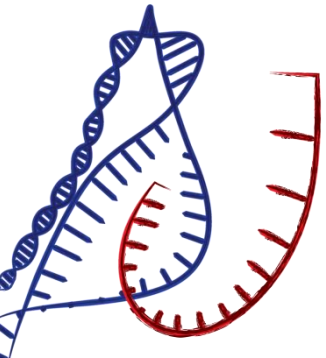
Interpretation of Quantitative Data

Marc Vaudel

Center for Medical Genetics and Molecular Medicine,
Haukeland University Hospital, Bergen, Norway

KG Jebsen Center for Diabetes Research, Department of Clinical Science,
University of Bergen, Norway



Foreword

Disclaimer:

- I am not an expert
- This is by no means exhaustive

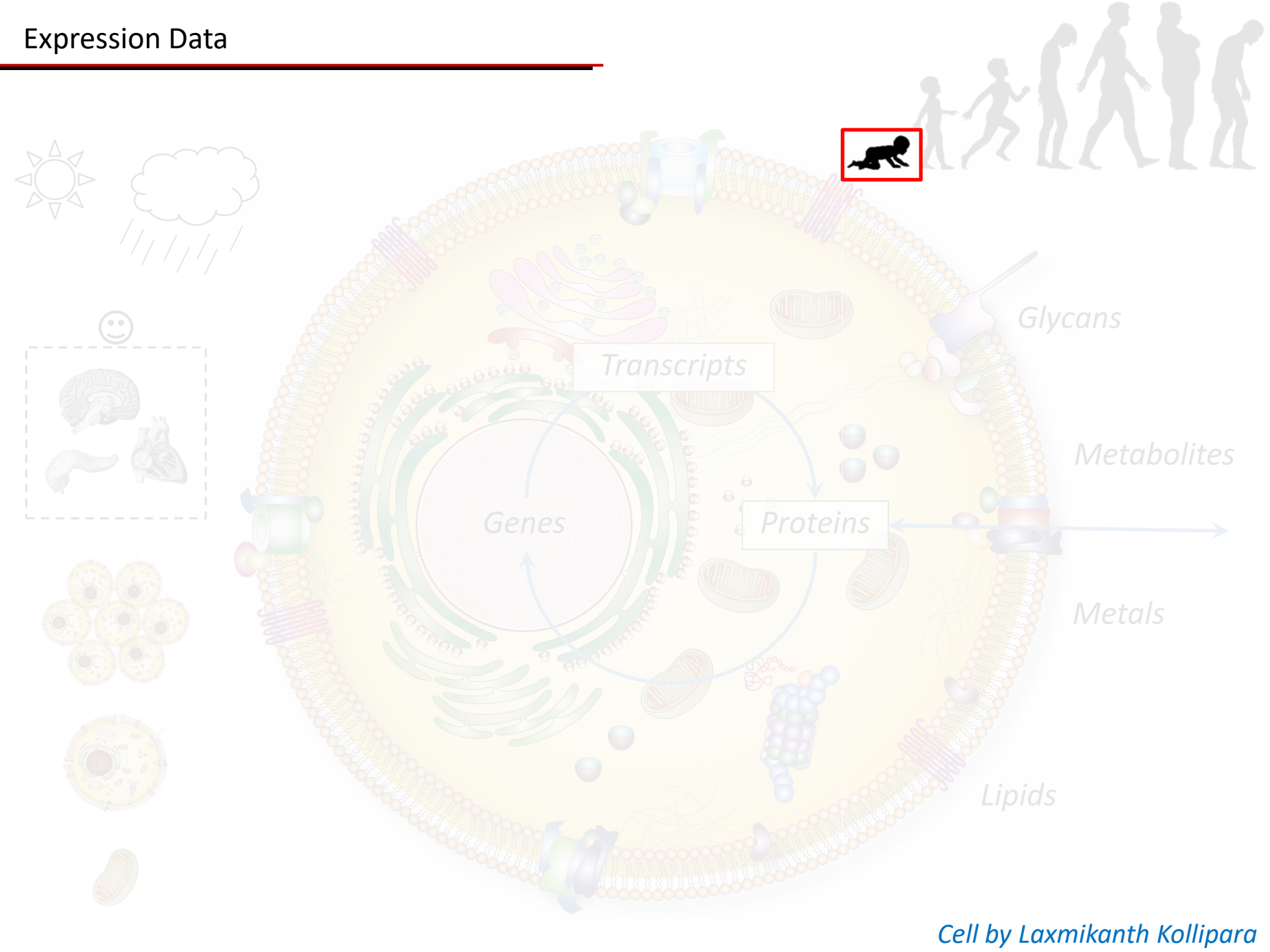
Take home message:

- There is a world beyond Excel/Graphpad/Sigmaplot
- How to get started

Warning:

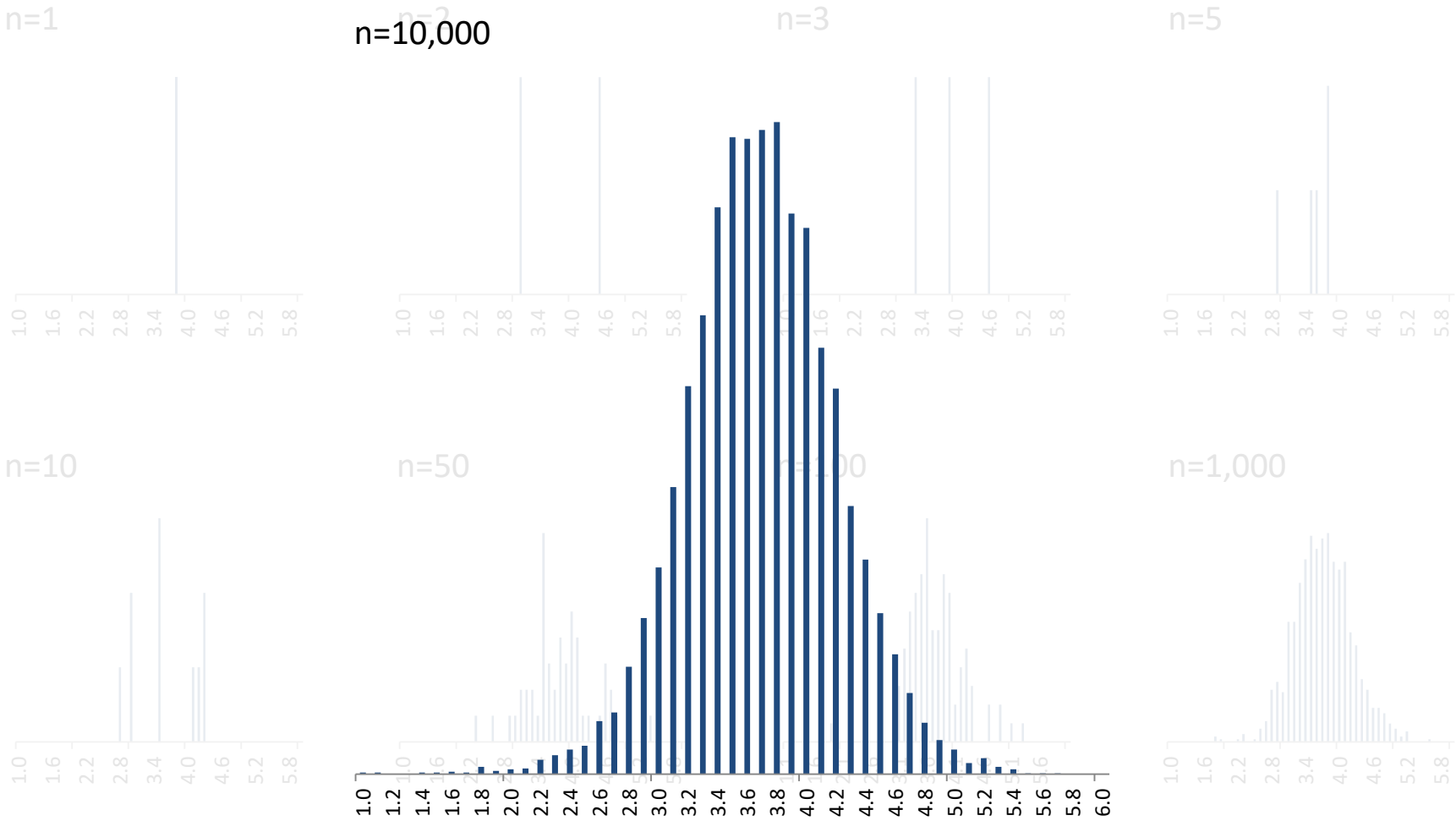
- It is a vast and dynamic field of research
- Frustration ahead

