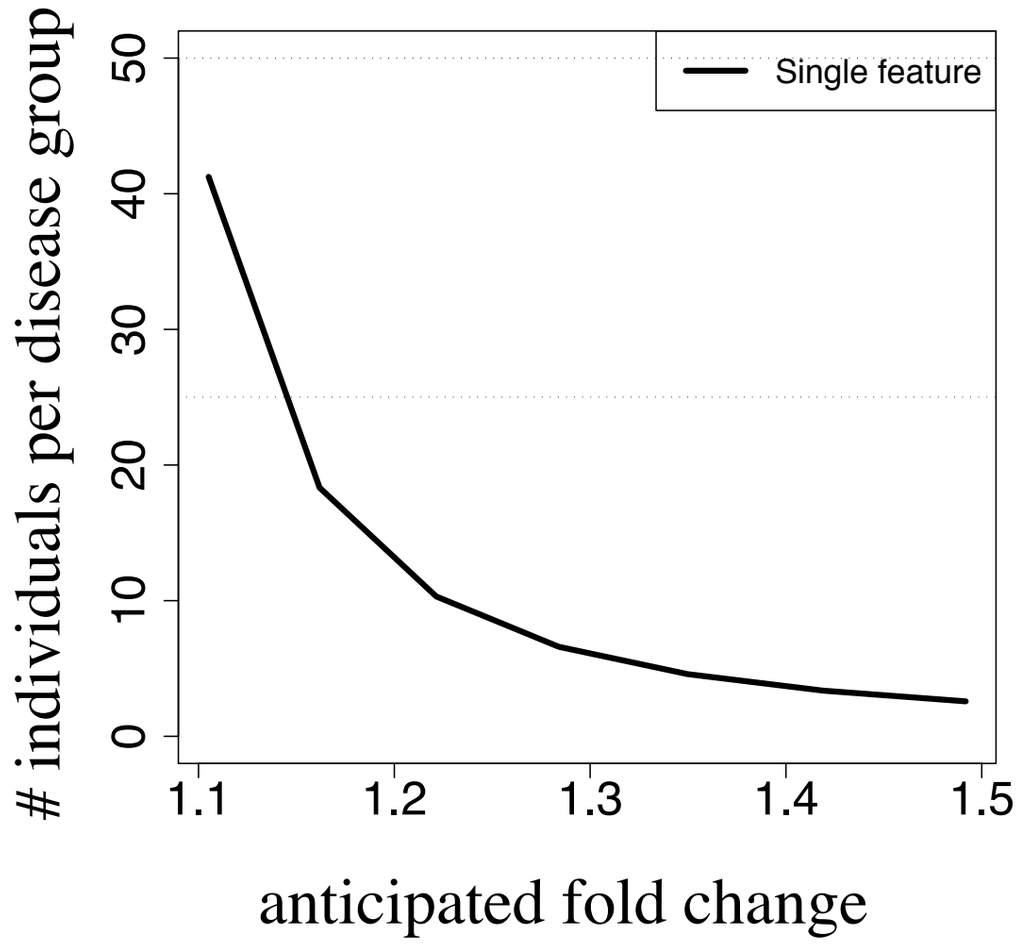
SO HOW MANY REPLICATES DO I NEED?

One protein, two-group comparison

Example: pilot study with diabetes patients.

A block-randomized design

If we only had one feature:



Conclusion:

The smaller the anticipated difference, the larger the sample size

SO HOW MANY REPLICATES DO I NEED?

Many proteins, two-group comparison

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

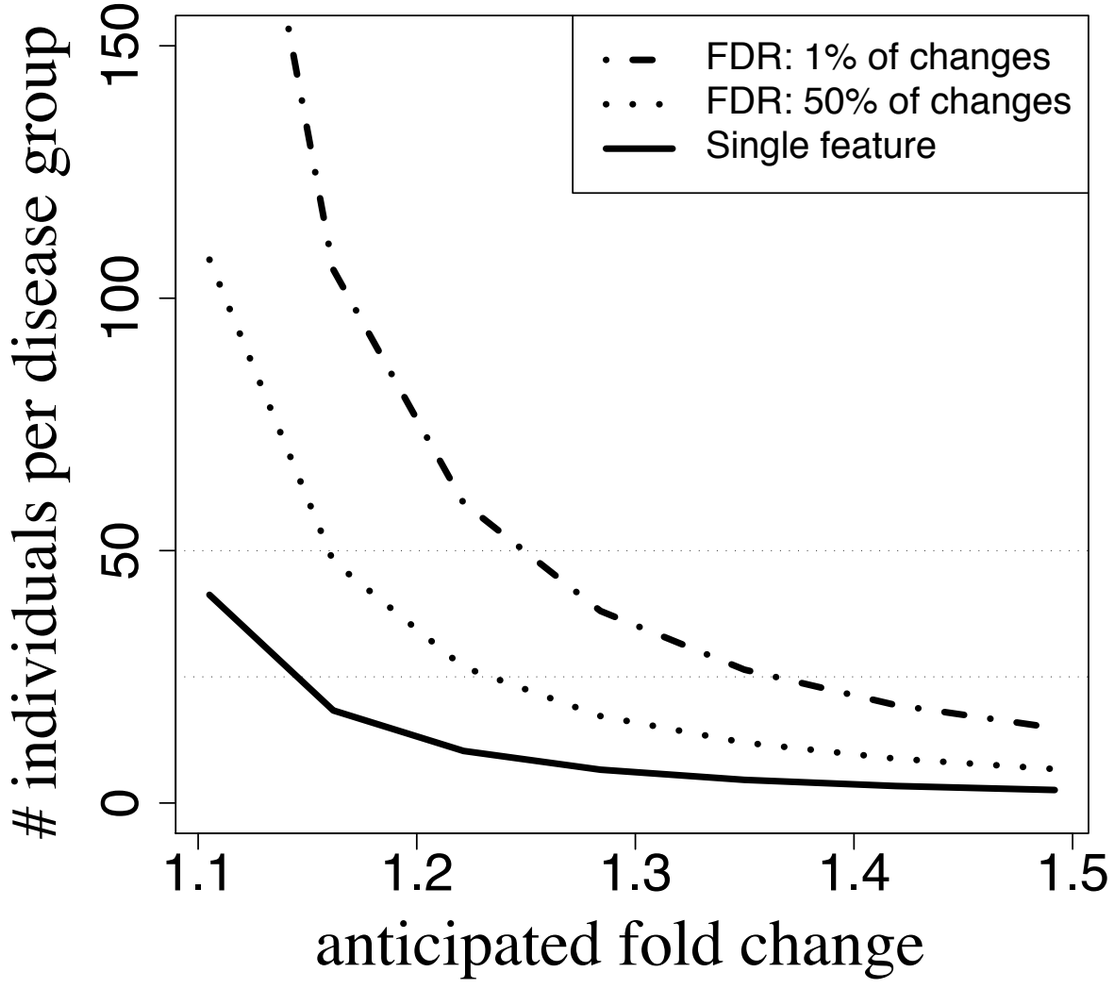
$$\alpha_{\text{ave}} \leq (\mathbf{1} - \beta)_{\text{ave}} \cdot \mathbf{q}\frac{\mathbf{1}}{\mathbf{1} + (\mathbf{1} - \mathbf{q}) \cdot \mathbf{m_0/m_1}};$$

- average probability of a false positive discovery

SO HOW MANY REPLICATES DO I NEED?

Many proteins, two-group comparison

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



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

Oberg and Vitek, *JPR*, 2009

COMPLEX DESIGNS: BIO AND TECH REPS

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	+	$\text{Indiv(Group)}_{j(i)} \sim N\left(0, \sigma^2_{Indiv}\right)$	+	$\text{Prep(Indiv)}_{k(ij)} \sim N\left(0, \sigma^2_{Prep}\right)$	+	$\text{Error}_{l(ijk)} \sim N\left(0, \sigma^2_{Error}\right)$

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

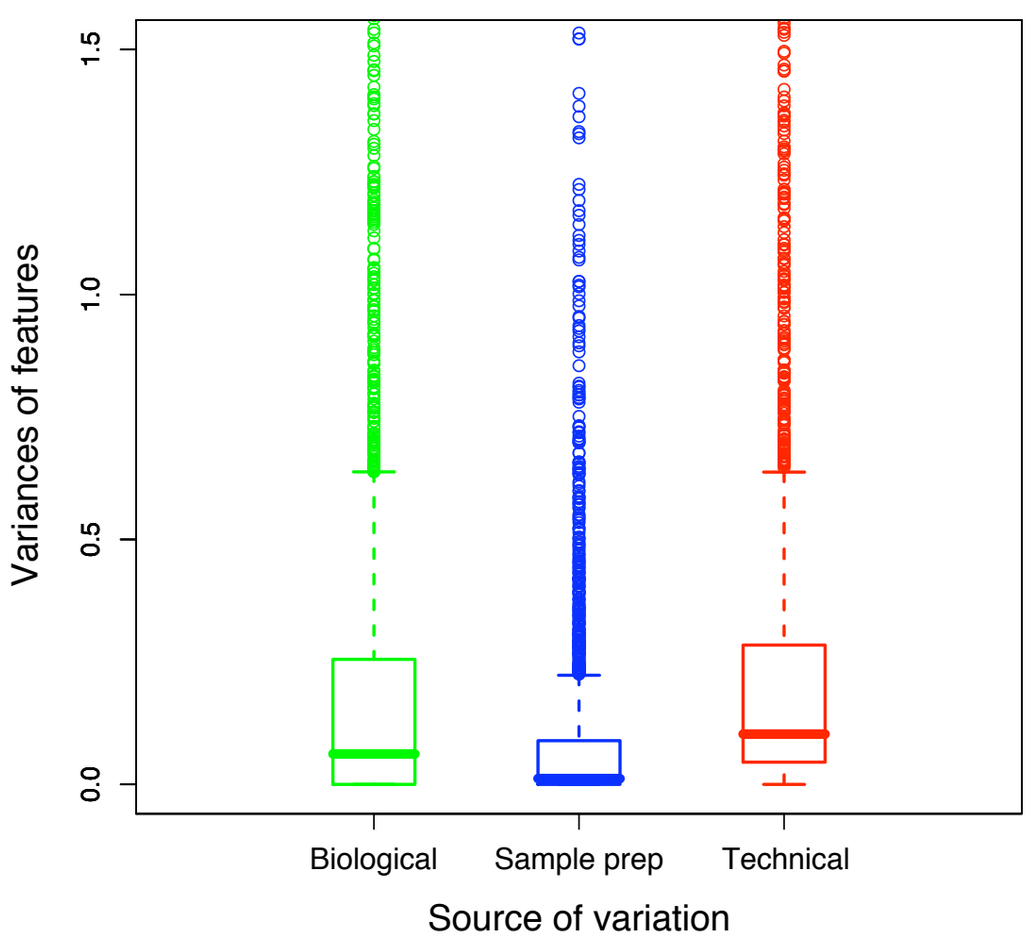
I: # individuals per disease group

J: # sample preps

K: # replicate runs

A pilot experiment

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



COMPLEX DESIGNS: BIO AND TECH REPS

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	+	$\text{Indiv(Group)}_{j(i)} \sim N\left(0, \sigma^2_{\text{Indiv}}\right)$	+	$\text{Prep(Indiv)}_{k(ij)} \sim N\left(0, \sigma^2_{\text{Prep}}\right)$	+	$\text{Error}_{l(ijk)} \sim N\left(0, \sigma^2_{\text{Error}}\right)$

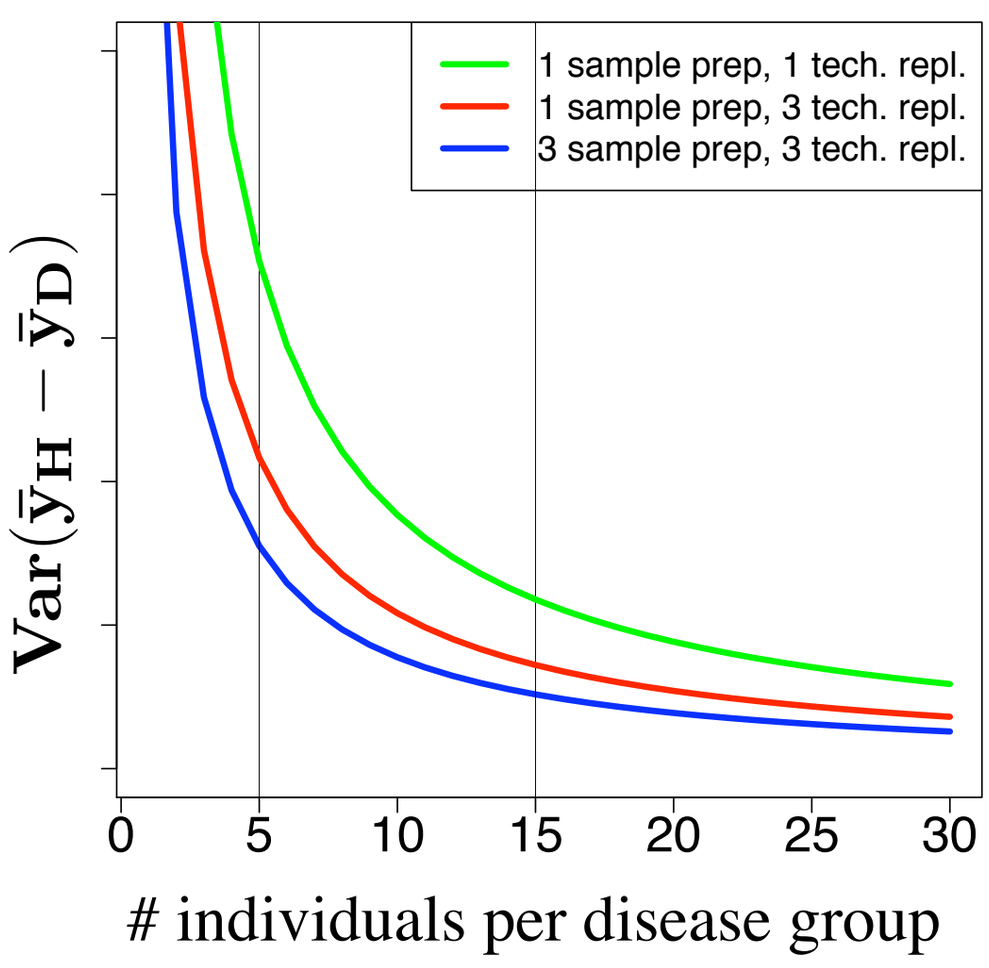
$$\text{Var}(\bar{y}_H - \bar{y}_D) = 2\left(\frac{\sigma^2_{\text{Indiv}}}{I} + \frac{\sigma^2_{\text{Prep}}}{IJ} + \frac{\sigma^2_{\text{Error}}}{IJK}\right)$$

