

Introduction to Modern Statistics

Wenbin Guo
Bioinformatics, UCLA
wbguo@ucla.edu
2025 Winter

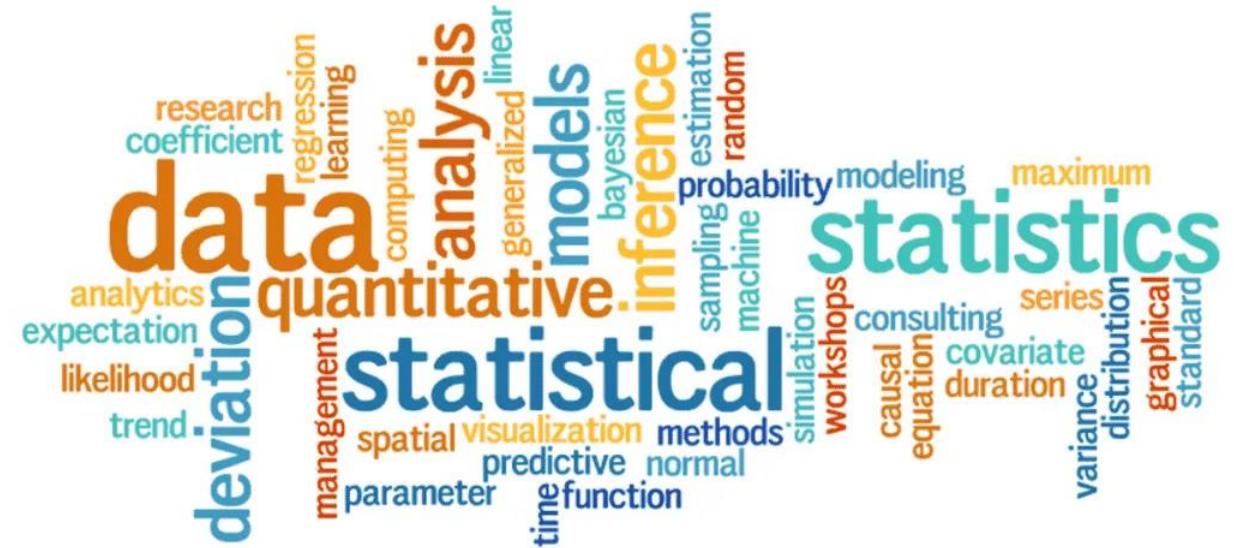
Notation of the slides

- Code or Pseudo-Code chunk starts with "➤", e.g.
➤ print("Hello world!")
- Link is underlined
- Important terminology is in **bold** font
- Practice comes with



Agenda

- Day 1: Probability and Statistics basics
 - Uncertainty; Probability; Distribution
 - Descriptive statistics
- Day 2: Inference
 - Hypothesis testing and p -values
 - Permutation test and bootstrap
 - False discovery rate control
- Day 3: Modeling
 - Regression techniques
 - Model selection



Day 3: Modeling

Wenbin Guo
Bioinformatics IDP, UCLA
wbguo@ucla.edu
2025 Winter

Overview

Time

- 2-hour workshop (45min + 45min + practice/Q&A)

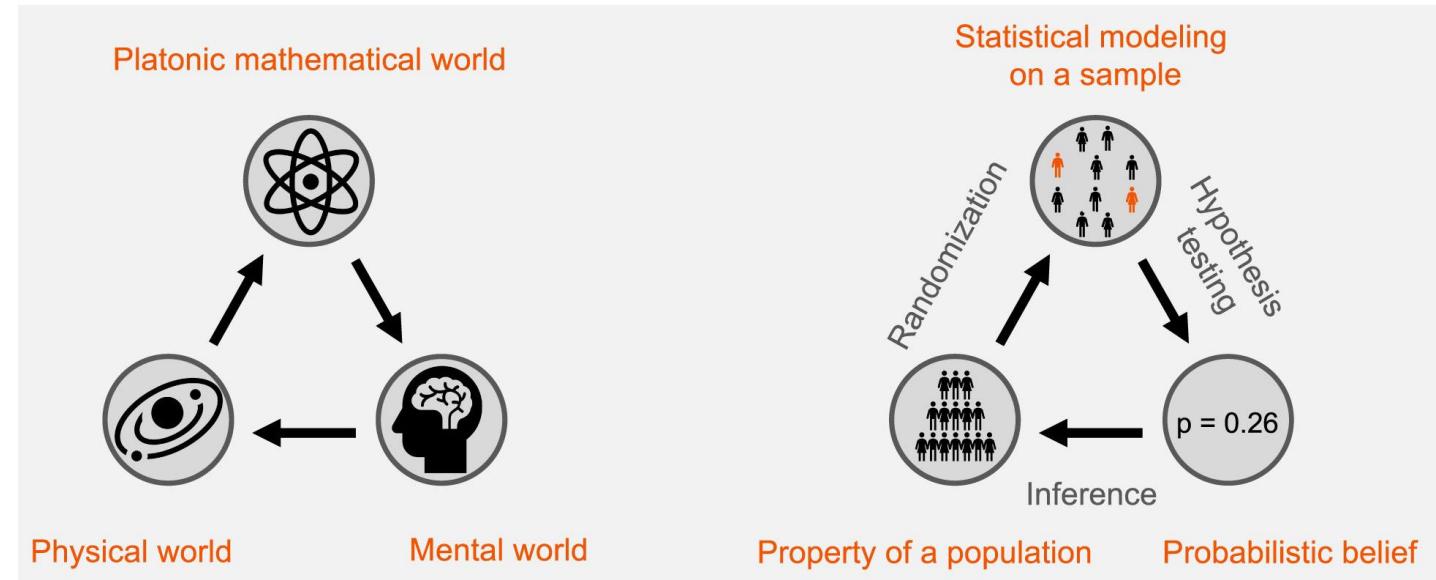
Topics

- ❑ Likelihood and Maximum likelihood estimate
- ❑ Regression techniques
 - ❑ Linear
 - ❑ Logistic
 - ❑ Local
 - ❑ Penalized
- ❑ Model selection
- ❑ Statistical fallacy

Summary – Day1&2

Introduction to probability and statistics

- ❑ Uncertainty
- ❑ Probability
- ❑ Distributions
- ❑ Descriptive statistics
- ❑ Inferential statistics



Summary – Day1&2

Foundations of statistics

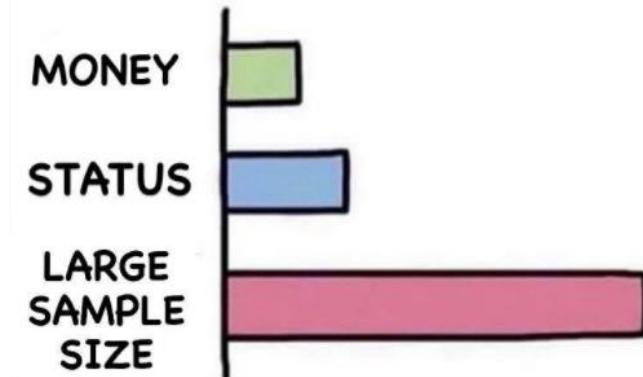
- ❑ Population & Samples
- ❑ Law of Large Numbers (LLN)
- ❑ Central Limit Theorem (CLT)



Statsystem
20h ·

...

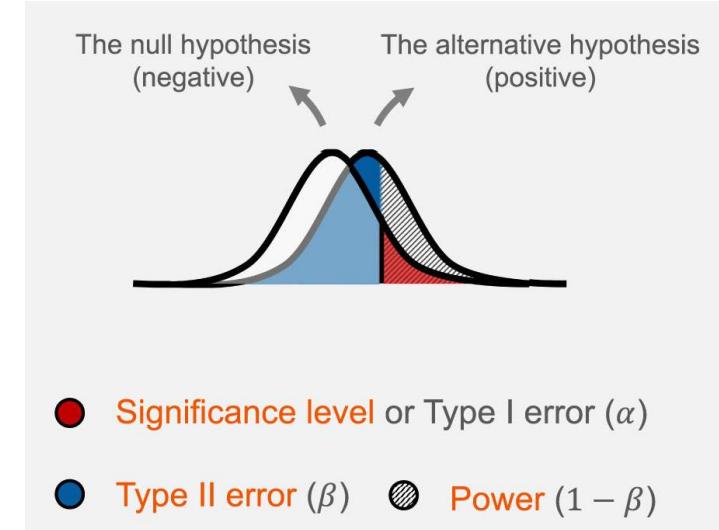
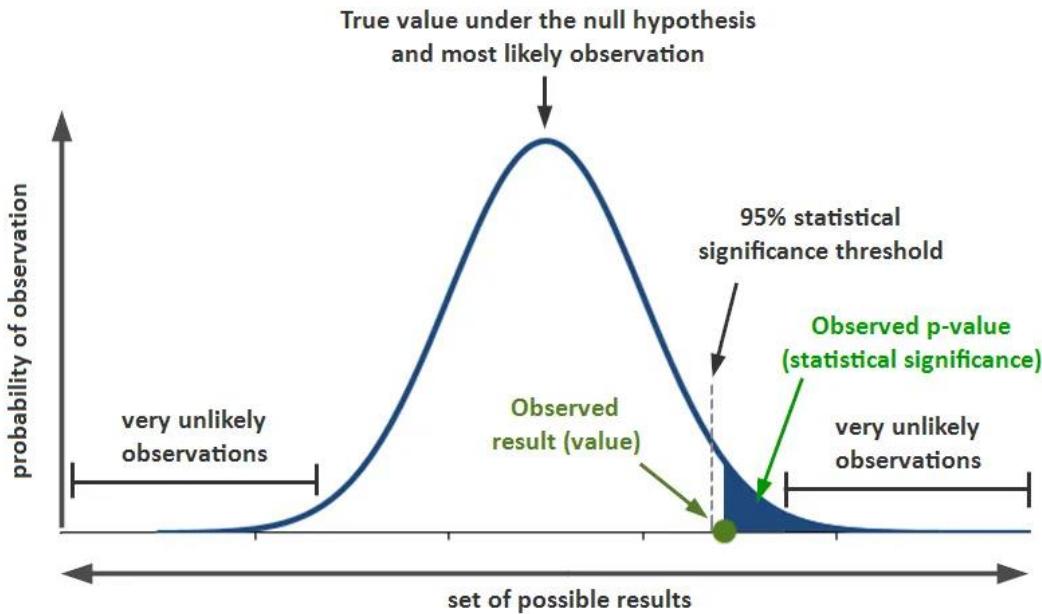
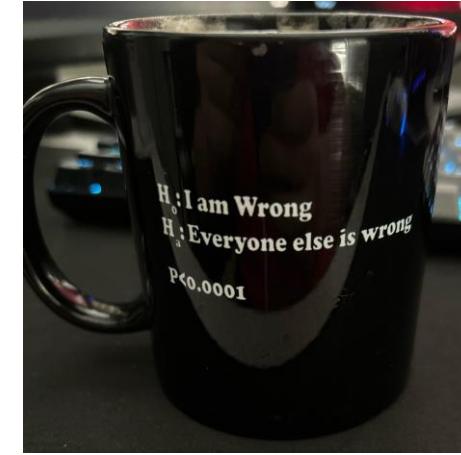
WHAT GIVES PEOPLE
FEELINGS OF POWER