Expression Data

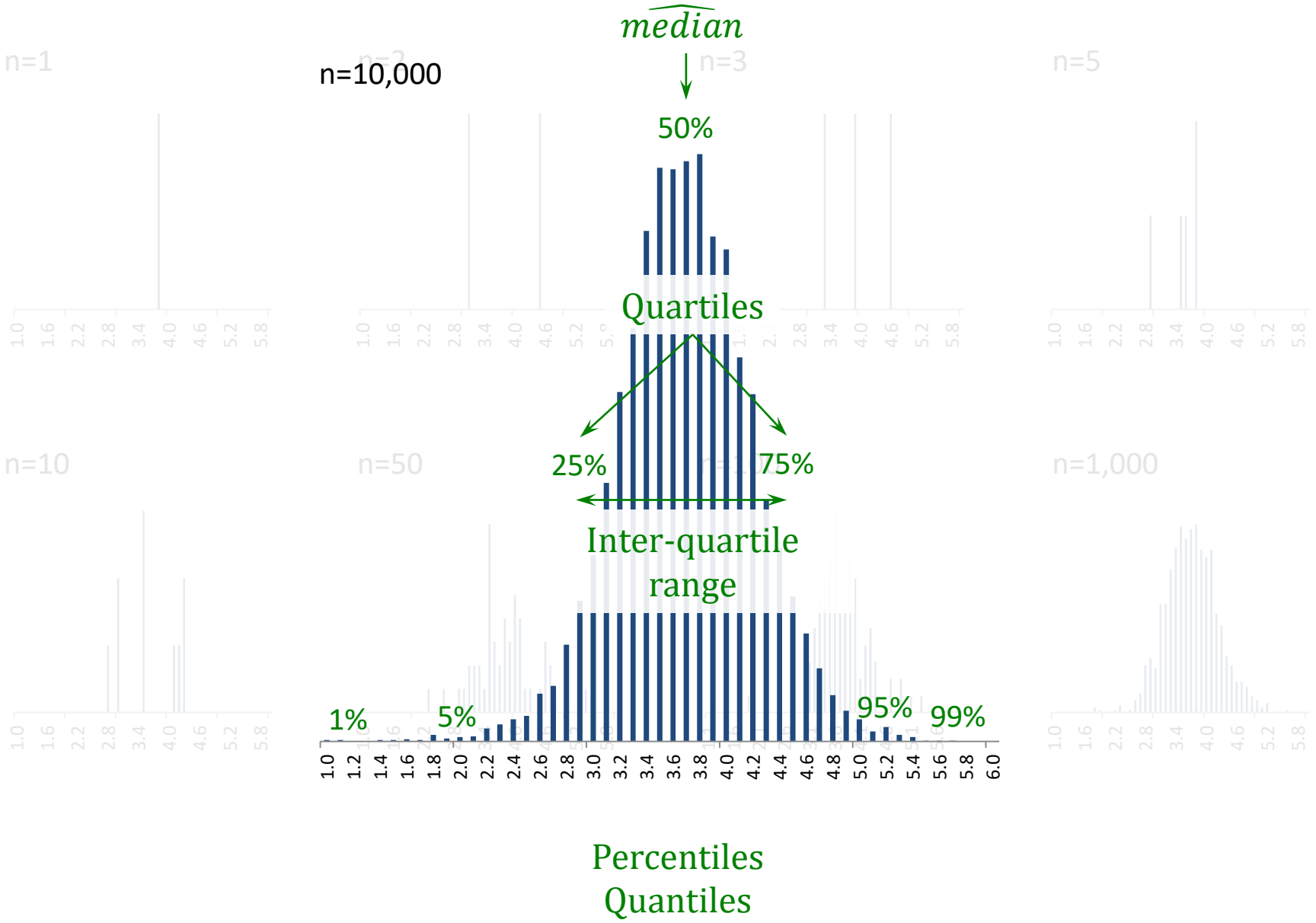


Cell by Laxmikanth Kollipara

Birth Weight in Norway



Population

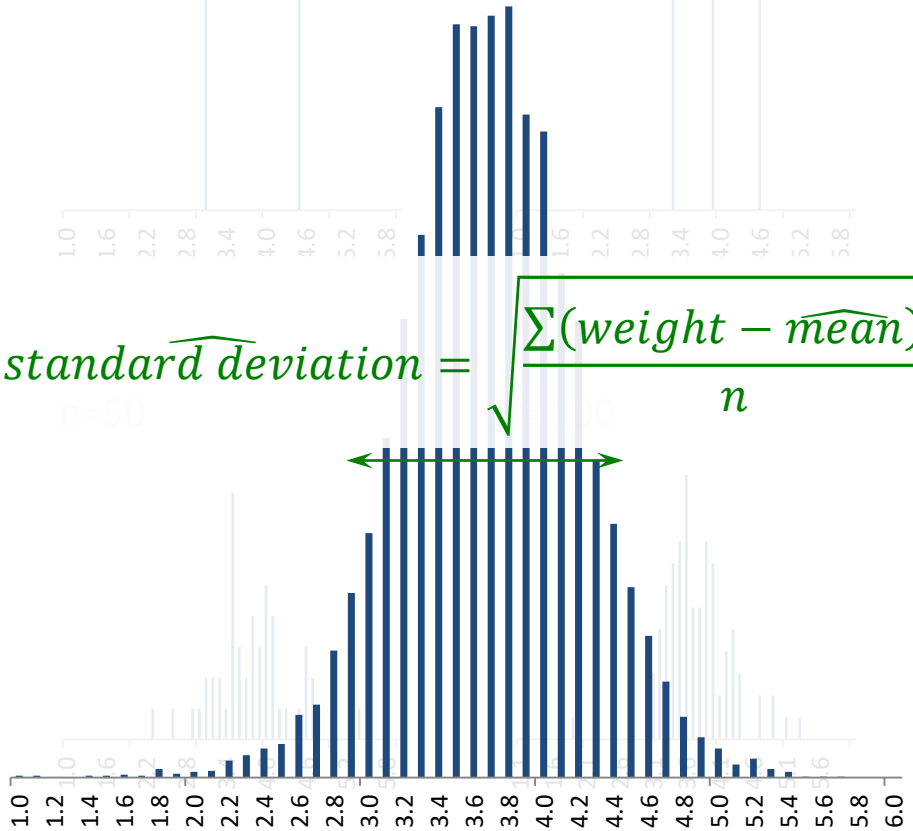

n=1
n=2
n=3
n=5
n=10
n=50
n=1,000
n=10,000
median
50%
Quartiles
25%
75%
Inter-quartile range
1%
5%
95%
99%
1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0 3.2 3.4 3.6 3.8 4.0 4.2 4.4 4.6 4.8 5.0 5.2 5.4 5.6 5.8 6.0
Percentiles
Quantiles

Population

n=10,000

$$\widehat{mean} = \frac{\sum weight}{n}$$

$$\widehat{standard\ deviation} = \sqrt{\frac{\sum (weight - \widehat{mean})^2}{n}}$$



Population

$\{2 \quad 3 \quad 10\}$ $\widehat{mean} =$ $\widehat{median} =$

$\{1 \quad 2 \quad 3\}$ $\widehat{mean} =$ $\widehat{median} =$

$\{0 \quad 2 \quad 6 \quad 8 \quad 12 \quad 14\}$ $\widehat{mean} =$ $\widehat{median} =$

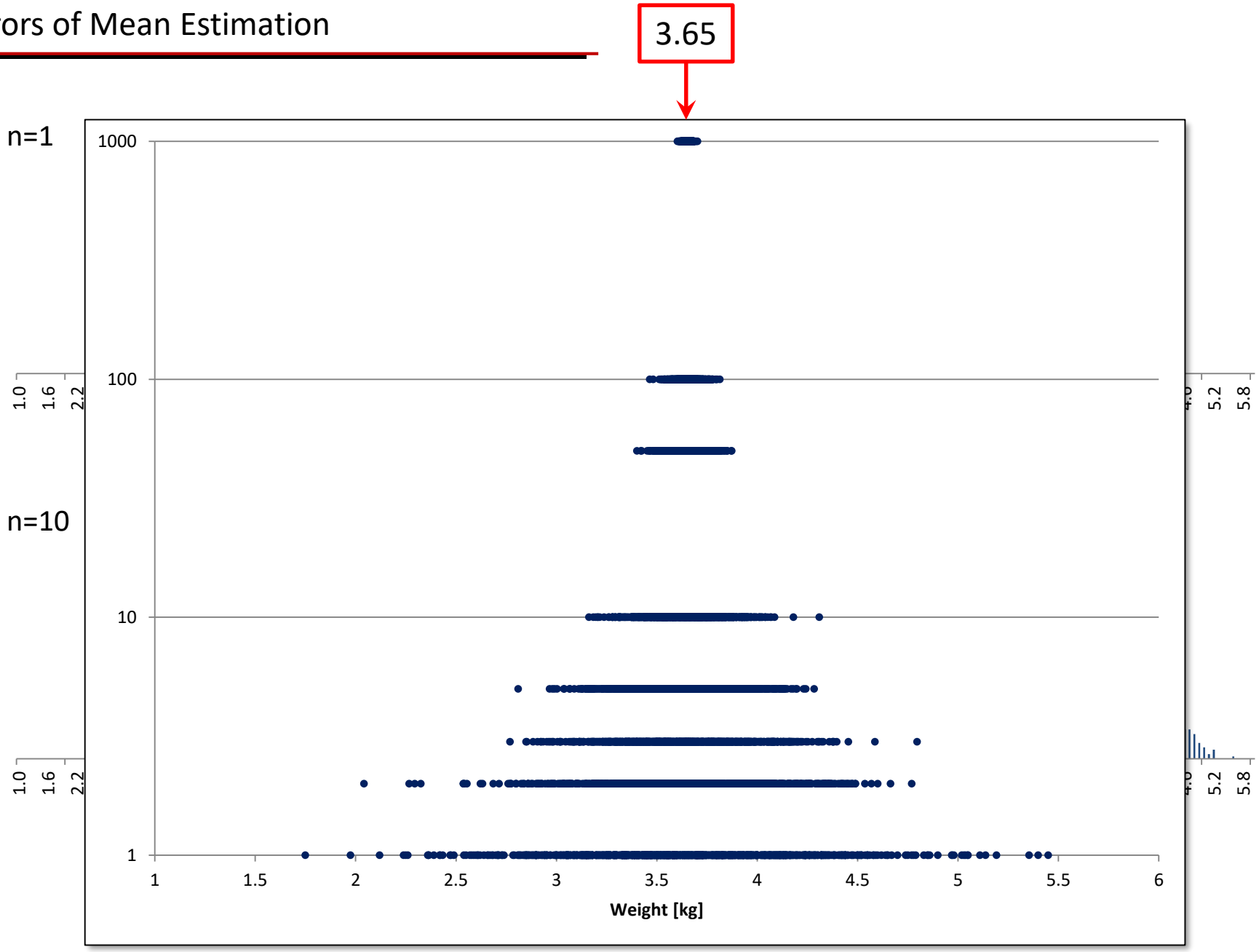
Population

$\{2 \quad 3 \quad 10\}$ $\widehat{mean} = 7.5$ $\widehat{median} = 3$

$\{1 \quad 2 \quad 3\}$ $\widehat{mean} = 2$ $\widehat{median} = 2$

$\{0 \quad 2 \quad 6 \quad 8 \quad 12 \quad 14\}$ $\widehat{mean} = 7$ $\widehat{median} = 7$

Errors of Mean Estimation

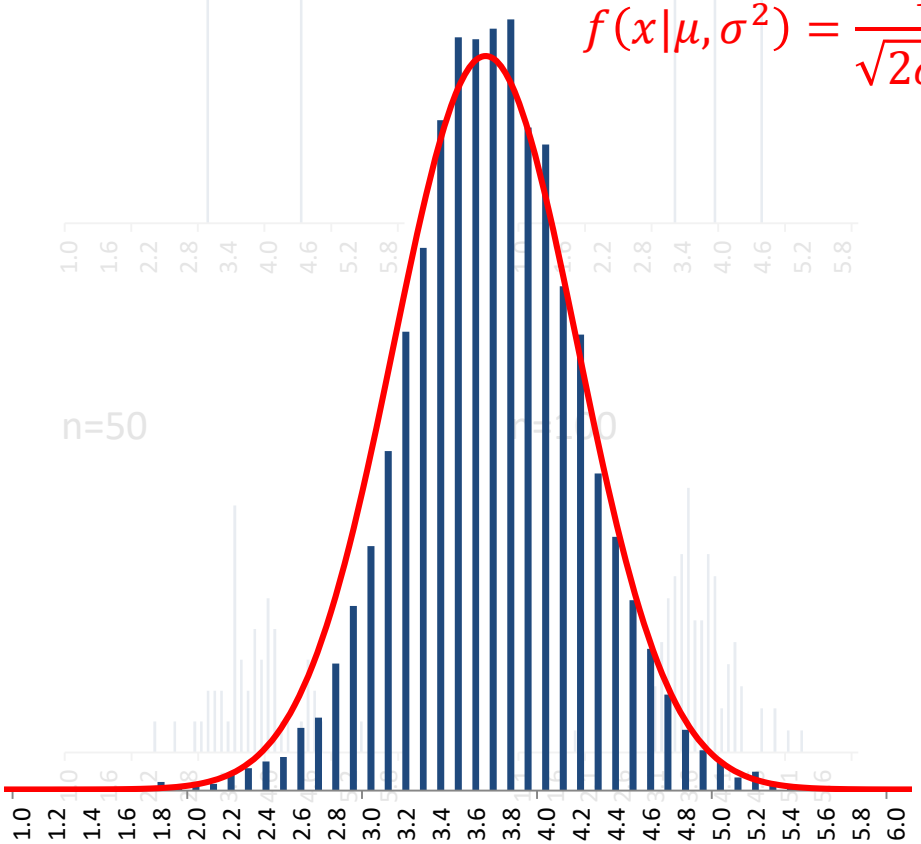


Central limit theorem: averages of random variables drawn from independent distributions distribute normally at high number of variables.