I: # individuals per disease group

J: # sample preps

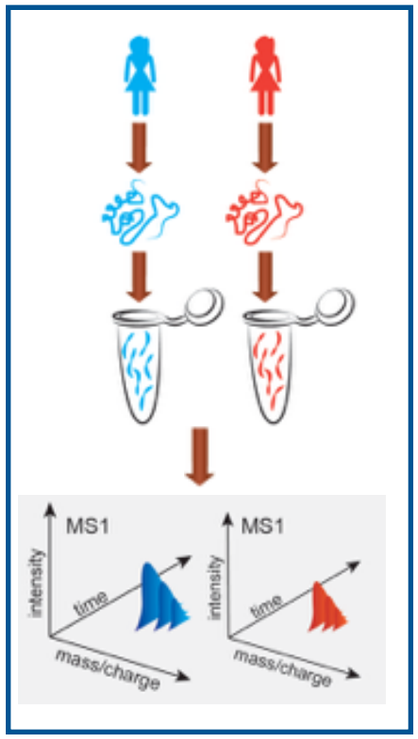
K: # replicate runs

Conclusion:
Maximize the number of biological replicates

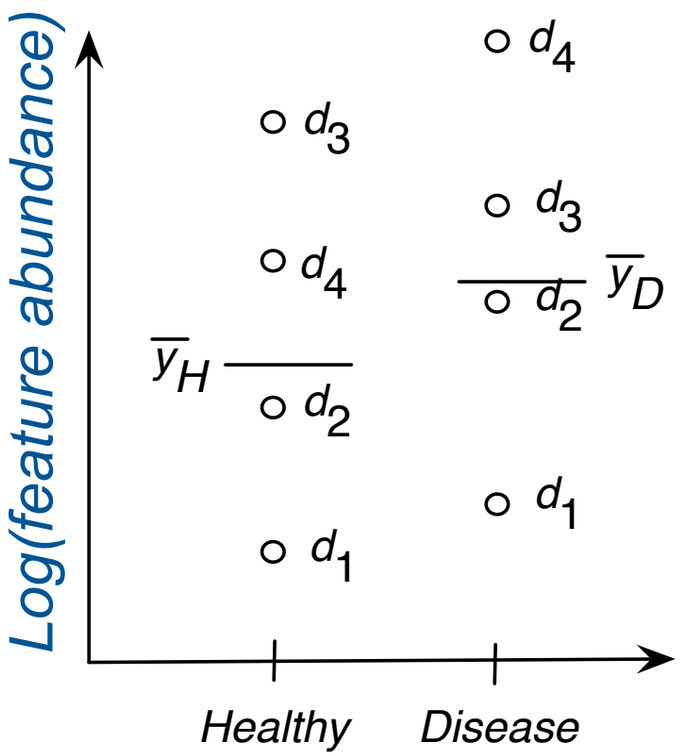


COMPLEX DESIGNS

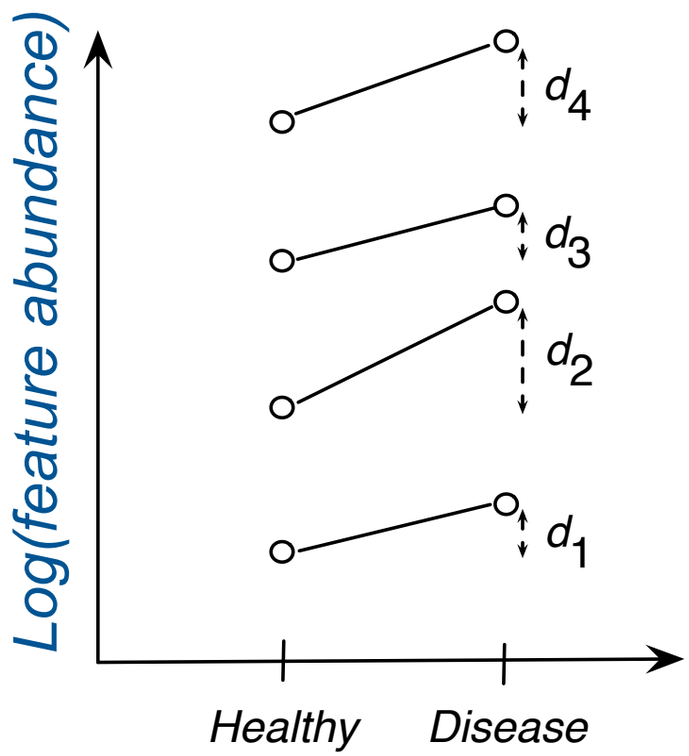
Blocking helps reduce the sample size



(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

COMPLEX DESIGNS

Blocking helps most 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	+	$\text{Block}_k \sim N(0, \sigma^2_{Block})$	+	$\text{Indiv(Group)}_{j(i)} \sim N(0, \sigma^2_{Indiv})$	+	$\text{Error}_{l(ijk)} \sim N(0, \sigma^2_{Error})$

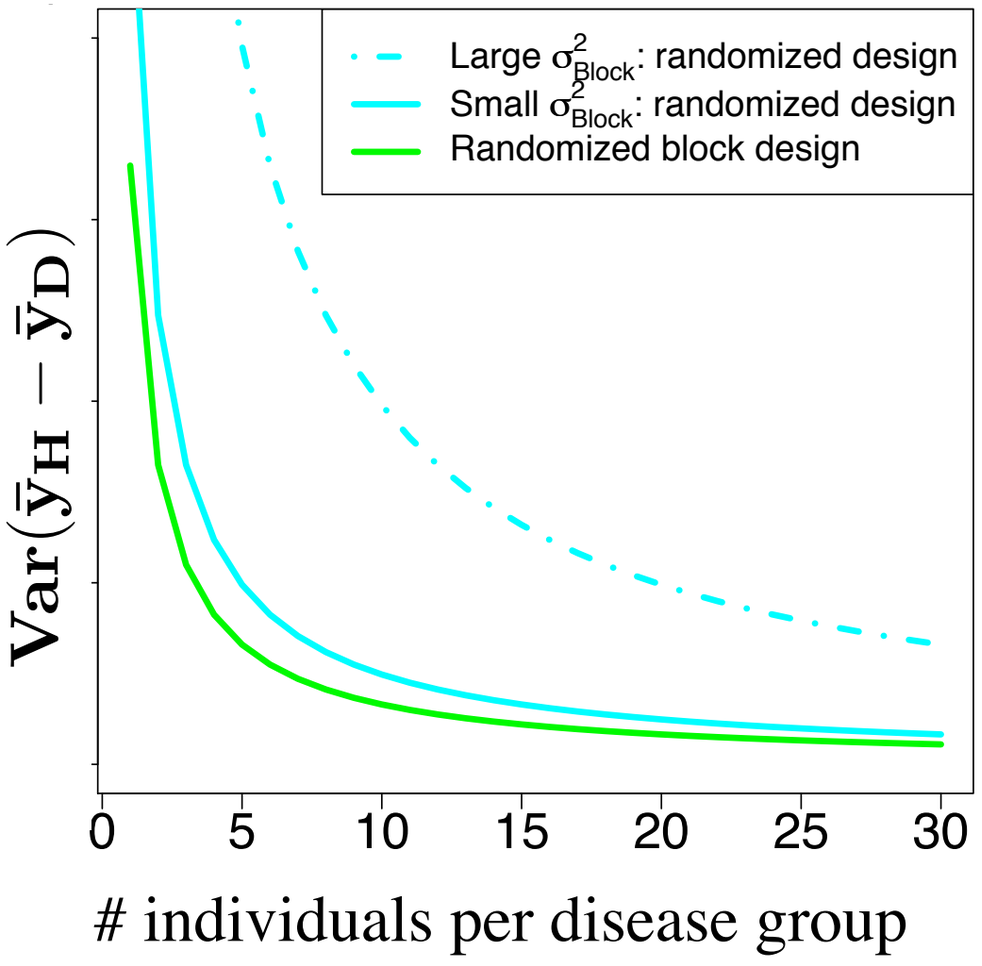
A completely randomized design

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

SO HOW MANY REPLICATES DO I NEED?

One protein, blocking

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:

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

where $\mathbf{z}_{1-\beta}$ and $\mathbf{z}_{1-\alpha/2}$ 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

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

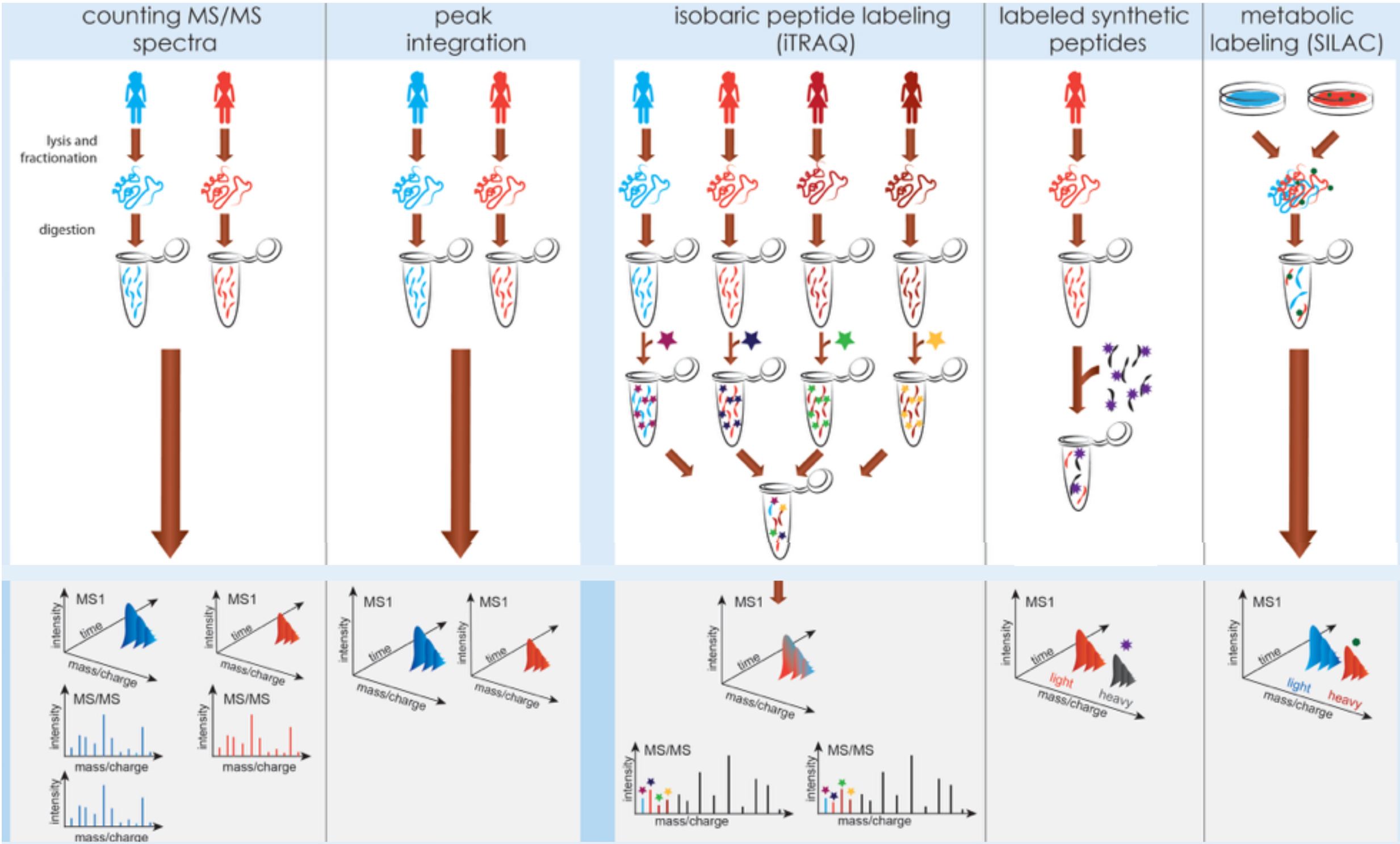
→ solve for the number of individuals I

COMPLEX DESIGNS: MULTIPLEXING

Label-free ***Label-based***

Sample preparation

Global LC-MS/MS



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

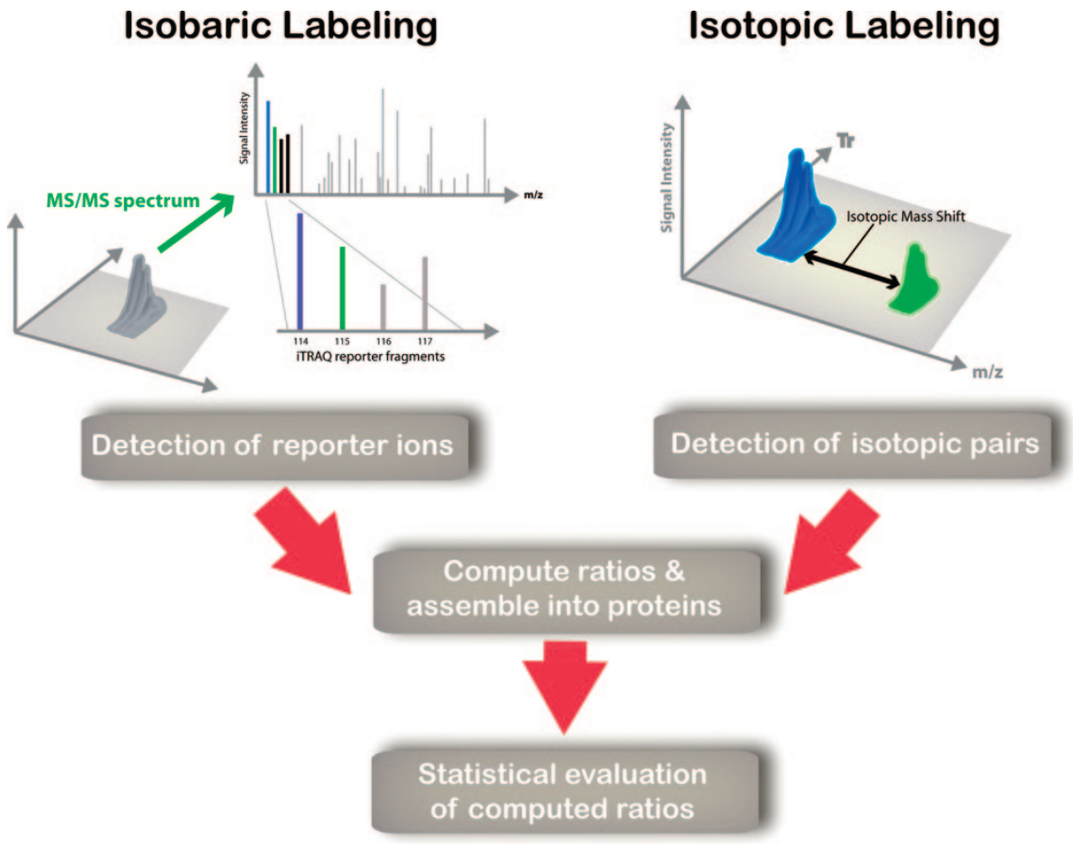
MULTIPLEXING: RUN = BLOCK

Balanced incomplete block design

Each sample appears with samples from other groups an equal number of times

Example: 5 groups, 4 labels, run=block

Disease group	Replicate set 1 Block 1	Block 2	Block 3	Block 4	Block 5	⋯
D_1	X	X	X	X		⋯
D_2	X	X	X		X	⋯
D_3	X	X		X	X	⋯
D_4	X		X	X	X	⋯
D_5		X	X	X	X	⋯