Summary – Day1&2

Hypothesis testing

- ❑ Statistical tests
- ❑ p -value
- ❑ Decision errors



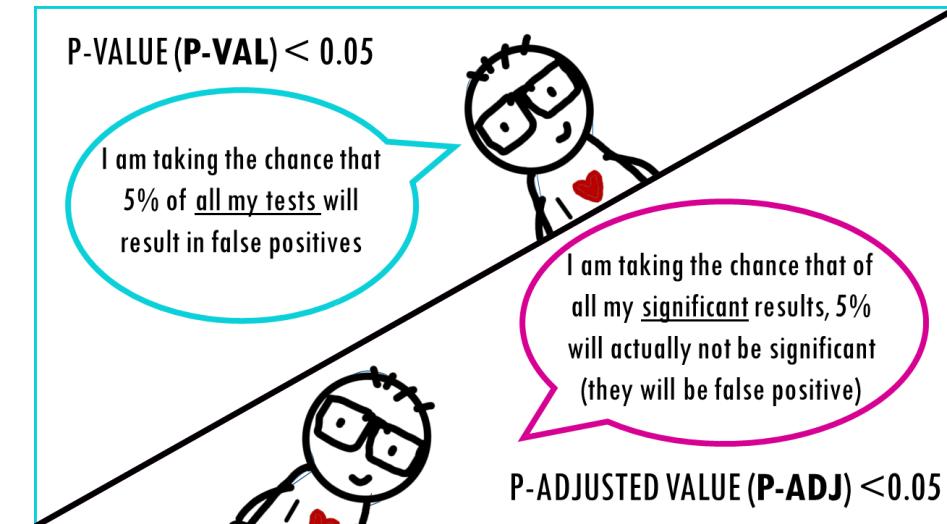
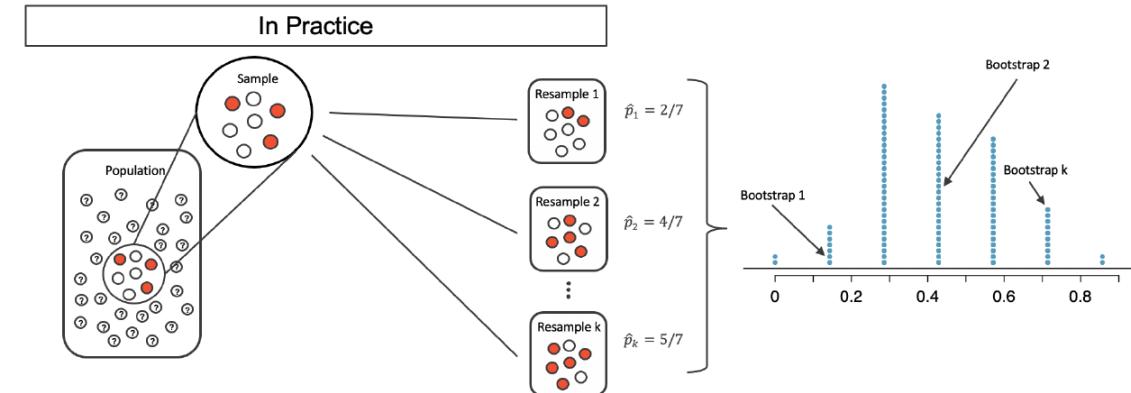
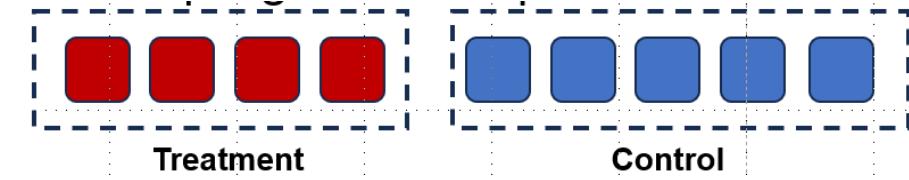
Summary – Day1&2

Computation aided inference

- ❑ Permutation test
- ❑ Bootstrap

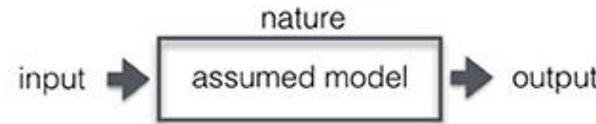
Multiple test correction

- ❑ Bonferroni correction
- ❑ Benjamini-Hochberg

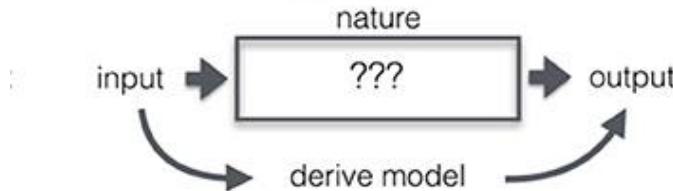


Inference vs. prediction: different focuses

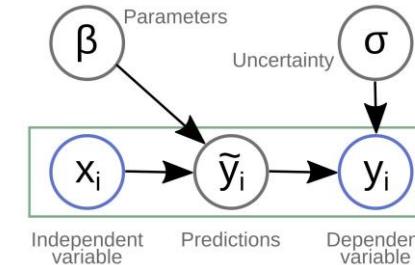
Inference: Understanding relationships, estimating parameters, and testing hypotheses.



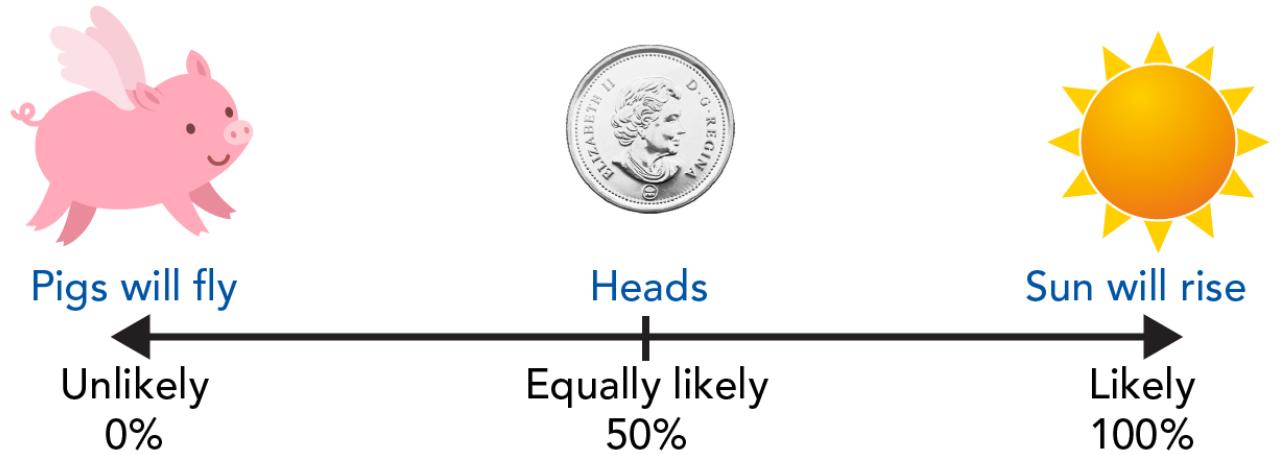
Prediction: Accurately forecasting or predicting outcomes for new data.



In linear model



	Inference	Prediction
Overview	<p>A diagram showing the relationship between parameters β and σ, and variables x_i, \tilde{y}_i, and y_i. The \tilde{y}_i node is highlighted with a red circle. The label "Inferences" is written above the β node. The label "Predictions" is written below the \tilde{y}_i node.</p>	<p>A diagram showing the relationship between parameters β and σ, and variables x_i, \tilde{y}_i, and y_i. The \tilde{y}_i node is highlighted with a large red circle. The label "Predictions" is written below the \tilde{y}_i node.</p>
Goal	Understanding	Forecasting
Focus	Parameter estimation	Model accuracy
Task	Hypothesis testing	Classification or regression
Evaluation	Statistical significance	Predictive performance (AUC, MSE)



Likelihood

“It is likely that unlikely things should happen” – Aristotle

Consider a toy example

Flip a coin 10 times and record the outcome D



All heads

What's the probability of observing such data when $\theta = 0.5$?

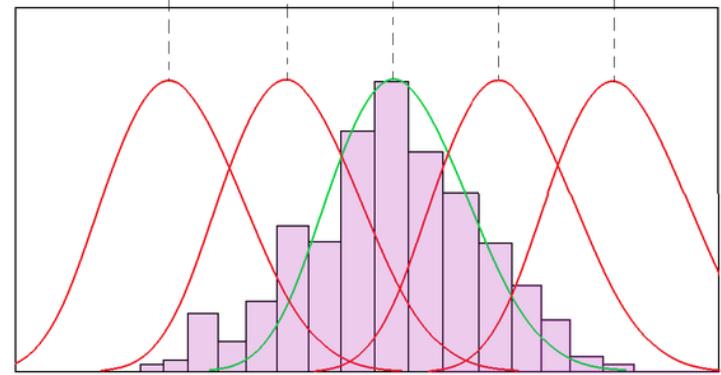
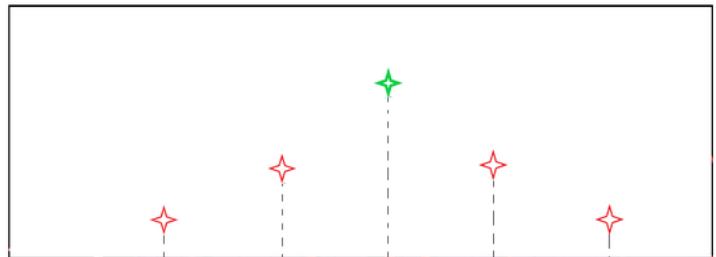
$$P(D | \theta = 0.5) = (\frac{1}{2})^{10} < 0.001$$

So, we encounter such a small probability event just by chance?