Birth Weight in Norway

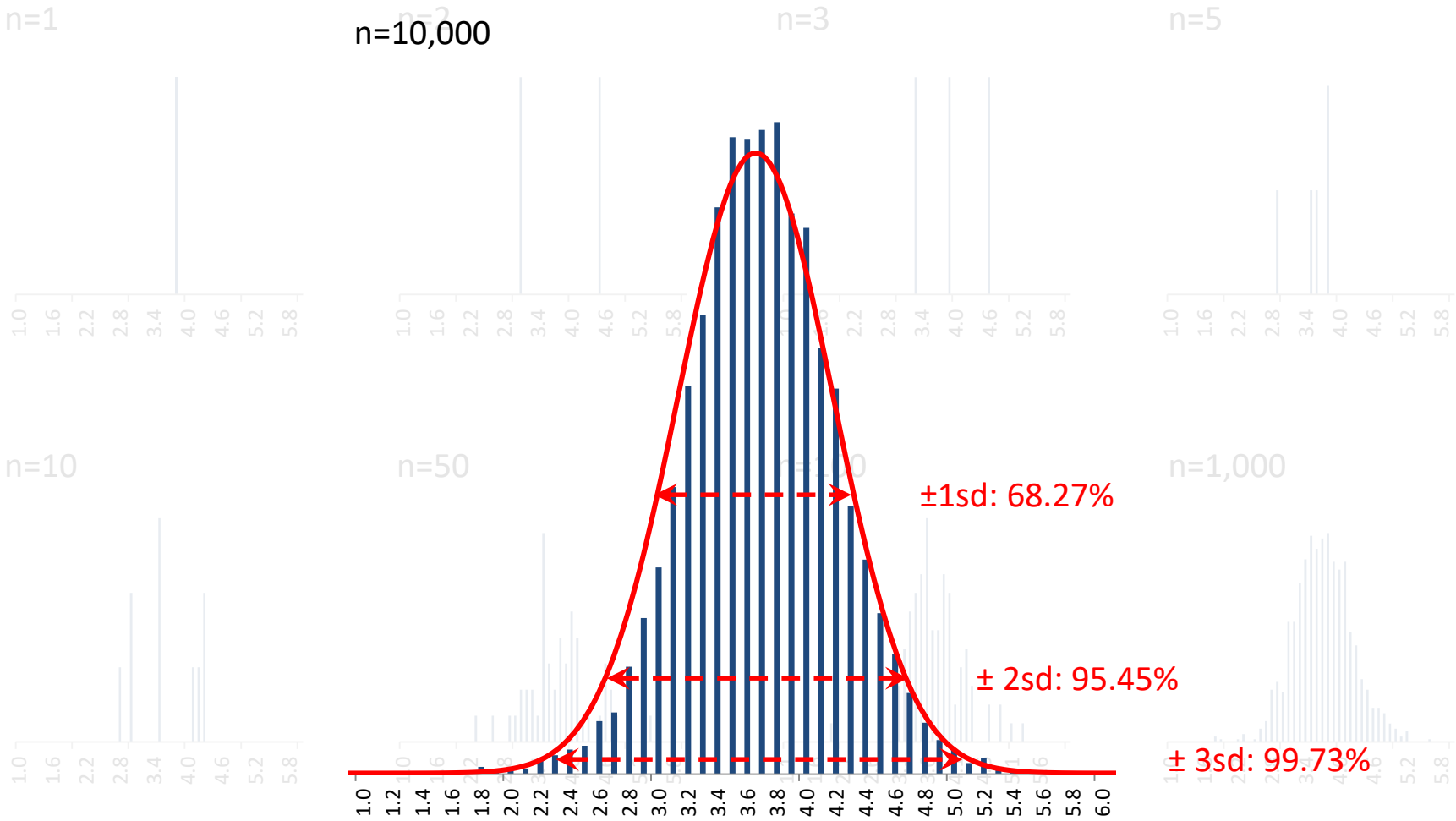
n=10,000

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



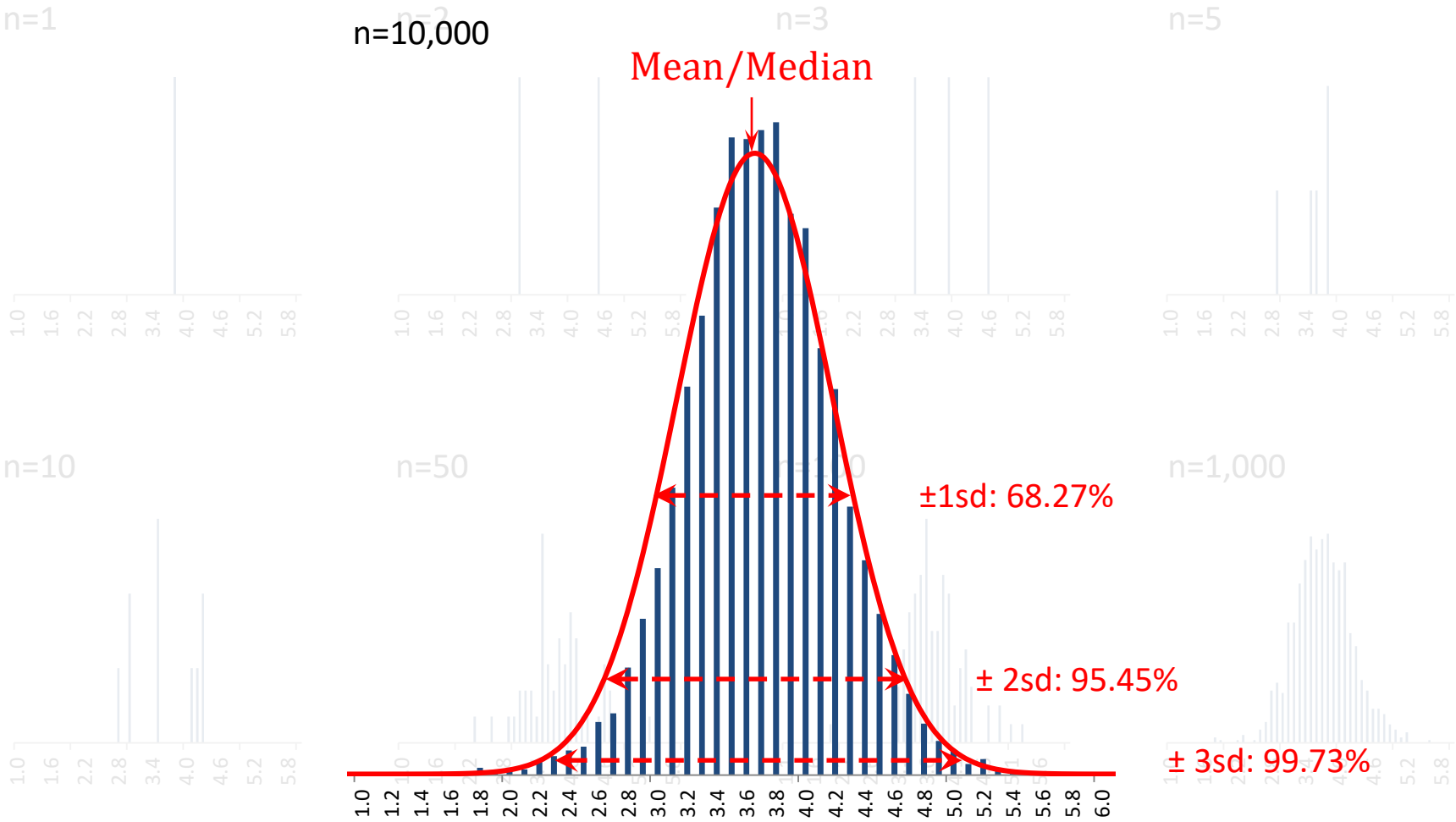
In practice: a well defined good approximation to many distributions.