Example: 5 groups, 2 labels, run=block

Disease group	Replicate set 1 Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8	Block 9	Block 10	⋯
D_1	X_{L_1}	X_{L_2}	X_{L_1}	X_{L_2}							⋯
D_2	X_{L_2}				X_{L_1}	X_{L_2}	X_{L_1}				⋯
D_3		X_{L_1}			X_{L_2}			X_{L_1}	X_{L_2}		⋯
D_4			X_{L_2}			X_{L_1}		X_{L_2}		X_{L_1}	⋯
D_5				X_{L_1}			X_{L_2}		X_{L_1}	X_{L_2}	⋯

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

COMPLEX DESIGNS: MULTIPLEXING

Find allocations that minimize variance

A block-randomized design

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

Balanced incomplete block design

$$\mathbf{Var}(\hat{\mathbf{D}}_\mathbf{1} - \hat{\mathbf{D}}_\mathbf{2}) = \mathbf{2}\frac{\mathbf{n_b}}{\mathbf{n_g n_p n_s}}\left(\sigma^2_{\mathbf{Indiv}} + \sigma^2_{\mathbf{Error}}\right)$$

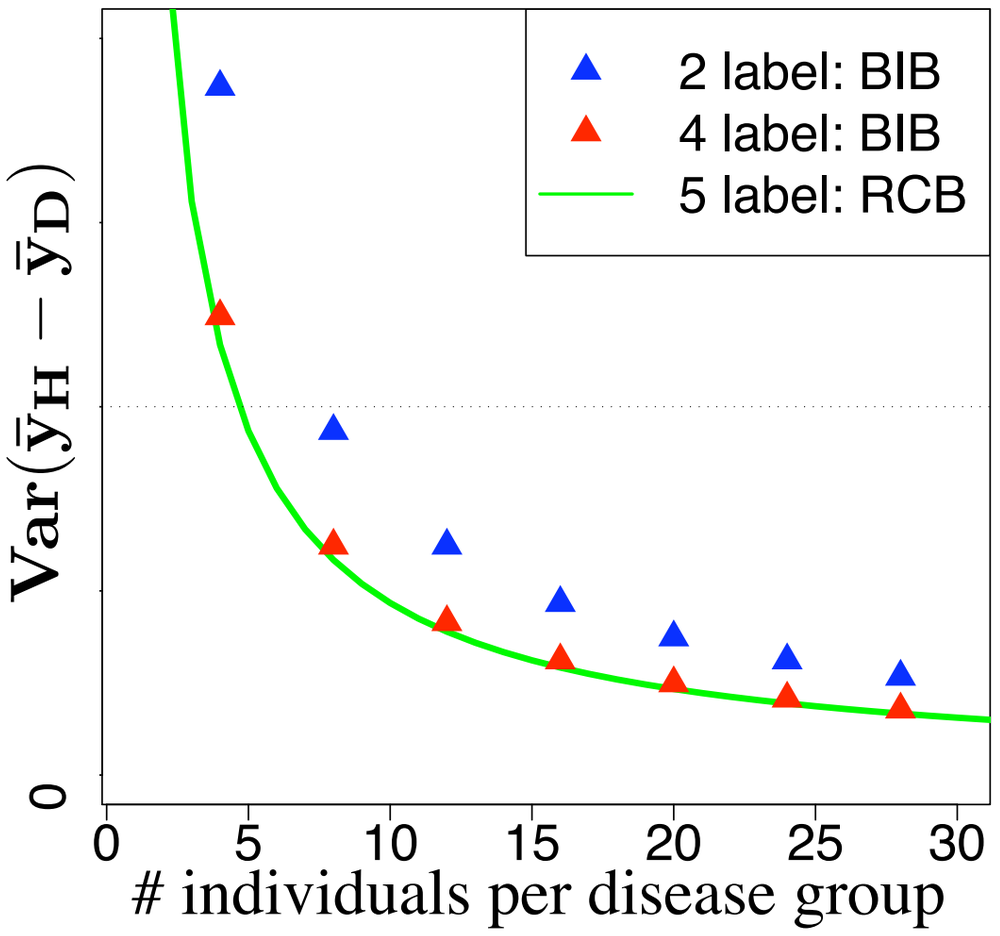
I: # individuals per disease group

$\mathbf{n_b}$, $\mathbf{n_g}$, $\mathbf{n_p}$, $\mathbf{n_s}$ - counts of paired allocations

Conclusion:
Limited sample size options
Larger blocks are more efficient
(assuming same variances in all workflows)

A pilot experiment

- 3 groups of cardiovascular disease
- iTRAQ workflow



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

COMPLEX DESIGN: POOLING

Helpful in theory; useful for a subset of research objectives

- Goal: pool biological specimens from a same group prior to the experiment
  - Insufficient biological material
  - Need to gain efficiency

- Strategy
  - Cannot pool all samples from a group (will not be able to assess the biol. variation)
  - Can pool multiple disjoint subsets of subjects

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

COMPLEX DESIGN: POOLING

Helpful in theory; useful for a subset of research objectives

- If pooled samples are viewed as averages of biological material, then
  - For a CRD with 1 tech. rep.:

$$Var(\hat{\mu}_1 - \hat{\mu}_2) = 2\left(\frac{\sigma^2_{Indiv} + \sigma^2_{Prep} + \sigma^2_{Error}}{I}\right).$$

  - Pool of r specimens:

$$Var(\hat{\mu}_1 - \hat{\mu}_2) = 2\left(\frac{\sigma^2_{Indiv}}{I\ r} + \frac{\sigma^2_{Prep} + \sigma^2_{Error}}{I}\right).$$

COMPLEX DESIGN: POOLING

Helpful in theory; useful for a subset of research objectives

- Disadvantage:
  - Potential bias: pipetting errors cause unequal contributions of specimens to the pool
  - Difficult to detect outlying or contaminated specimens
  - Downstream analysis limited to testing (i.e. no prediction/classification for individual subjects)
- Required assumption: biological averaging
  - Model-based averaging on the log scale
  - Physical mixing on the original scale
  - Assume that the two scales are equivalent (generally not true)

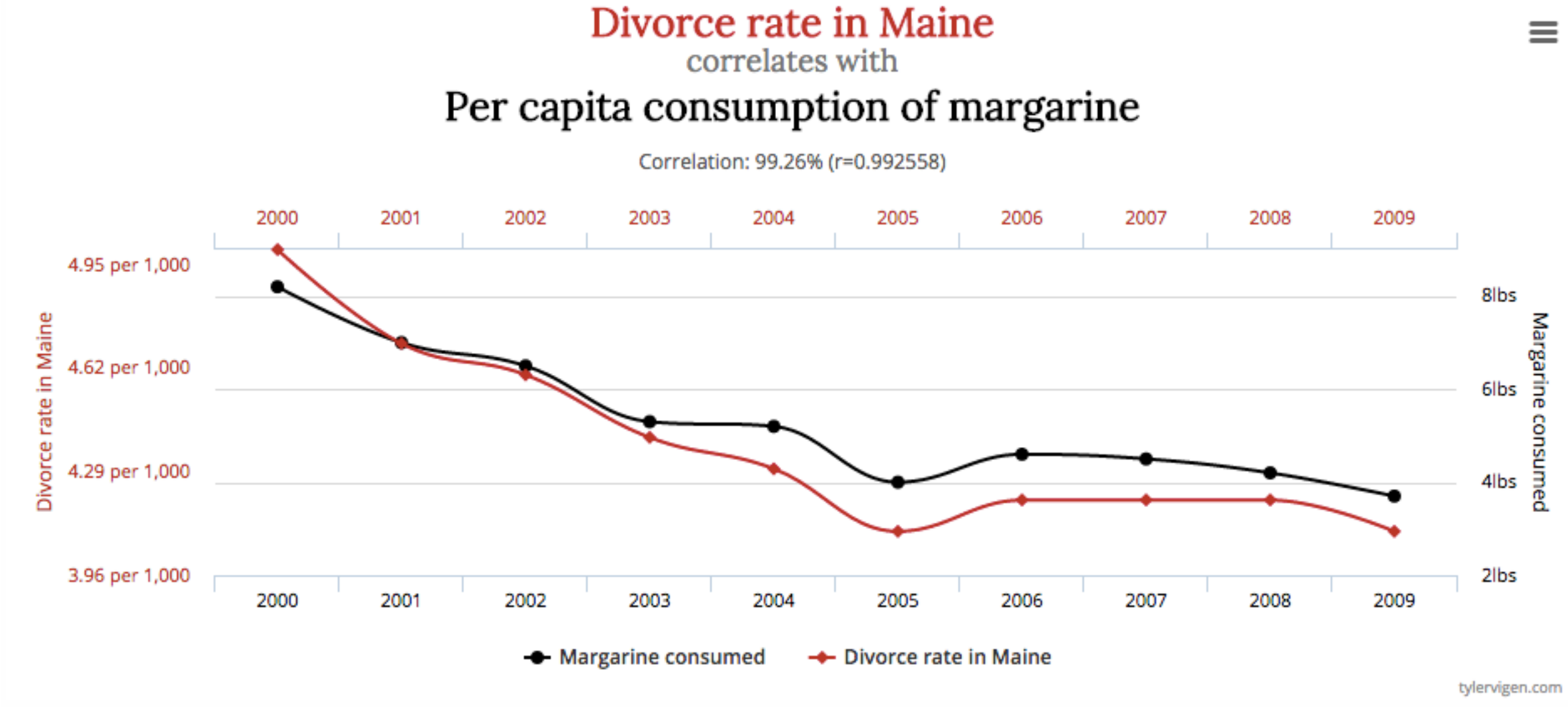
OUTLINE

- So how many replicates do I need?
  - Design of complex experiments
- Associations and calibration
  - Correlation and linear regression
- Statistical analysis of count data
  - Comparing proportions and spectral counts

ASSOCIATIONS AND CALIBRATIONS

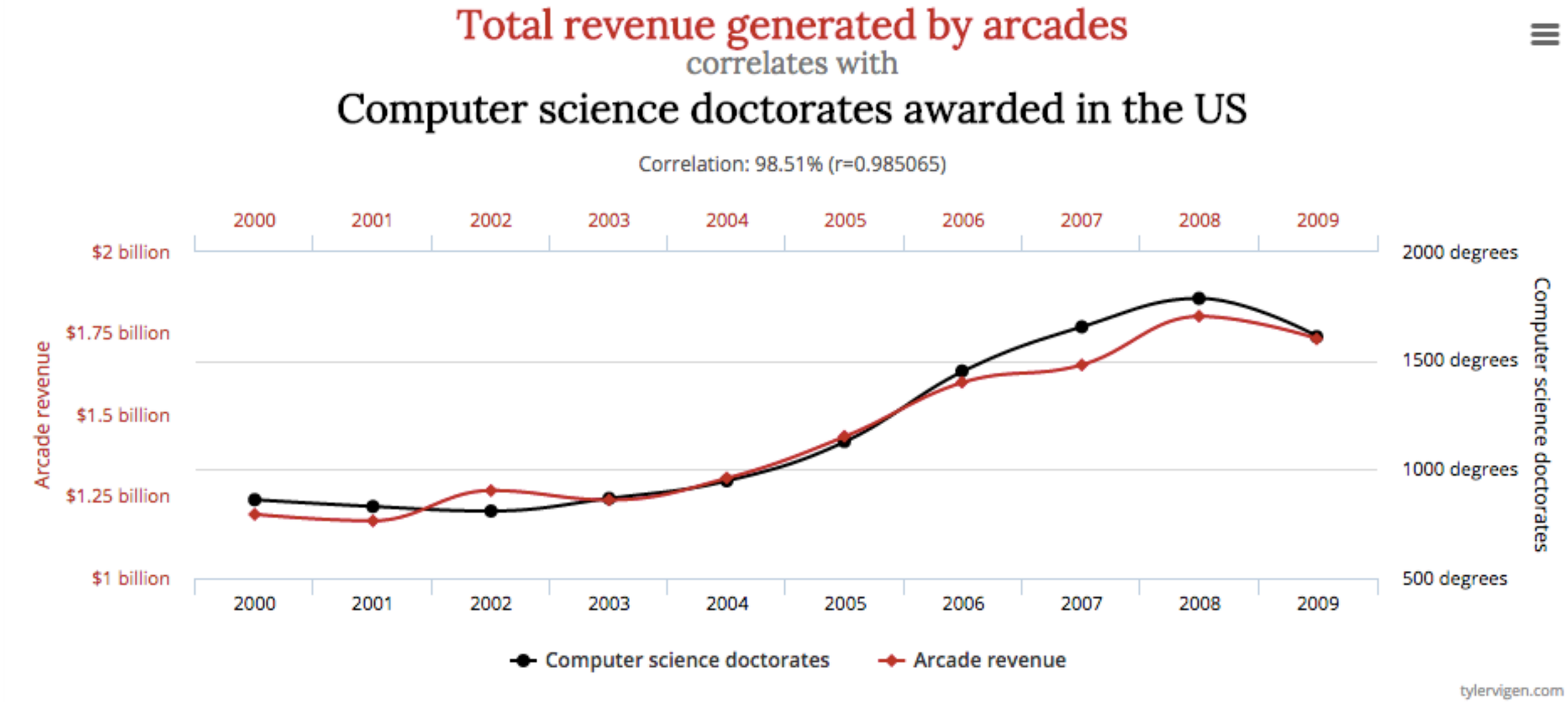
- Associations
  - do two proteins co-vary across samples?
- Calibrations
  - does protein abundance reflect the true abundance?