What if $\theta \neq 0.5$? (we are questioning the fairness of the coin)



Maximum likelihood estimate plot



Multiple PDFs over the random sample histogram plot

Maximum likelihood estimate (MLE)

With model parameters θ and observed data D ,

Define the likelihood function

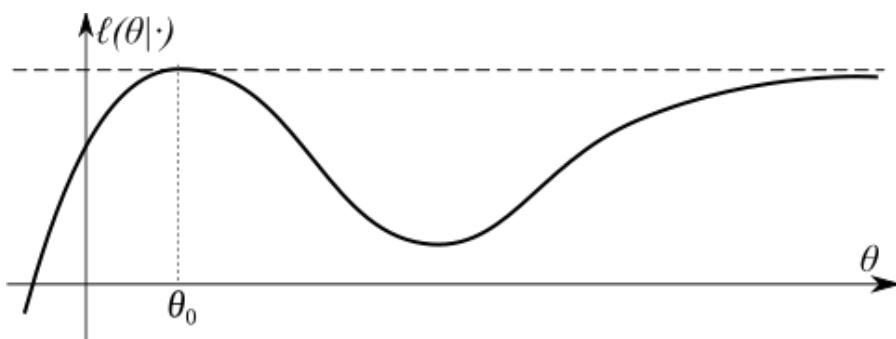
$$L(\theta) = L(\theta; D) = f(D | \theta)$$

The goal of MLE is to find $\hat{\theta}$ such that

$$\hat{\theta} = \operatorname{argmax} L(\theta; D)$$



Ronald Fisher
(1890- 1962)



a genius who almost single-handedly created the foundations for modern statistical science

Likelihood for Frequentist & Bayesian

Probability
(mathematics)

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Everyone uses Bayes' formula when the prior $P(H)$ is known.

Bayesian path

Statistics
(art)

$$P_{\text{Posterior}}(H|D) = \frac{P(D|H)P_{\text{prior}}(H)}{P(D)}$$

Bayesians require a prior, so they develop one from the best information they have.

Frequentist path

$$\text{Likelihood } L(H; D) = P(D|H)$$

Without a known prior frequentists draw inferences from just the likelihood function.

Likelihood-Ratio test (LRT)

10.21 Definition. Consider testing

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \notin \Theta_0.$$

The likelihood ratio statistic is

$$\lambda = 2 \log \left(\frac{\sup_{\theta \in \Theta} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)} \right) = 2 \log \left(\frac{\mathcal{L}(\hat{\theta})}{\mathcal{L}(\hat{\theta}_0)} \right)$$

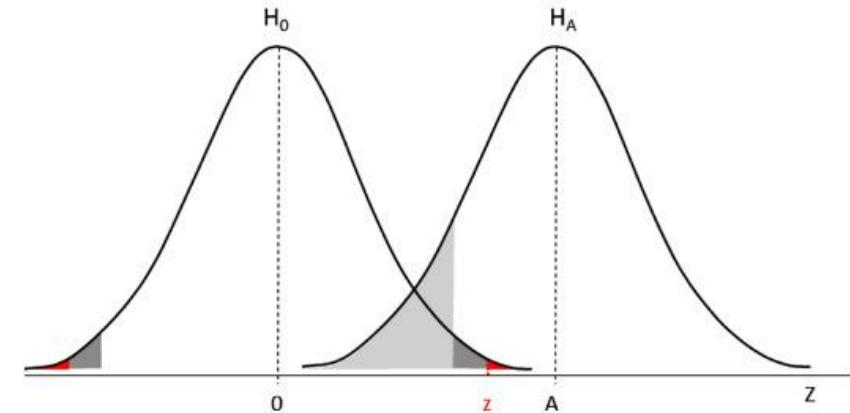
where $\hat{\theta}$ is the MLE and $\hat{\theta}_0$ is the MLE when θ is restricted to lie in Θ_0 .

When H_0 is true, λ follows a chi-square distribution

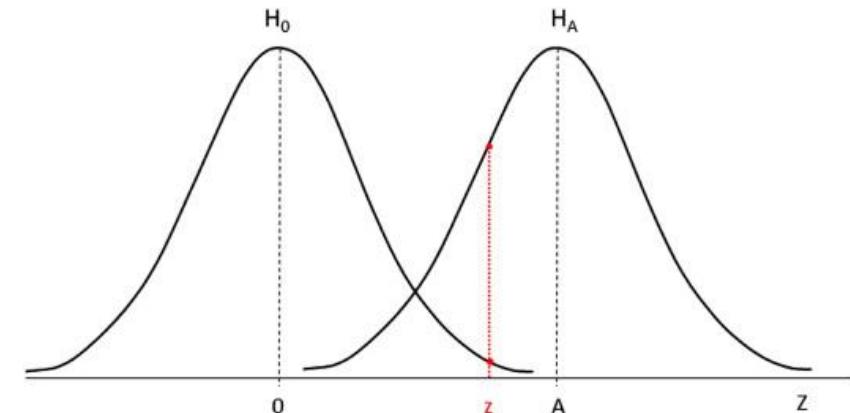
$$\lambda \sim \chi_d^2$$

Where d is the dimension difference between H_0 and H_1

a) Significance test and p -value



b) Likelihood ratio



([Nyeman-Pearson Lemma](#) further proved it's the most powerful test)

Let's do some practice!

➤ git clone <https://github.com/wbvguo/qcbio-Intro2ModernStats.git>



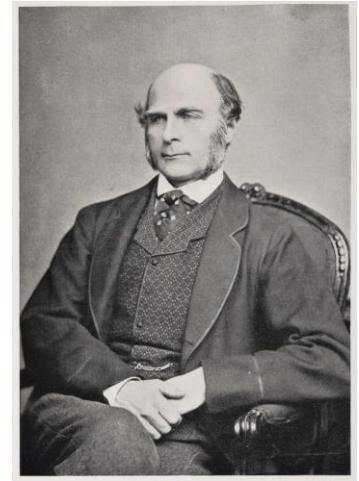
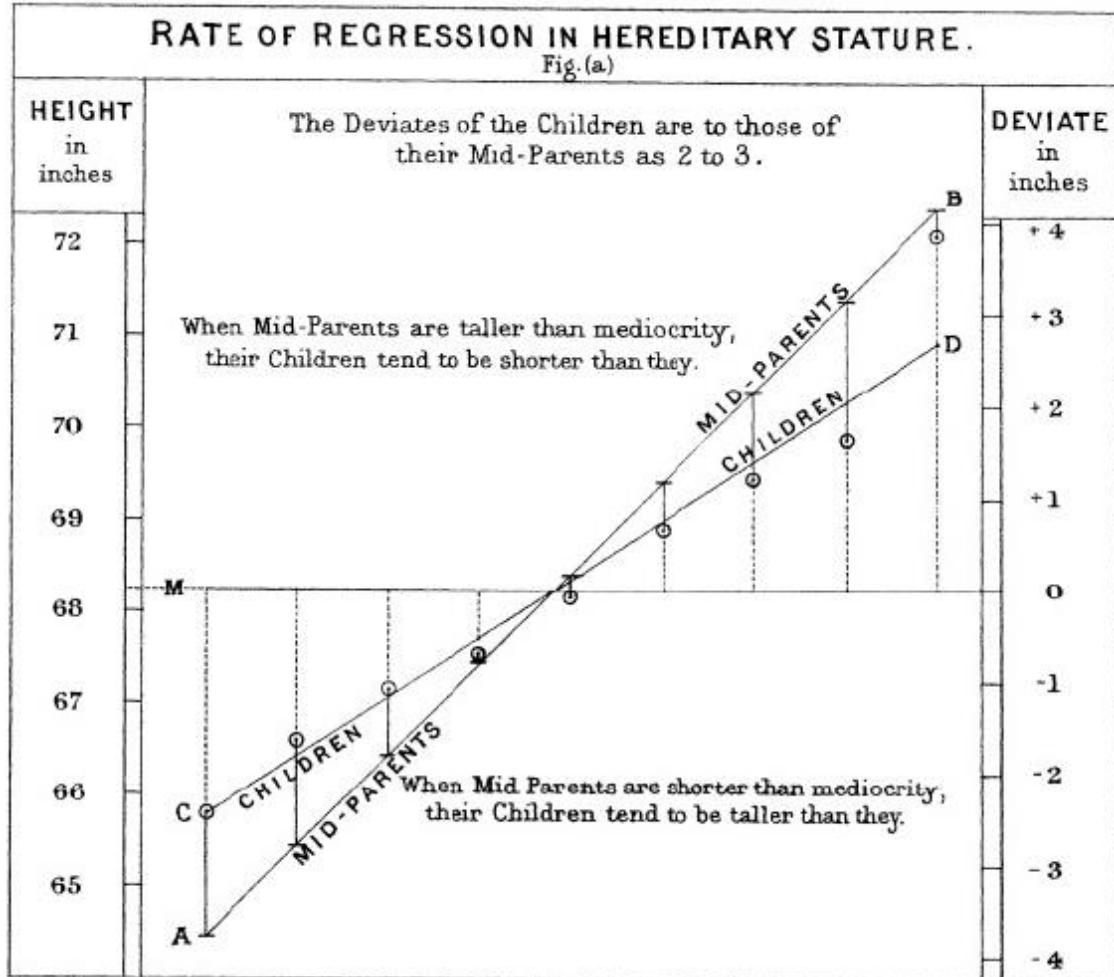
Regression

Chaos finds order in the mean.



The history of regression

"the average regression of the offspring is a constant fraction of their respective mid-parental deviations"



Francis Galton
(1822-1911)

Regression towards the mean

Simple linear regression

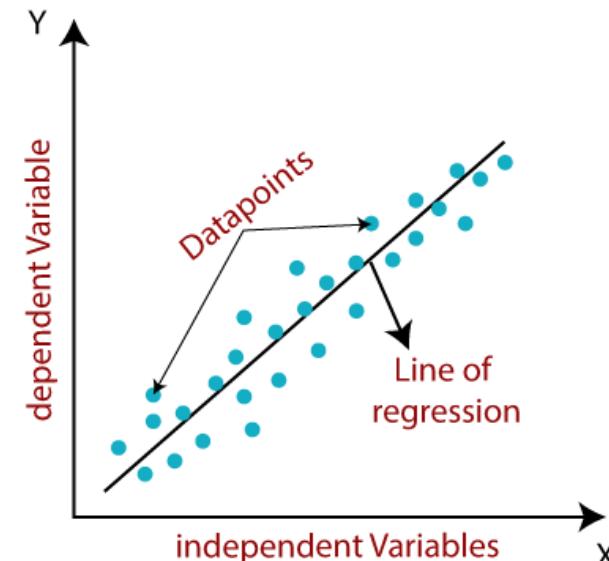
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Annotations for the equation:

- Dependent Variable → Y_i
- Population Y intercept → β_0
- Population Slope Coefficient → β_1
- Independent Variable → X_i
- Random Error term → ε_i
- Linear component → $\beta_0 + \beta_1 X_i$
- Random Error component → ε_i

How can we estimate β_0 and β_1 ?

