Introduction to Statistics and Experimental Design

Reuben Thomas
Gladstone Bioinformatics Core
2/23/2021

Historical figures

Ibn Sina

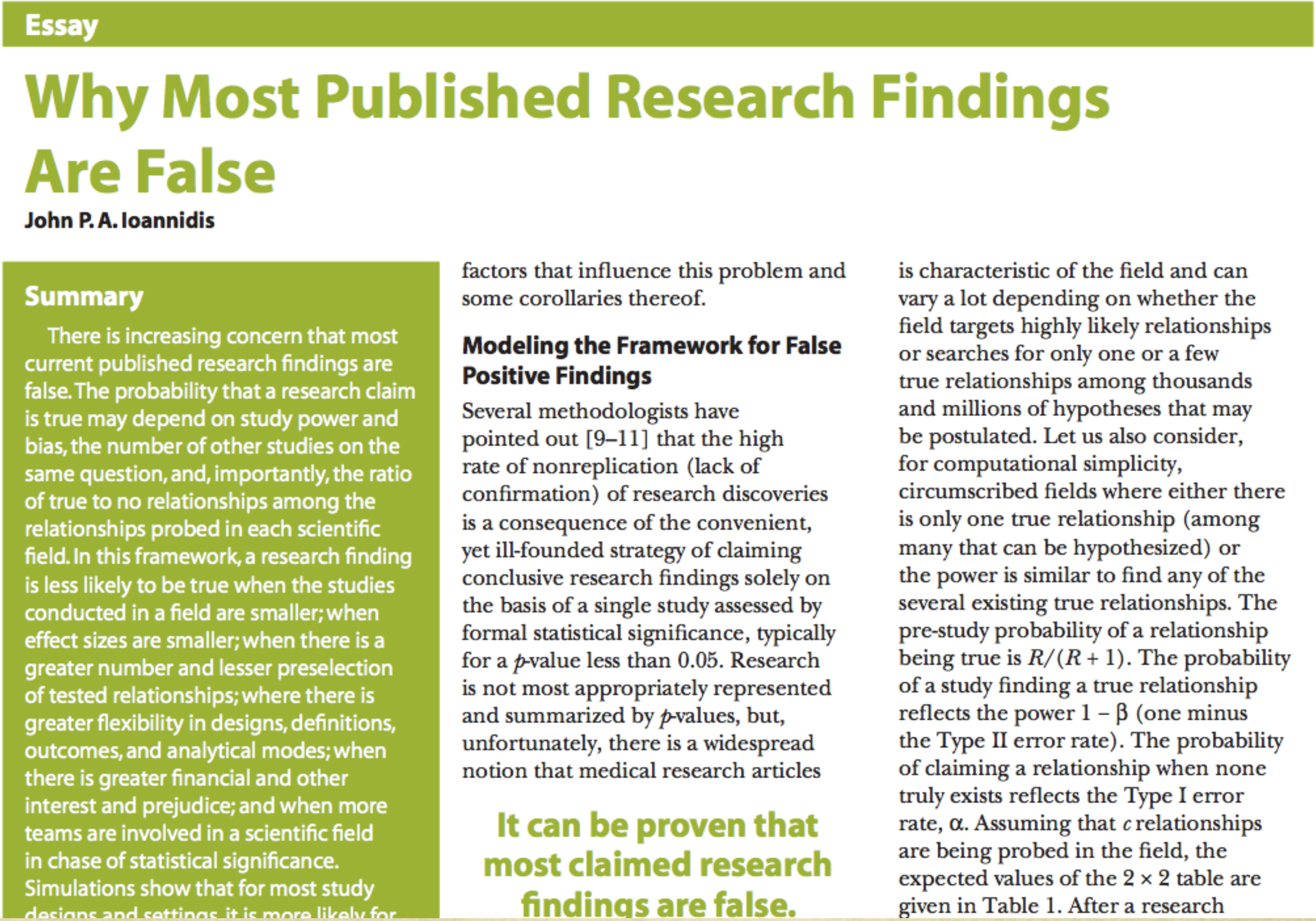
1000 C.E., Cannon of Medicine

Ronald Fisher

1920 CE, Design of Experiments

http://www.newworldencyclopedia.org/entry/Ronald_Fisher

https://en.wikipedia.org/wiki/Avicenna#/media/File:Avicenna_Portrait_on_Silver_Vase_-_Museum_at_BuAli_Sina_(Avicenna)_Mausoleum_-_Hamadan_-_Western_Iran_(7423560860).jpg



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

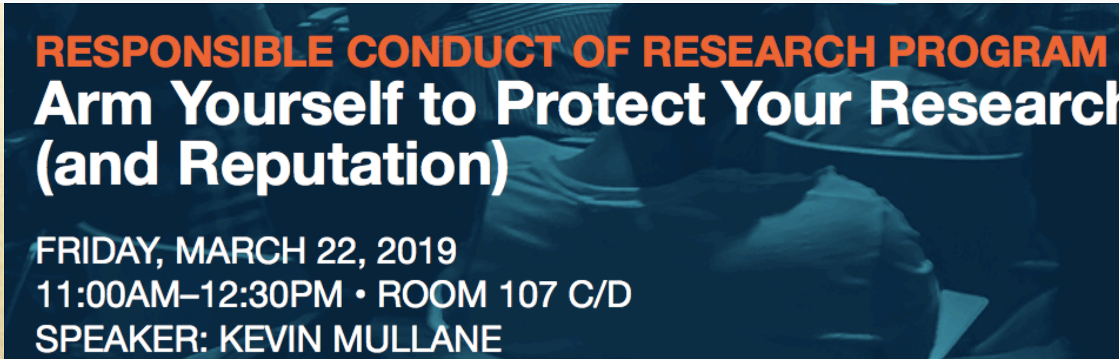
is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Medical research has a credibility problem

- Estimated that ~75% published research findings cannot be reproduced
- ~$28 billion per year (nearly half of the annual non-clinical research budget in the US) is wasted on attempts to reproduce published studies
- Only a small percentage are due to overt fraud (intentional fabrication)
- Most are what are considered “detrimental research practices”
- Patient lives placed at risk

Thanks: Kevin Mullane
Director, Corporate Liaison & Ventures
Corporate Ventures and Translation
Gladstone Institutes



RESPONSIBLE CONDUCT OF RESEARCH PROGRAM

Arm Yourself to Protect Your Research (and Reputation)

FRIDAY, MARCH 22, 2019
11:00AM–12:30PM • ROOM 107 C/D
SPEAKER: KEVIN MULLANE

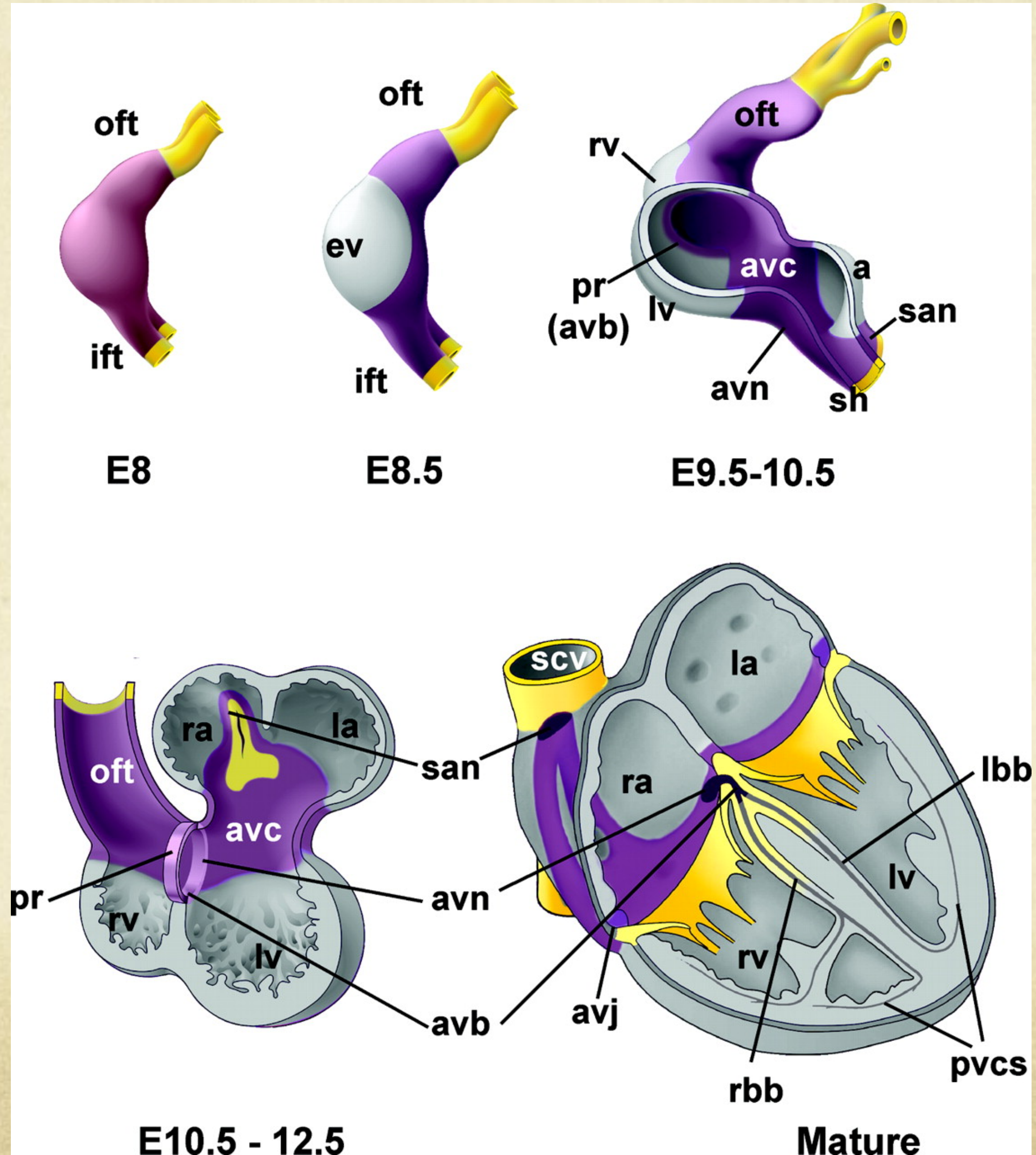
Motivation for use of Statistics

- We would like to make scientific claims that are as **generalizable** as possible
- However,
  - Empirical data are inherently noisy
  - Resources are limited
- Enter **Statistics!**

About this course

- Very basic introduction to concepts underpinning statistics and experimental design
- Goal is to get you thinking a little deeper about the data you want to generate and use to make claims about
- No prerequisites
- Please interrupt with questions and comments!

Experiment: Gene controlling developing heart



Seven pillars of statistical wisdom

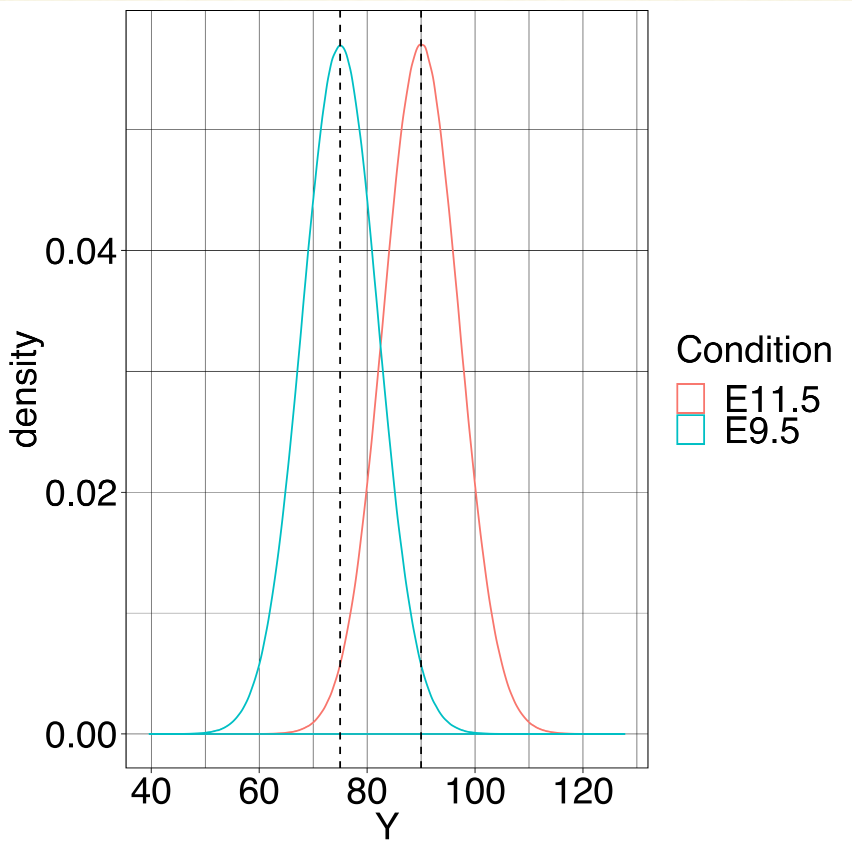
- Aggregation
- Information
- Inter-comparison
- Likelihood
- Regression
- Residuals

https://commons.wikimedia.org/wiki/File:Seven_Pillars_2008_e5.jpg

- Experimental design

Stephen M. Stigler. 2016. Seven Pillars of Statistical Wisdom. Harvard University Press

1. Aggregation: one number to capture an entire distribution



Target population

- All subjects/units that we want base our claims/conclusions on
  - The cardiac tissue of all mice at embryonic stage E9.5
  - All children below 5 years old who are diagnosed with autism