SPURIOUS ASSOCIATIONS ABOUND



tylervigen.com/spurious-correlations

SPURIOUS ASSOCIATIONS ABOUND

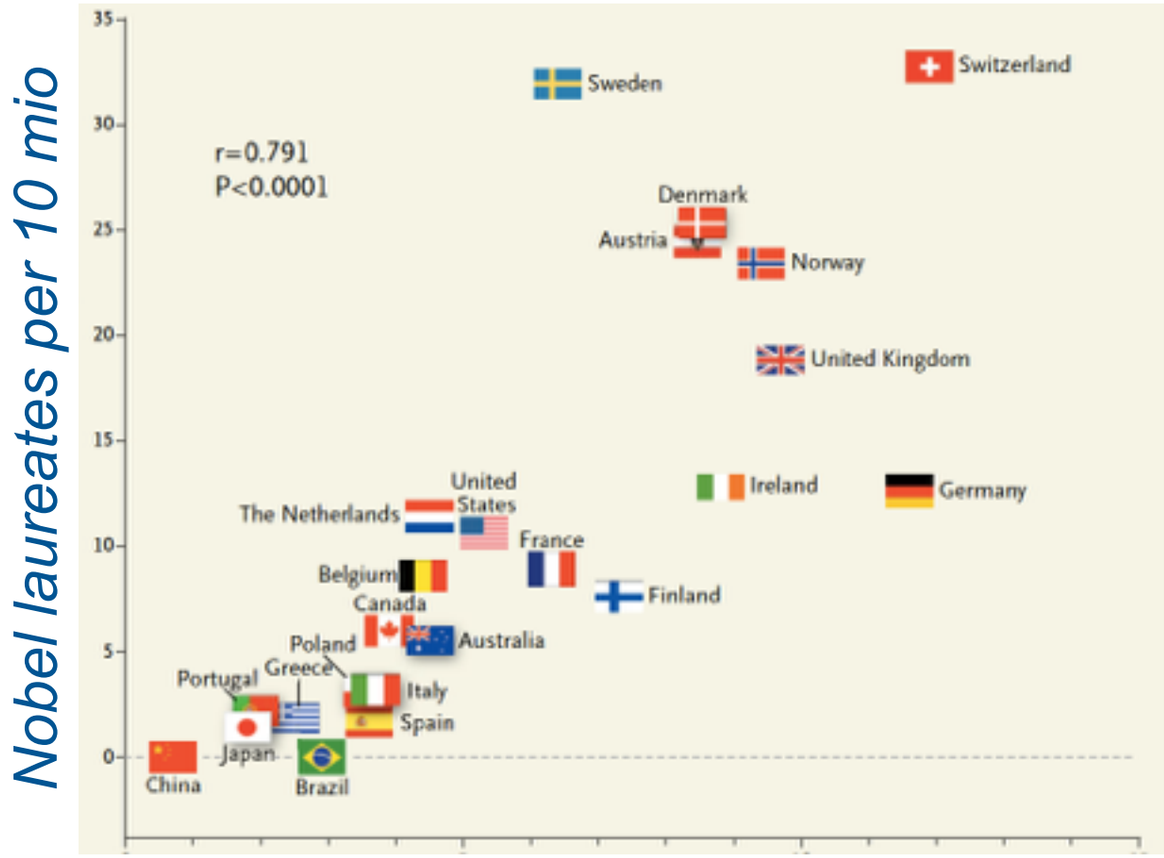


tylervigen.com/spurious-correlations

SPURIOUS ASSOCIATIONS ABOUND

Easy to dismiss when we understand the context

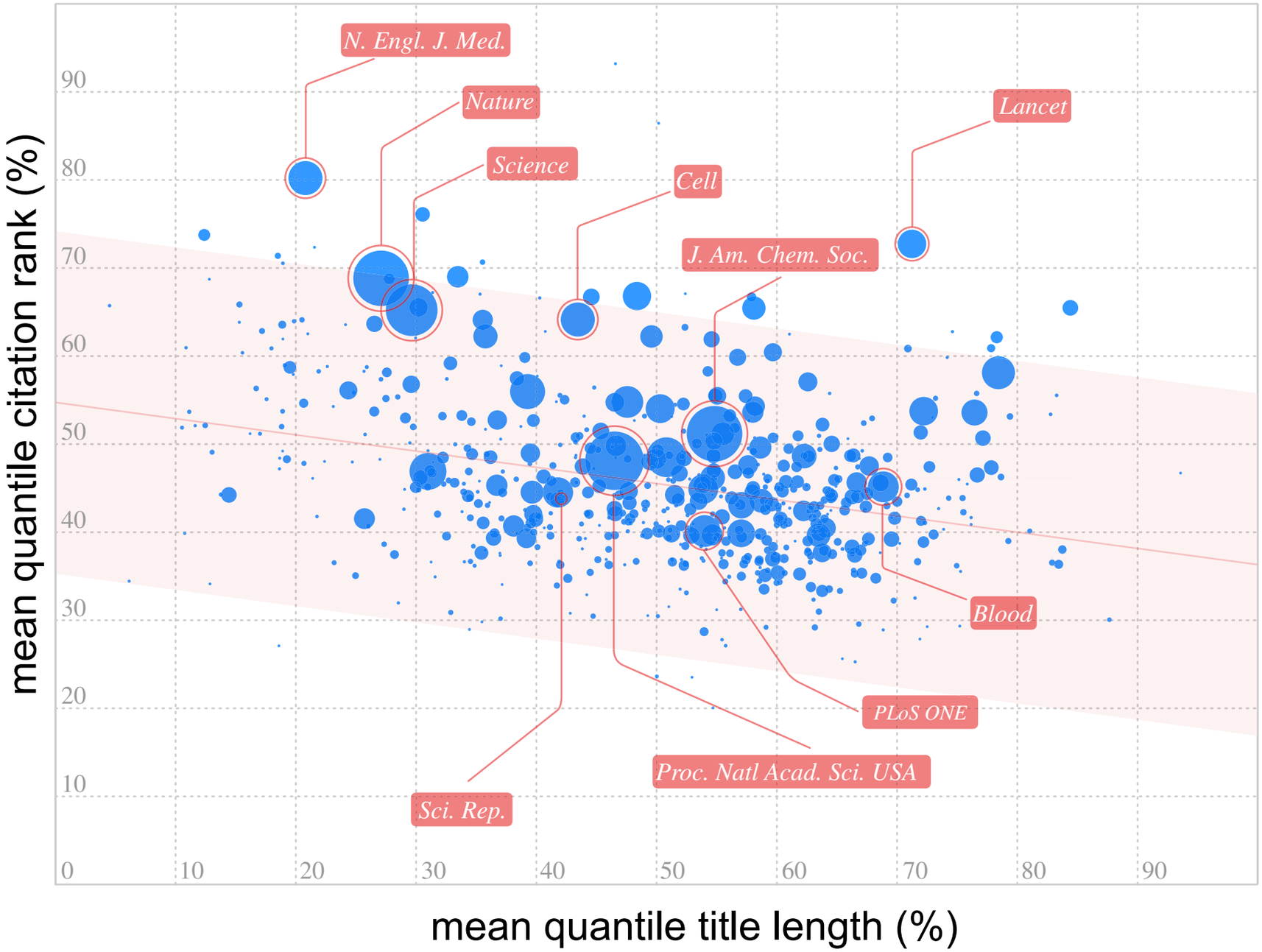
- Premier medical journal
  - *Nobel prize is related to cognitive ability*
  - *flavanols (organic molecules present in chocolate) are linked to cognitive ability*
- Technical flows
  - *Nobel prize winners between 1900-2011*
  - *Chocolate consumption after 2002*
  - *Countries with many Nobel prizes have a high Human Development Index and high per capita income*



New England Journal of Medicine, 367:1562 (2012)
A. Jogalekar, Scientific American, 2012

SPURIOUS ASSOCIATIONS ABOUND

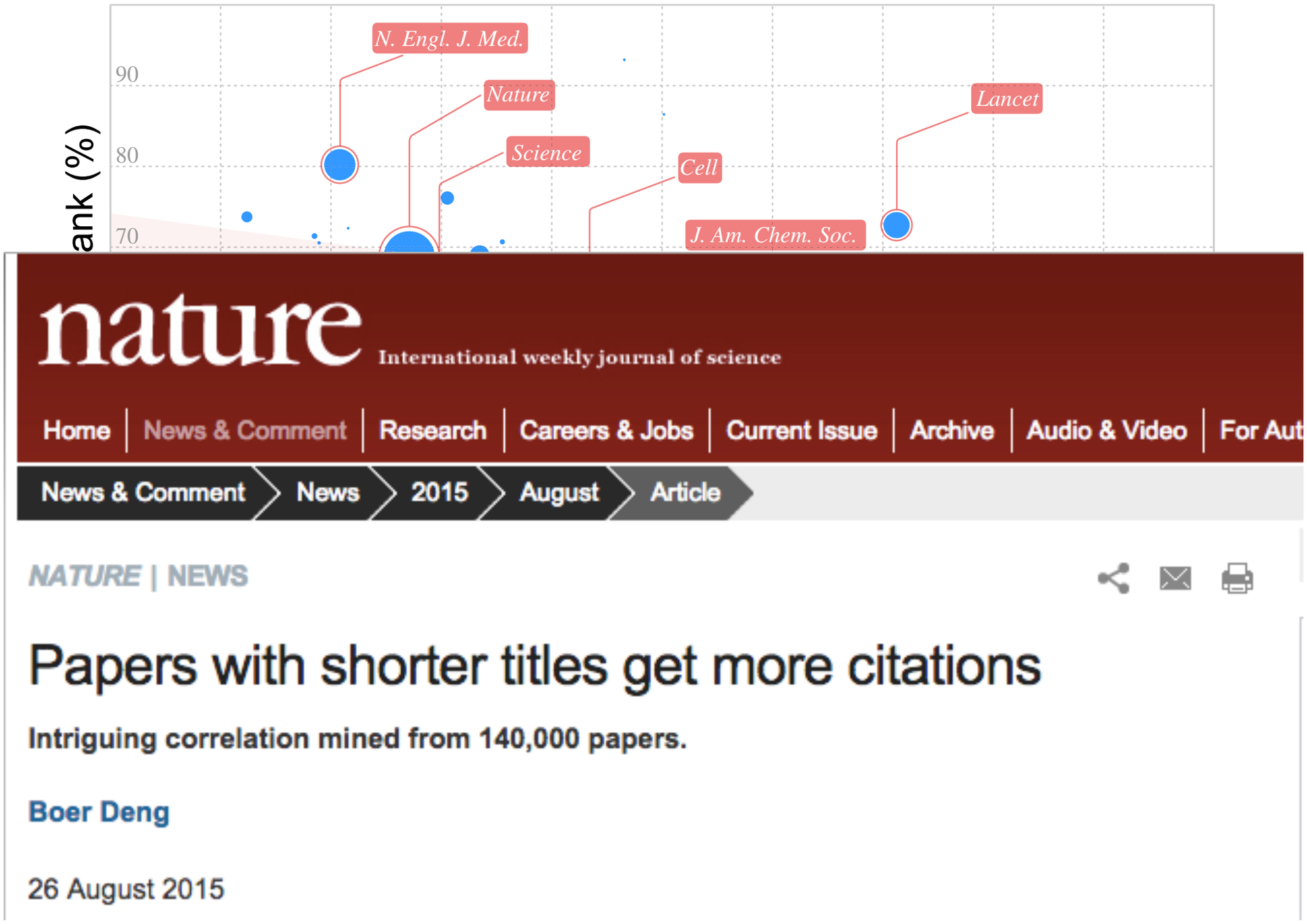
Length of title negatively correlates with number of citations