- Ordinary Least Square (OLS)
- Maximum Likelihood Estimate (MLE)



OLS derivation

Rational: minimize the fitting error (loss function)

Define the loss

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 .$$

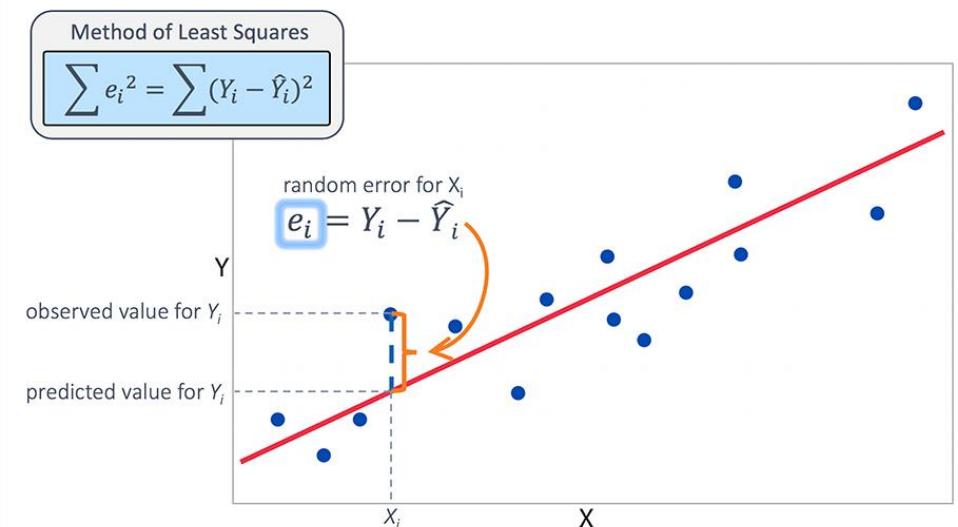
Take the derivative and set to 0

$$0 = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$0 = -2 \sum_{i=1}^n (x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2)$$

Solve the equation

$$\left. \begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned} \right\}$$



$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad \text{where } \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \in \mathbb{R}^{n \times 2}$$

MLE derivation

$$y = \beta_0 + \beta_1 x + \epsilon$$

$$\epsilon \sim N(0, \delta^2)$$

Rational: maximize the likelihood

$$\begin{aligned} p(y|\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^n p(y_i|\beta_0, \beta_1, \sigma^2) \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \cdot \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \right] \end{aligned}$$

Log-Likelihood:

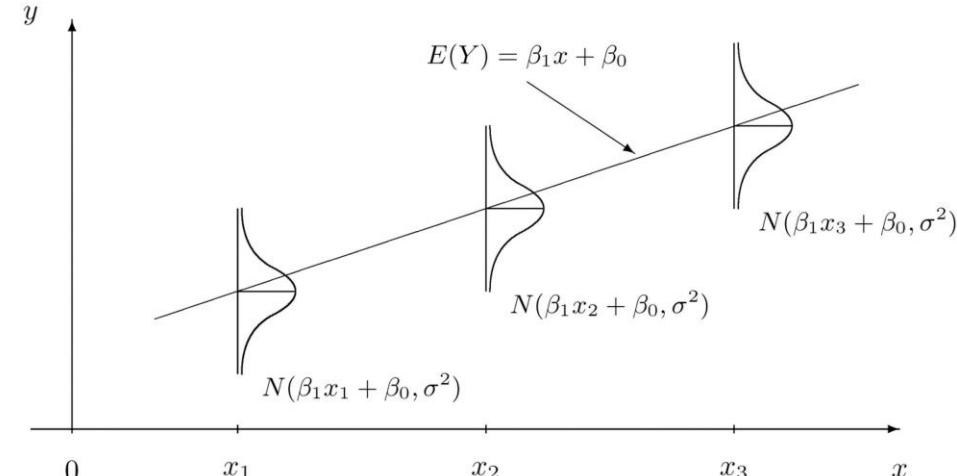
$$\begin{aligned} \text{LL}(\beta_0, \beta_1, \sigma^2) &= \log p(y|\beta_0, \beta_1, \sigma^2) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

Take the derivative w.r.t each parameter and set to 0

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$



OLS and MLE are equivalent in this setting

Inference

$$y = \beta_0 + \beta_1 x + \epsilon$$
$$\epsilon \sim N(0, \delta^2)$$



Since $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$, we can derive $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$

Now, we have estimates $\hat{\beta}$ from the data, so we can test them

Hypothesis testing:

$$H_0 : \beta_j = 0; \quad H_1 : \beta_j \neq 0$$

The t -statistic:

$$T = \frac{\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{j^{\text{th}} \text{ diagonal entry of } (\mathbf{X}^T \mathbf{X})^{-1}}}$$

For the denominator, we use the *plug-in estimator* (replacing the true value by the estimator).

Prediction

After we obtain the estimator, we can use it to predict for new values

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Residual:

$$e_i = y_i - \hat{y}_i$$

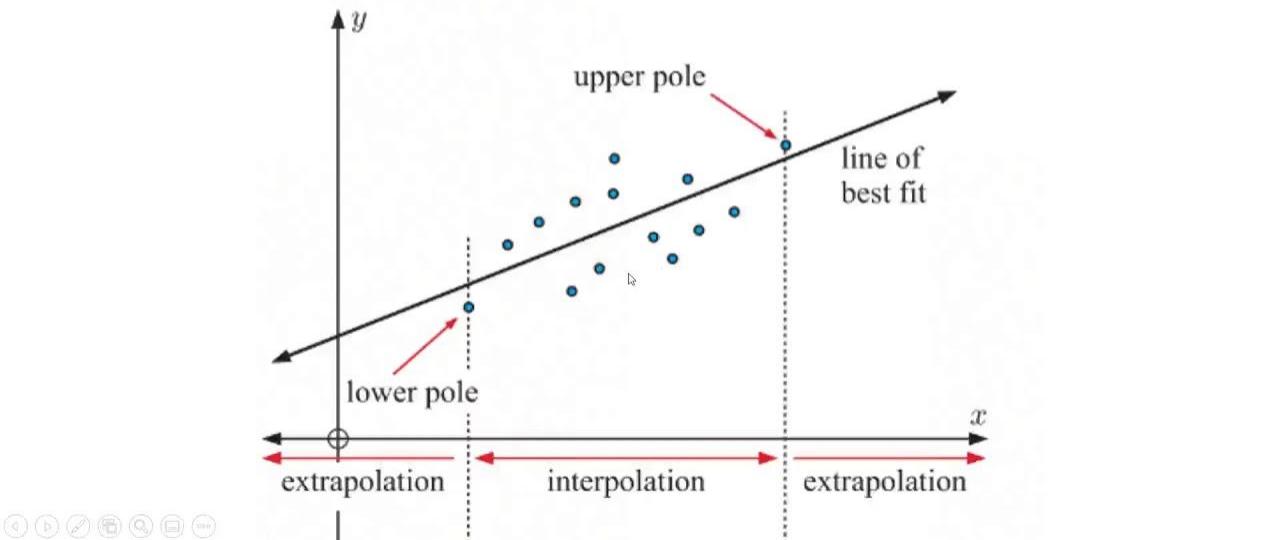
Coefficient of determination:

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = r^2$$

Interpolation / Extrapolation

In between the points = reliable

Outside the points = unreliable



$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

The sample correlation

Multiple regression

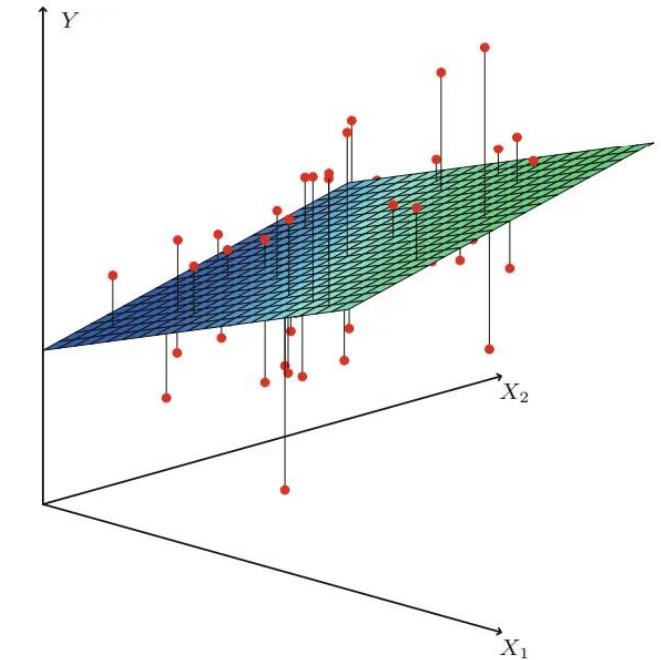
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$
$$\epsilon \sim N(0, \delta^2)$$

Parameter estimation:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Inference

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$$



Consistent with the simple linear regression

Question: what would happen when $p \gg n$?

Logistic regression

$$\text{logit}(p_i) = \beta_0 + \sum_{j=1}^k \beta_j x_{ij}$$

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

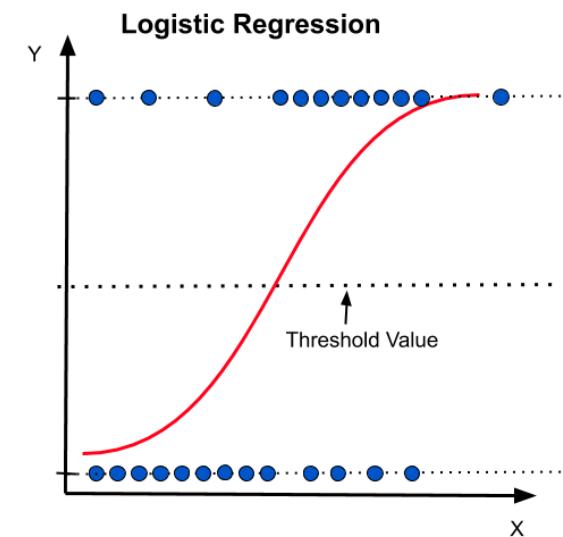
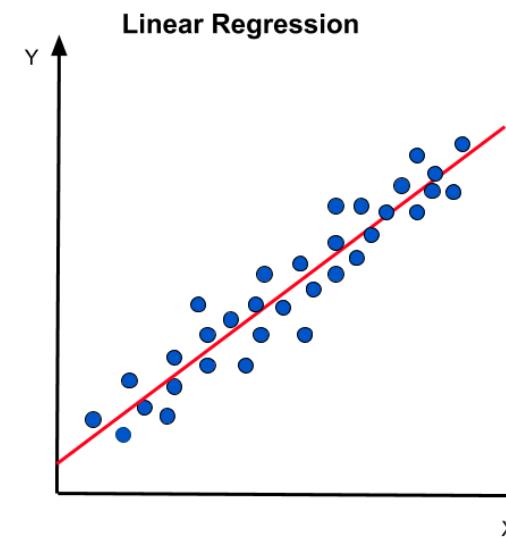
Logit transformation

$$Y_i | X_i = x_i \sim \text{Bernoulli}(p_i)$$

Likelihood:

$$\mathcal{L}(\beta) = \prod_{i=1}^n p_i(\beta)^{Y_i} (1 - p_i(\beta))^{1-Y_i}$$

MLE can be obtained using numerical approach



Local regression

LOcal regrESSion (LOESS): non-parametric Model

$$Y_i = \mu(x_i) + \epsilon_i$$

$$\mu(x_i) \approx \beta_0 + \beta_1(x_i - x) + \dots + \beta_p(x_i - x)^p$$

Loss function

$$\sum_{i=1}^n w_i(x)(Y_i - \beta_0 - \beta_1(x_i - x) - \dots - \beta_p(x_i - x)^p)^2$$