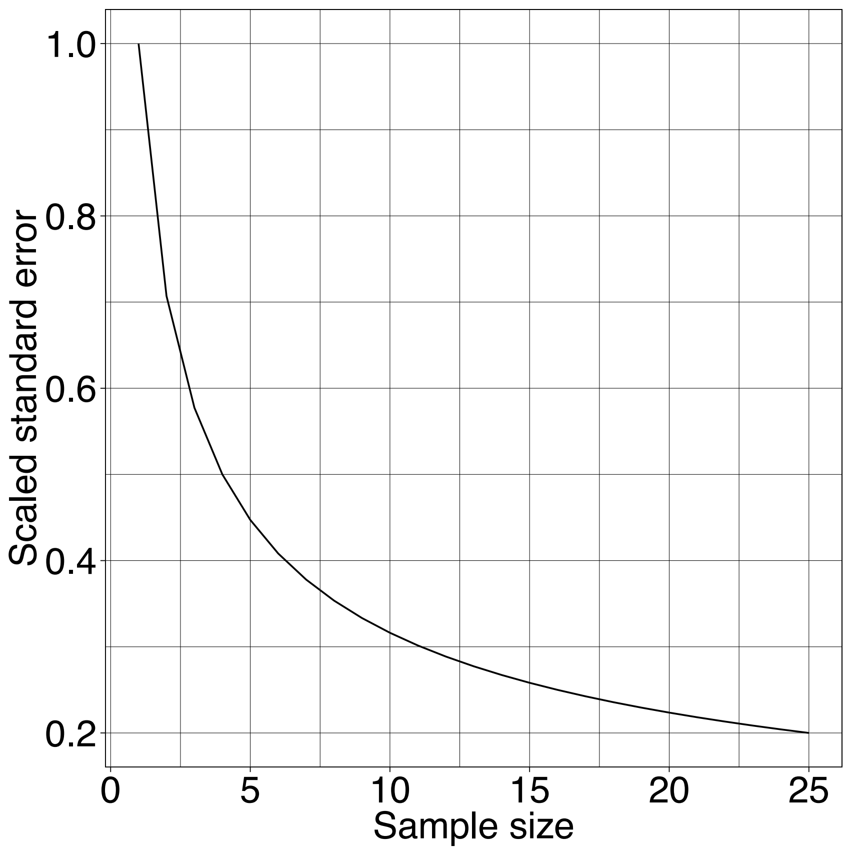
Seven pillars of statistical wisdom

- Aggregation
- **Information**
- Inter-comparison
- Likelihood
- Regression
- Residuals
- Experimental design

https://commons.wikimedia.org/wiki/File:Seven_Pillars_2008_e5.jpg

Stephen M. Stigler. 2016. Seven Pillars of Statistical Wisdom. Harvard University Press

2. Information on aggregate measure: rate of gain decreases with increasing sample size



Seven pillars of statistical wisdom

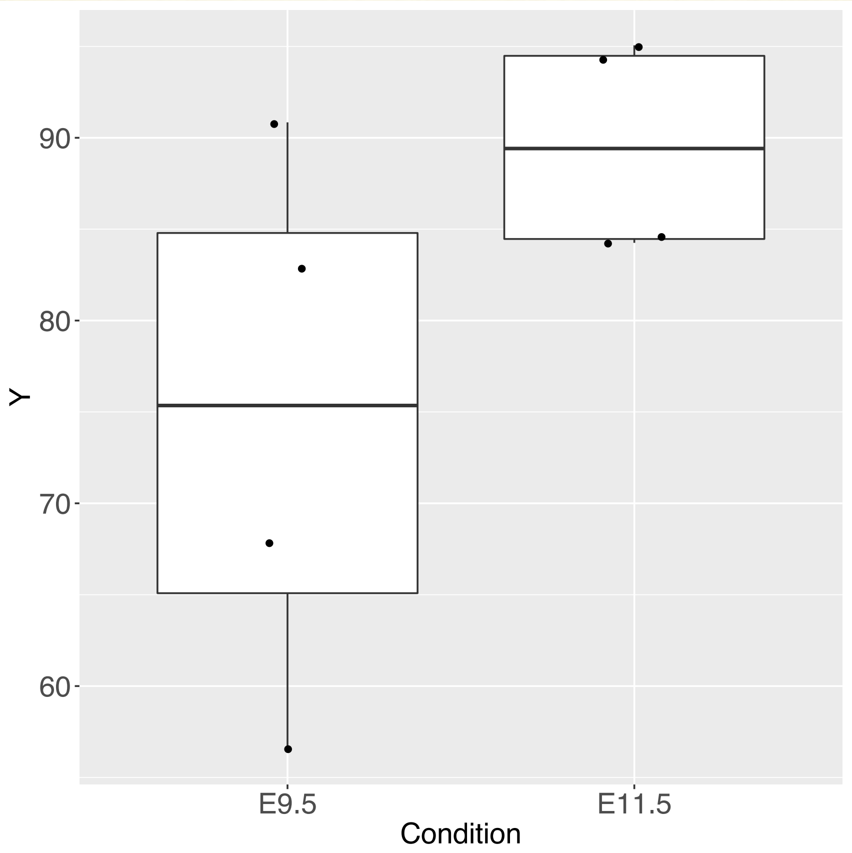
- Aggregation
- Information
- **Inter-comparison**
- Likelihood
- Regression
- Residuals
- Experimental design

https://commons.wikimedia.org/wiki/File:Seven_Pillars_2008_e5.jpg

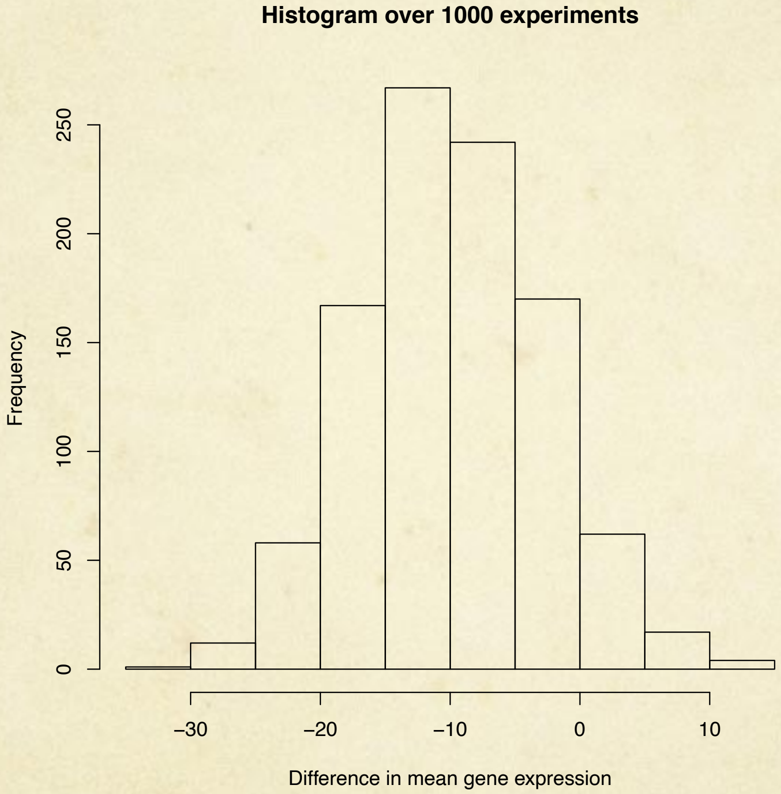
Stephen M. Stigler. 2016. Seven Pillars of Statistical Wisdom. Harvard University Press

3. **Inter-comparison**: with limited data can make conclusions applicable to larger target population

Is gene differentially expressed between the two developmental time-points?

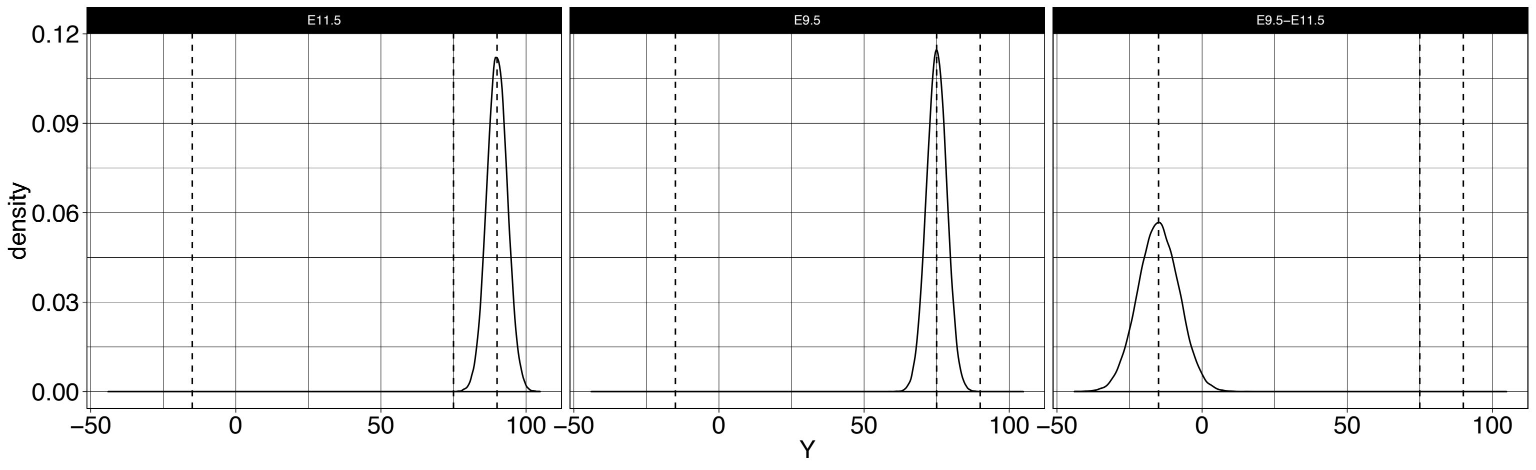


Convince a skeptic: Repeat this experiment 1000 times



Central limit theorem allows us to estimate the variation of the location of the distribution

$E11.5:\mathrm{N}ormal\left(90,\frac{7}{\sqrt{4}}\right)$ $E9.5:\mathrm{N}ormal\left(75,\frac{7}{\sqrt{4}}\right)$ $E9.5-E11.5:\mathrm{N}ormal\left(75-90,\frac{7+7}{\sqrt{4}}\right)$



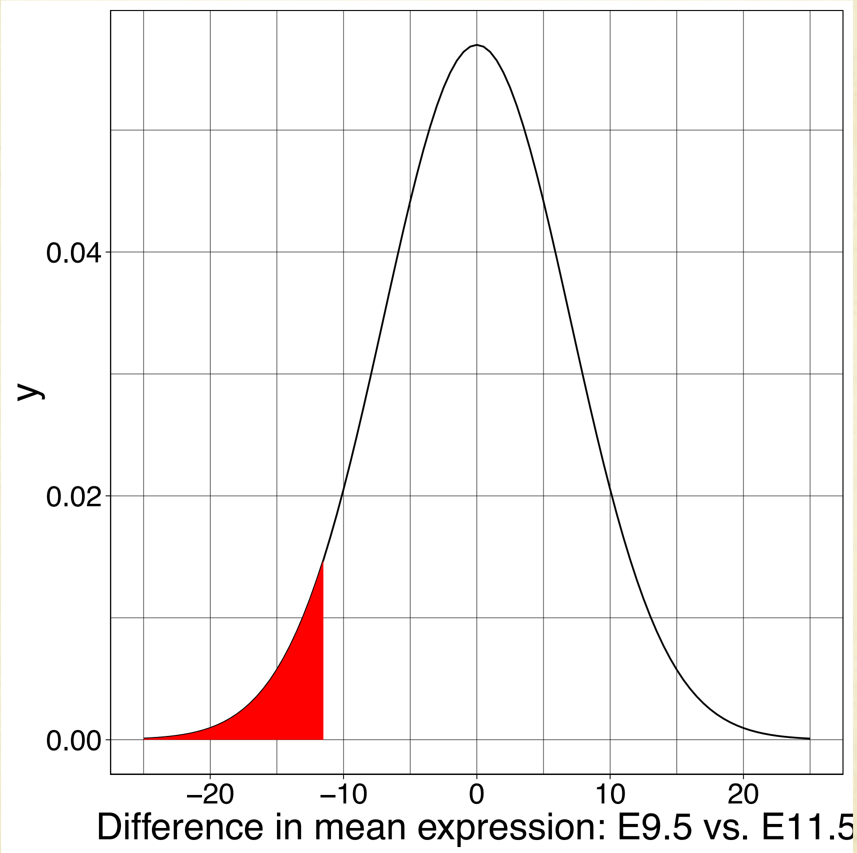
Two conclusions from the data

- *Conclusion 1*: The difference is interesting, biologically meaningful – PI happy, start writing manuscript, plan further experiments.

- *Conclusion 2*: Skeptical viewpoint, there is no difference, or unable to conclude that there is one – back to the drawing board.

- All statistical hypothesis testing is based on the latter the skeptical viewpoint

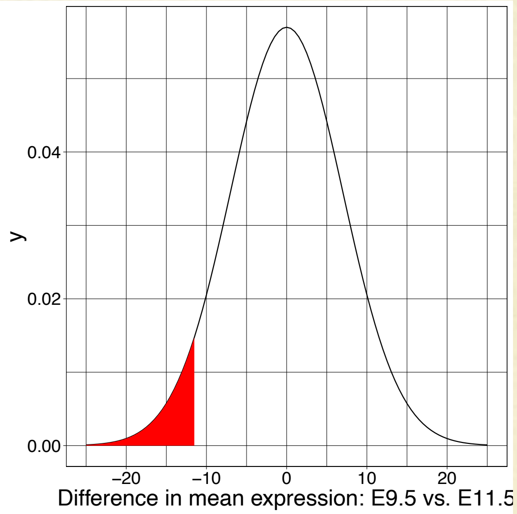
Theoretical distribution of difference in means