A. Letchford et al.,Royal Society Publishing, 2015

SPURIOUS ASSOCIATIONS ABOUND

Length of title negatively correlates with number of citations



A. Letchford et al.,Royal Society Publishing, 2015

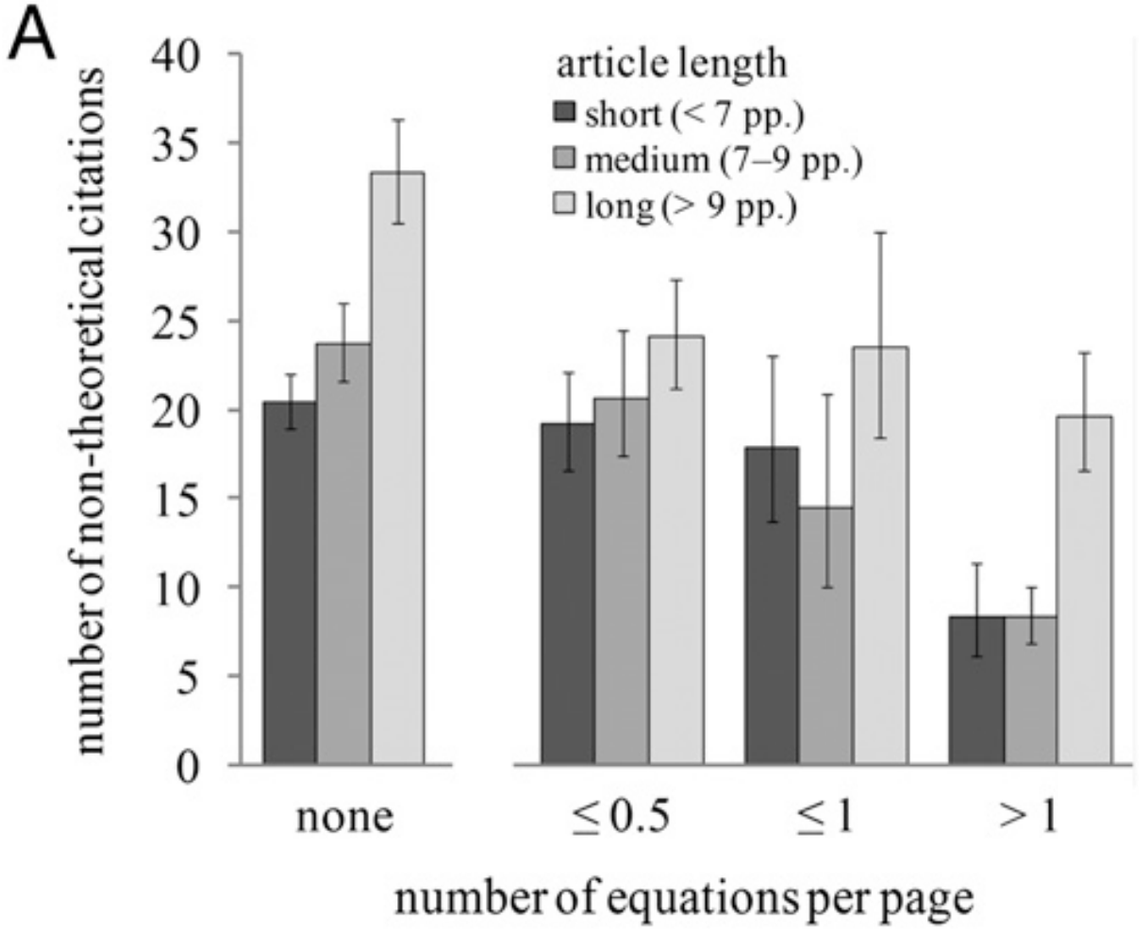
PNAS

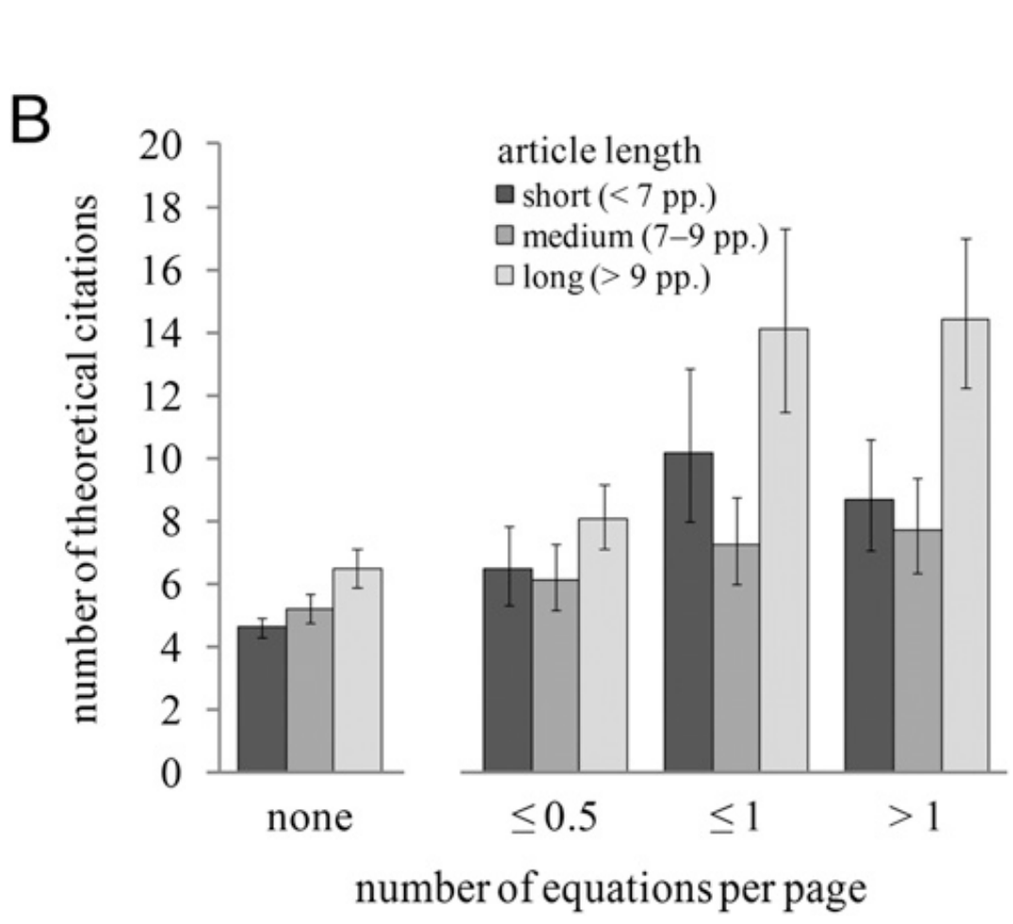
Heavy use of equations impedes communication among biologists

Tim W. Fawcett[1] and Andrew D. Higginson

School of Biological Sciences, University of Bristol, Bristol BS8 1UG, United Kingdom

Edited[†] by Robert M. May, University of Oxford, Oxford, United Kingdom, and approved June 6, 2012 (received for review April 4, 2012)





SIMPLE LINEAR REGRESSION

Correlation between

$x_1, \ldots, x_n$ and $y_1, \ldots, y_n$

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

Linear regression for

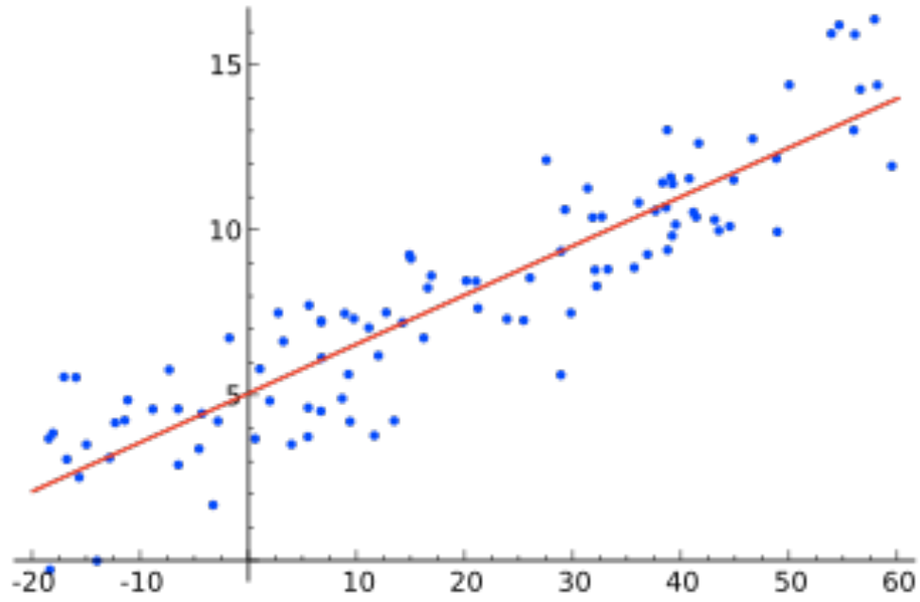
$x_1, \ldots, x_n$ and $y_1, \ldots, y_n$

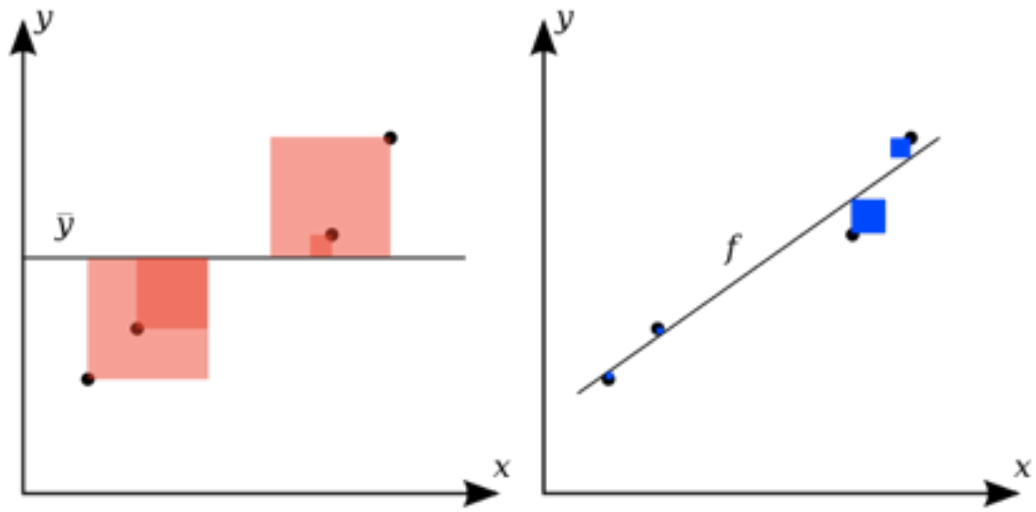
$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\operatorname{Cov}[x, y]}{\operatorname{Var}[x]}$$

$$= r_{xy}\frac{s_y}{s_x},$$



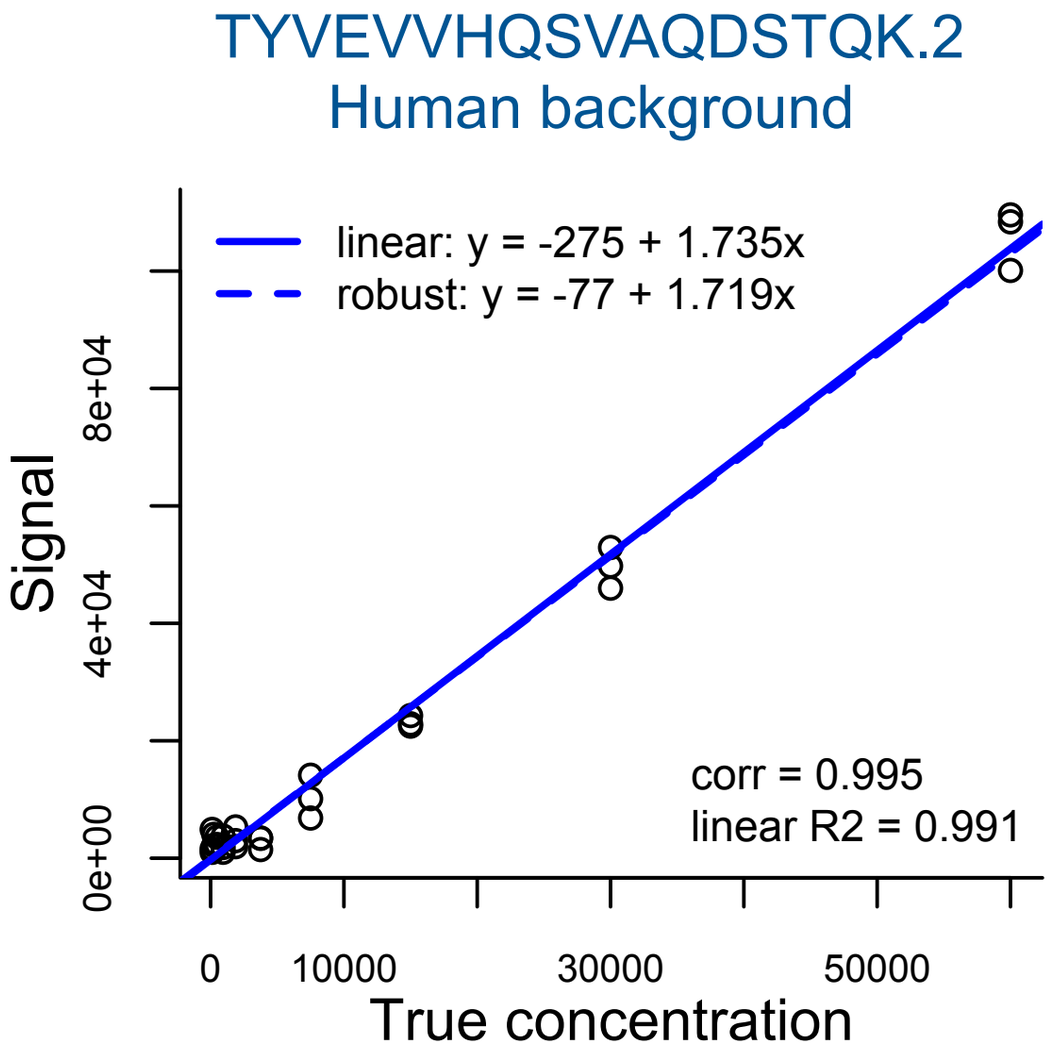


$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

$$R^2 = r_{xy}^2$$

https://en.wikipedia.org/wiki/Simple_linear_regression
https://en.wikipedia.org/wiki/Coefficient_of_determination

EXAMPLE: CALIBRATION



- Motivating example
  - OpenSWATH (Röst *et al*, *Nature Biotech*, 2014)
  - 387 peptide ions
  - 10 log-spaced concentrations
  - water, yeast, human backgrounds
- Goal: performance evaluation
  - Detect known concentrations
  - Quantify background noise
  - Quantify the slope
  - Quantify LoD and LoQ

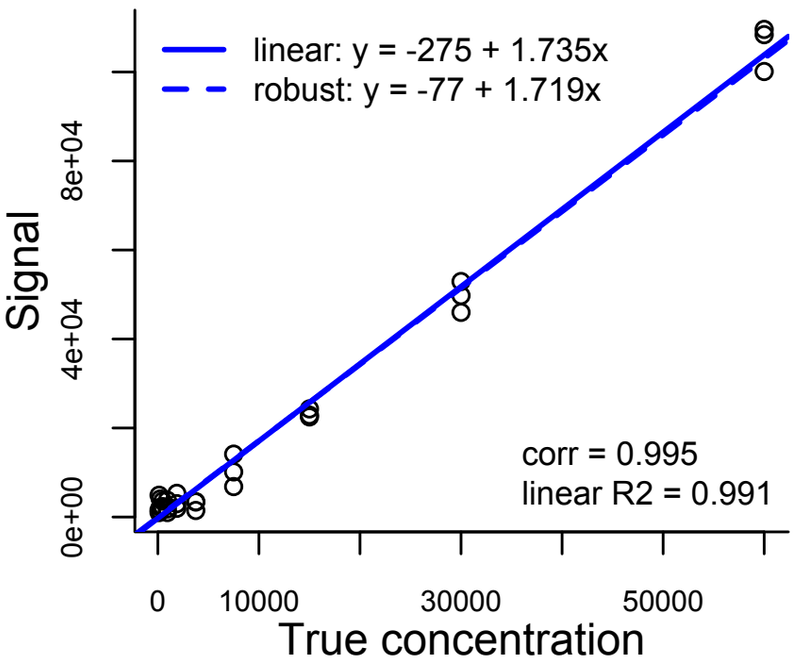
- From the graph
  - Linear relationship is theoretically plausible
  - High correlation, high R^2
  - Ordinary least squares fit agrees with robust fit

→ False perception of a good agreement

PROBLEM 1

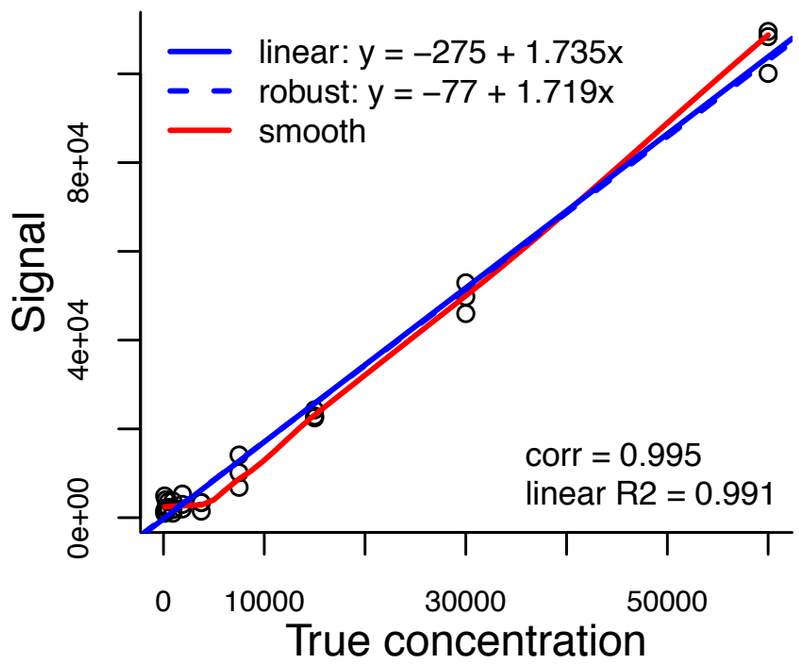
Linear fit may may uninterpretable, or wrong