Where $w_i(x) = W\left(\frac{x_i - x}{h}\right)$ is a weight function (kernel)

Choice 1: Type of model

- Linear regression
- Degree 2 polynomial
- Degree 3 polynomial

Choice 2: Weighting scheme

- Normal density
- Other schemes (called kernels)

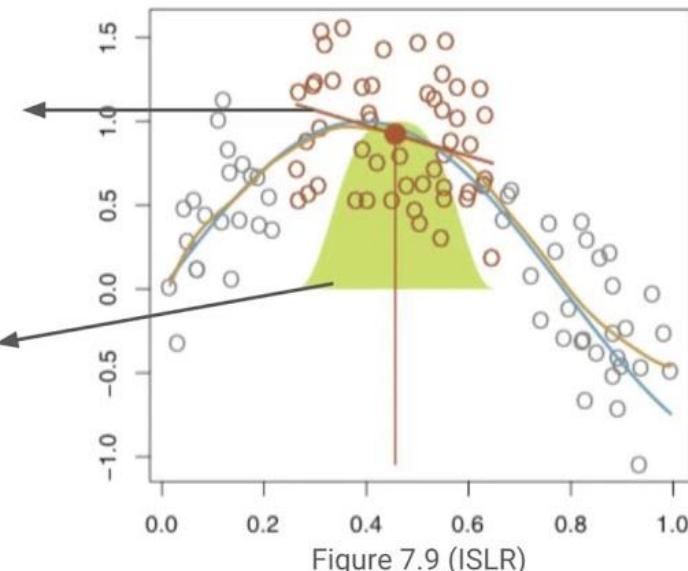


Figure 7.9 (ISLR)

Overfitting: Von Neumann's elephant

"I remember my friend Johnny von Neumann used to say, with four parameters I can fit an elephant, and with five I can make him wiggle his trunk." - Enrico Fermi



John von Neumann
(1903-1957)

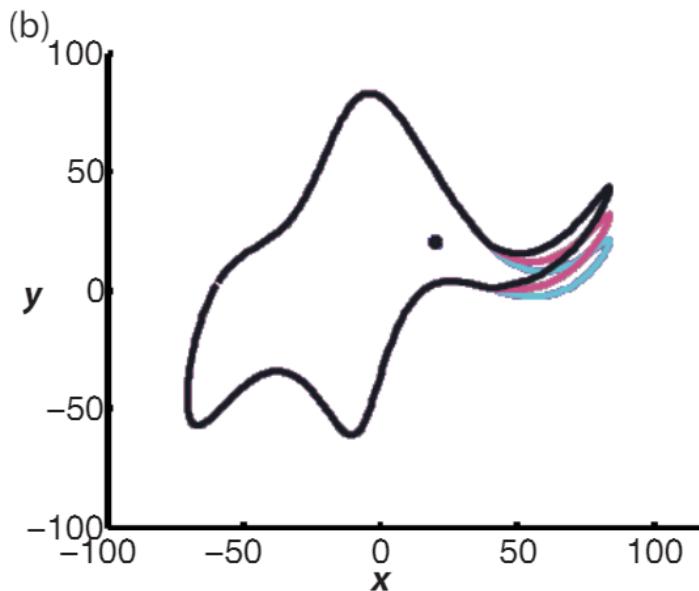


Table I. The five complex parameters p_1, \dots, p_5 that encode the elephant including its wiggling trunk.

Parameter	Real part	Imaginary part
$p_1=50-30i$	$B_1^x=50$	$B_1^y=-30$
$p_2=18+8i$	$B_2^x=18$	$B_2^y=8$
$p_3=12-10i$	$A_3^x=12$	$B_3^y=-10$
$p_4=-14-60i$	$A_4^x=-14$	$A_4^y=-60$
$p_5=40+20i$	Wiggle coeff.=40	$x_{\text{eye}}=y_{\text{eye}}=20$

Penalized regression

- Lasso (L1 penalty)

$$J(\boldsymbol{\theta}) = \text{MSE}(\boldsymbol{\theta}) + \alpha \sum_{i=1}^n |\theta_i|$$

- Ridge (L2 penalty)

$$J(\boldsymbol{\theta}) = \text{MSE}(\boldsymbol{\theta}) + \alpha \frac{1}{2} \sum_{i=1}^n \theta_i^2$$

This forces the learning algorithm to not only fit the data but also keep the model weights as small as possible!

- Elastic-net

$$J(\boldsymbol{\theta}) = \text{MSE}(\boldsymbol{\theta}) + r\alpha \sum_{i=1}^n |\theta_i| + \frac{1-r}{2}\alpha \sum_{i=1}^n \theta_i^2$$

Penalized regression

- Lasso (L1 penalty)

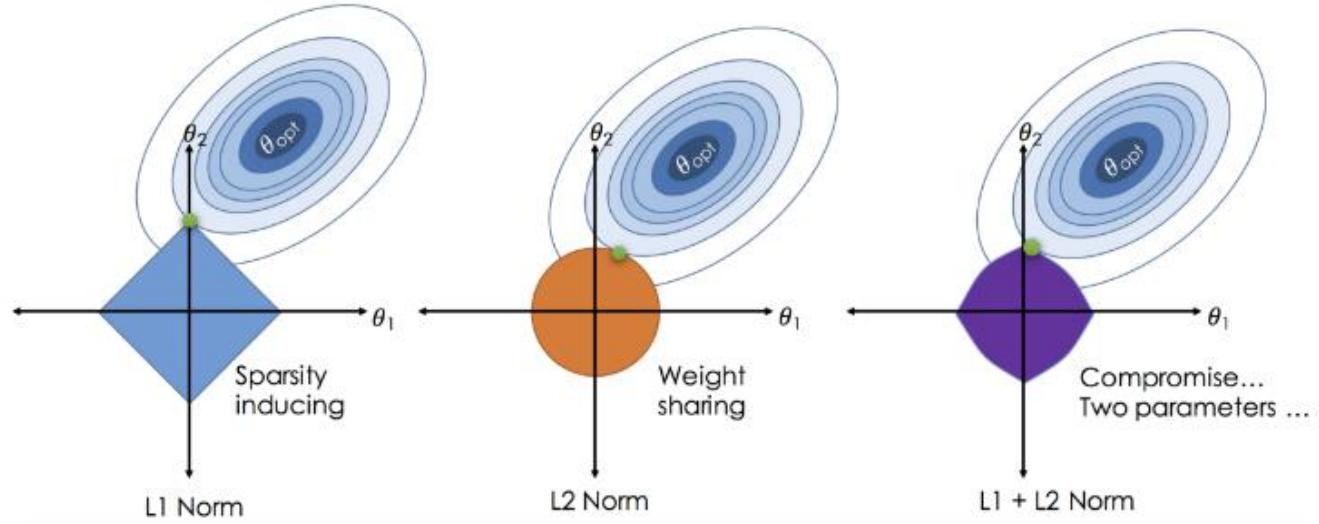
$$J(\boldsymbol{\theta}) = \text{MSE}(\boldsymbol{\theta}) + \alpha \sum_{i=1}^n |\theta_i|$$

- Ridge (L2 penalty)

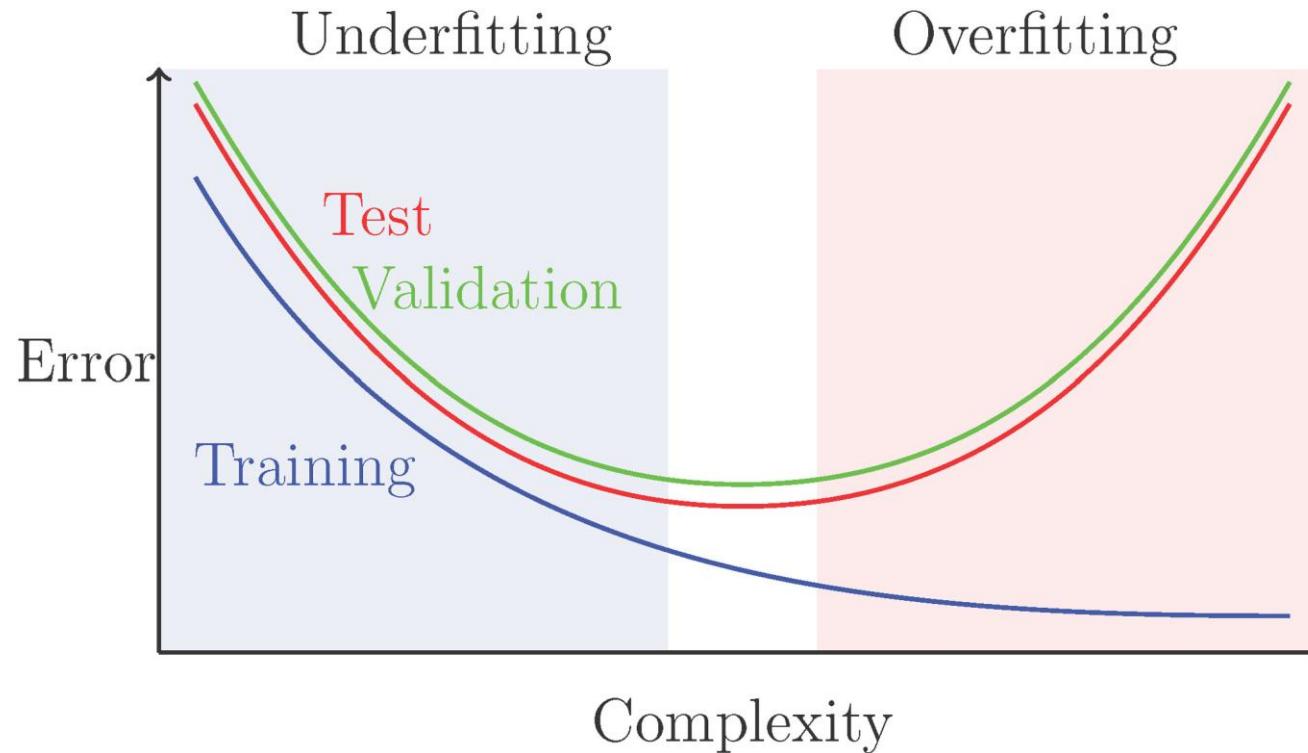
$$J(\boldsymbol{\theta}) = \text{MSE}(\boldsymbol{\theta}) + \alpha \frac{1}{2} \sum_{i=1}^n \theta_i^2$$