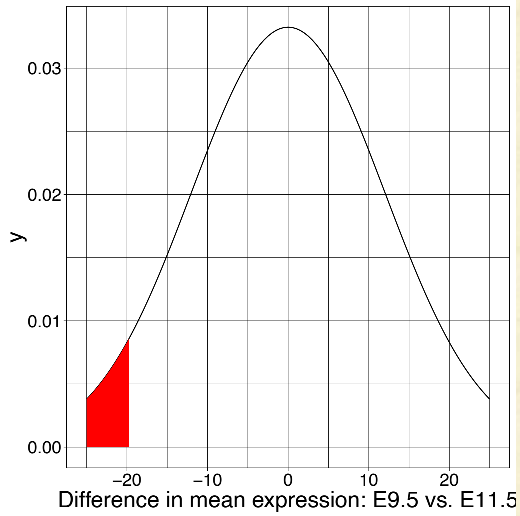
Type I error and p-value

Alter underlying variation

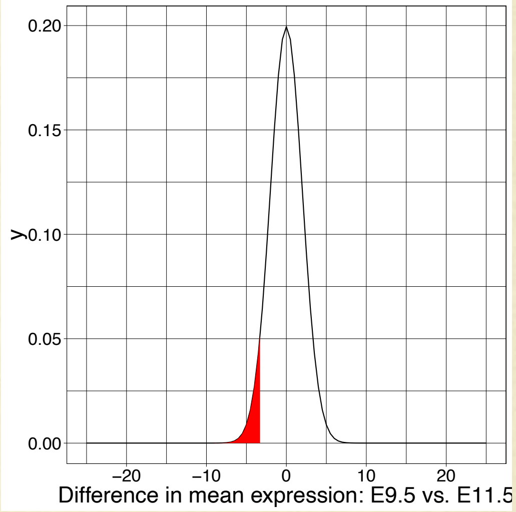
sd=14



sd=24

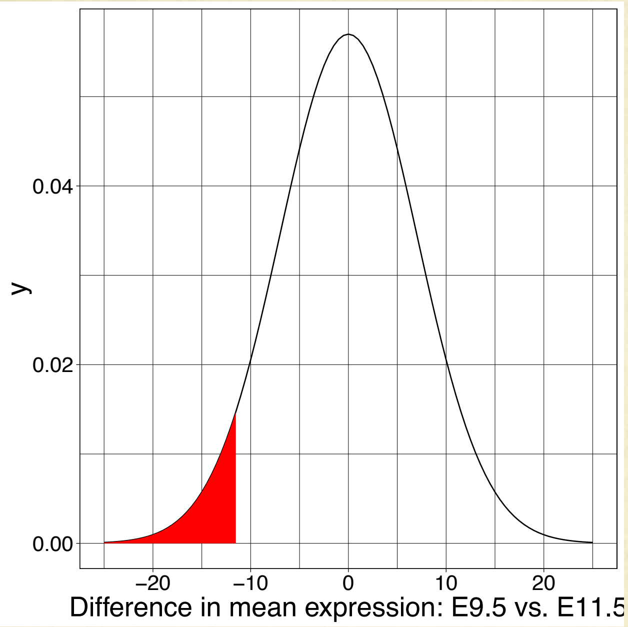


sd=4

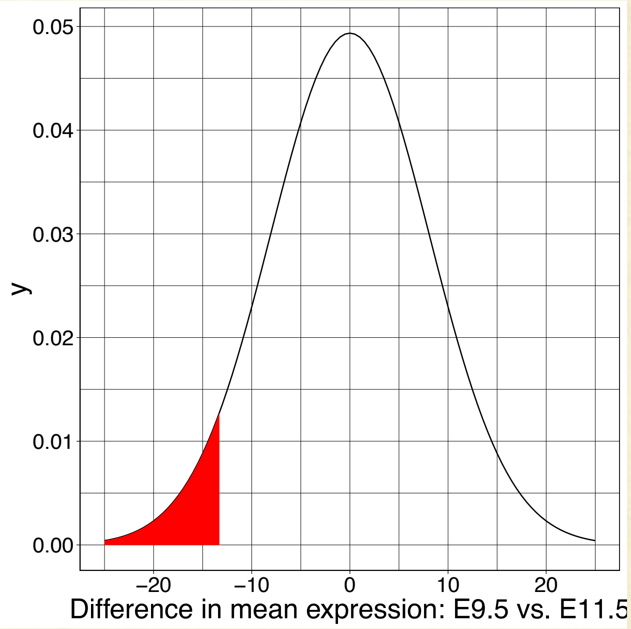


Alter the number of replicates

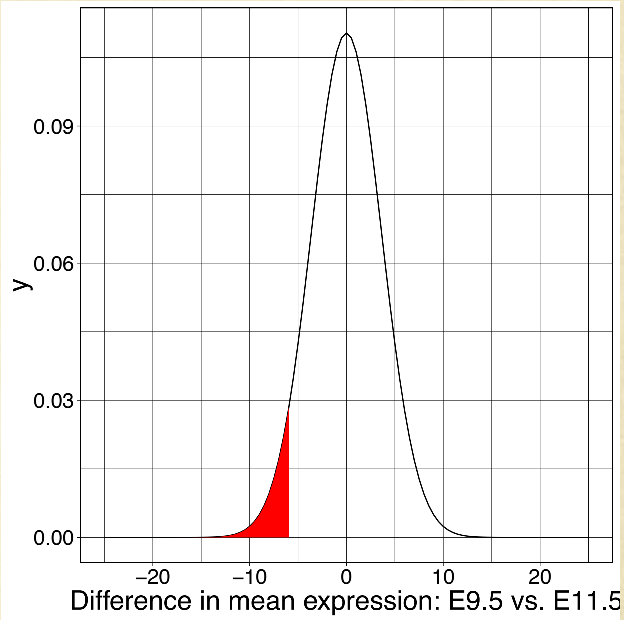
n=4



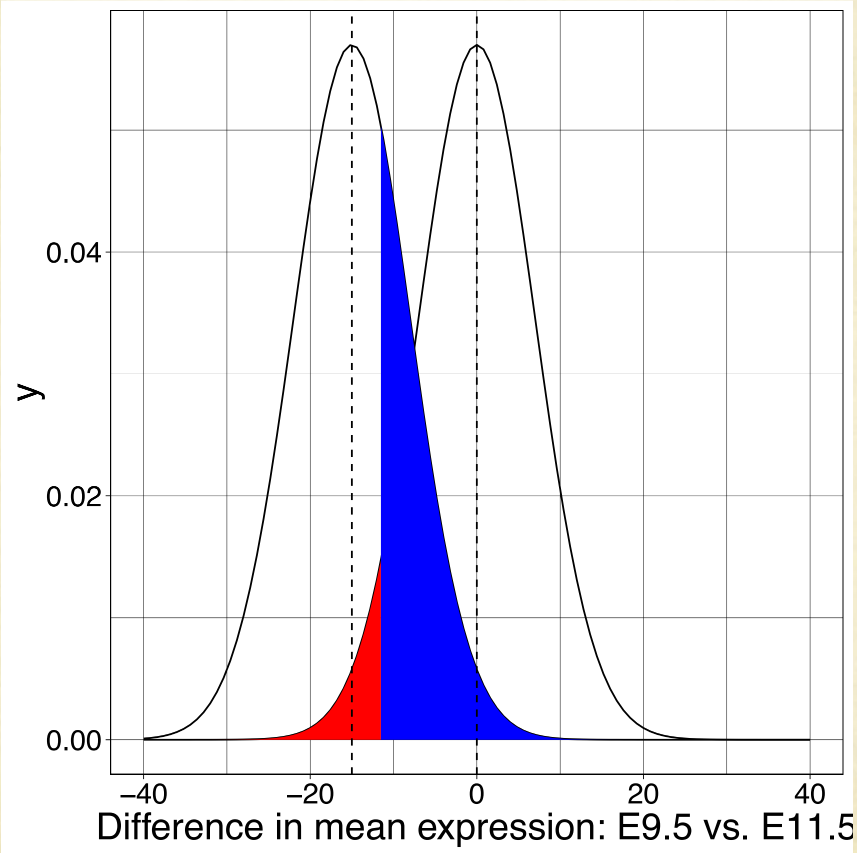
n=3



n=15



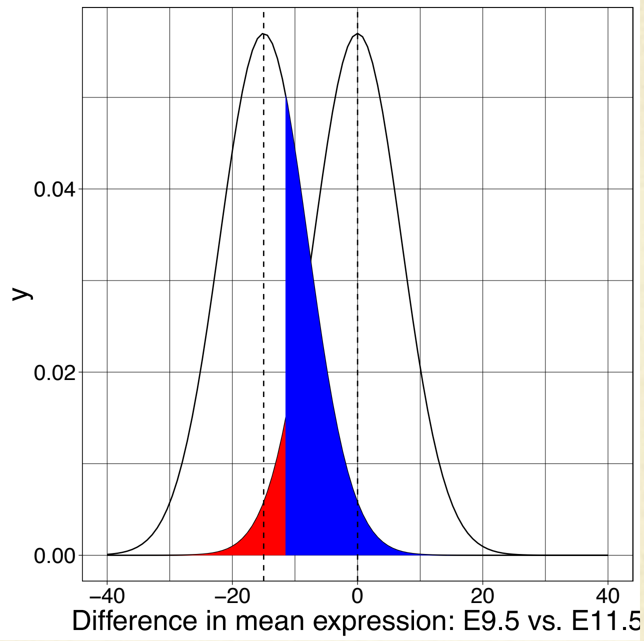
Power to detect a difference of means of -15



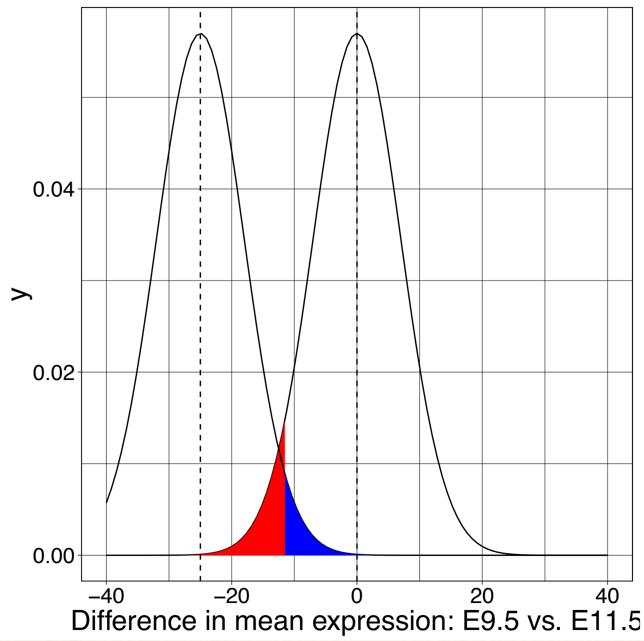
Type I and Type II error

Power to detect varying levels of difference in mean differences

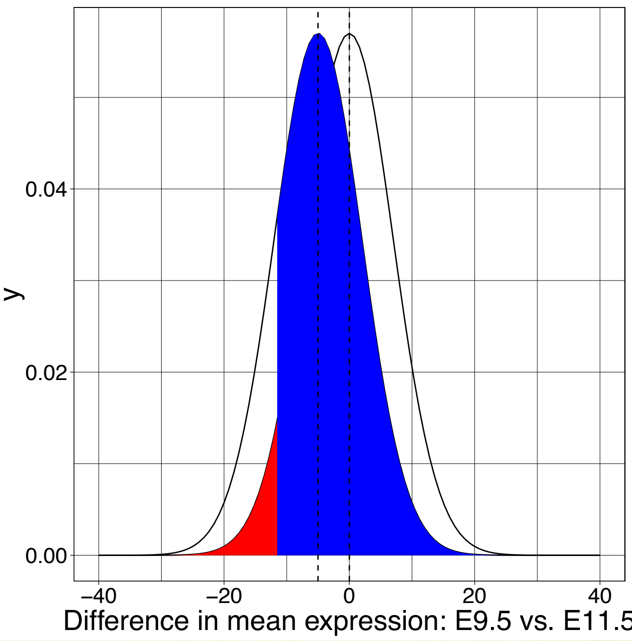
Mean diff = -15



Mean diff = -25



Mean diff = -5



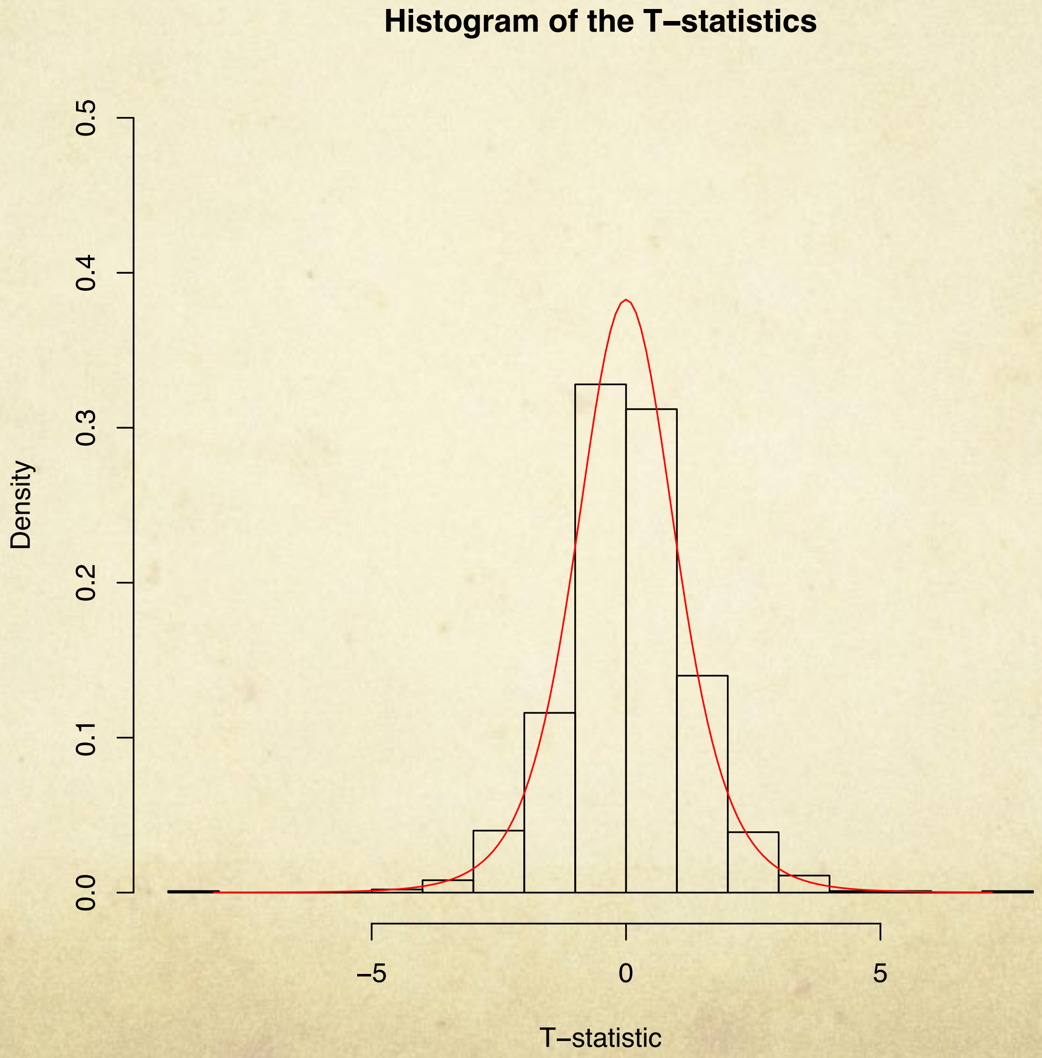
Terminology for Hypothesis Testing

- Response variables, predictor variables
- Type of variable: Continuous and categorical
  - What are the variables whose association we are interested in estimating?
  - What types are these variables?