All concentrations

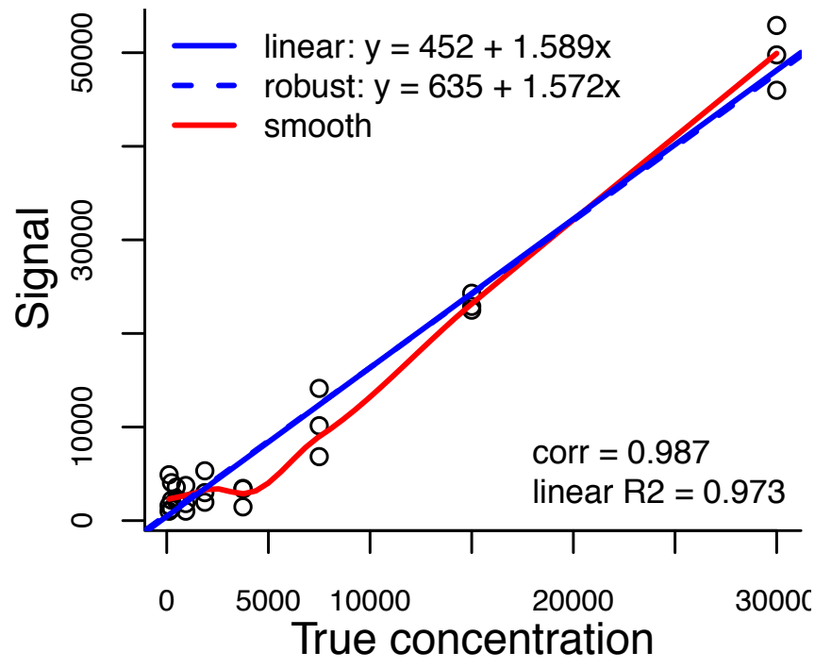


Negative intercept (no interpretation)

All concentrations, smoothed



No top concentration, smoothed



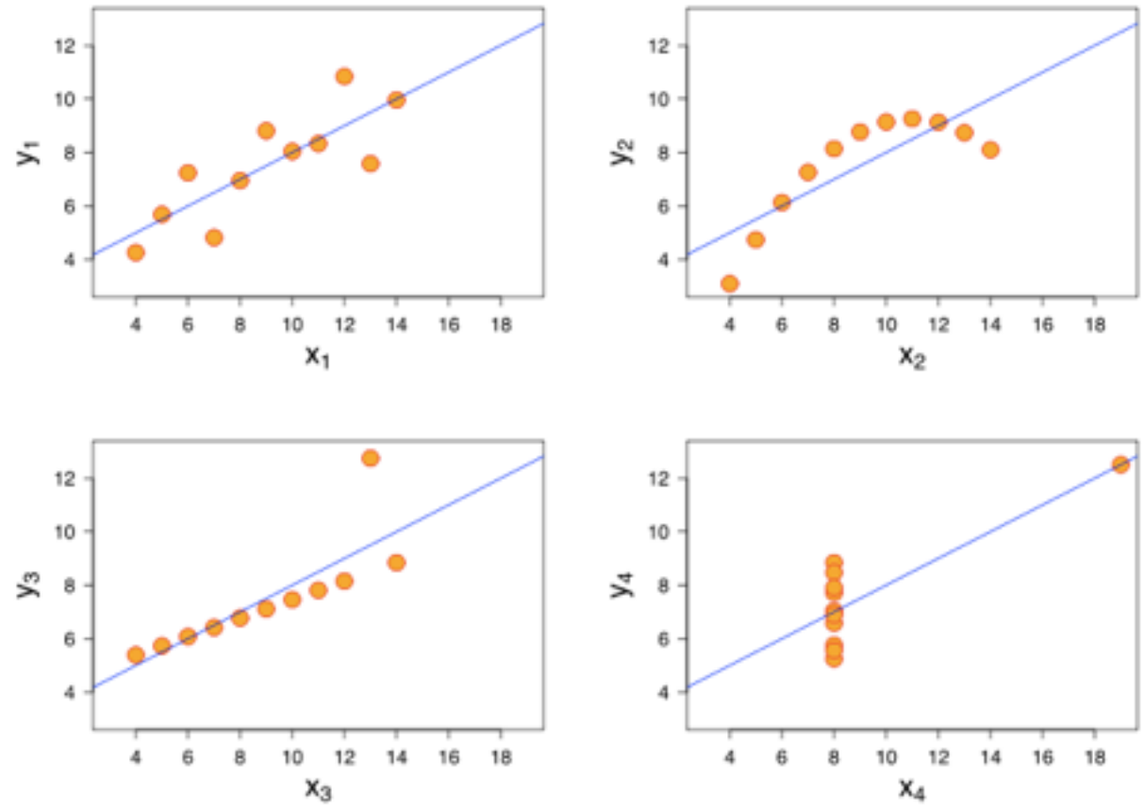
Perception of linearity depends on the # of concentrations

Correlation, slope and R^2 do not quantify linearity

PROBLEM 1

Linear fit may may uninterpretable, or wrong

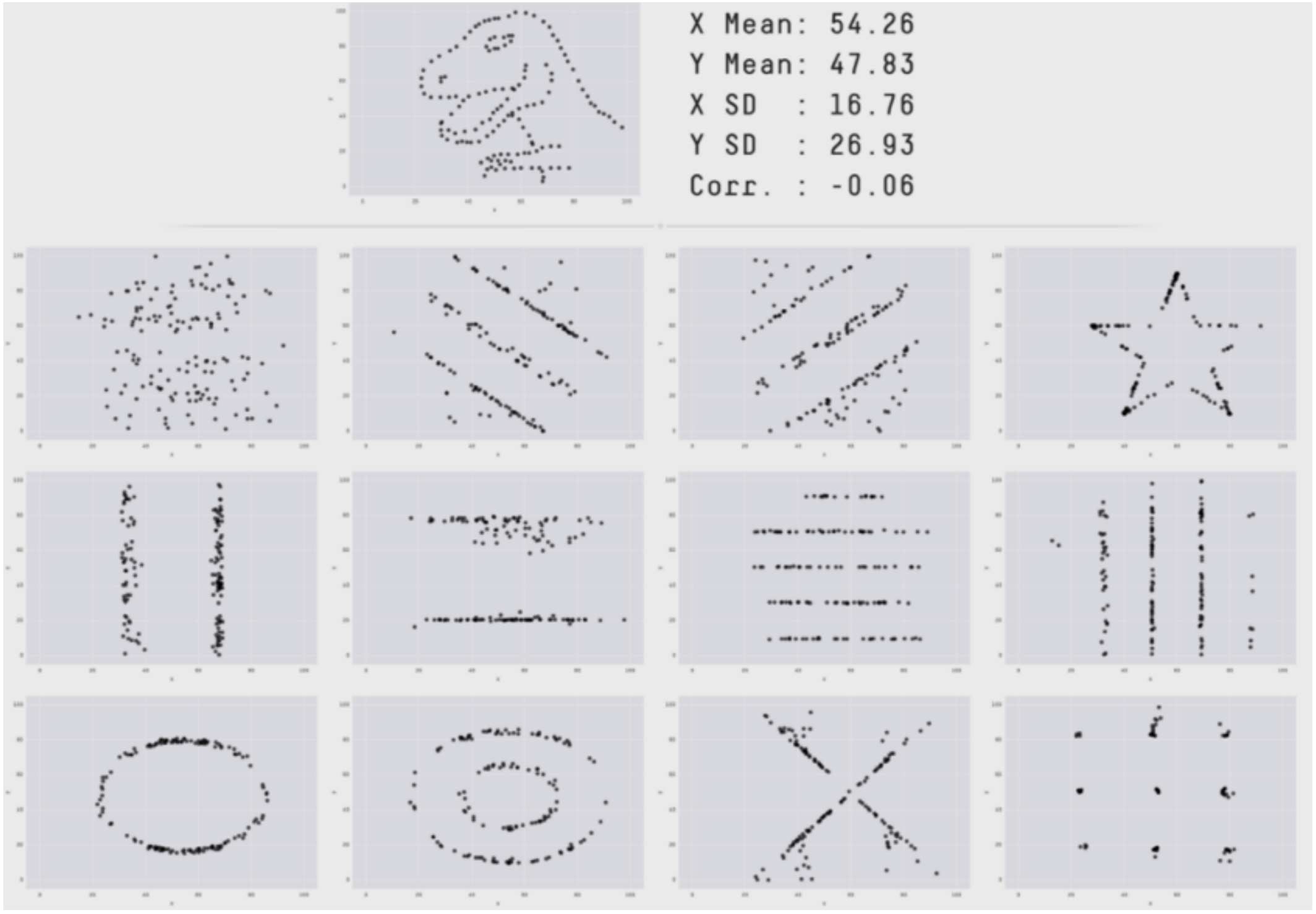
Property	Value
Mean of *x*	9
Sample variance of *x*	11
Mean of *y*	7.50
Sample variance of *y*	4.125
Correlation between *x* and *y*	0.816
Linear regression line	$y = 3.00 + 0.500x$



Anscombe quartet: all datasets have same means, variances, correlations and R^2

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

DATASAURUS

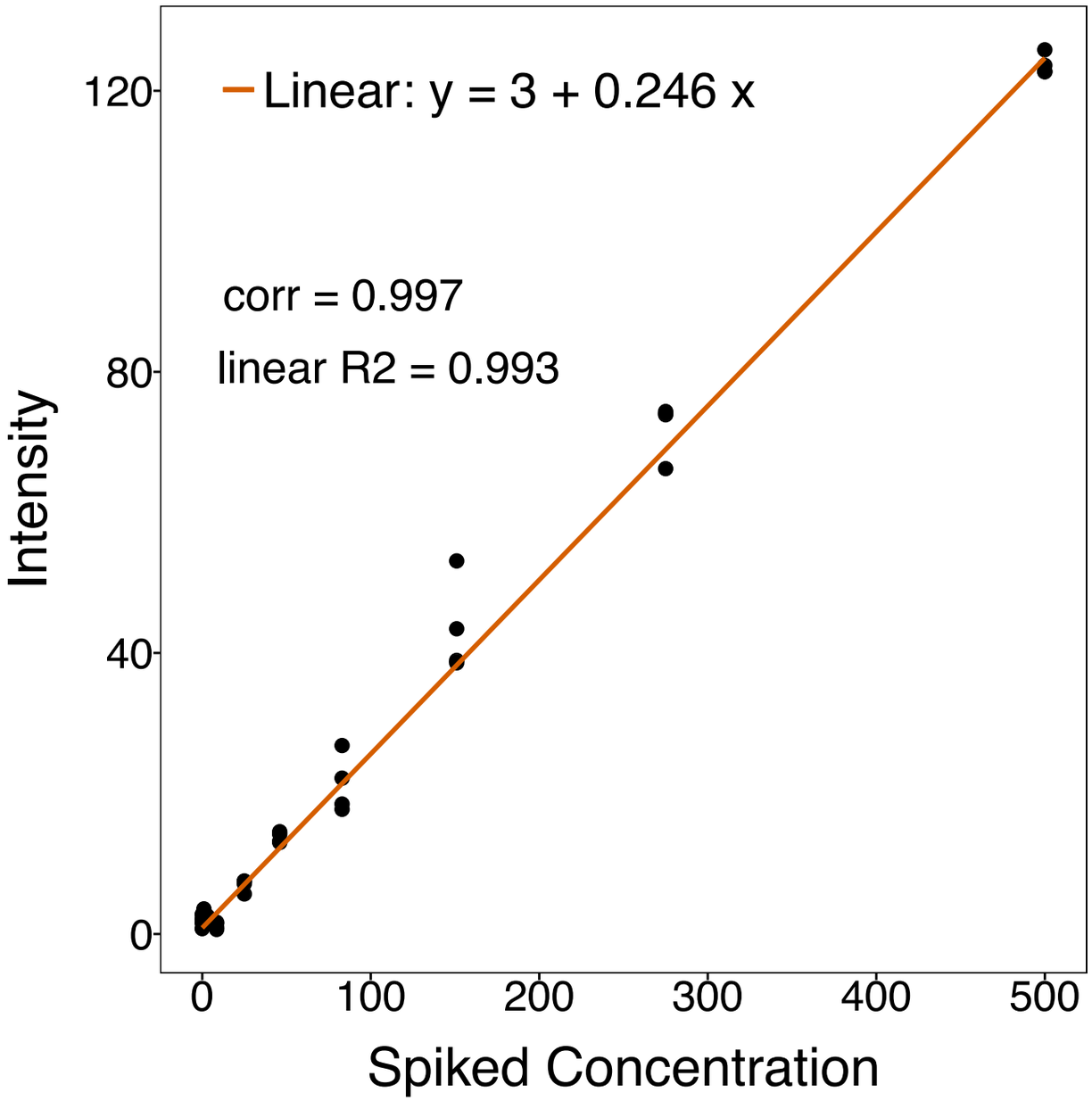


All datasets have same means, variances, correlations and R^2

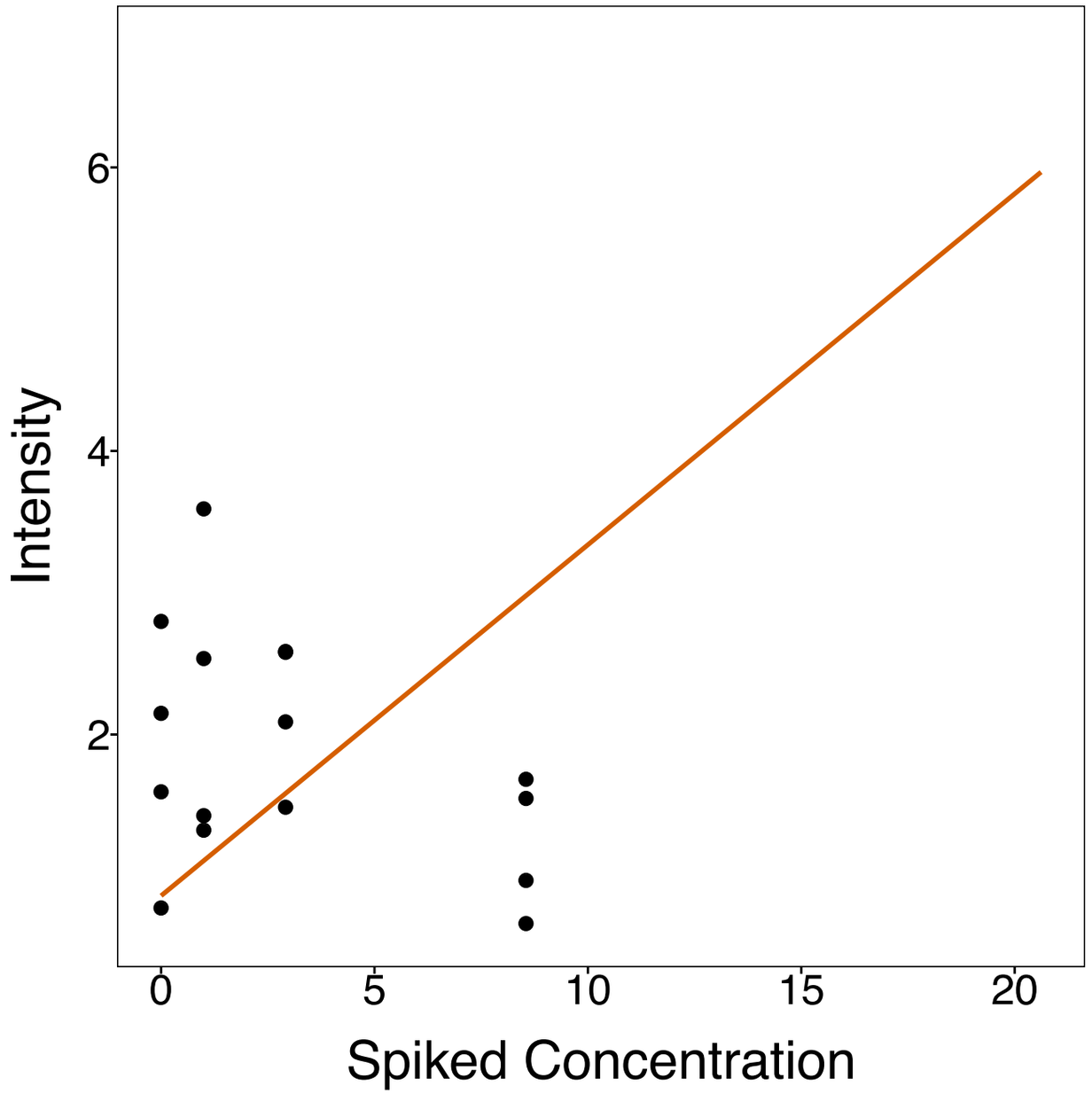
https://www.autodeskresearch.com/publications/samestats