- Elastic-net

$$J(\boldsymbol{\theta}) = \text{MSE}(\boldsymbol{\theta}) + r\alpha \sum_{i=1}^n |\theta_i| + \frac{1-r}{2}\alpha \sum_{i=1}^n \theta_i^2$$

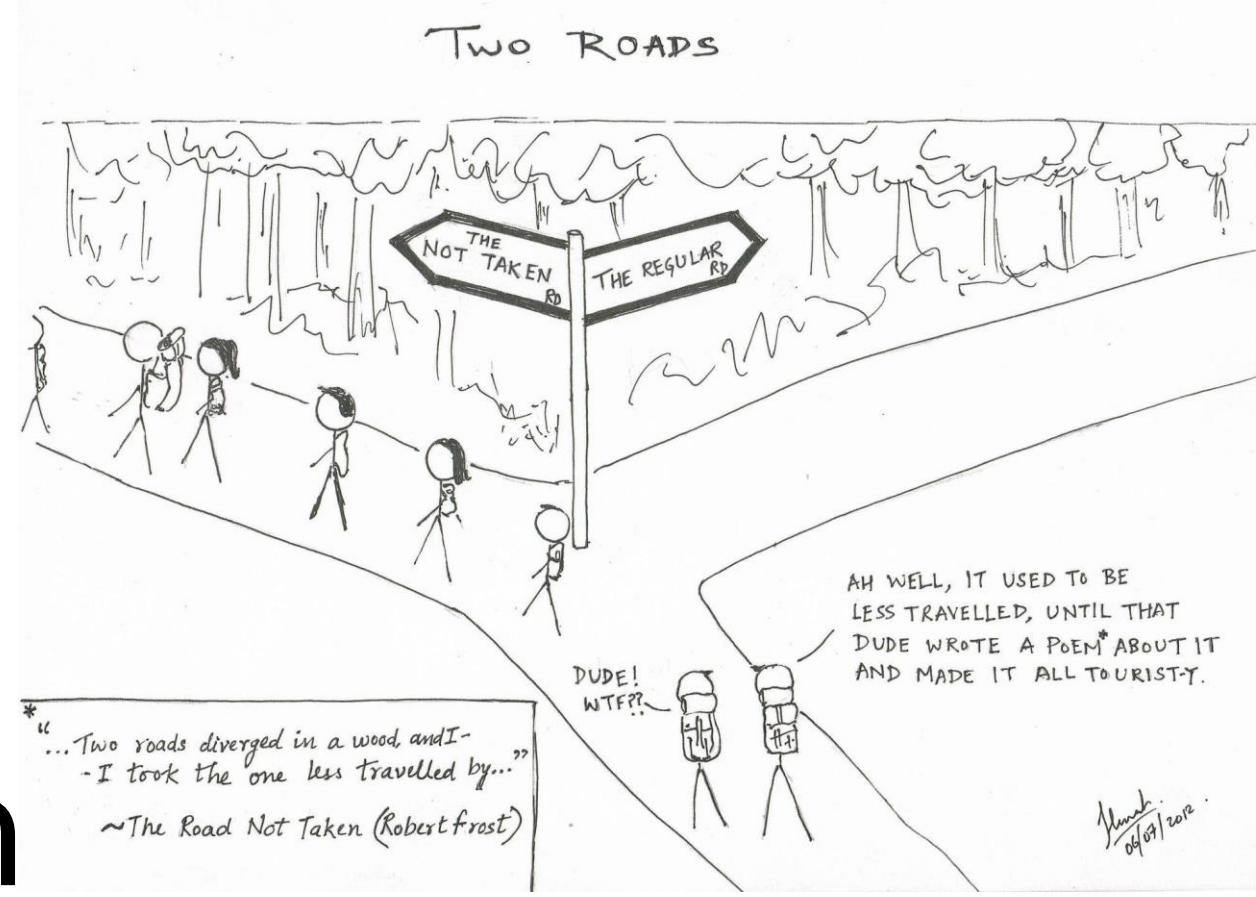


The need for penalization



Model selection

“The choices we make dictate the life we lead.” – William Shakespeare



Model selection in history

In modeling planet movement

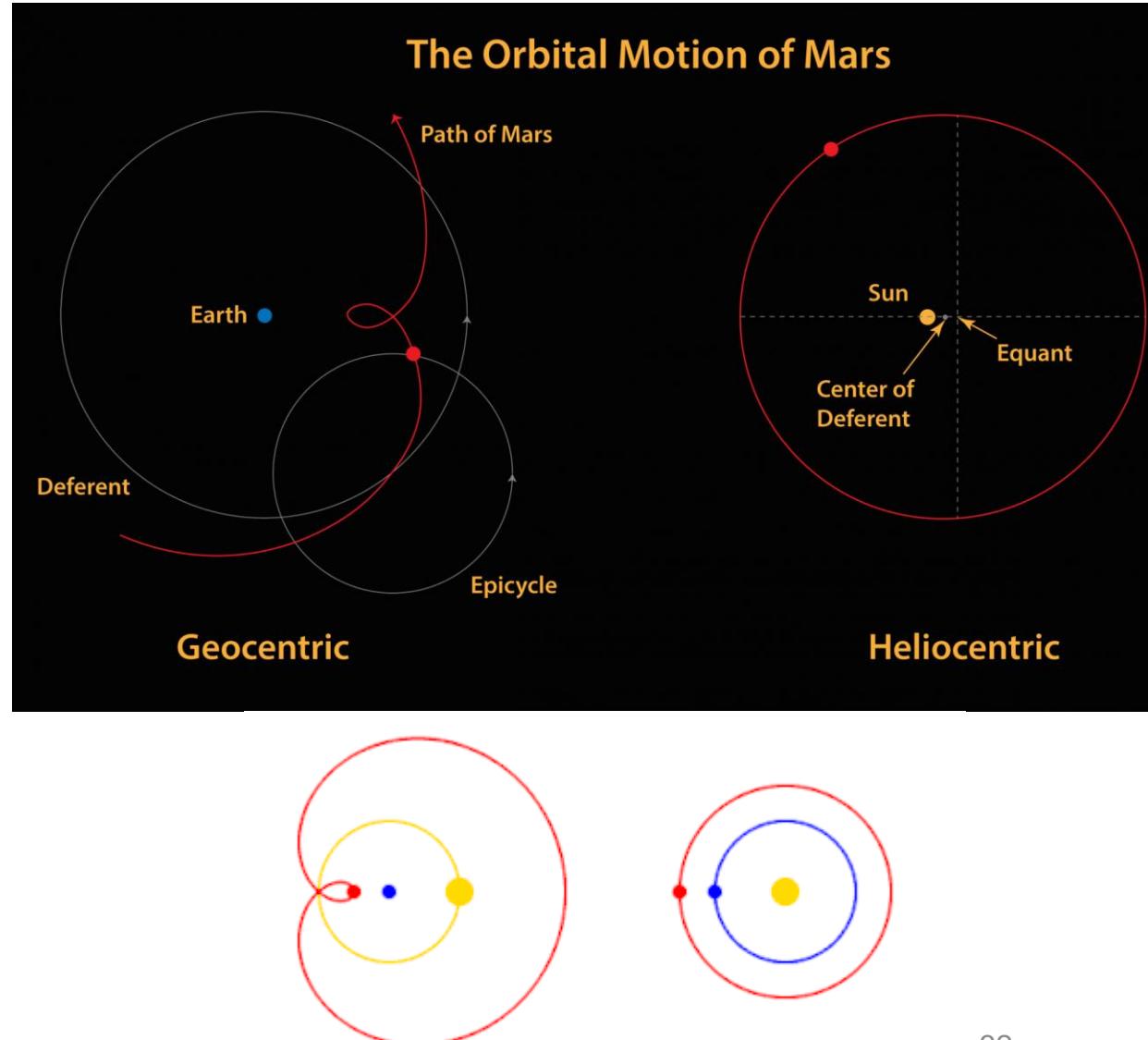
- Apollonius of Perga: epicycle

$$z(t) = \sum_{k=1}^d r_k e^{i\omega_k t}$$

- Johannes Kepler: elliptical orbits

$$r = \frac{a(1 - e^2)}{1 + e \cos \theta}$$

Occam's Razor: favor parsimonious model



Model selection – probabilistic approach

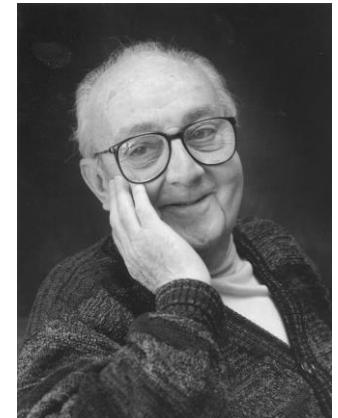
trade-off between the **goodness of fit** of the model and the **simplicity** of the model

- ❑ Akaike information criterion (AIC)

$$AIC = 2k - 2 \ln(\hat{L})$$

- ❑ Bayesian information criterion (BIC)

$$BIC = k \ln(n) - 2 \ln(\hat{L}).$$



George Box
(1919-2013)

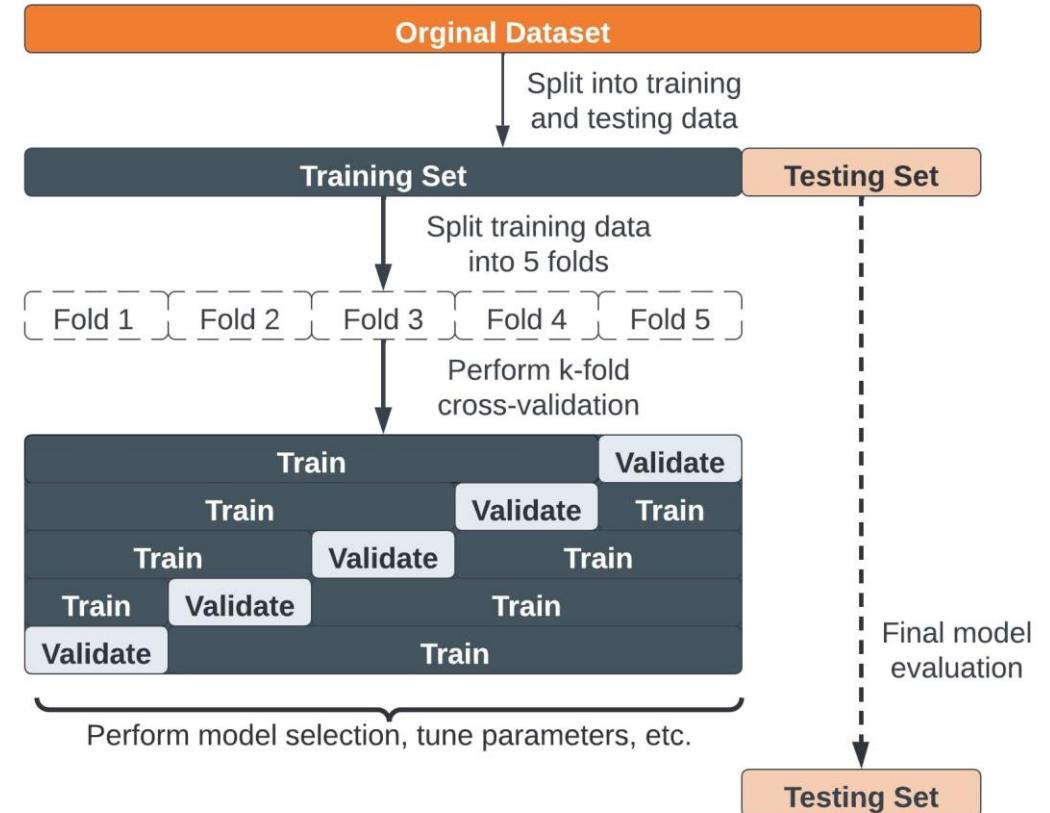
“All models are wrong, but some are useful.”