Birth Weight in Norway

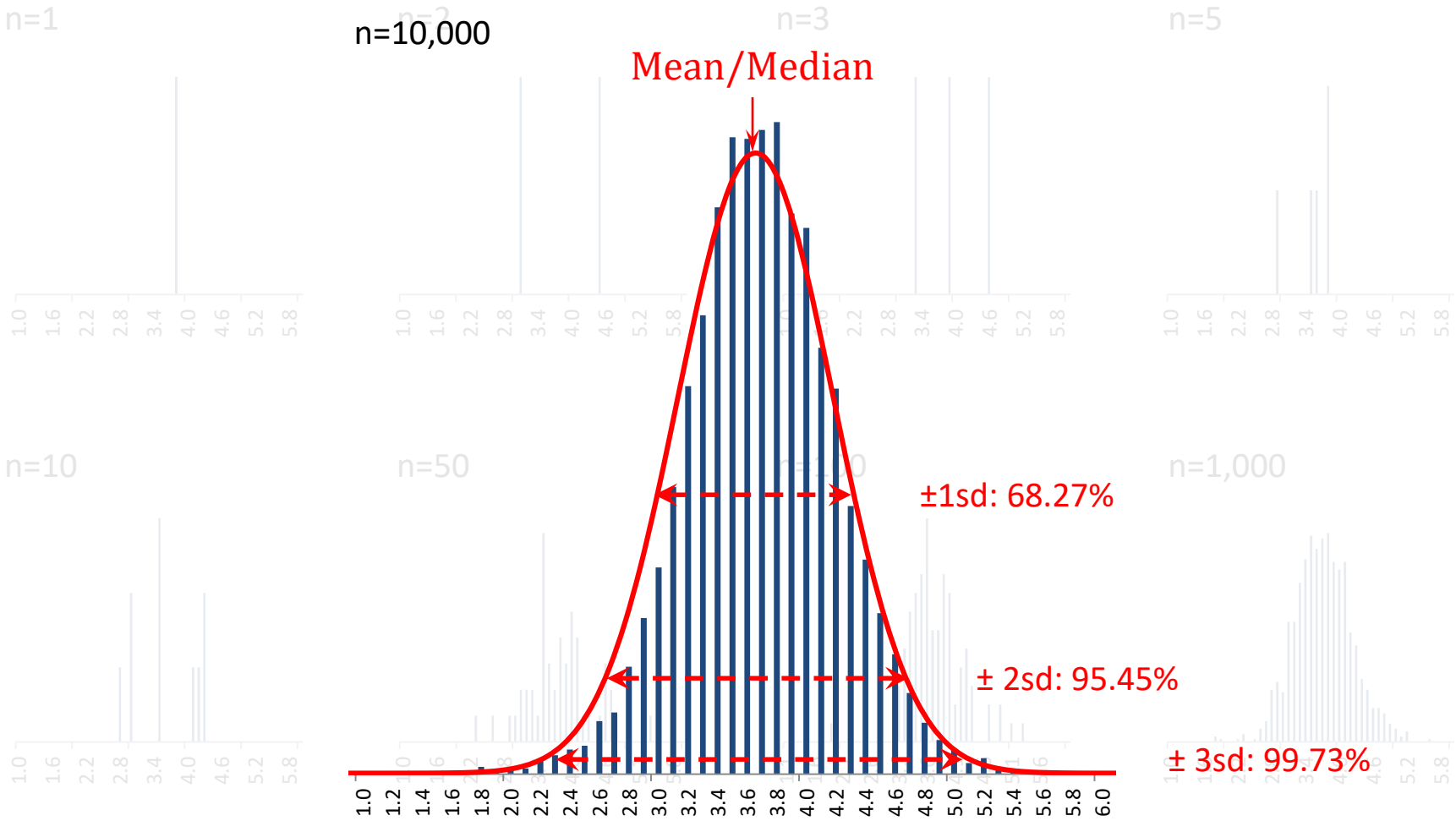


68-95-99.7 Rule

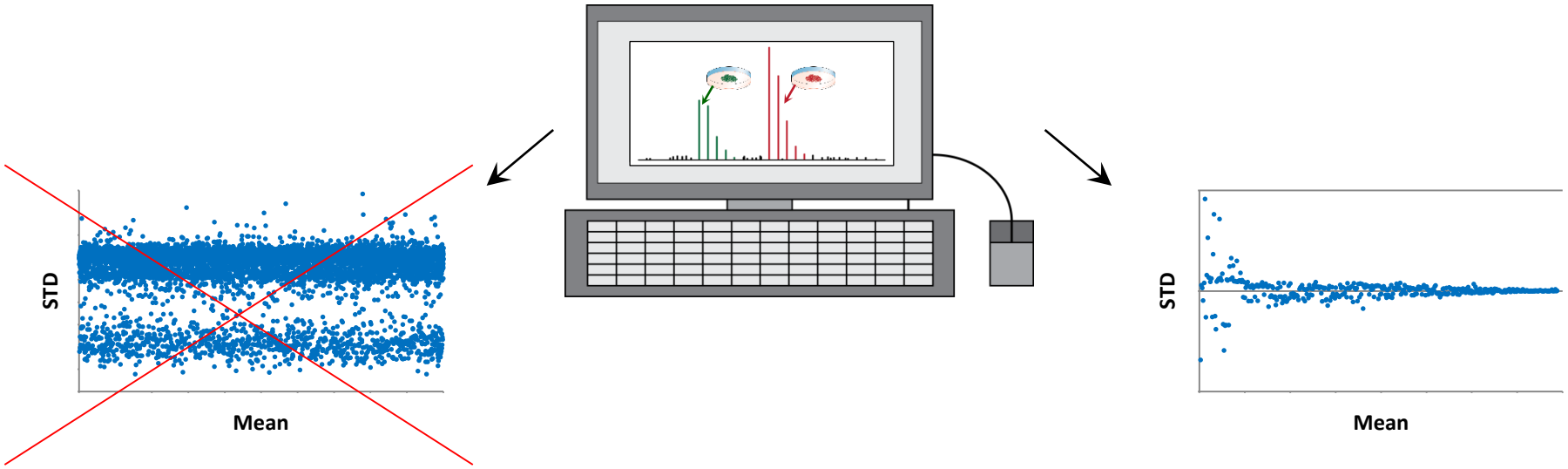
Birth Weight in Norway



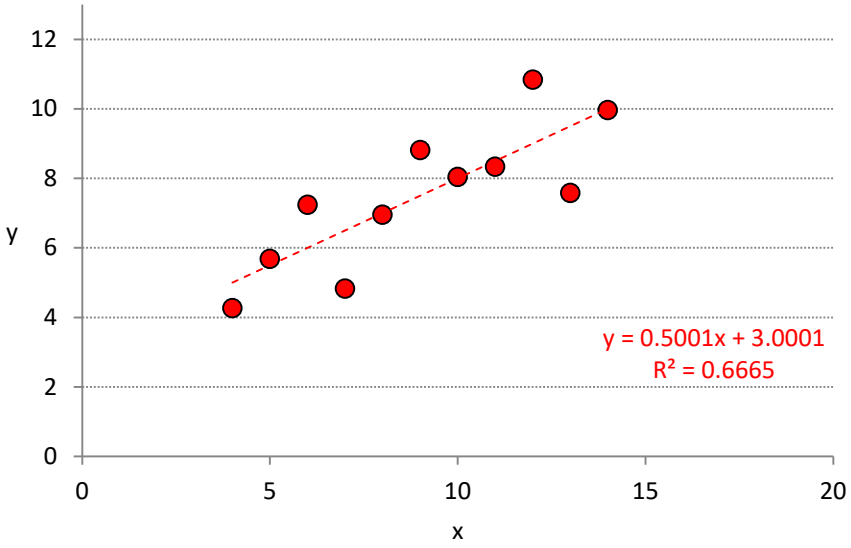
Birth Weight in Norway

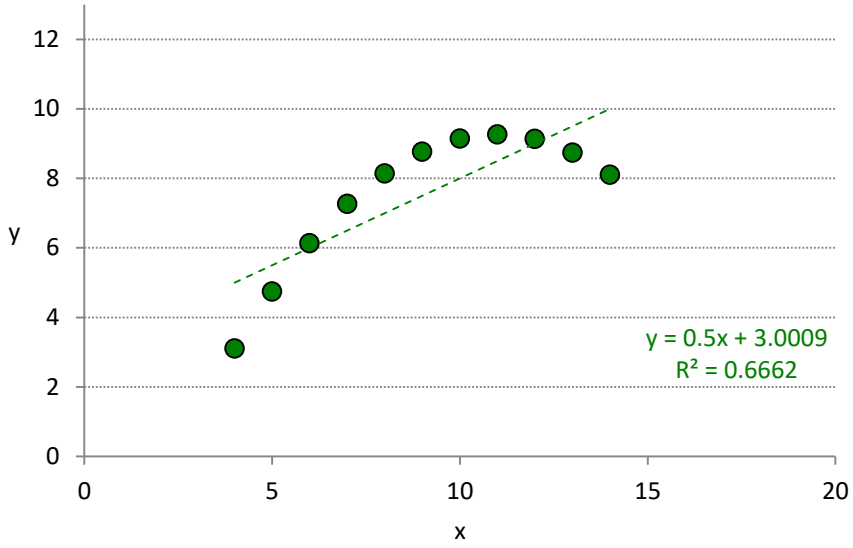


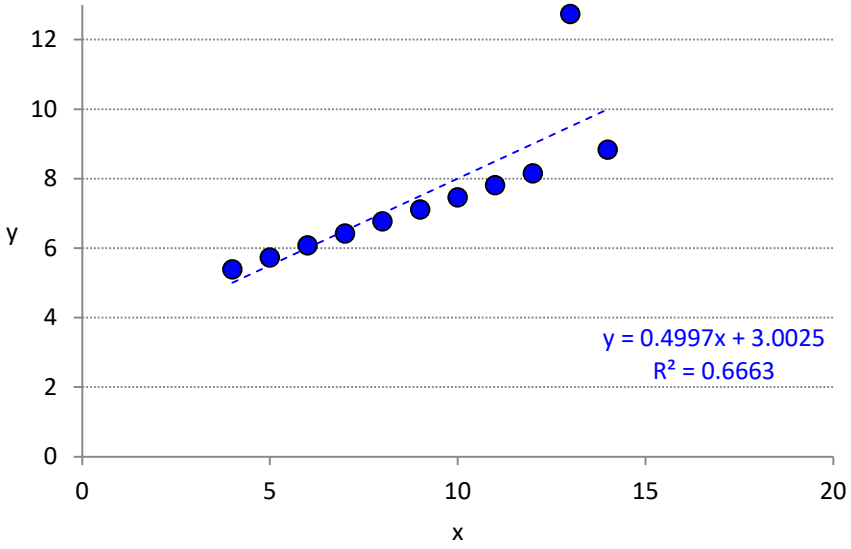
The first thing to do when analysing data?

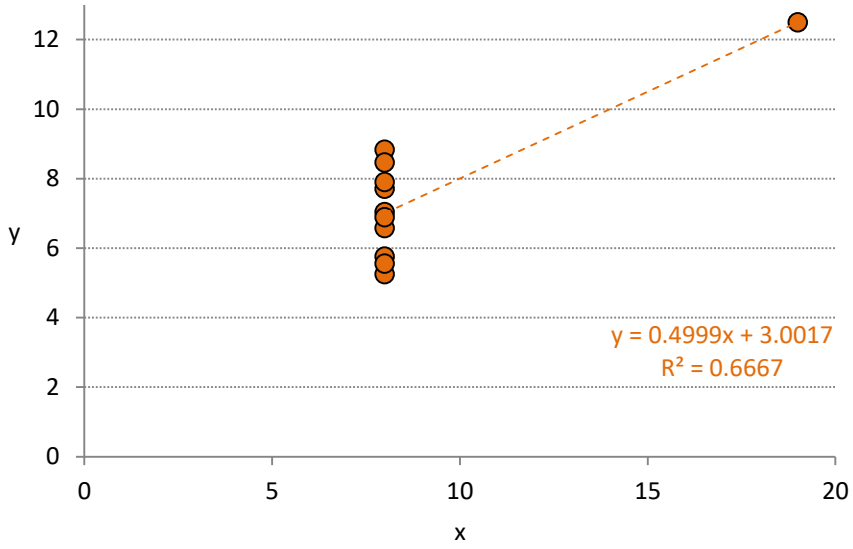


Data Exploration



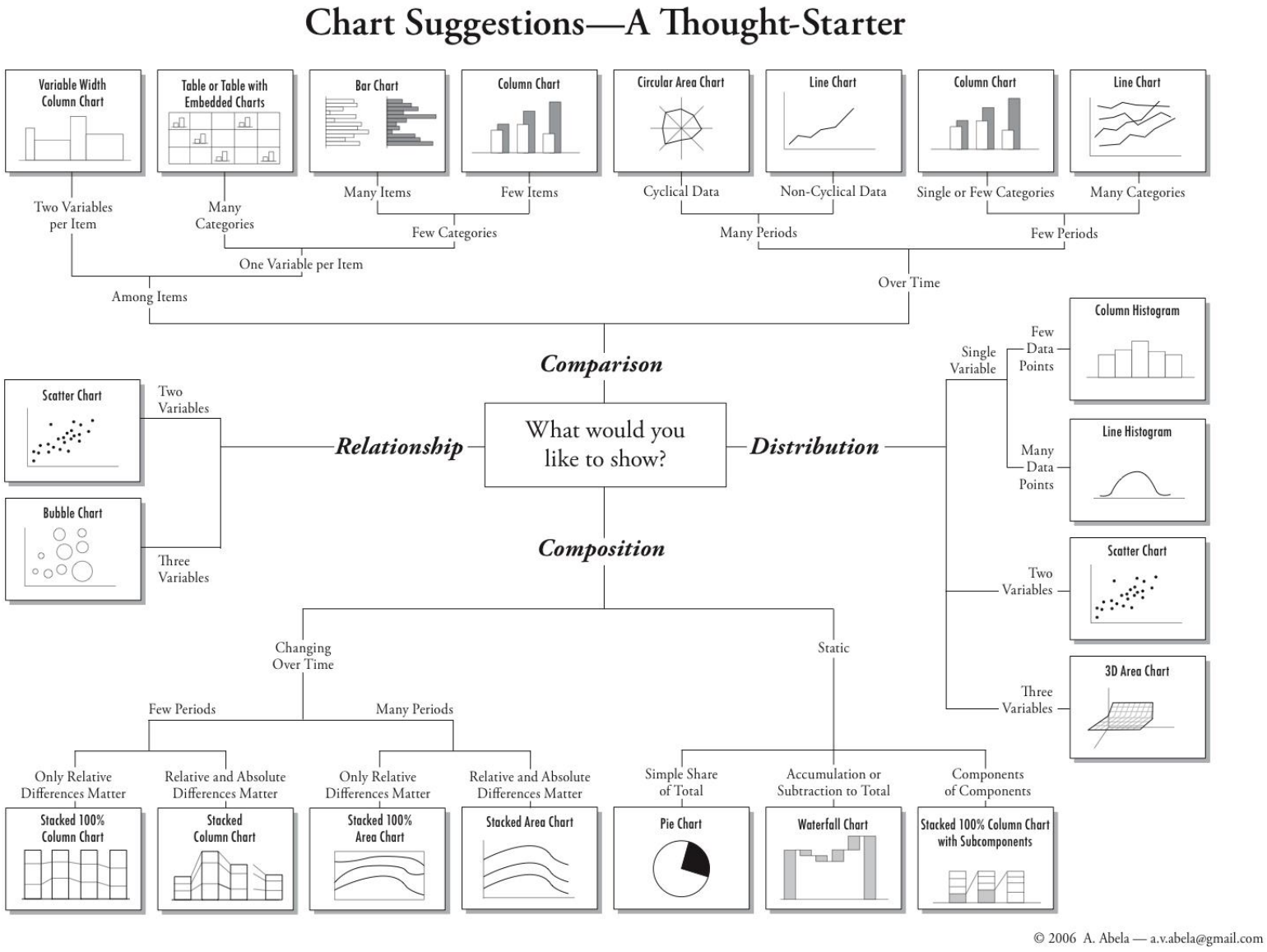




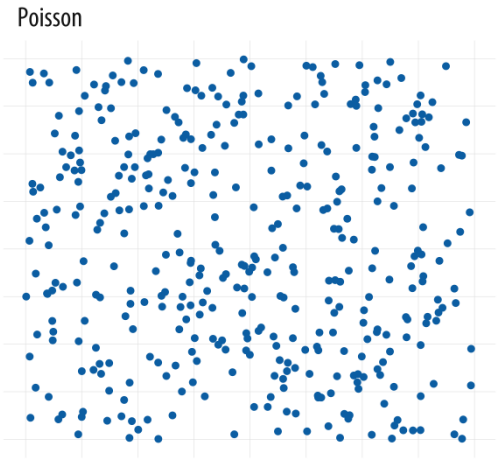


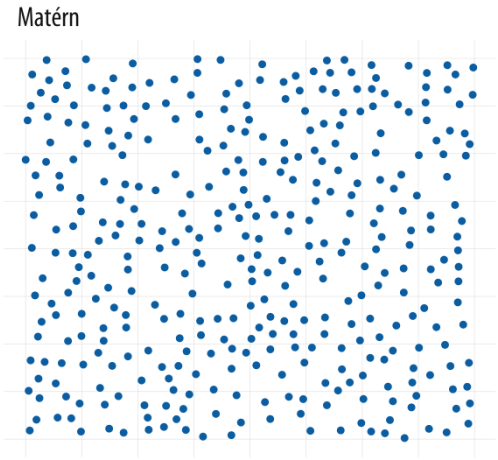
Anscombe, The American Statistician, 1973

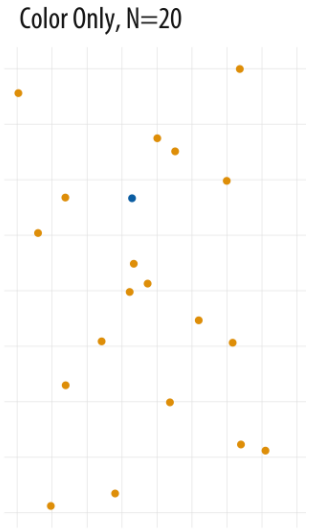
Plot, plot, and plot. As unbiased as possible