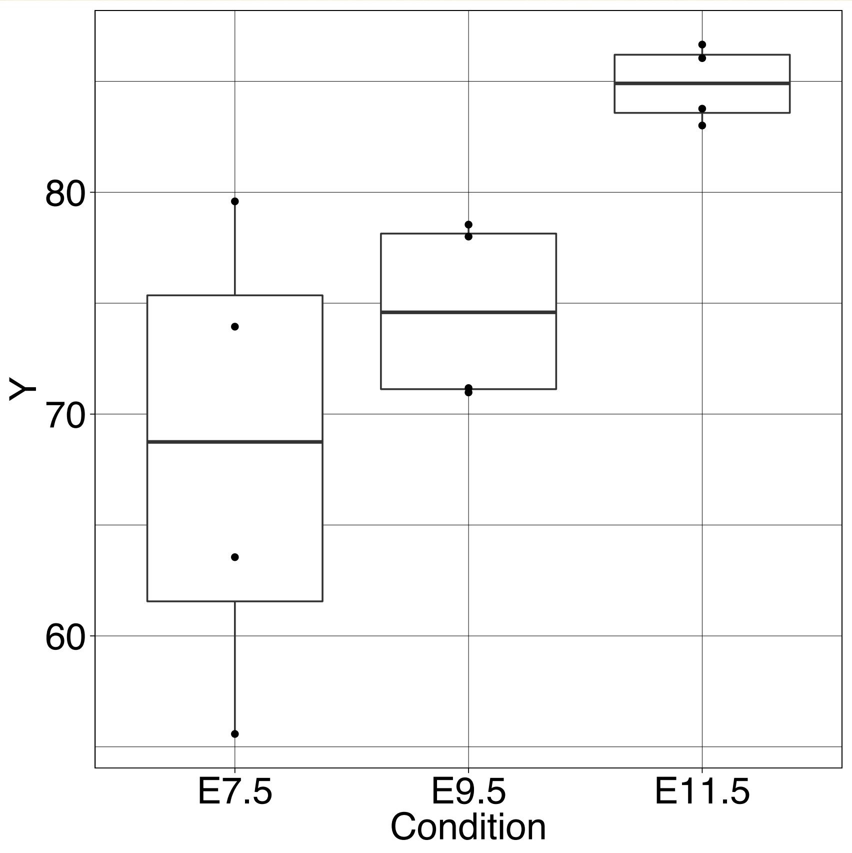
Z/T-statistic

$$Z = \frac{mean\left(Y_{E9.5}\right) - mean\left(Y_{E11.5}\right)}{sd(Y)\sqrt{\frac{1}{n} + \frac{1}{n}}}$$

T-statistic and sampling distribution



Continuous response and categorical predictor



Y: gene expression

X: development time

One-way ANOVA – F-statistics

Two categorical variables

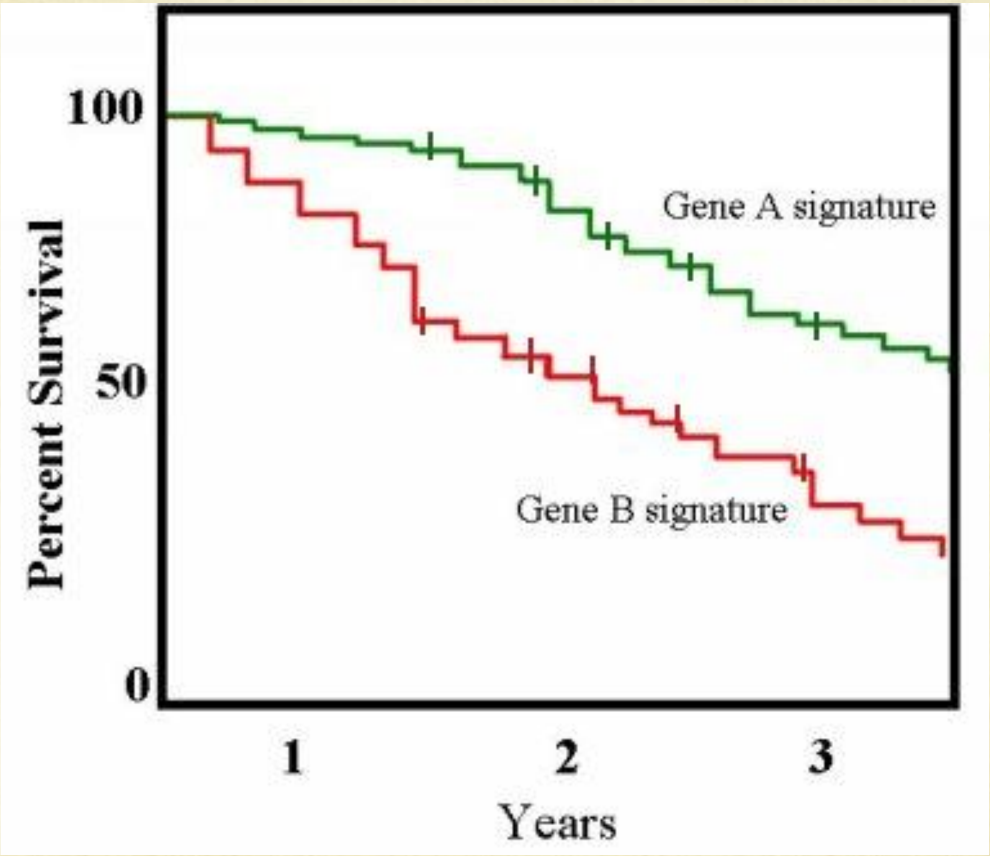
	In TGF-b signaling pathway	Not in TGF-b signaling pathway
Differentially expressed	20	980
Not differential expressed	80	18920

Y1: gene differentially expressed or not
Y2: gene in TGF-b signaling pathway or not

Odds ratio, Chi-square statistics

Continuous response with a categorical variable

https://commons.wikimedia.org/wiki/File:Km_plot.jpg

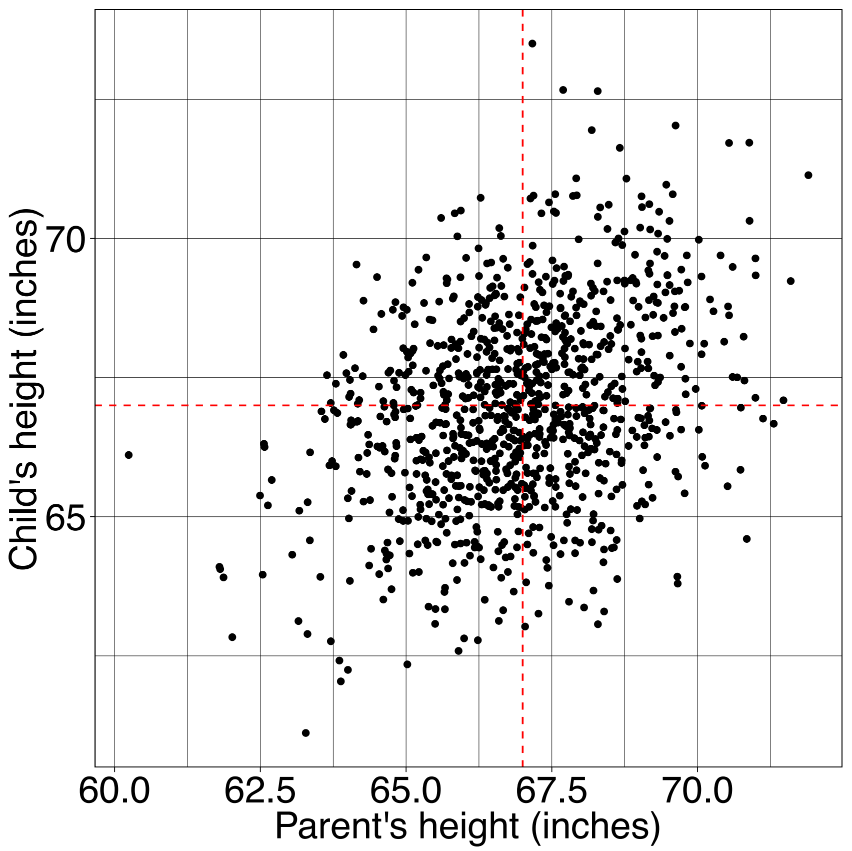


Y: survival time in years

X: Gene signature

Hazard ratio, logrank test

Continuous response against a continuous variable



Y: Child's height

X: Parent's height

Slope, linear regression

Seven pillars of statistical wisdom

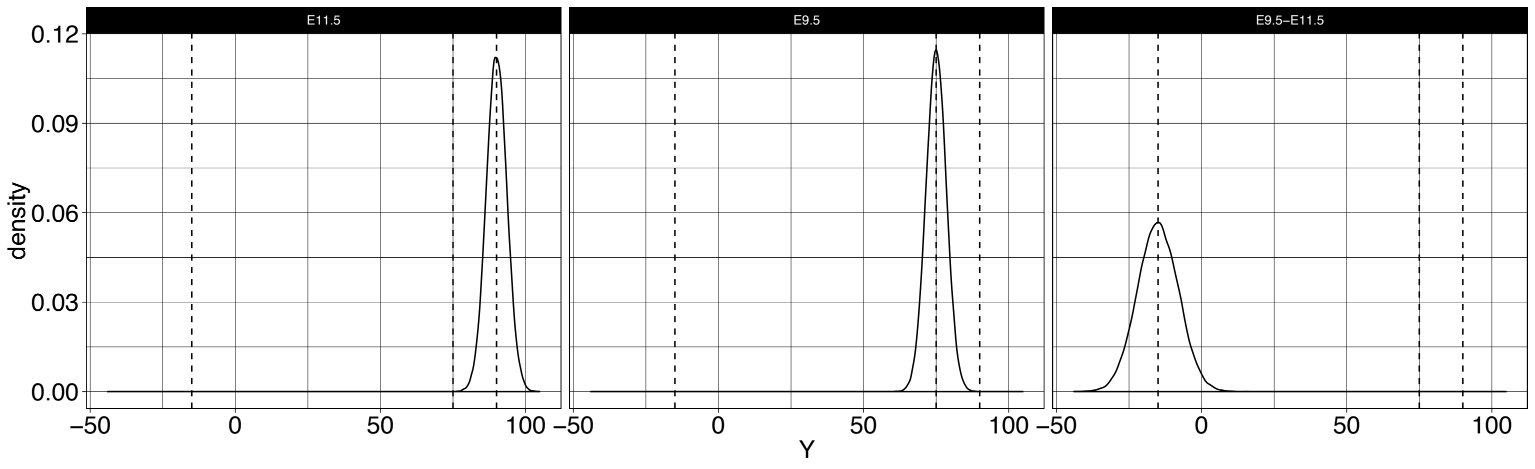
- Aggregation
- Information
- **Inter-comparison**
- **Likelihood**
- Regression
- Residuals
- Experimental design

https://commons.wikimedia.org/wiki/File:Seven_Pillars_2008_e5.jpg

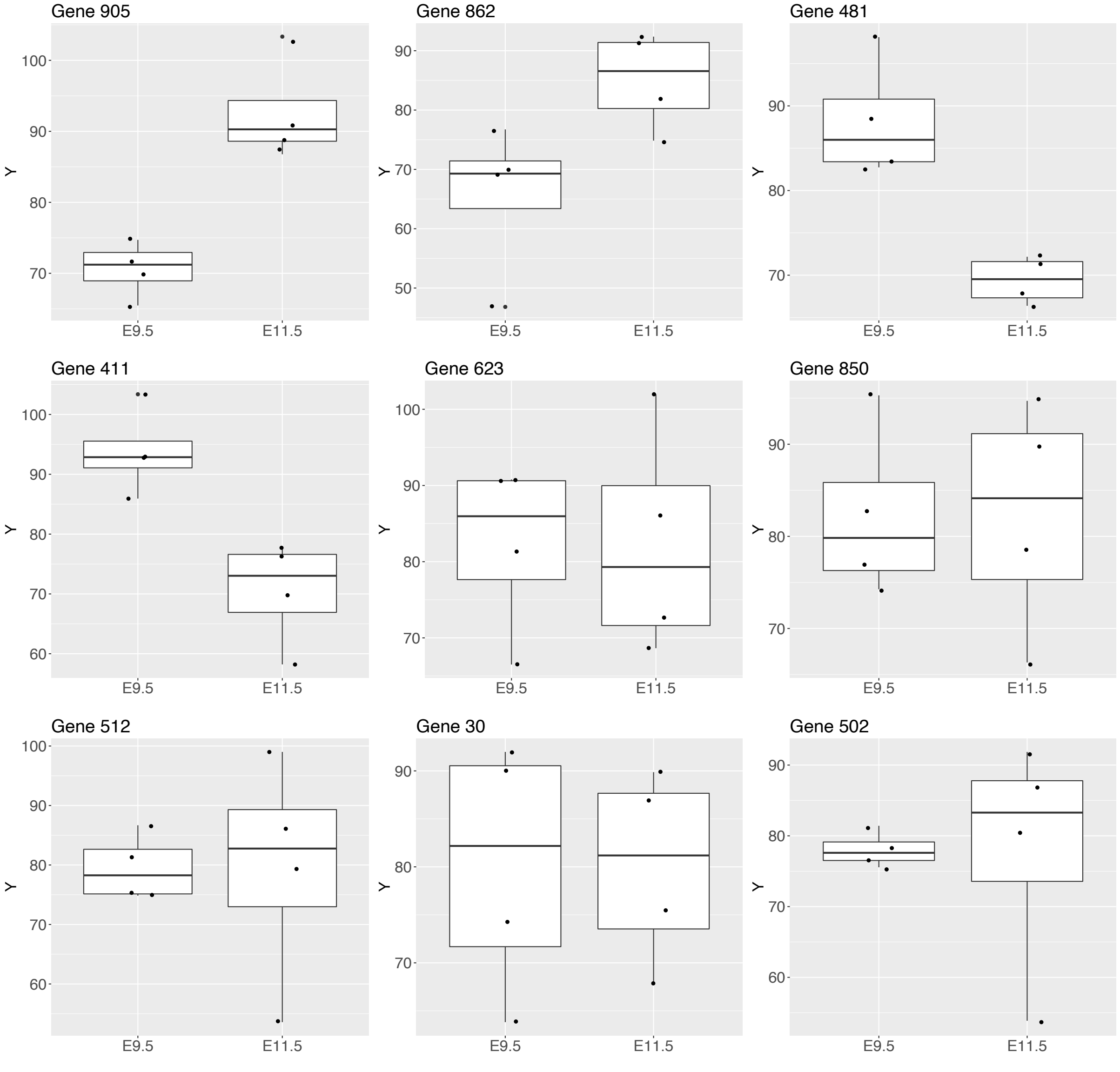
Stephen M. Stigler. 2016. Seven Pillars of Statistical Wisdom. Harvard University Press