MORE EXAMPLES

High R^2 does not always mean good fit



Zoom out

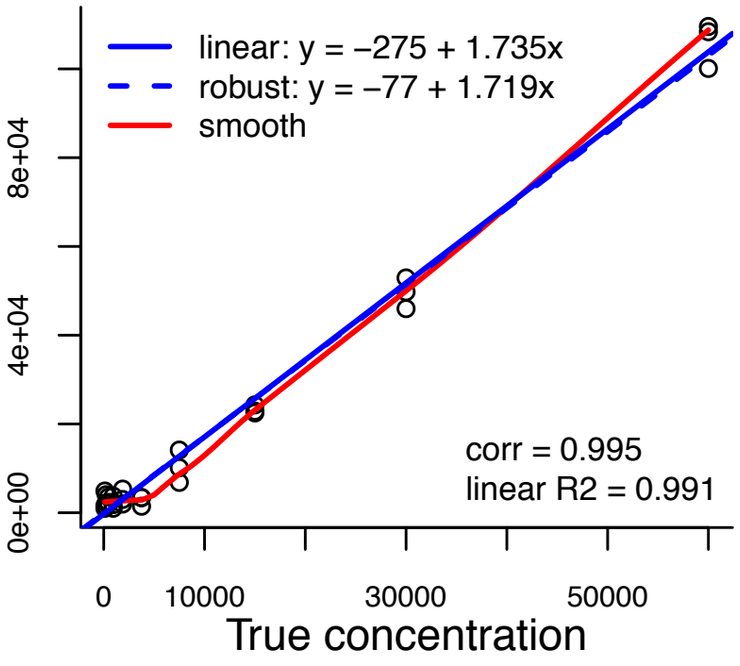


Zoom in

Calibration experiment, CPTAC

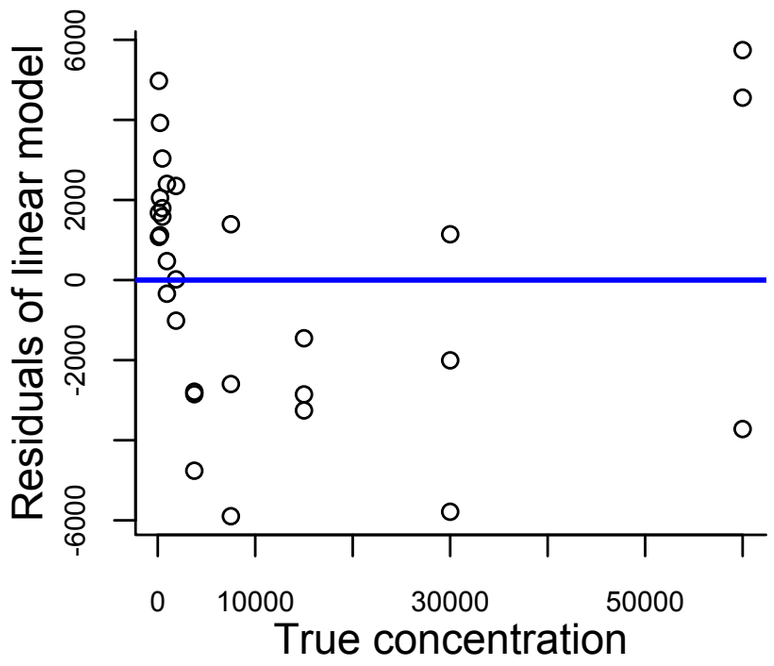
PROBLEM 2
Unequal leverage, unequal variance

All concentrations



High concentrations are influential

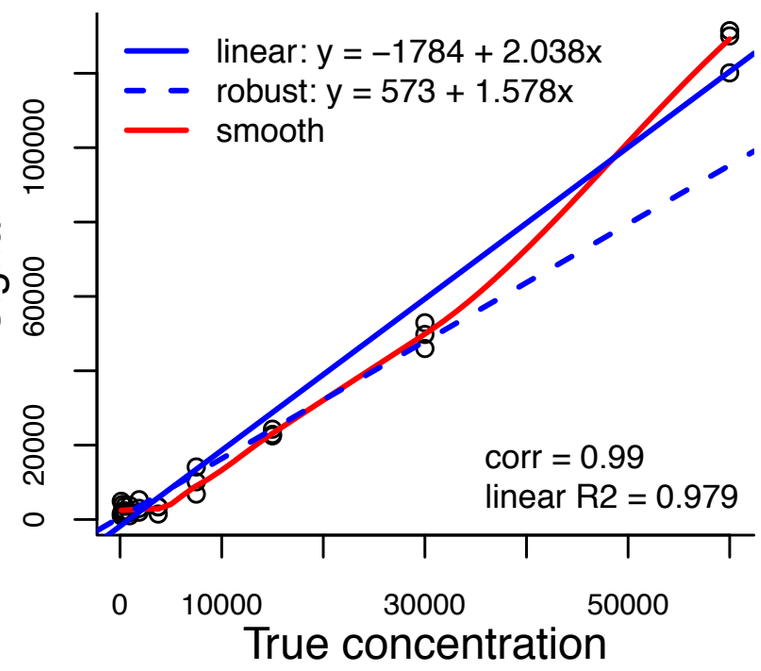
Residuals of the linear model fit



High concentrations have more variance

Low concentrations have worse fit

Intensities at max concentration up by 20%



Moving the intensities produces a similarly good fit

The fit is unduly affected by highly variable values

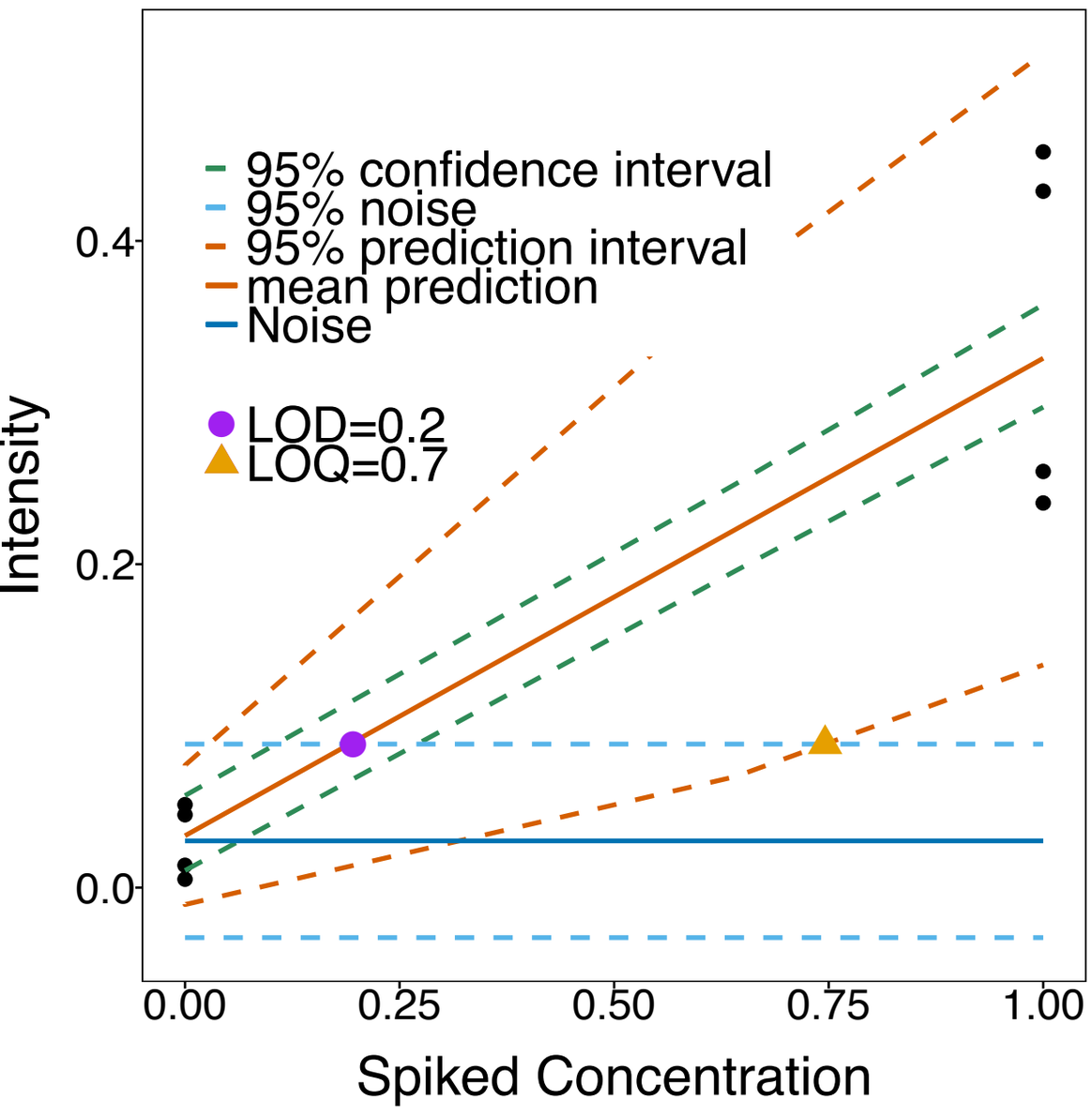
Poor fit at low concentrations is unnoticed