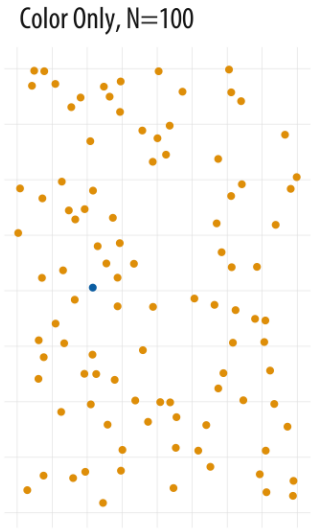


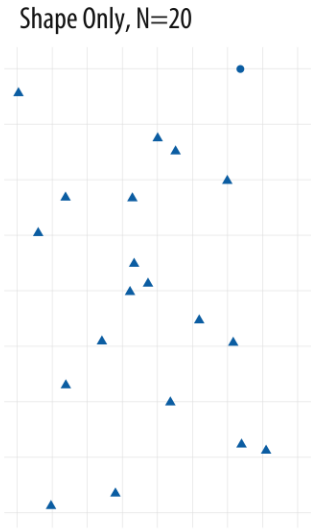
Perceptio

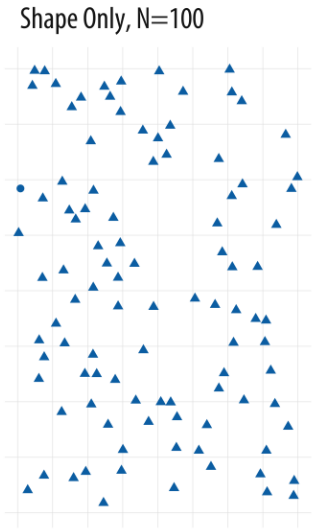

Poisson

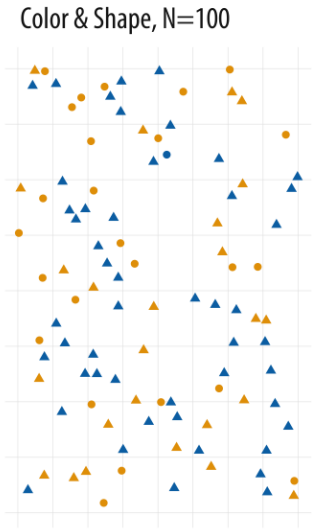

Matérn


Color Only, N=20


Color Only, N=100

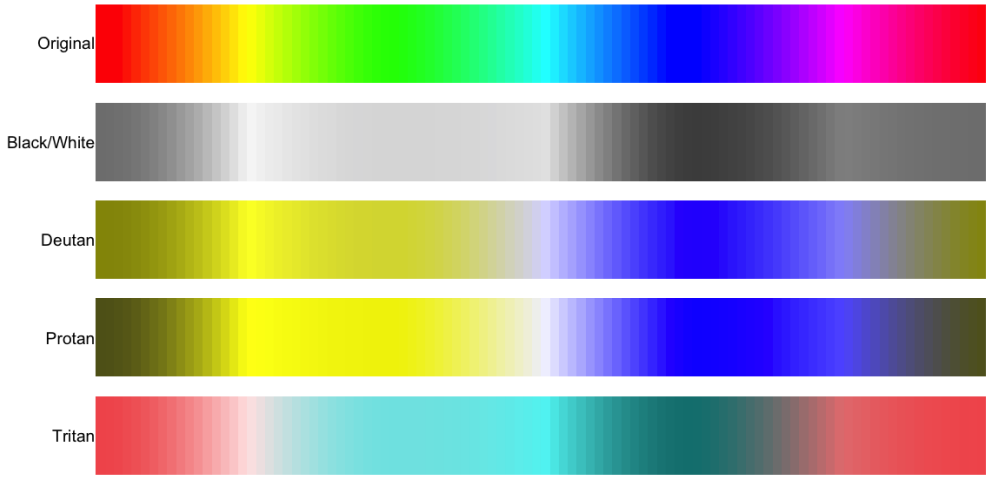

Shape Only, N=20

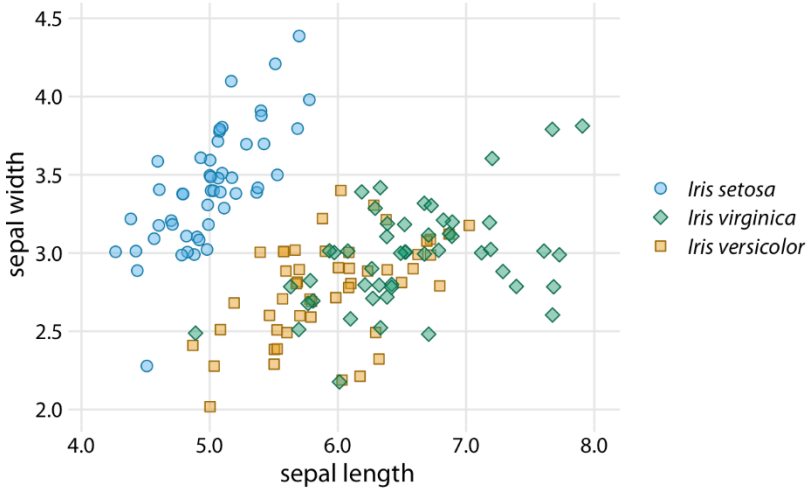

Shape Only, N=100


Color & Shape, N=100

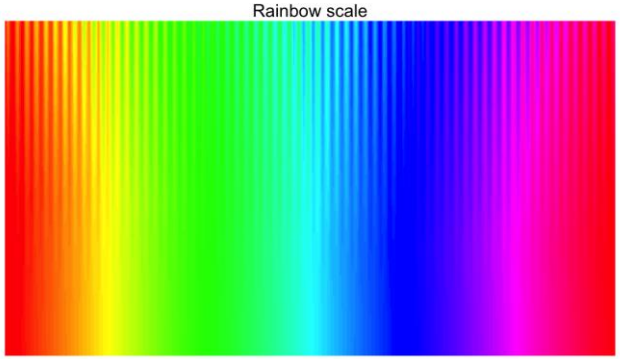
The difficult choice of colors

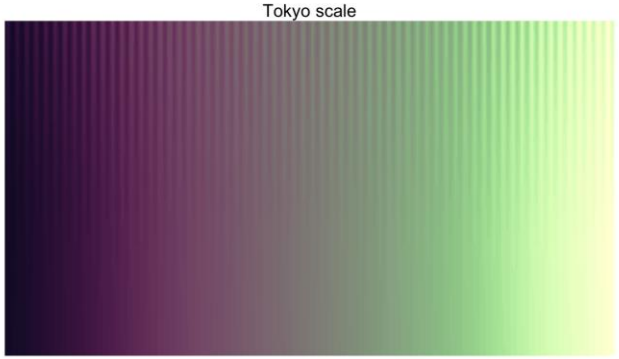
Color perception





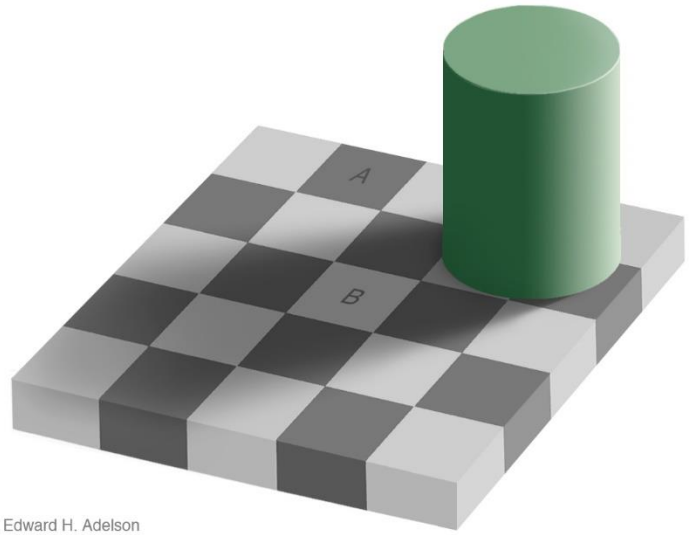
Color discrimination



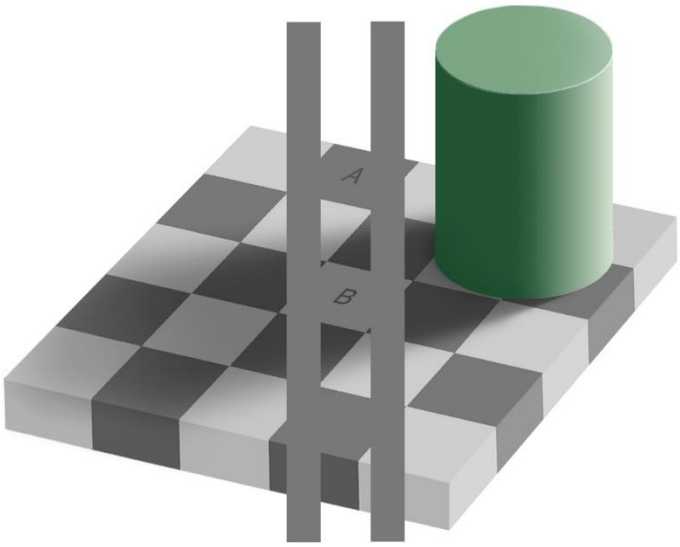


data-imaginist.com, serialmentor.com/dataviz & scico

Perception in dissemination