Model selection – resampling approach

Cross-validation (CV)

Data split

- A **training** set is used to train the machine learning model(s) during development.
- A **validation** set is used to estimate the generalization error of the model created from the training set for the purpose of model selection.
- A **test** set is used to estimate the generalization error of the final model.

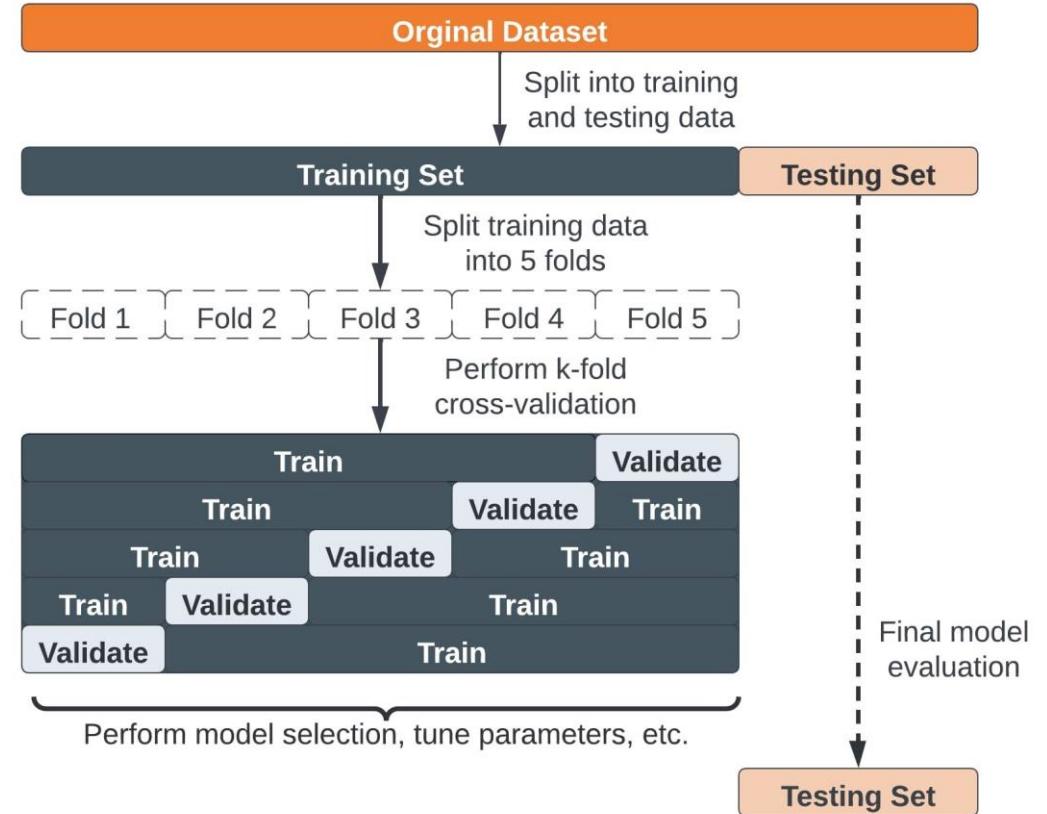


Model selection – resampling approach

Cross-validation (CV)

Procedure

- ❑ Split data into portions.
- ❑ Train model on a subset of the portions.
- ❑ Test model on the remaining subsets of the data.
- ❑ Repeat steps 2-3 until the model has been trained and tested on the entire dataset.
- ❑ Average the model performance across all iterations of testing to get the total model performance.



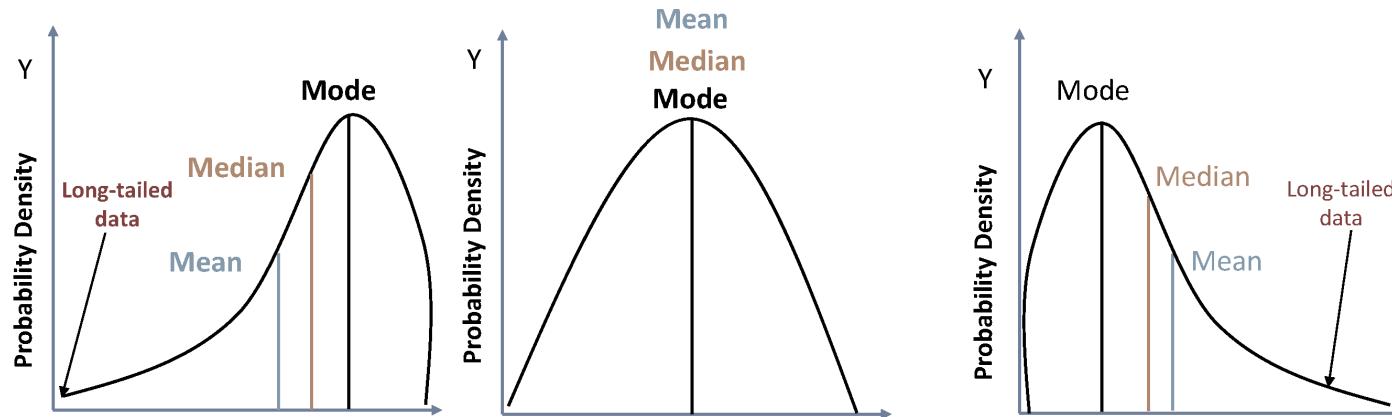
Statistical fallacy

May you have a clear mind and sharp eyes.



Flaw of averages

The average can be a **poor representation** for the samples
(due to skewness, outliers, etc.)



Before learning statistics:

- I am standing next to a guy with citations more than 250,000 😳

After learning statistics:

- Our **average** citations are more than 125,000! 😎



Photo with Dr. [Heng Li](#)
RECOMB 2024@MIT

Stability isn't always good

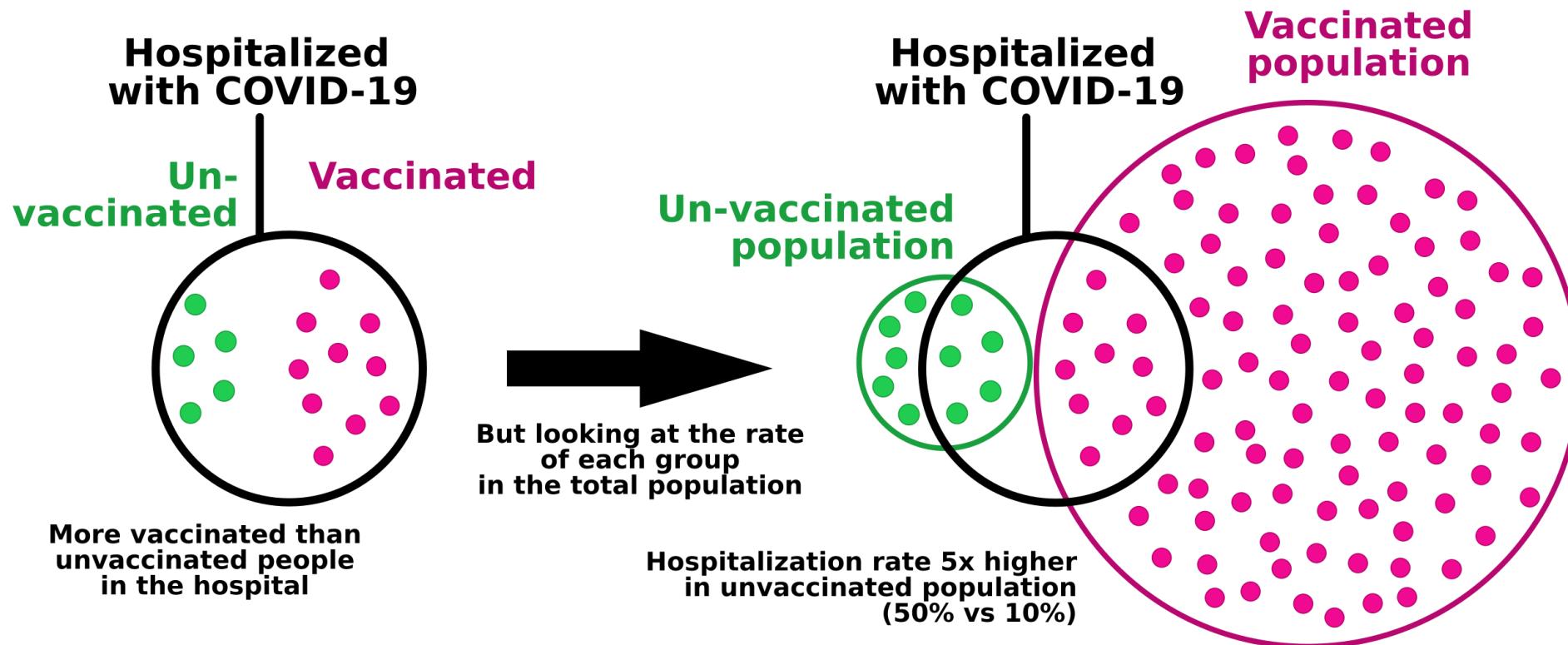
Stability only reflects **variation**, not **mean**

His condition is very stable

- He can be just fine and recovering smoothly
- He can be seriously ill and in ICU every day