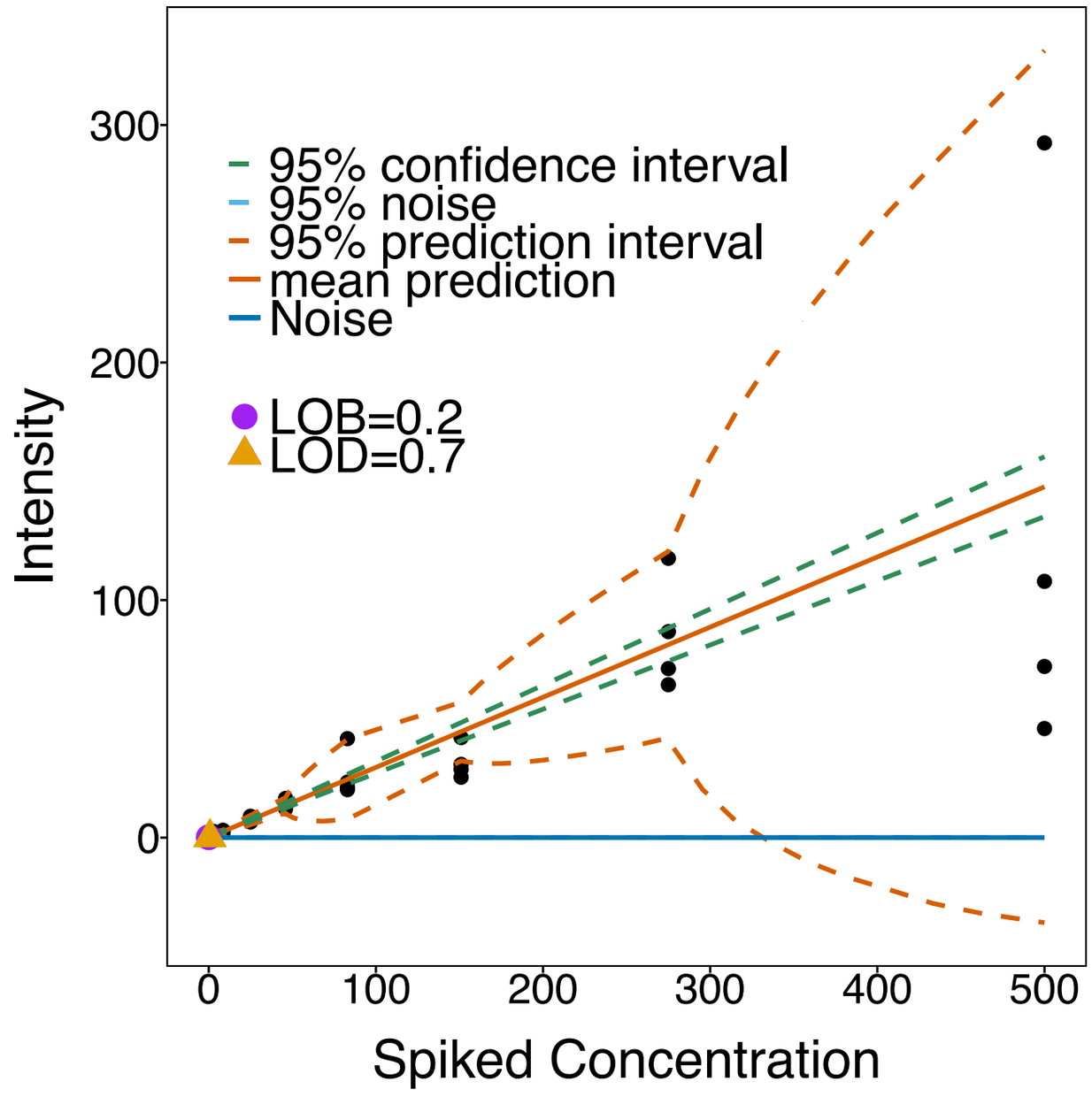
CORRELATION

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

CONFIDENCE VS PREDICTION

Prediction intervals should be used for determining LoD and LoQ





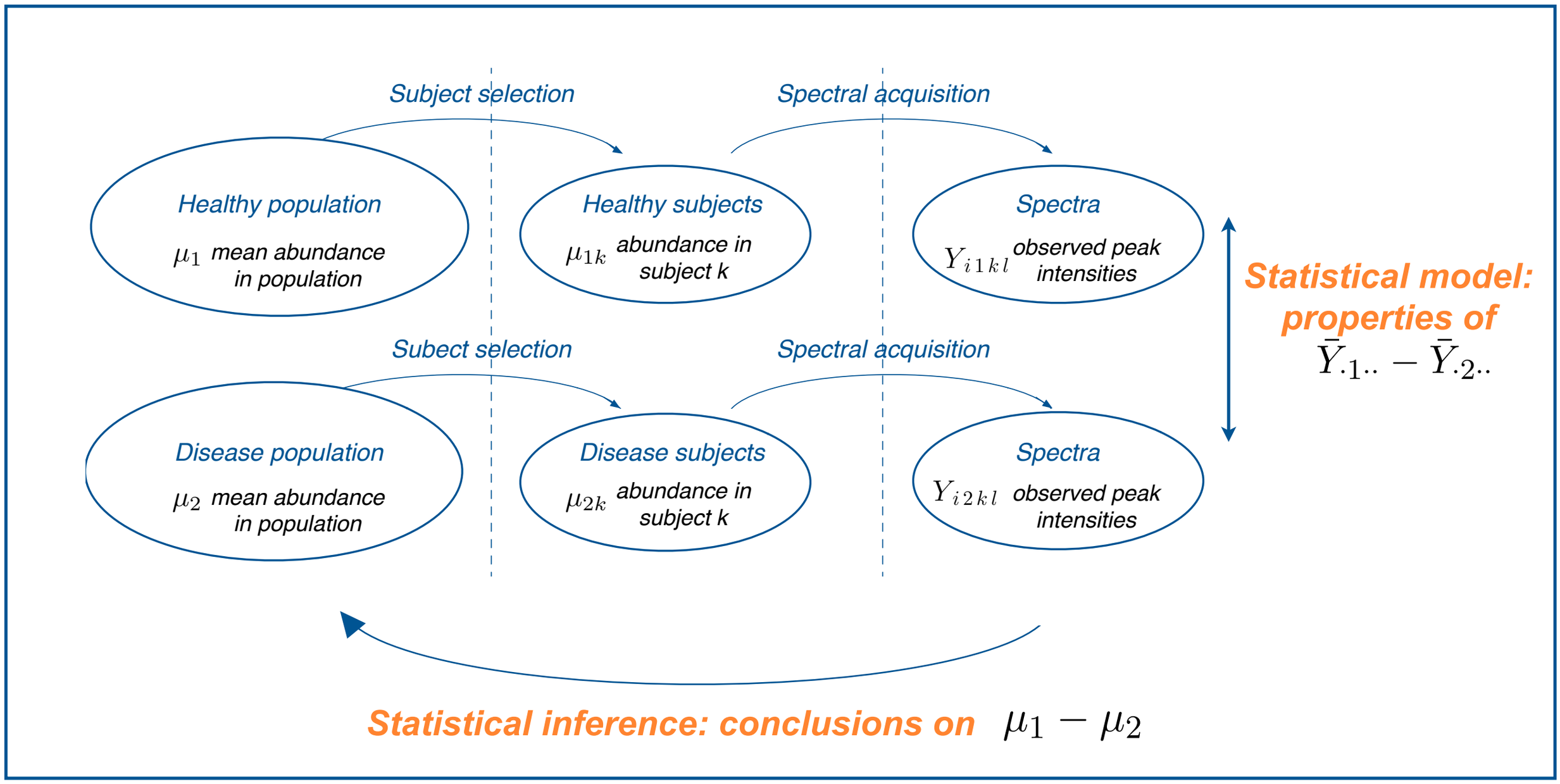
Confidence interval: $\left[\hat{y} - t_{\alpha/1}^{n-2} \cdot \sqrt{\text{var}\{\hat{y}\}},\ \hat{y} + t_{\alpha/1}^{n-2} \cdot \sqrt{\text{var}\{\hat{y}\}}\right]$

Prediction interval: $\left[\hat{y} - t_{\alpha/1}^{n-2} \cdot \sqrt{\text{var}\{\hat{y}\} + s^2},\ \hat{y} + t_{\alpha/1}^{n-2} \cdot \sqrt{\text{var}\{\hat{y}\} + s^2}\right]$

OUTLINE

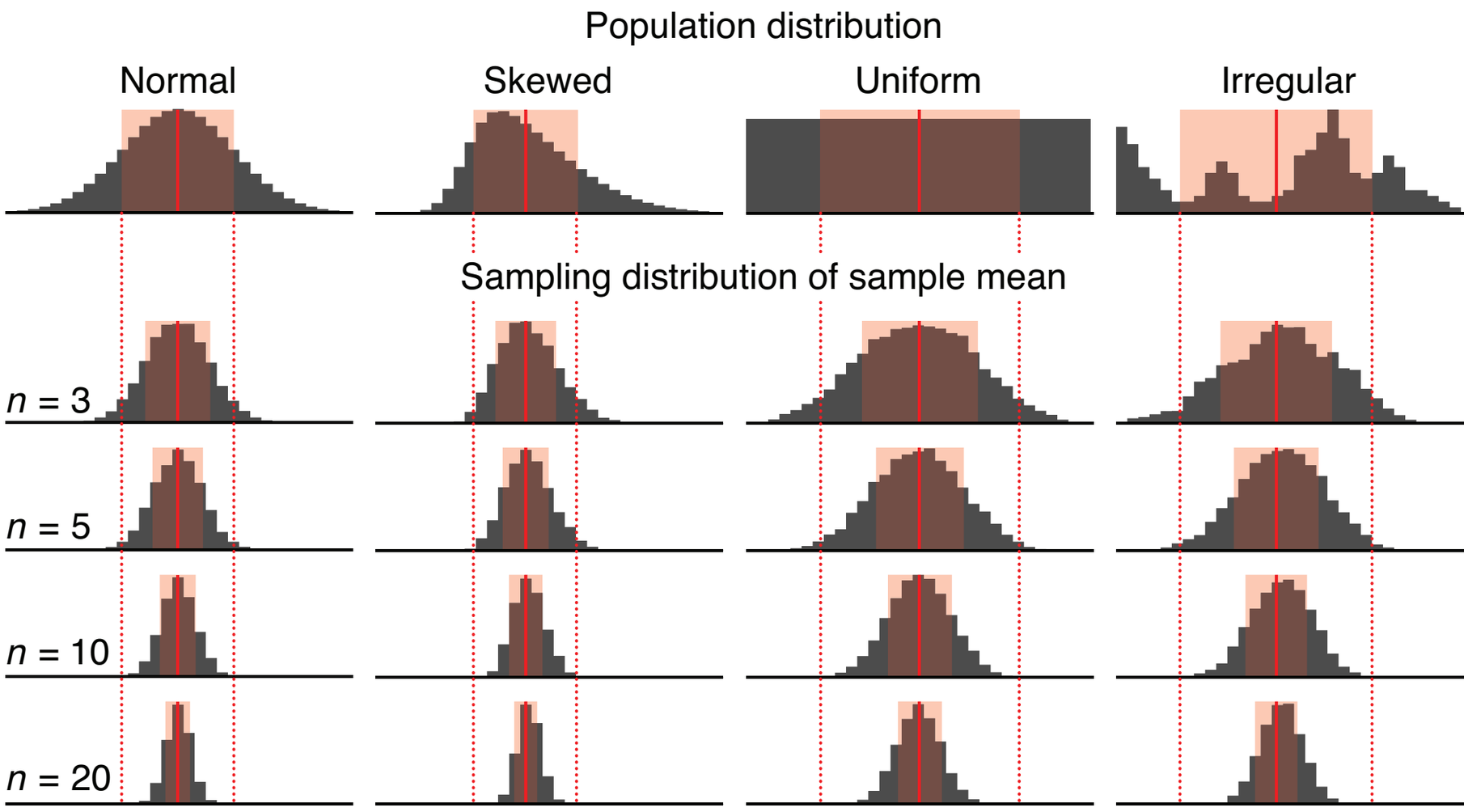
- So how many replicates do I need?
  - Design of complex experiments
- Associations between two measurements
  - Correlation and linear regression
- Statistical analysis of count data
  - Comparing proportions and spectral counts

RECALL THE GOAL OF GROUP COMPARISON



What if measurements are binary?
(E.g., up or down?)

RECALL THE CENTRAL LIMIT THEOREM



Probability distribution of the data

Repeatedly selecting n data points and calculating means

Averages of binary variables will approach Normal distribution as sample size increases

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

ANALYZING PROPORTIONS

Hypothesis testing and confidence interval

Population proportions

H0: 'status quo', no change in abundance, $\pi_1 - \pi_2 = 0$

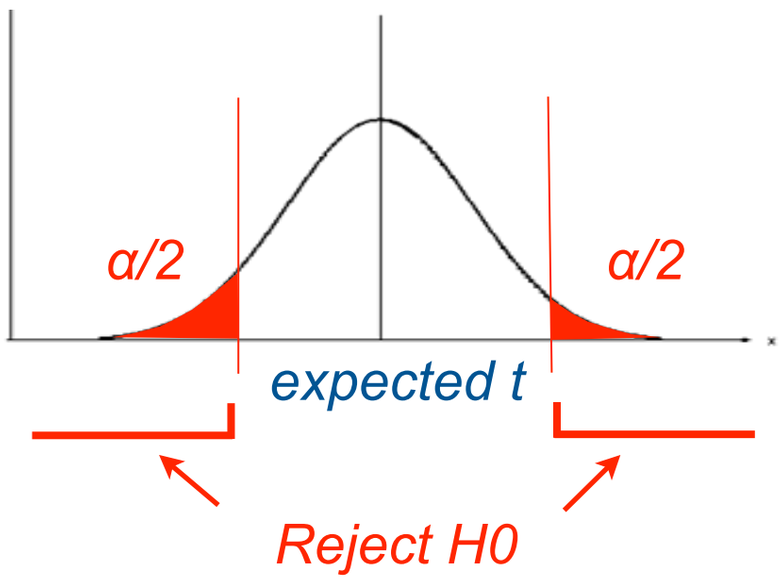
Ha: change in abundance, $\pi_1 - \pi_2 \neq 0$

Proportions in each group

$$\textbf{observed t} = \frac{\textbf{difference of group means}}{\textbf{estimate of variation}} = \frac{p_1 - p2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Distribution of the score if H0 is true ~ Normal (0,1)

α = False Positive Rate

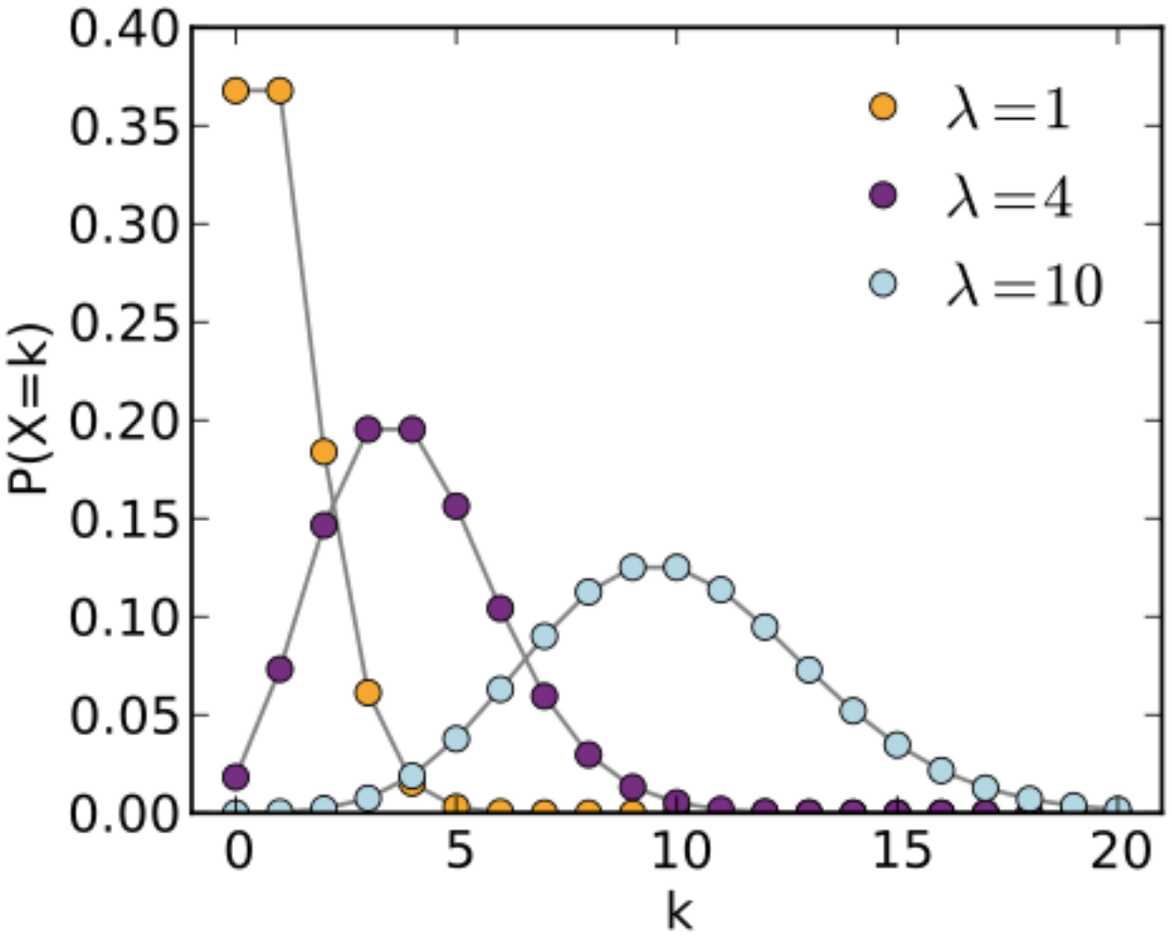


Confidence interval for difference of proportions

$$p_1 - p_2 \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

CATEGORICAL MEASUREMENTS

Comparing counts of MS/MS spectra



Poisson distribution

Negative Binomial distribution

https://en.wikipedia.org/wiki/Poisson_distribution
https://en.wikipedia.org/wiki/Negative_binomial_distribution

Northeastern University

THE COLLEGE OF SCIENCE AND
THE BARNETT INSTITUTE OF CHEMICAL AND BIOLOGICAL ANALYSIS

PRESENTS

The Barry L. Karger Medal Lectures and Award

MAY 15-16, 2017

CURRY STUDENT CENTER
NORTHEASTERN UNIVERSITY
346 HUNTINGTON AVENUE
BOSTON, MASSACHUSETTS

WITH DISTINGUISHED LECTURER

DR. RUDOLF AEBERSOLD

Head of Department of Biology
Institute of Molecular Systems Biology
ETH Zurich, Zurich, Switzerland

SCHEDULE OF EVENTS

MONDAY, MAY 15, 2017

3:30 P.M. LECTURE RECEPTION

4:00 P.M. KARGER MEDAL PUBLIC LECTURE
BALLROOM, CURRY STUDENT CENTER
"The Proteome in Context"

TUESDAY, MAY 16, 2017

9:30 A.M. CONTINENTAL BREAKFAST

10:00 A.M. KARGER MEDAL TECHNICAL LECTURE
MCLEOD SUITES, CURRY STUDENT CENTER
"SWATH-MS: Principles, current state and new developments"

calendar.northeastern.edu/event/the_barry_l_karger_medal_lectures_and_award