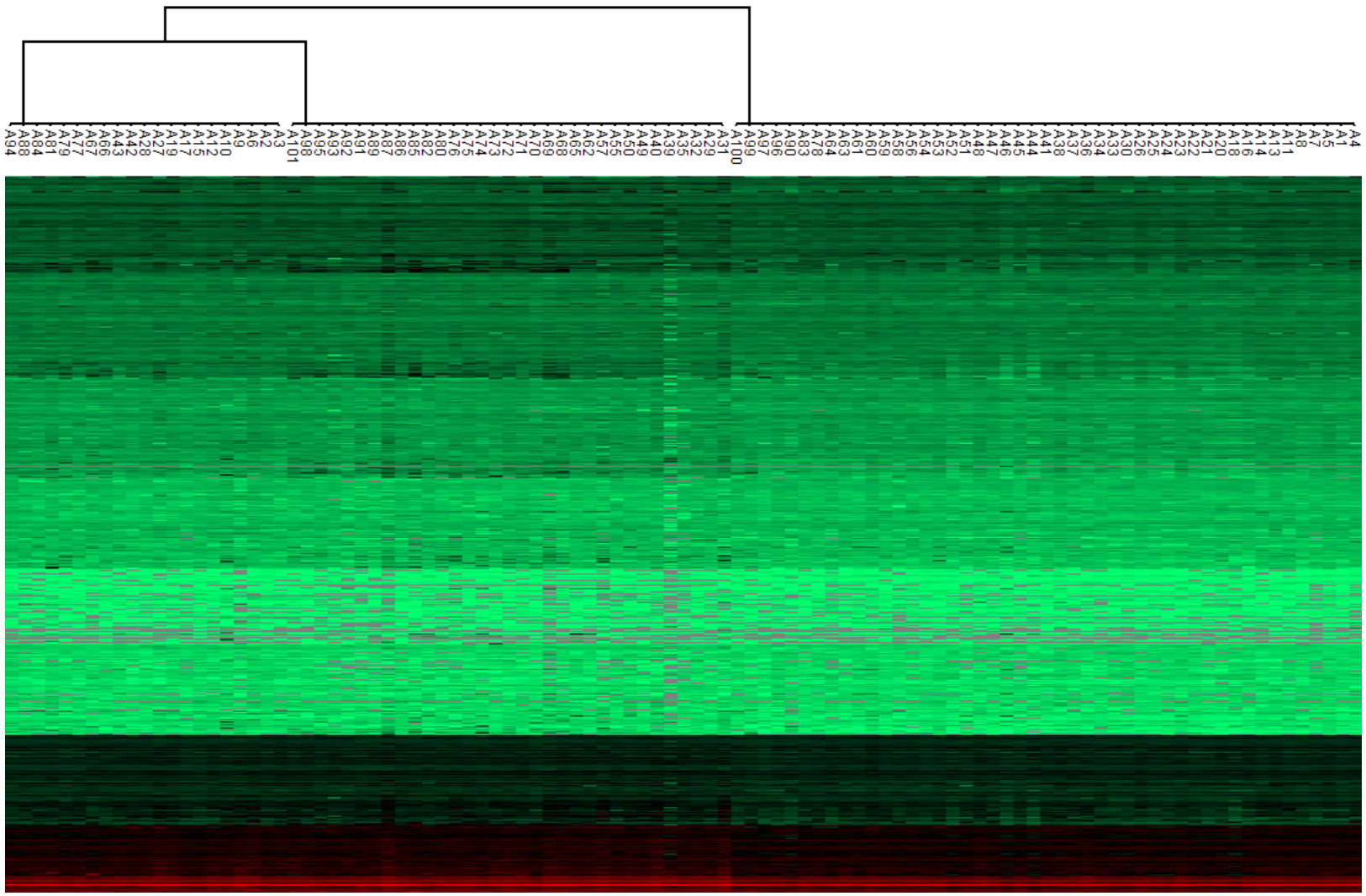


Edward H. Adelson



Edward H. Adelson & michaelbach.de

Clustering

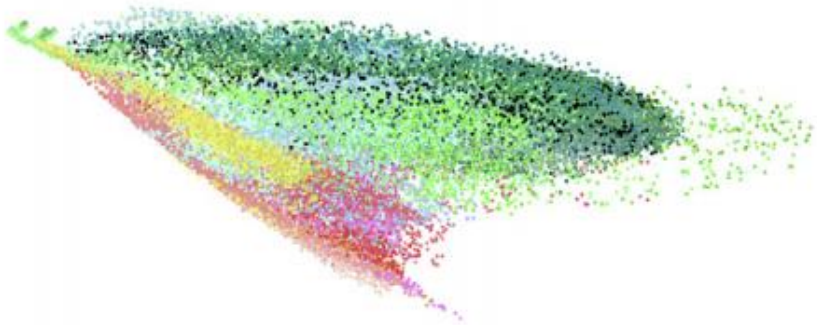


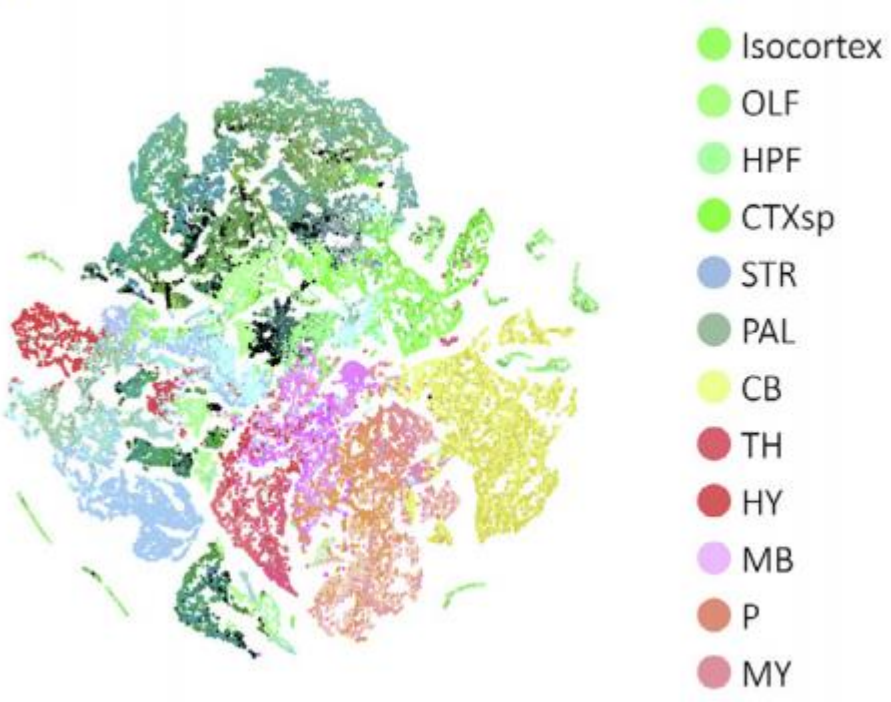
Opsahl et al., *Proteomics 2016*

Dimension reduction

PCA

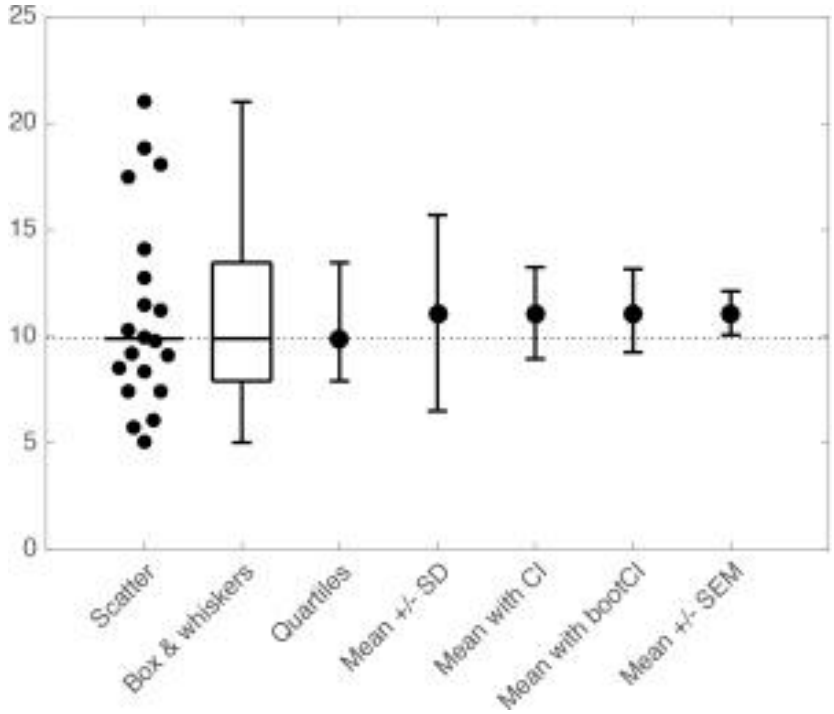
t-SNE

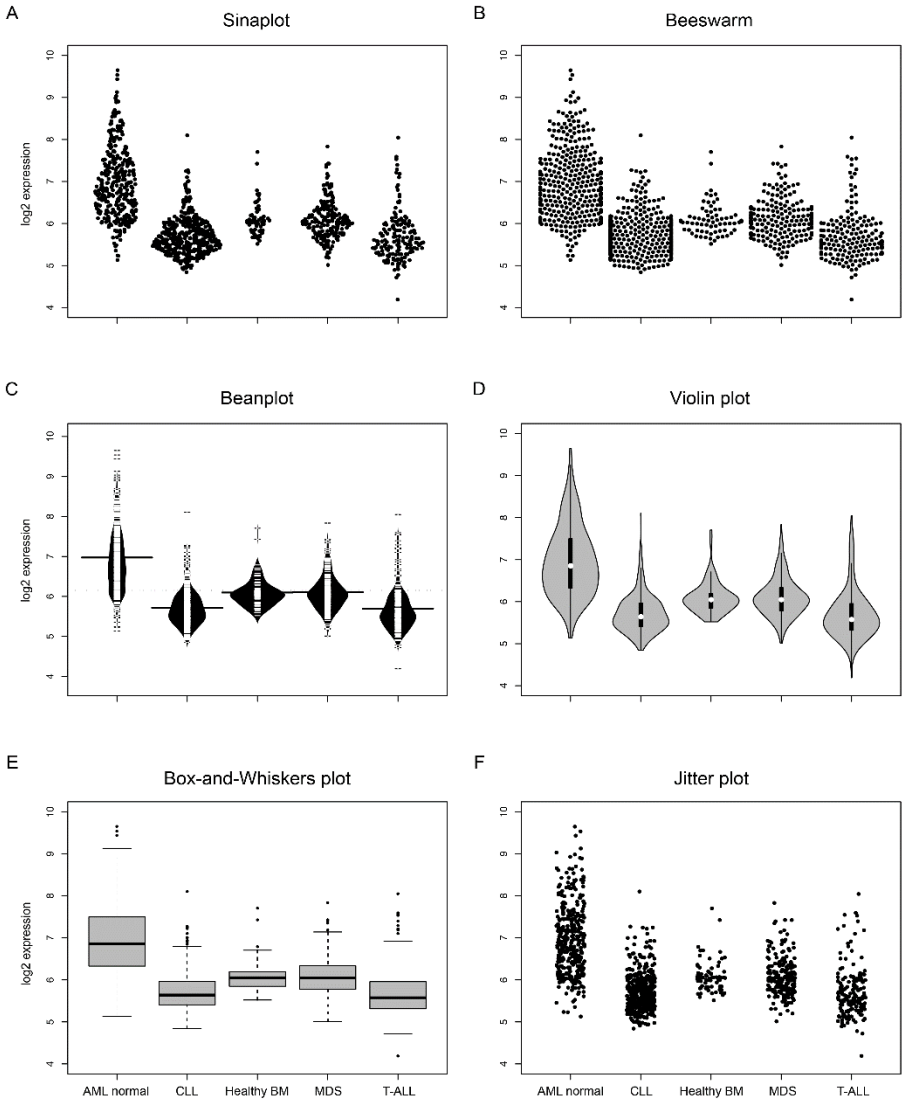




Van Maaten and Hiton, Journal of Machine Learning Research, 2008
illustration from Mhfouz et al., Methods, 2015

Plotting summary statistics

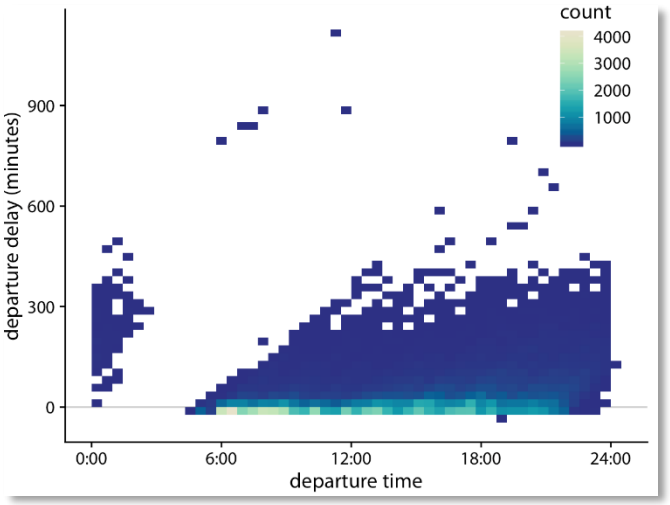


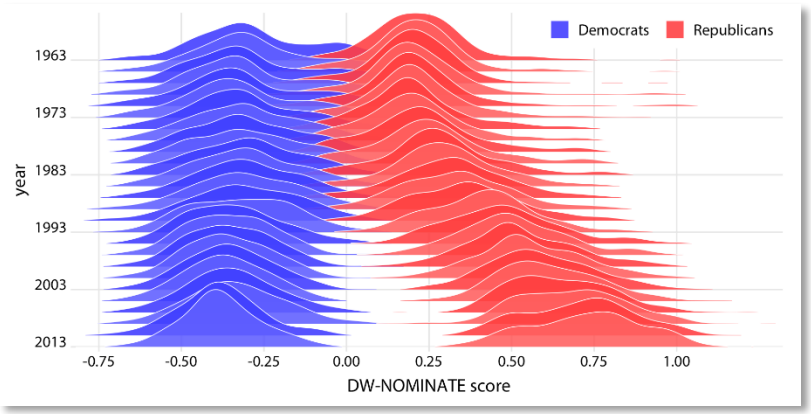


garstats.wordpress.com/2016/05/27/the-percentile-bootstrap

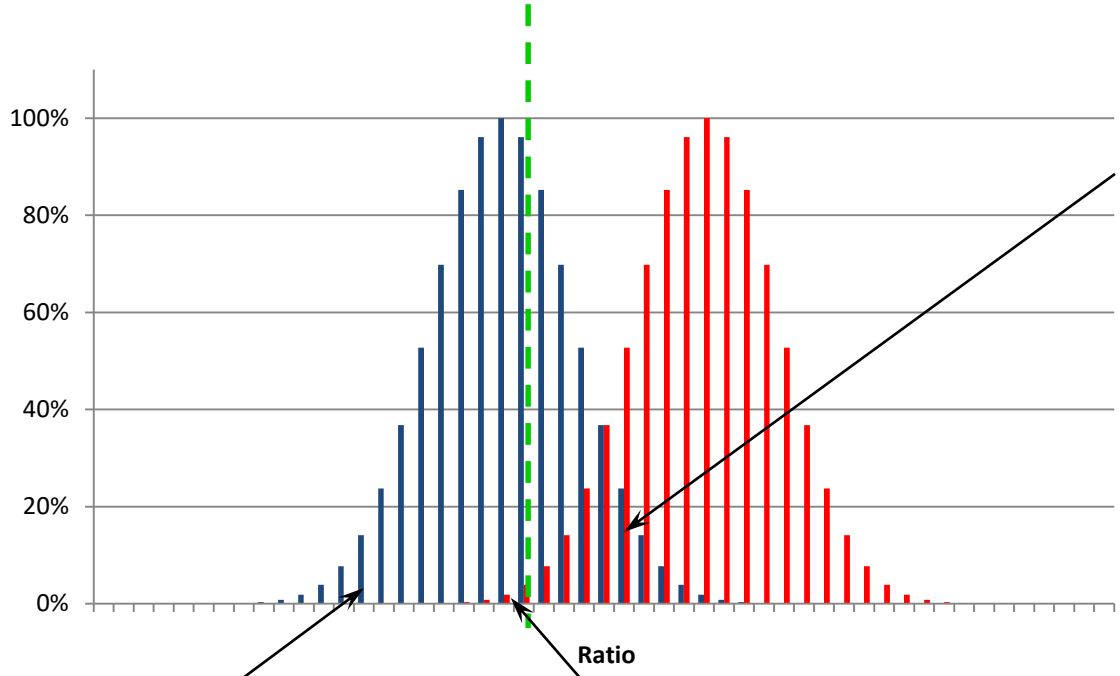
sinaplot (CRAN.R-project.org/package=sinaplot)