4. Likelihood: model variation in the location of data using probability

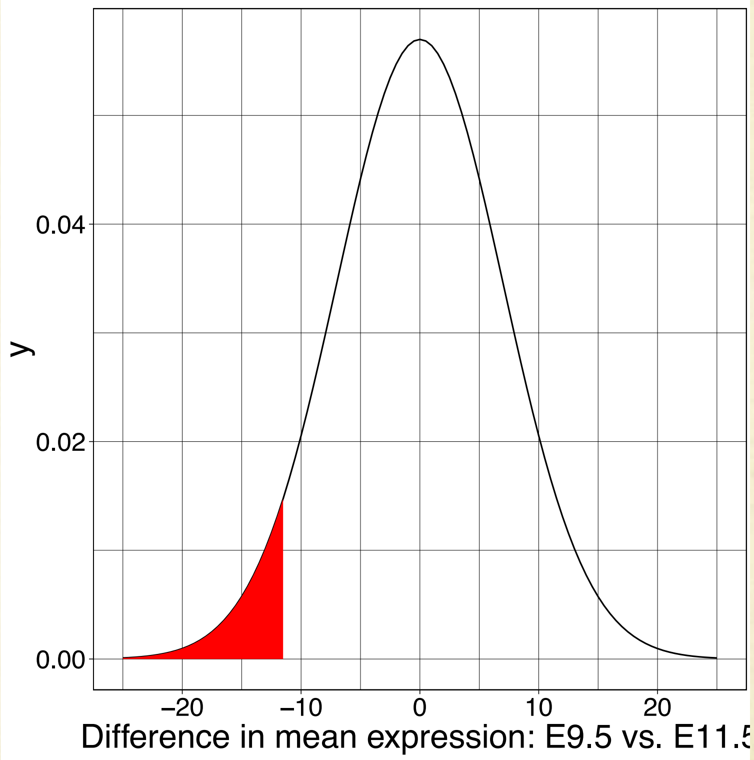
$E11.5 : \mathrm{N}ormal\left(90, \frac{7}{\sqrt{4}}\right)$ $E9.5 : \mathrm{N}ormal\left(75, \frac{7}{\sqrt{4}}\right)$ $E9.5 - E11.5 : \mathrm{N}ormal\left(75 - 90, \frac{7+7}{\sqrt{4}}\right)$



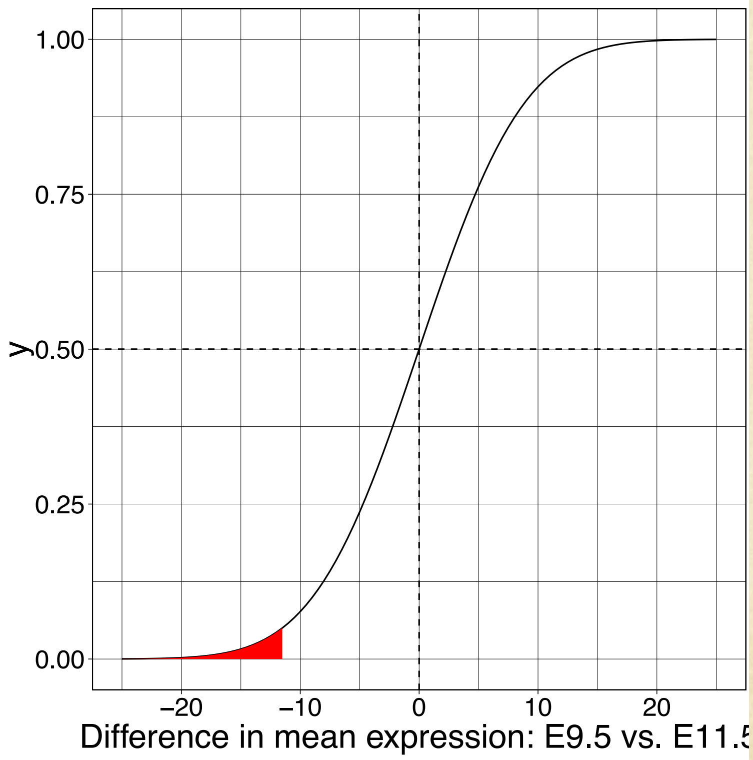
Testing for differences in expression of multiple genes



Density

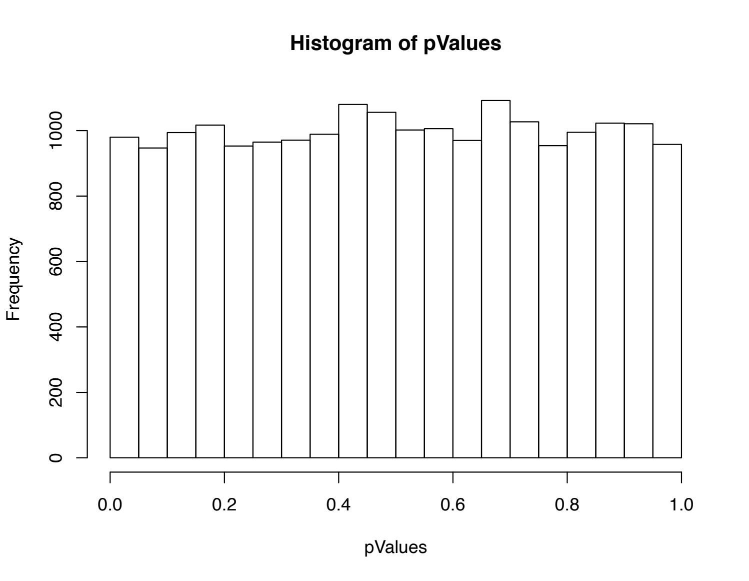


Distribution

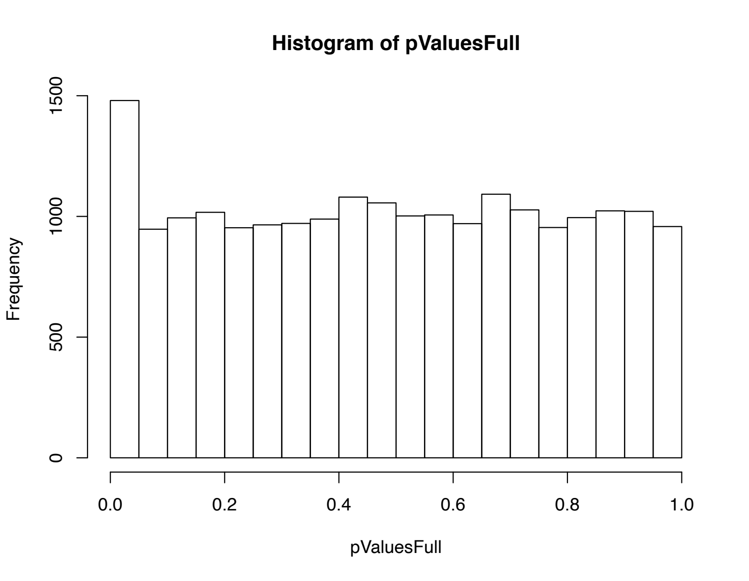


Look at distribution of p-values

No real differences



Possible differences



Multiple testing procedures: Holm, Benjamini-Hochberg

Seven pillars of statistical wisdom

- Aggregation
- Information
- Inter-comparison
- Likelihood
- **Regression**
- Residuals
- Experimental design

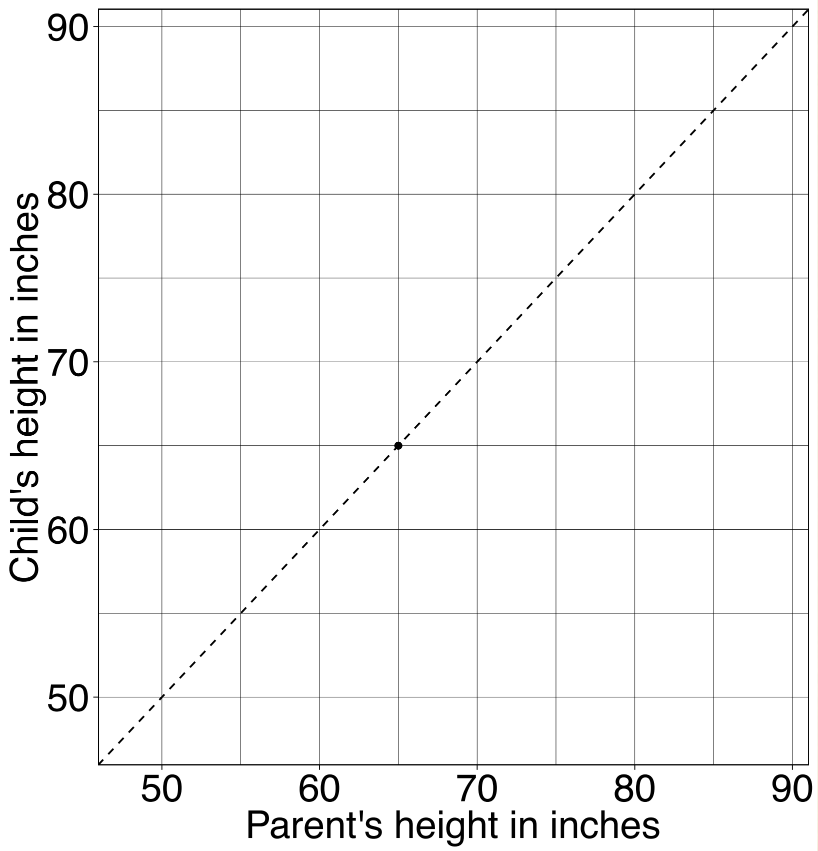
https://commons.wikimedia.org/wiki/File:Seven_Pillars_2008_e5.jpg

Stephen M. Stigler. 2016. Seven Pillars of Statistical Wisdom. Harvard University Press

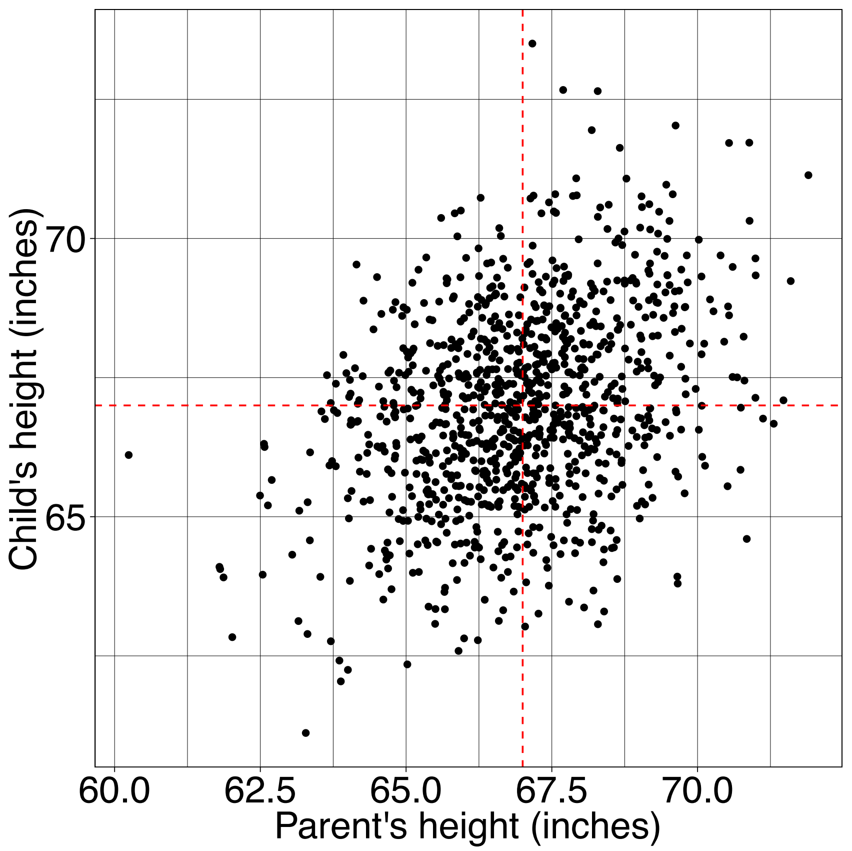
I have no faith in anything short of actual measurement and the Rule of Three

– Charles Darwin

Rule of three



5. Regression: associate multiple (noisy) factors with each other



Tall parents tend to have shorter children while tall children tend to have shorter parents

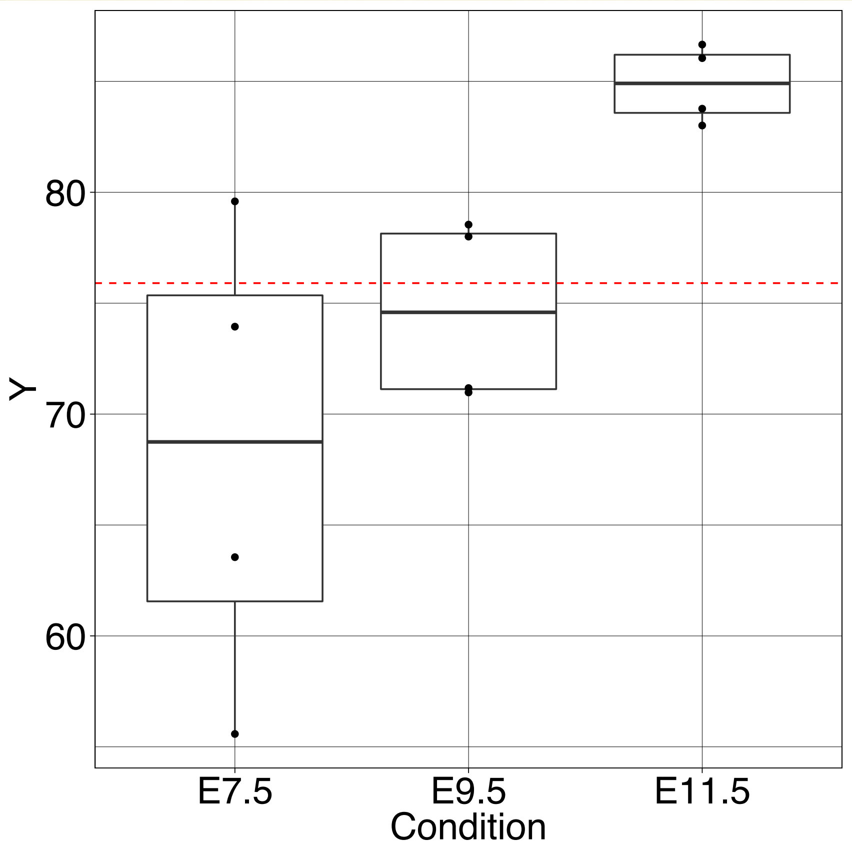
Seven pillars of statistical wisdom

- Aggregation
- Information
- Inter-comparison
- Likelihood
- Regression
- **Residuals**
- Experimental design

https://commons.wikimedia.org/wiki/File:Seven_Pillars_2008_e5.jpg

Stephen M. Stigler. 2016. Seven Pillars of Statistical Wisdom. Harvard University Press

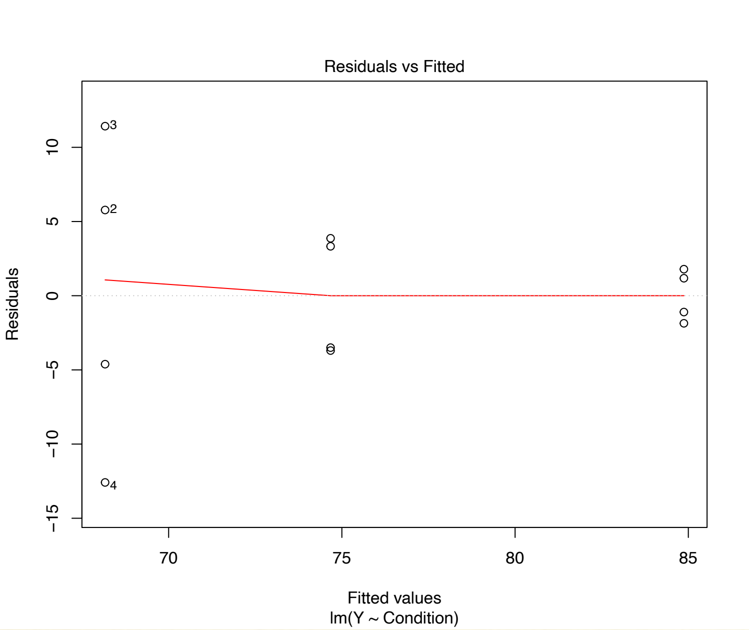
7. **Residual:** Variation left over after we have captured the known effects

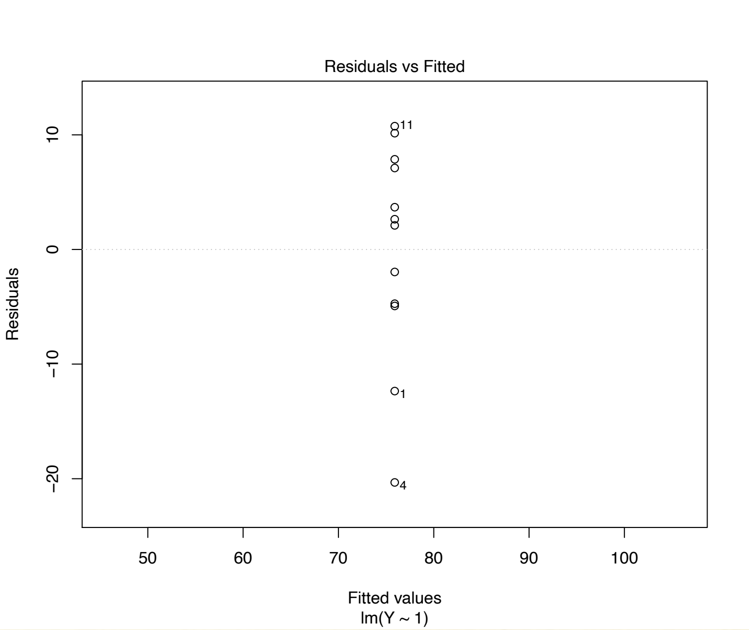


Residual: Predicted - Observed

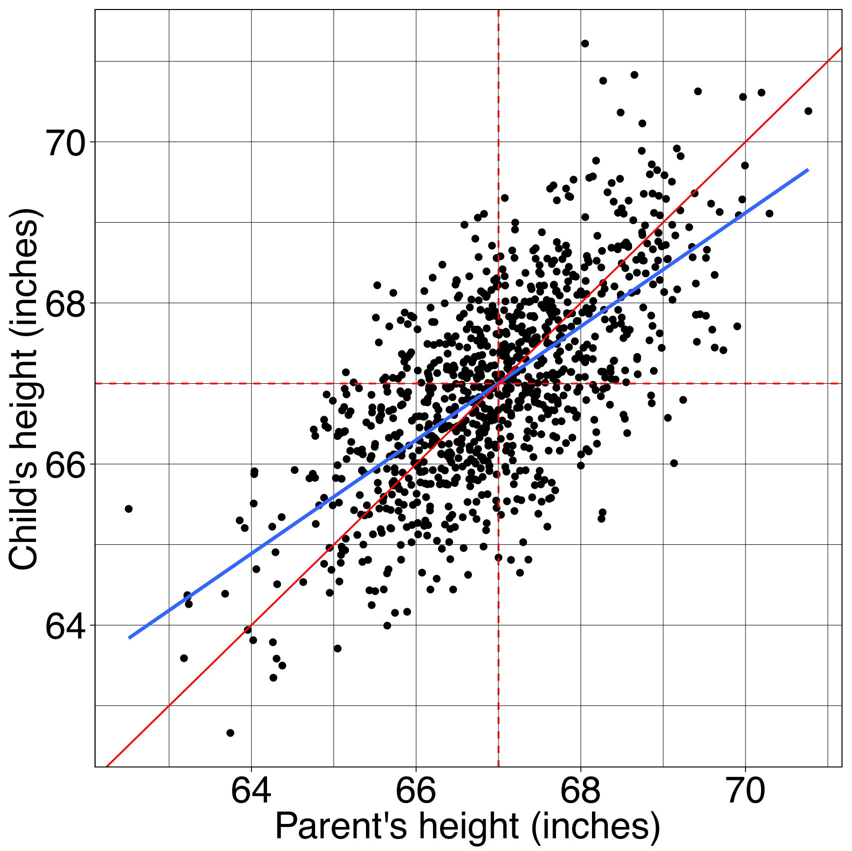
Full model

Mean model

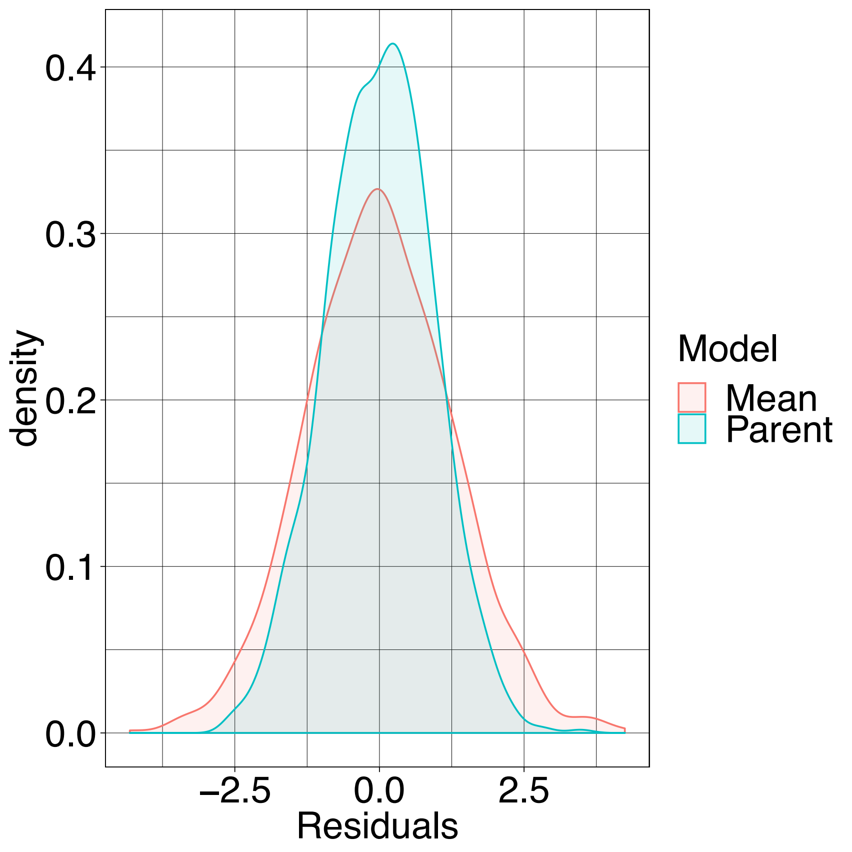




Predict child's height from parent's height



Distribution of error in predicting child's height



Seven pillars of statistical wisdom

- Aggregation
- Information
- Inter-comparison
- Likelihood
- Regression
- Residuals
- **Experimental design**

https://commons.wikimedia.org/wiki/File:Seven_Pillars_2008_e5.jpg

Stephen M. Stigler. 2016. Seven Pillars of Statistical Wisdom. Harvard University Press

7. **Design:** Capture effects of interest and avoid unwanted variation

- Identify the response and variable(**S**) of interest
- Identify target population that you want to base your claims on
- Identify factors that affect the response of interest
- Choose samples from target population
- **Randomly** assign samples across different levels of factors affecting response
- **Block out** variation that is not of interest by randomly assigning to levels of factors within a block

Which is better? Design 1 or Design 2?

	Design 1 – Sample prep date	Design 2 – Sample prep date
Sample_1_E9.5	Jan 9th, 2019	Jan 11th, 2019
Sample_2_E9.5	Jan 9th, 2019	Jan 9th, 2019
Sample_3_E9.5	Jan 9th, 2019	Jan 11th, 2019
Sample_4_E9.5	Jan 9th, 2019	Jan 9th, 2019
Sample_1_E11.5	Jan 11th, 2019	Jan 11th, 2019
Sample_2_E11.5	Jan 11th, 2019	Jan 9th, 2019
Sample_3_E11.5	Jan 11th, 2019	Jan 11th, 2019
Sample_4_E11.5	Jan 11th, 2019	Jan 9th, 2019

Which is better? Design 1 or Design 2?

	Design 1 – Gender	Design 2 - Gender
Sample_1_E9.5	Male	Male
Sample_2_E9.5	Male	Female
Sample_3_E9.5	Male	Male
Sample_4_E9.5	Male	Female
Sample_1_E11.5	Female	Male
Sample_2_E11.5	Female	Female
Sample_3_E11.5	Female	Male
Sample_4_E11.5	Female	Female

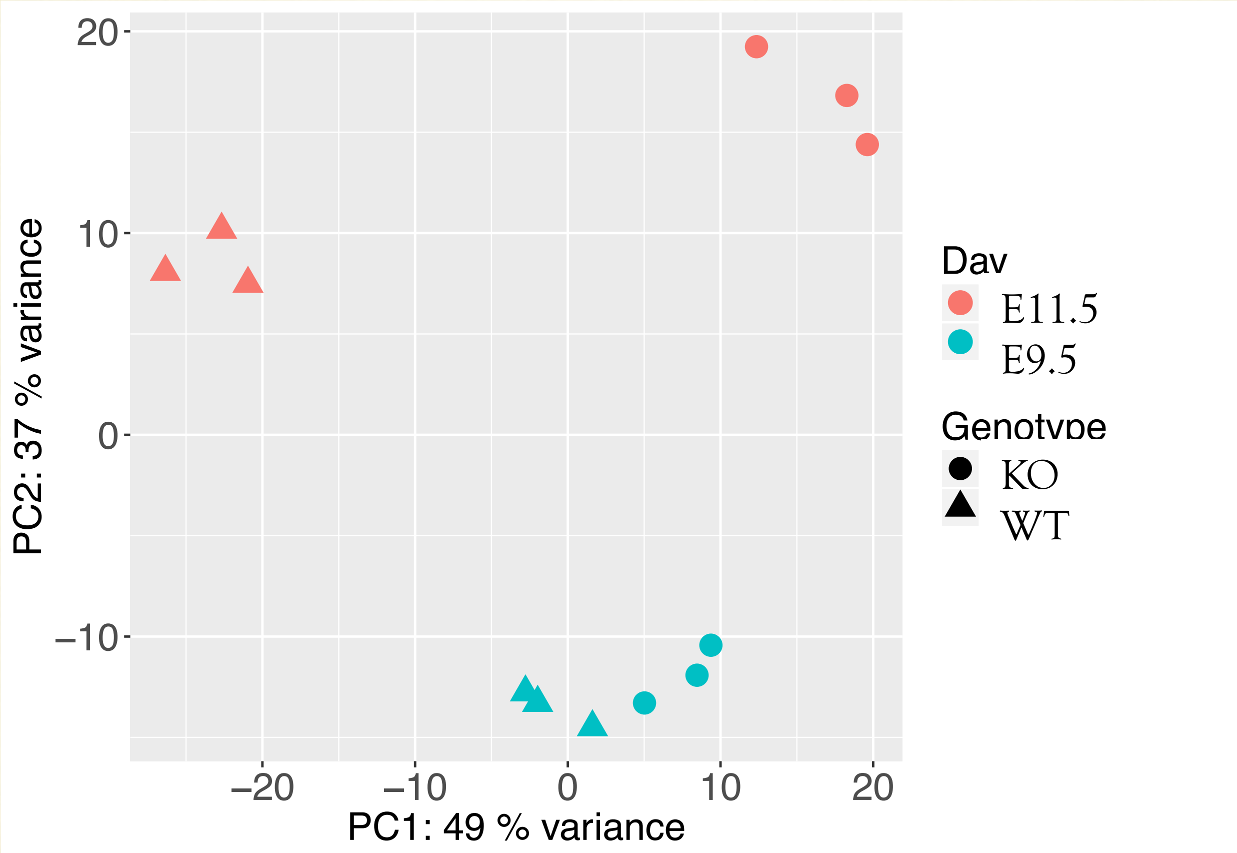
Which is better? Design 1 or Design 2?