Plotting summary statistics





Errors and Power



False Positive ☹

Type I

$\alpha = \Phi_{control}(ratio)$

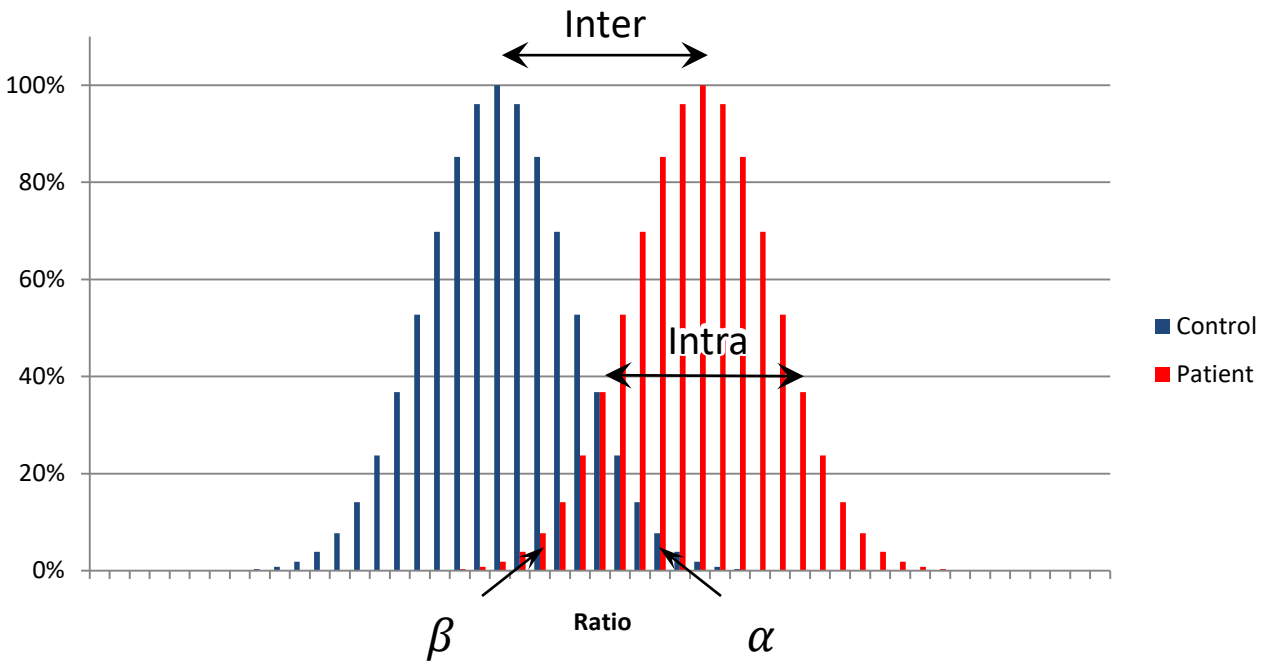
$power = 1 - \beta$

False Negative ☹

Type II

$\beta = \Phi_{patient}(ratio)$

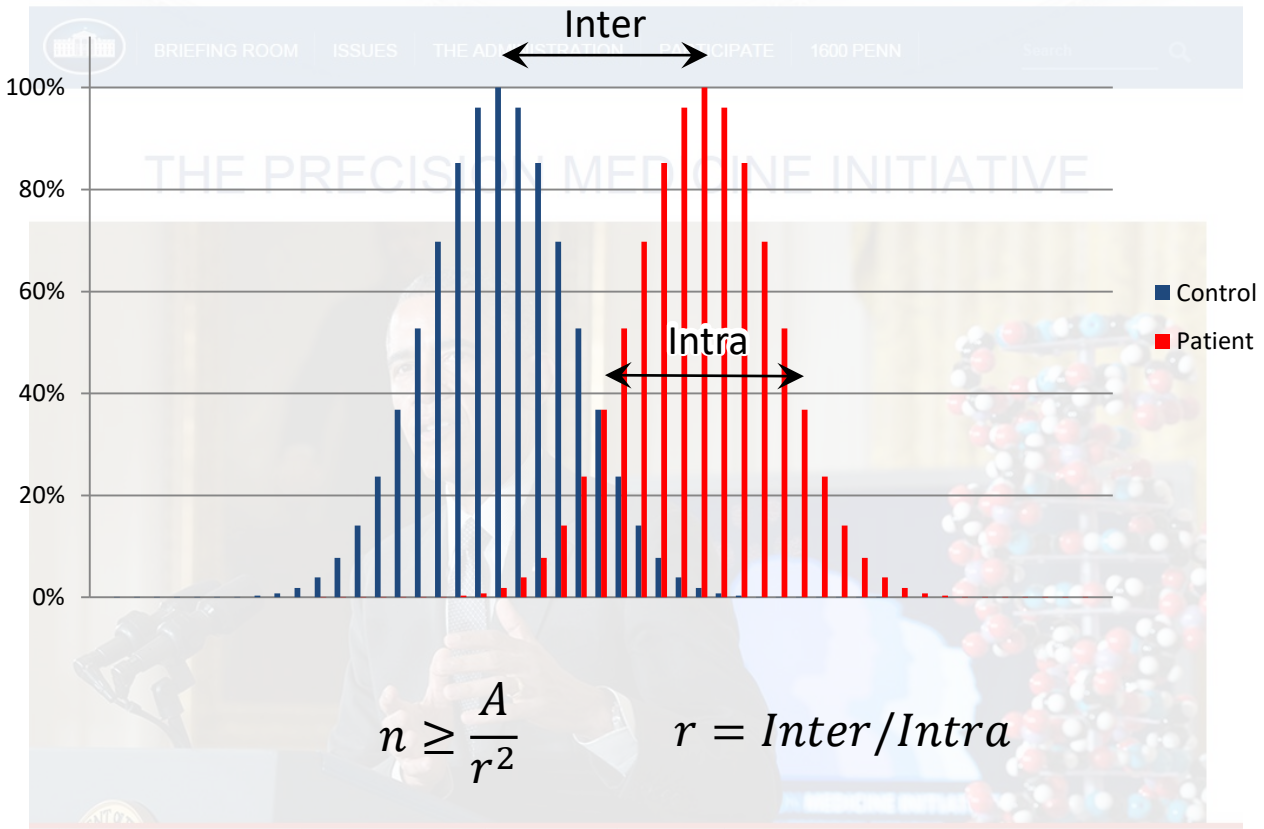
Multiple Variables Optimization



-> On normal distributions, Student's t-test:

$$n \geq \frac{2\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)+\Phi^{-1}(1-\beta)\right)^2}{\left(\frac{Inter}{Intra}\right)^2} \xrightarrow[r = Inter/Intra]{} n \geq \frac{A}{r^2}$$

Precision Medicine



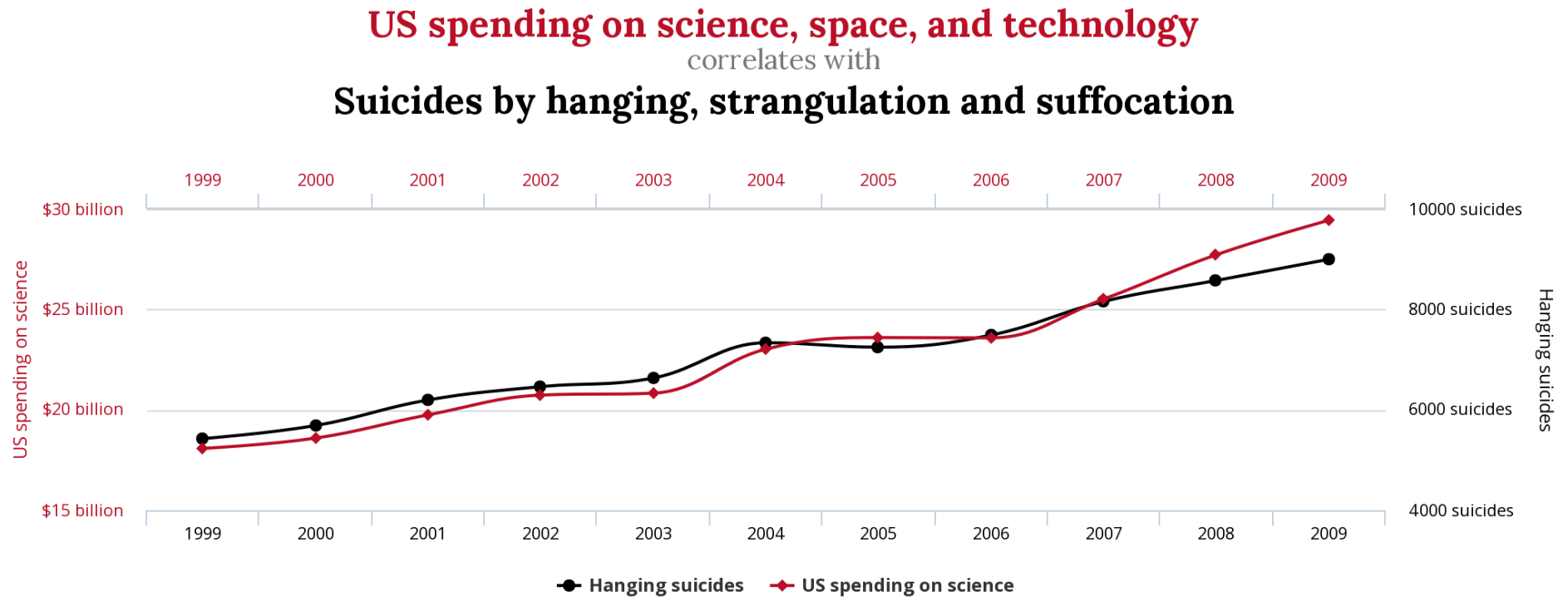
-> How can we improve the resolution of our experiments?

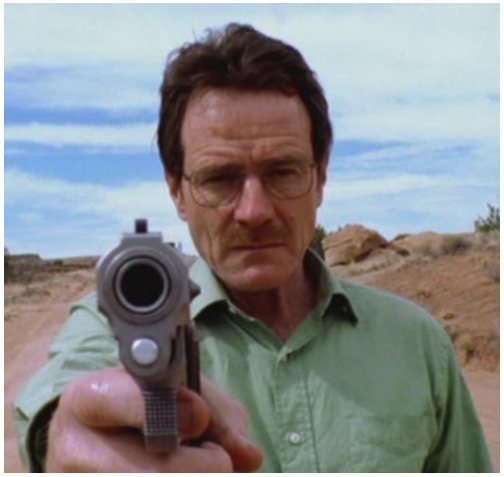
r decreases with noise

r decreases with filtering and smoothing

r increases with characterization

Multiple Hypothesis Testing

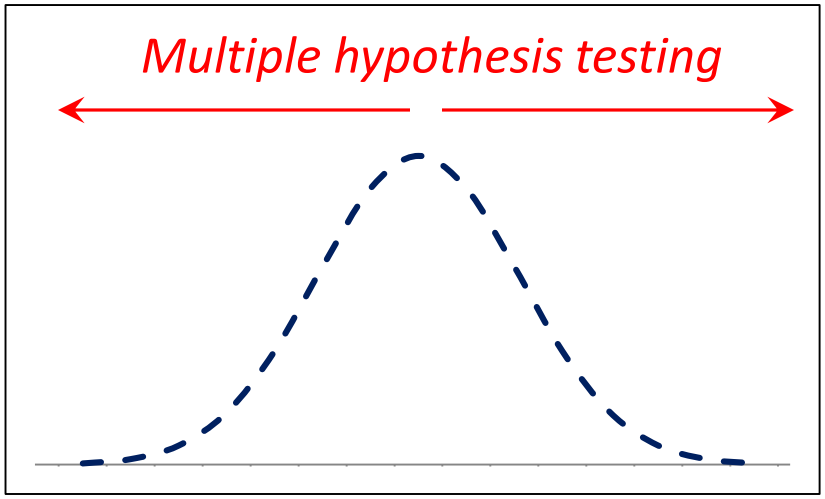


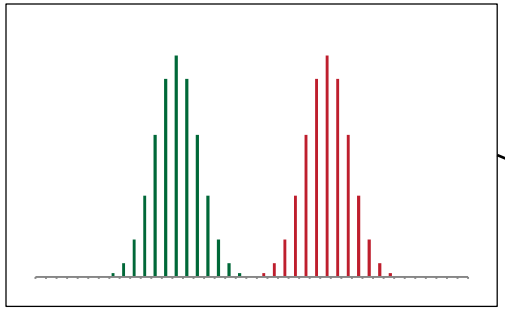


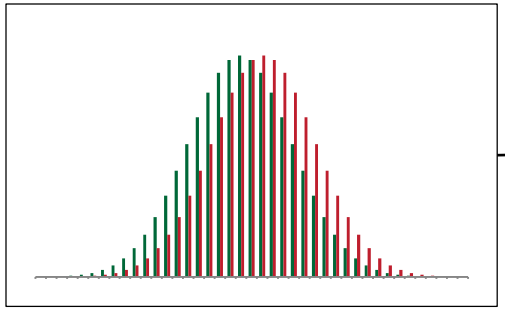
-> If you try long enough you will always find something

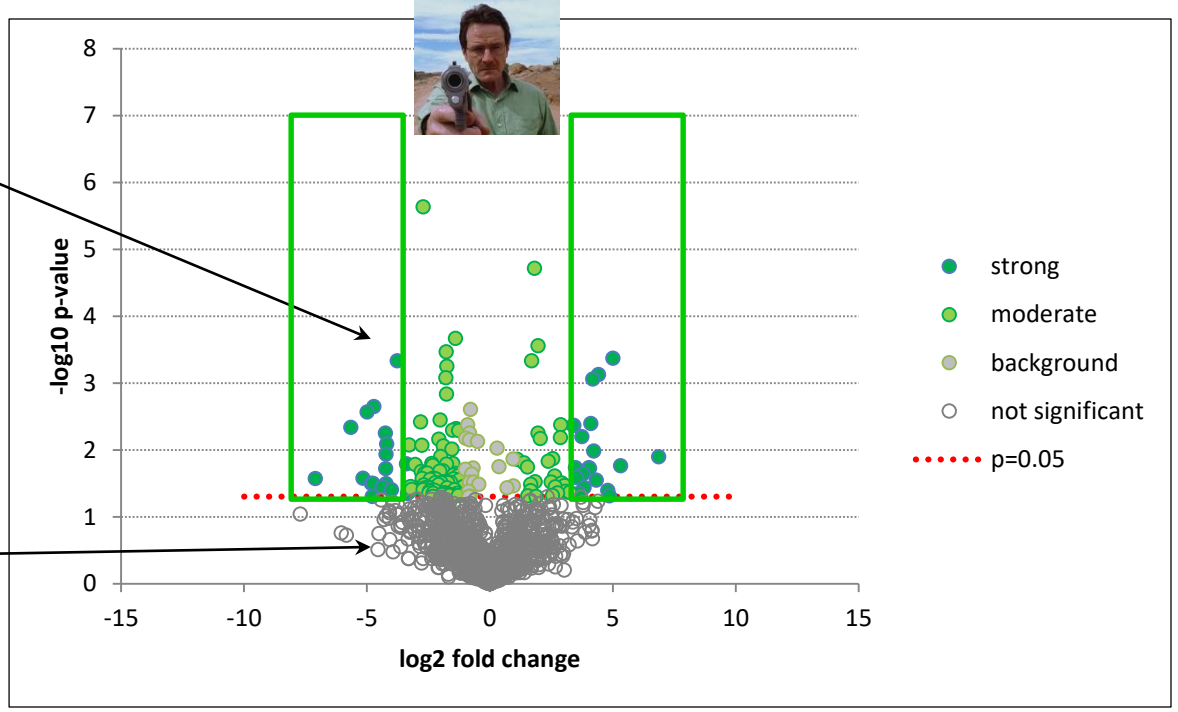
http://tylervigen.com/

Independent Hypothesis Weighting

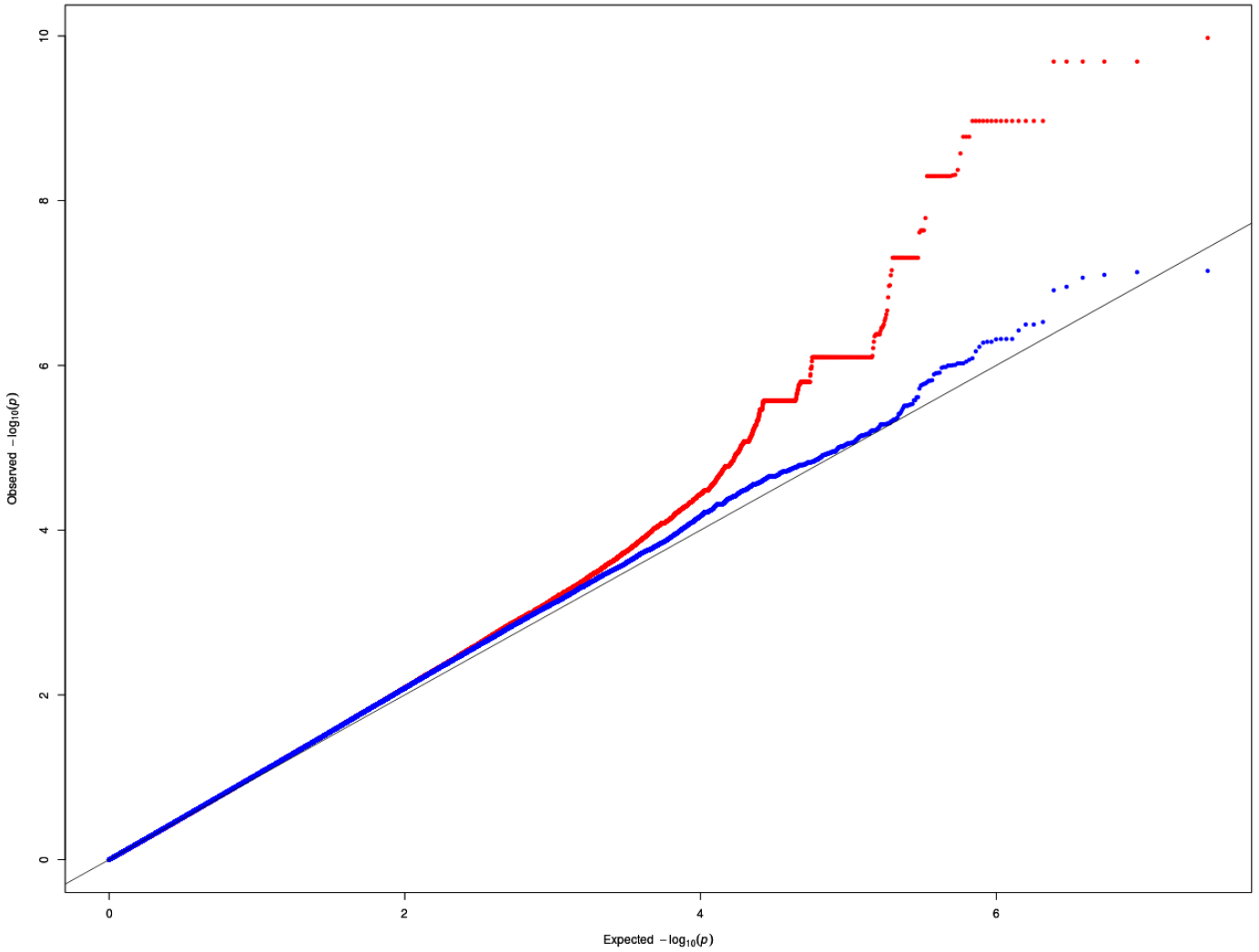




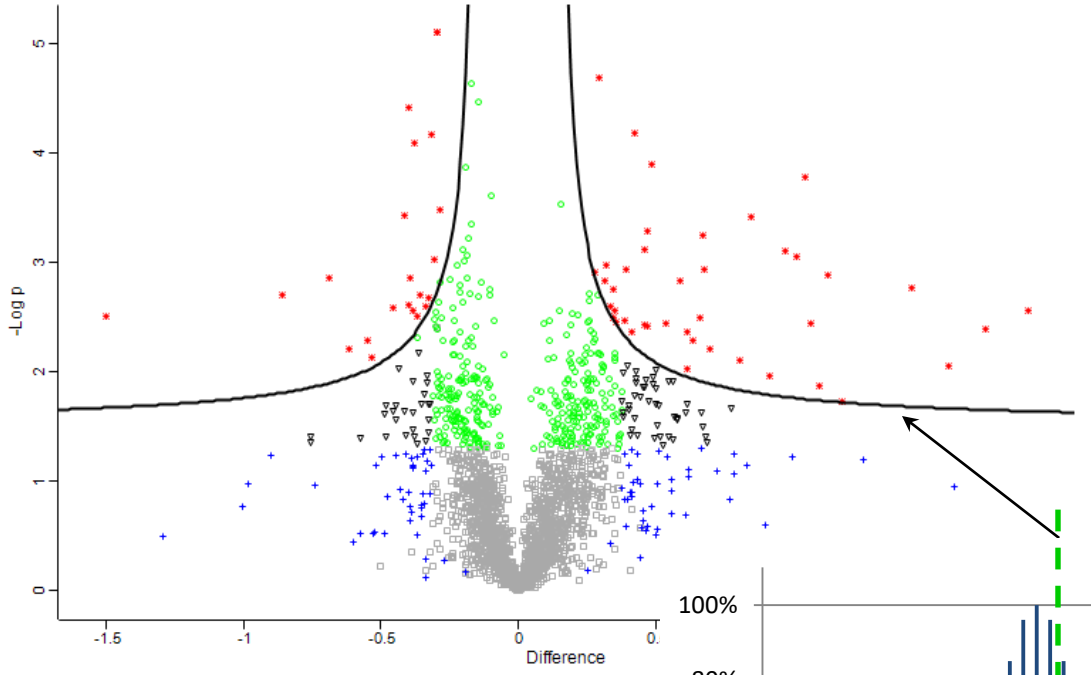


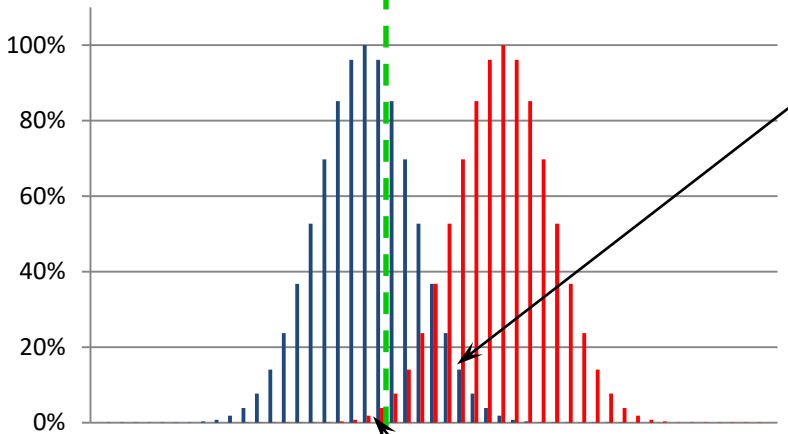


p-value Calibration



Differential Abundance FDR





False Positive ☹

False Negative ☹

Statistical Evaluation

n = 19,000

divorce	7.67 %	p-value < 0.002	5.96 %
Marital satisfaction	5.48 / 7	p-value < 0.001	5.64 / 7

"The P value was never meant to be used the way it's used today"

"Judge whether evidence was significant in the old-fashioned sense: worthy of a second look."

Nuzzo, Nature, 2014