	Design 1 – Sample prep date and Gender	Design 2 – Sample prep date and Gender
Sample_1_E9.5	Jan 11th, Male	Jan 11th, Male
Sample_2_E9.5	Jan 9th, Female	Jan 9th, Male
Sample_3_E9.5	Jan 11th, Female	Jan 11th, Female
Sample_4_E9.5	Jan 9th, Male	Jan 9th, Female
Sample_1_E11.5	Jan 11th, Male	Jan 11th, Male
Sample_2_E11.5	Jan 9th, Female	Jan 9th, Male
Sample_3_E11.5	Jan 11th, Male	Jan 11th, Female
Sample_4_E11.5	Jan 9th, Female	Jan 9th, Female

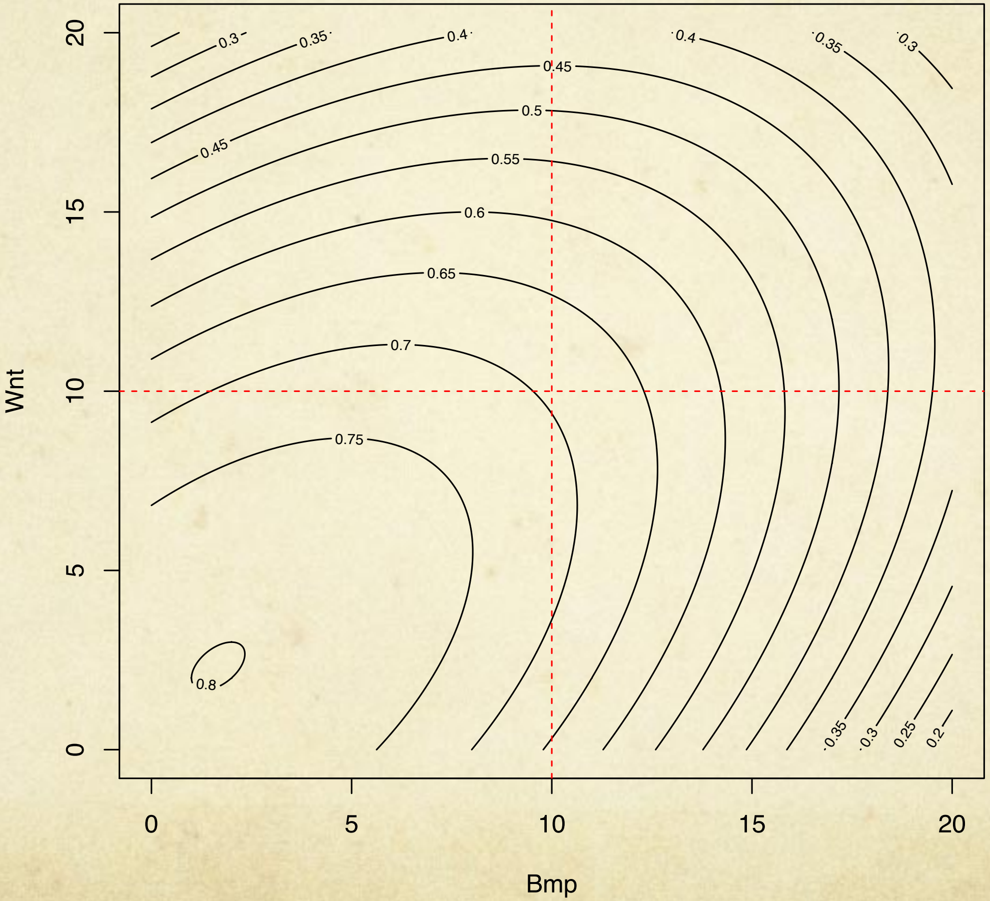
How many *n*? What do we need to perform statistical power calculations?

- What is the experimental design?
- Identify parameters of interest given experimental design – two variables models to more complex multivariate designs
- Test statistic for the parameters of interest
- Estimates of variation and correlation between variables of interest – use pilot data or publicly available data
- Sampling distribution of this test statistic
  - Check assumptions for the validity of the sampling distributions

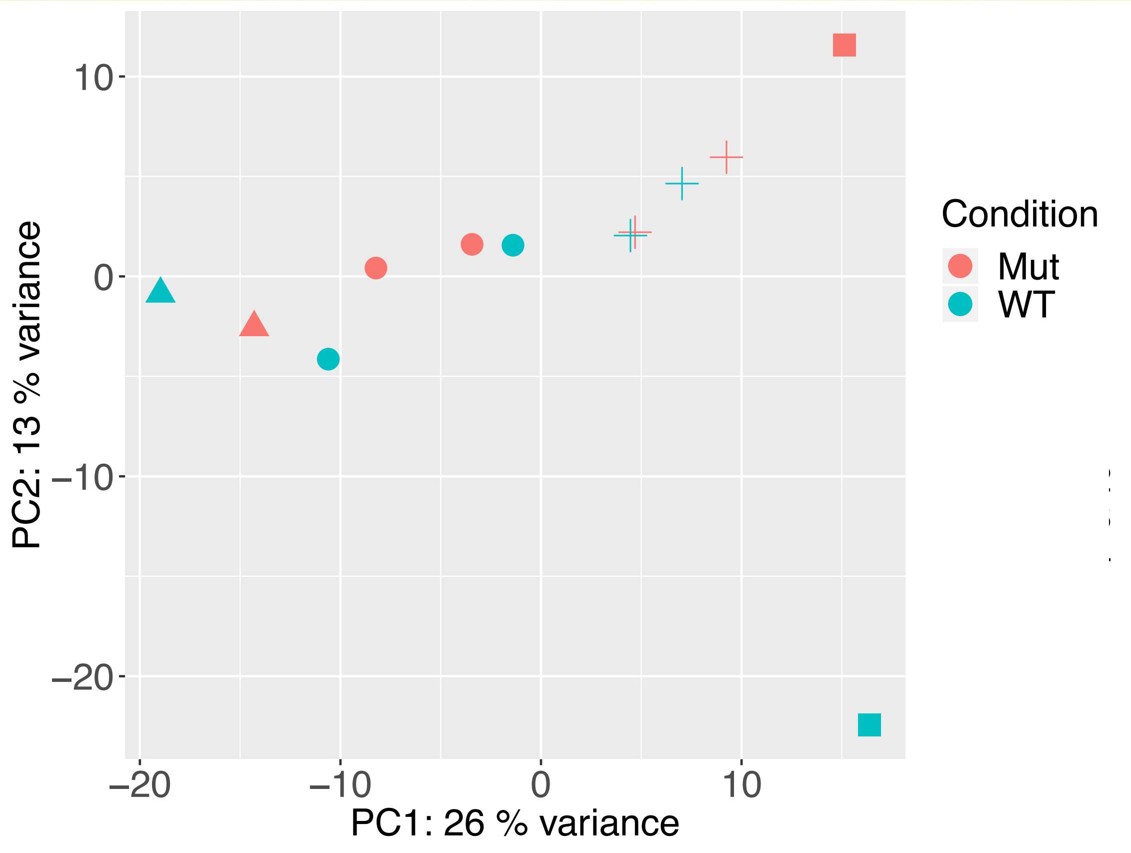
Genotype and development time effect on gene expression



Cellular reprogramming efficiency as a function of Wnt and Bmp levels



Genotype effect on gene expression



Litter effect dominates the variation

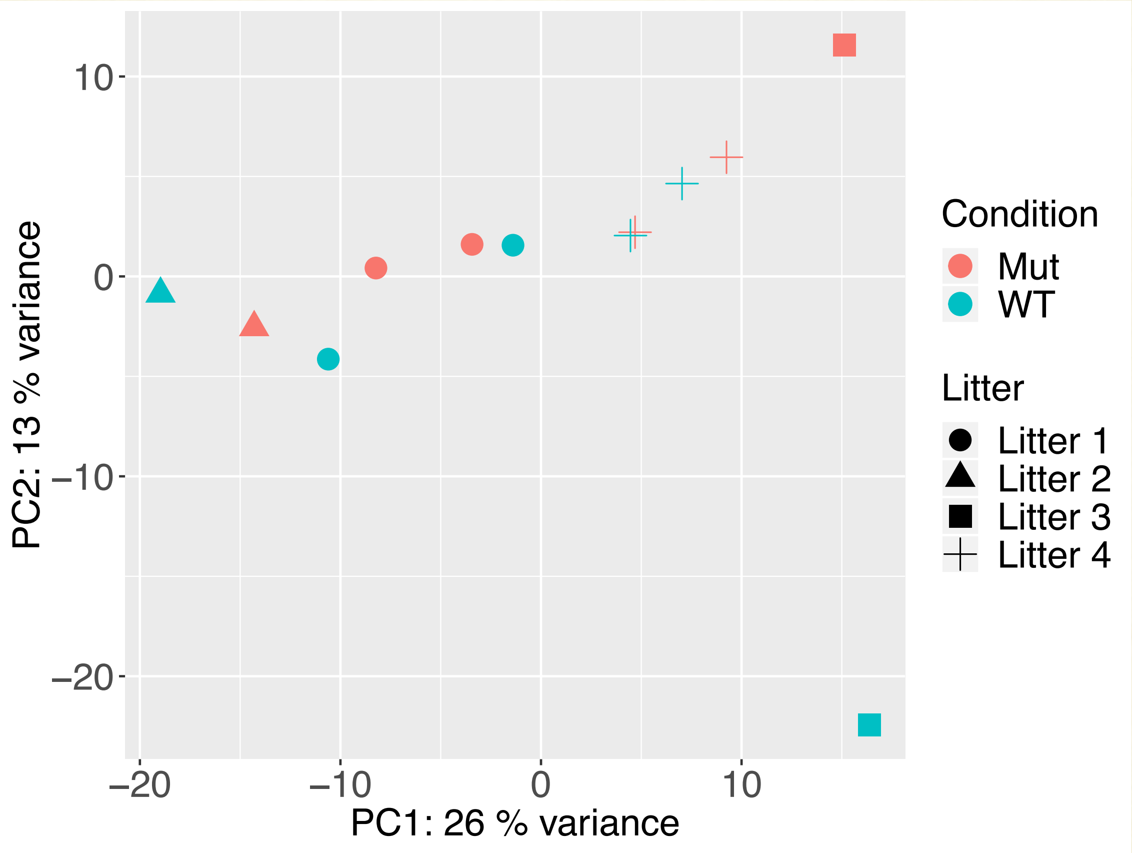
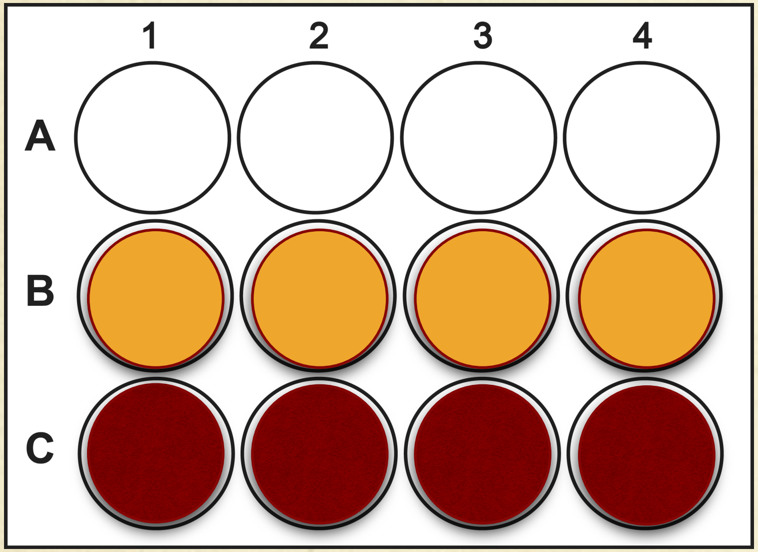
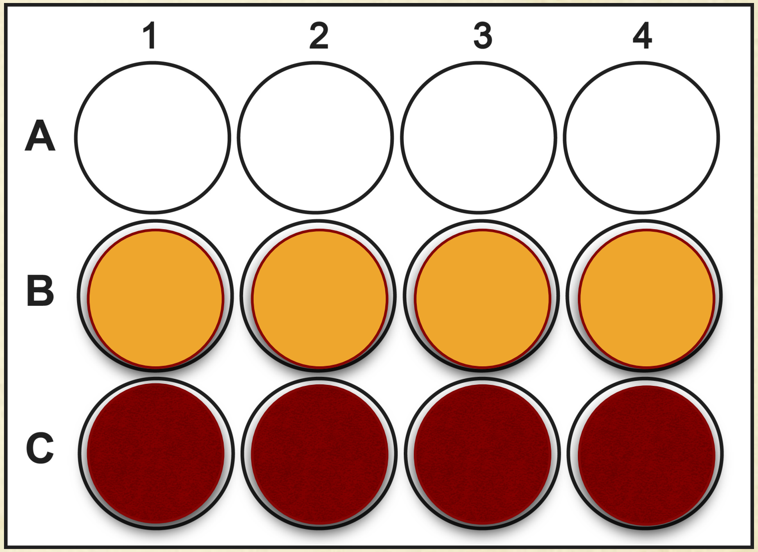


Plate design: Response over time

Before treatment with drug



After treatment with drug



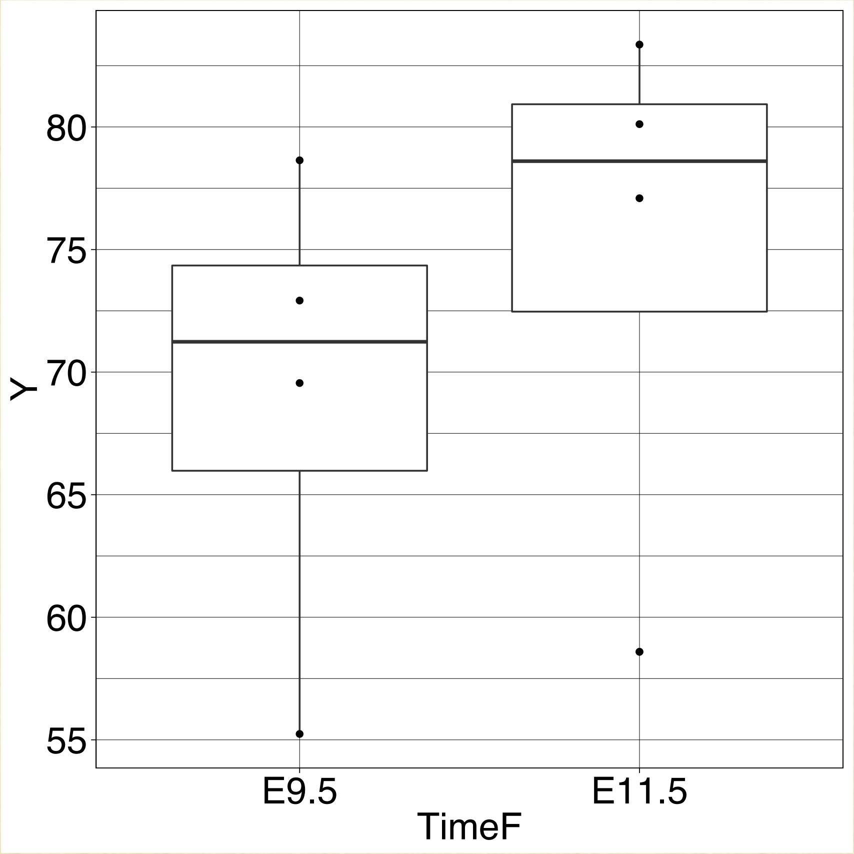
○ No drug

● Low dose

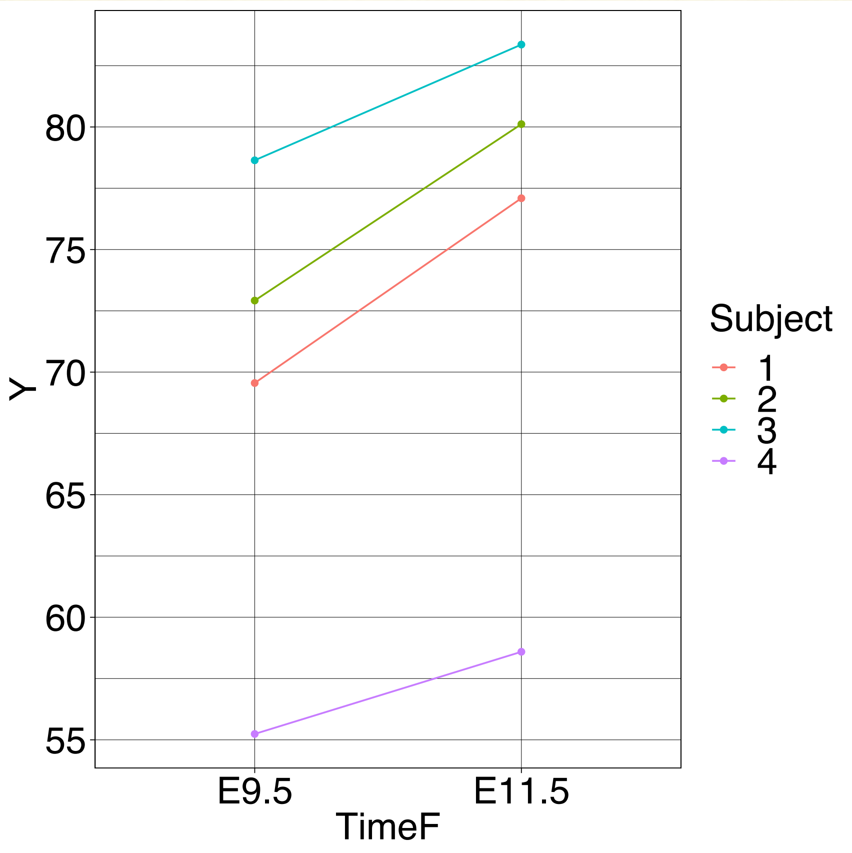
● High dose

http://www.cellsignet.com/media/plates/12.jpg

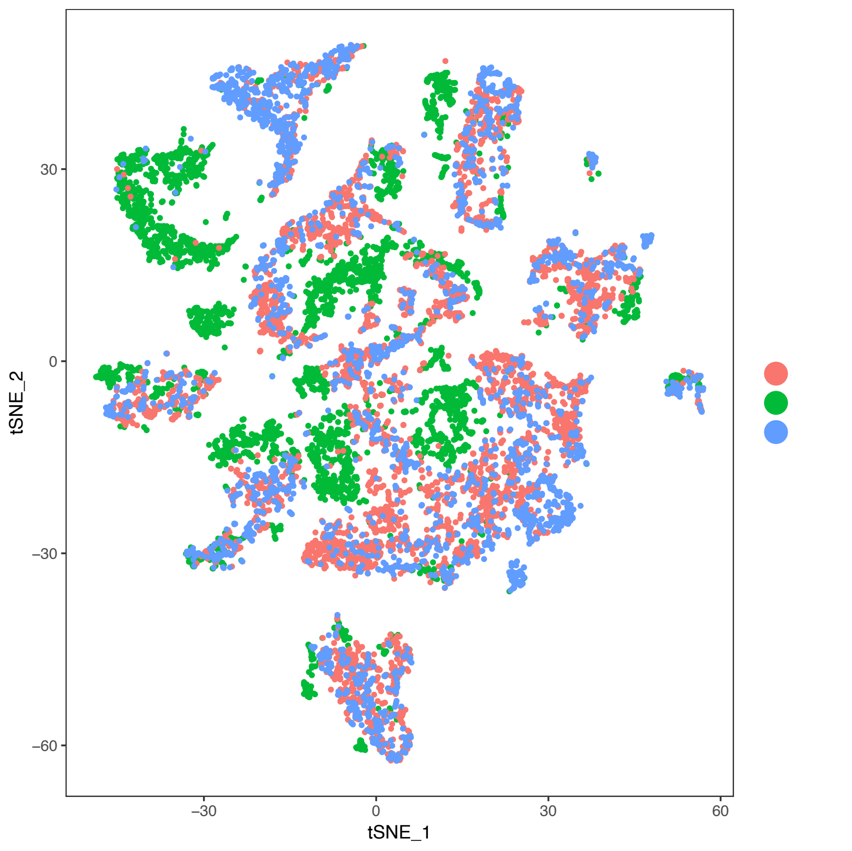
Gene expression association: smaller effect size



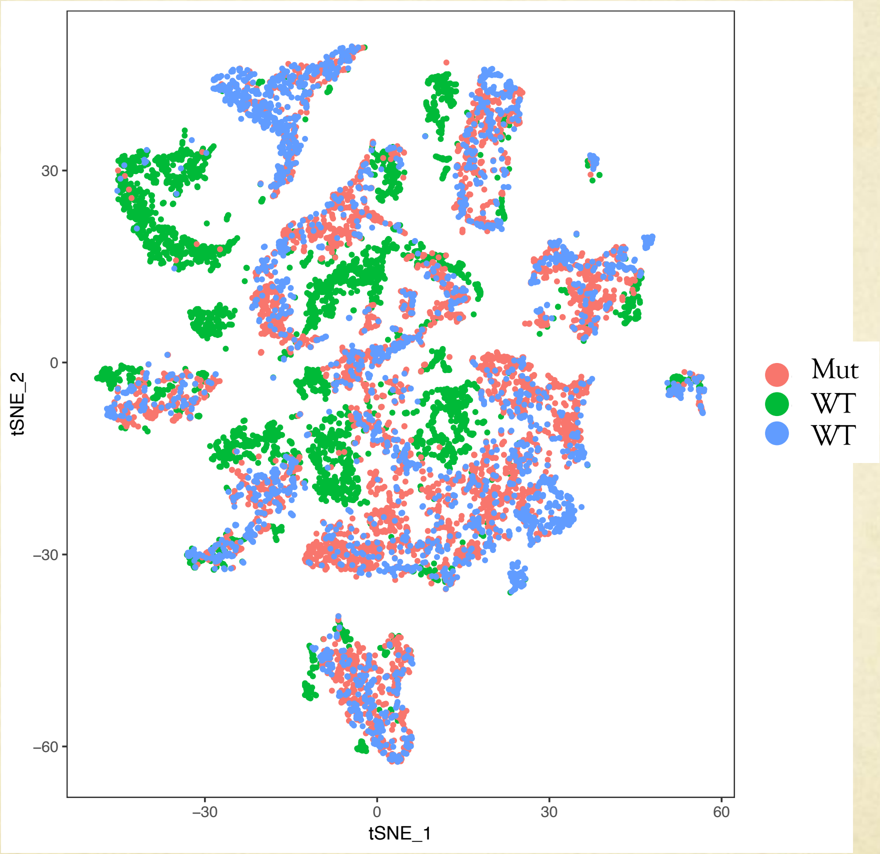
Longitudinal data design



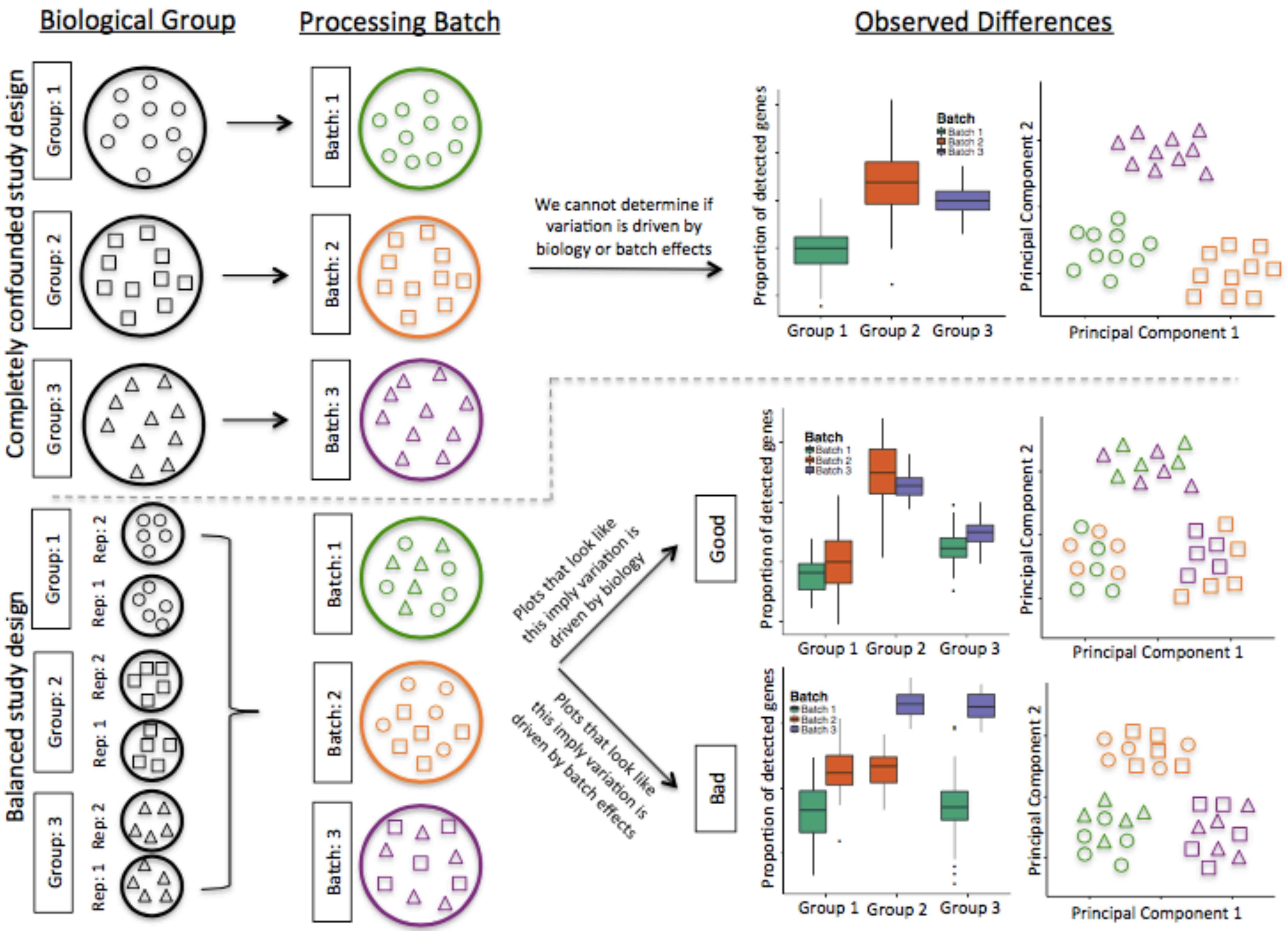
scRNA-seq data for 3 conditions



scRNA-seq data for 3 conditions



The Problem of Confounding Biological Variation and Batch Effects

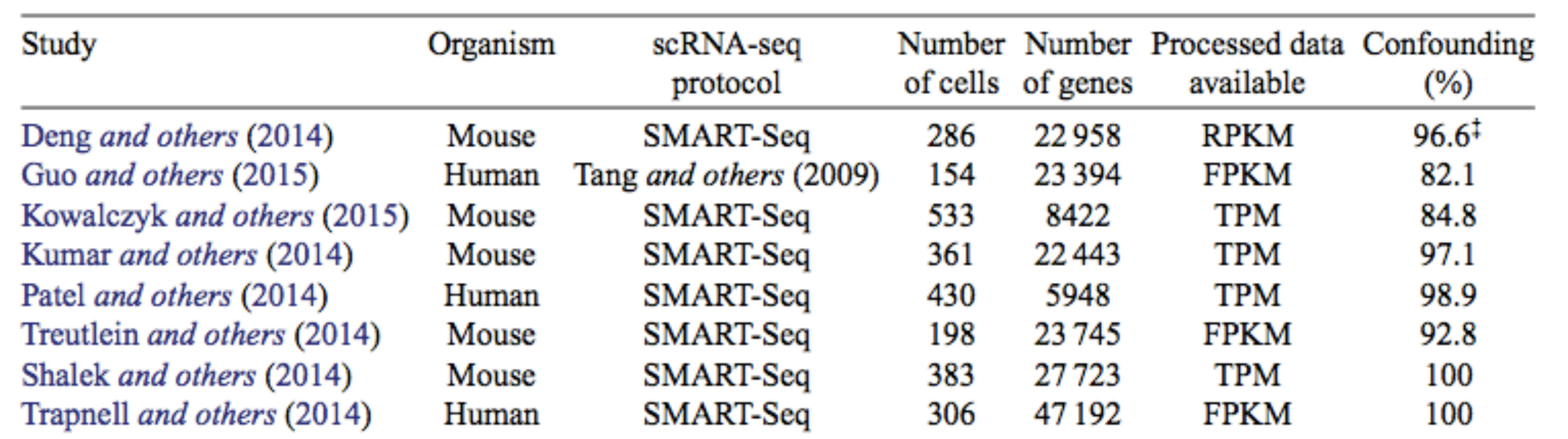


We cannot determine if variation is driven by biology or batch effects

Plots that look like this imply variation is driven by biology

Plots that look like this imply variation is driven by batch effects

Confounding in scRNA-seq data is a big problem



Study	Organism	scRNA-seq protocol	Number of cells	Number of genes	Processed data available	Confounding (%)
Deng *and others* (2014)	Mouse	SMART-Seq	286	22 958	RPKM	96.6‡
Guo *and others* (2015)	Human	Tang *and others* (2009)	154	23 394	FPKM	82.1
Kowalczyk *and others* (2015)	Mouse	SMART-Seq	533	8422	TPM	84.8
Kumar *and others* (2014)	Mouse	SMART-Seq	361	22 443	TPM	97.1
Patel *and others* (2014)	Human	SMART-Seq	430	5948	TPM	98.9
Treutlein *and others* (2014)	Mouse	SMART-Seq	198	23 745	FPKM	92.8
Shalek *and others* (2014)	Mouse	SMART-Seq	383	27 723	TPM	100
Trapnell *and others* (2014)	Human	SMART-Seq	306	47 192	FPKM	100

Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. Preprint available from:https://doi.org/10.1093/biostatistics/kxx053 (2017).

Seven pillars of statistical wisdom

- Aggregation
- Information
- Inter-comparison
- Likelihood
- Regression
- Residuals

https://commons.wikimedia.org/wiki/File:Seven_Pillars_2008_e5.jpg

- Experimental design

Stephen M. Stigler. 2016. Seven Pillars of Statistical Wisdom. Harvard University Press

Please give us feedback

- https://www.surveymonkey.com/r/RRTZPTC
- ~3 min