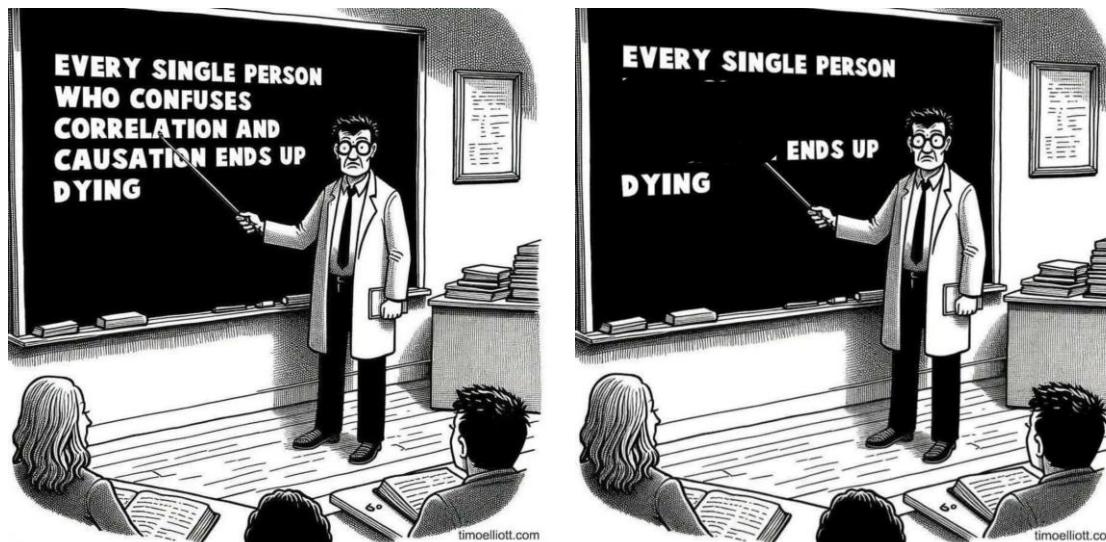
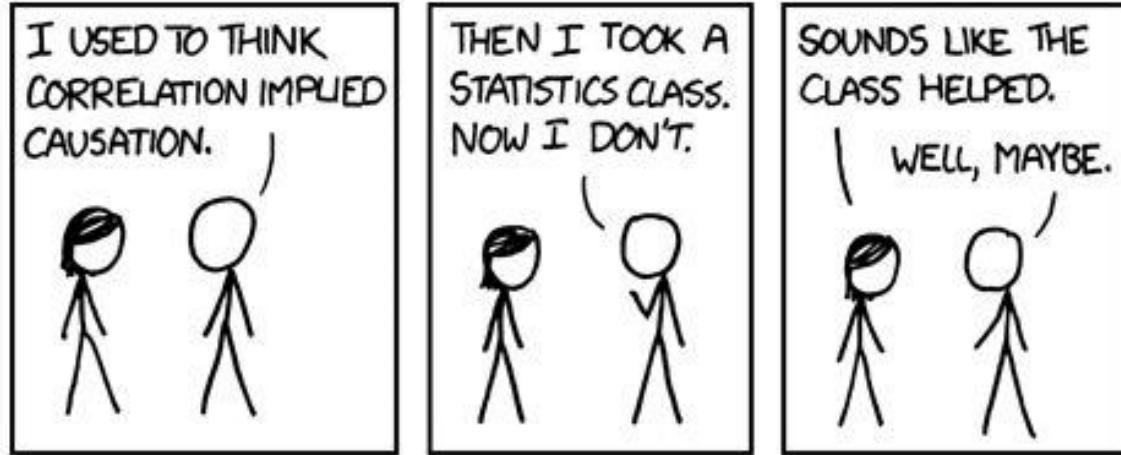


Base rate fallacy

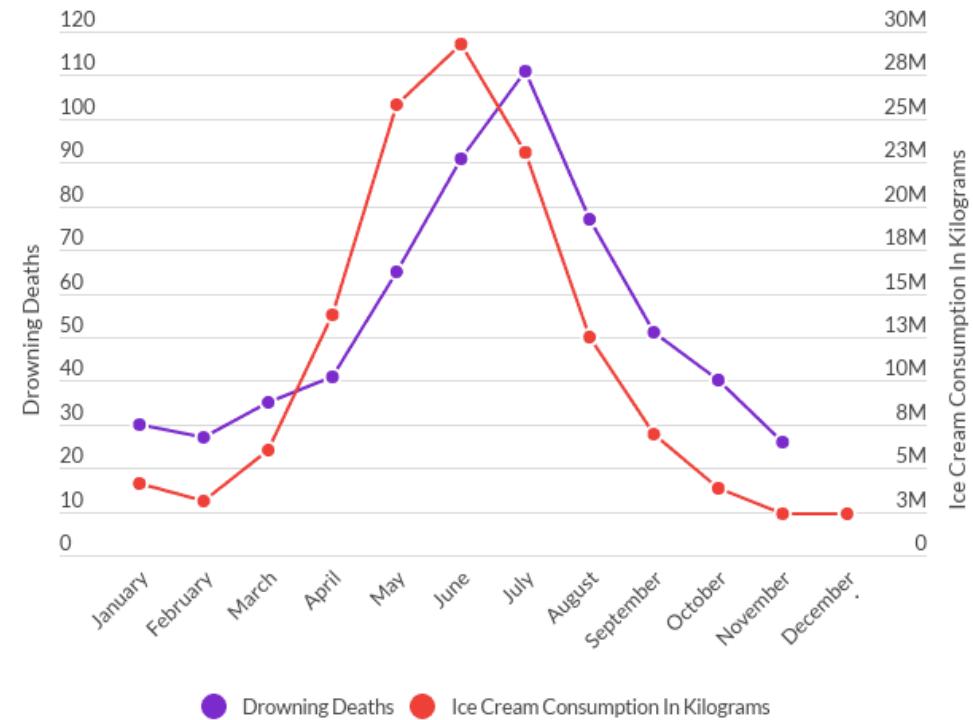


Also related to False positive paradox

Correlation doesn't imply causation



Drowning Deaths and Ice Cream Consumption by Month in Spain (2018)



Statista (2020)

Correlation doesn't imply causation (cont'd)



Bedtime in Preschool-Aged Children and Risk for Adolescent Obesity

Objective To obesity and wh
Study design analyzed. Healt reported their pre was observed t adolescent obe
Results One- after 8:00 p.m. similar regardle 23%, respectiv cent obesity wa times. This risk was not modified by maternal sensitivity ($P = .99$).

Conclusions Preschool-aged children with early weekday bedtimes were one-half as likely as children with late bedtimes to be obese as adolescents. Bedtimes are a modifiable routine that may help to prevent obesity. (*J Pediatr* 2016;176:17-22).

for adolescent
elopment were
16, mothers re-
child interaction
measured and
reference.
f had bedtimes
bedtimes were
0%, 16%, and
(CI) for adoles-
s with late bed-

What research article says

- Proving causality is hard and involves randomized controlled trials (or **causal inference**)
- Be cautious to conclude causality from correlations.

Letting Children Stay Up Late Leads To Overweight Teenagers

Controlling obesity early is key to preventing it later in life.



[PHOTO: MBI/ISTOCK]

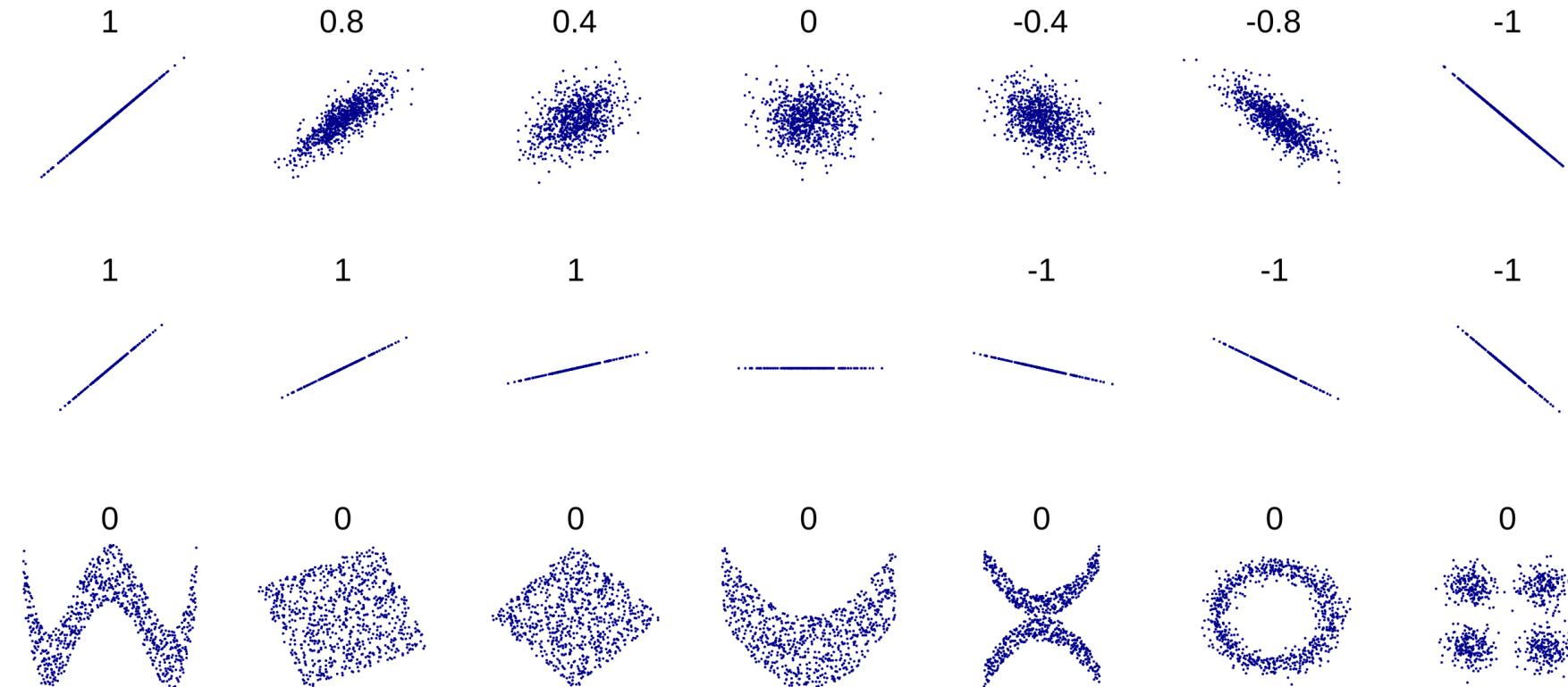
BY CHARLIE SORREL | 1 MINUTE READ

Hey pre-schoolers! If you're reading this after 8 p.m., then you're headed for a miserable time as a teenager, because you're going to get fat. **New research** out of the Ohio State University says that you should listen to your parents and go to bed early in order to avoid obesity when you get older.

What media says

No correlation doesn't imply independence

Correlation only quantifies **linear** relationship



Correlation is not necessarily transitive

- X positively correlates with Z
- Y positively correlates with Z

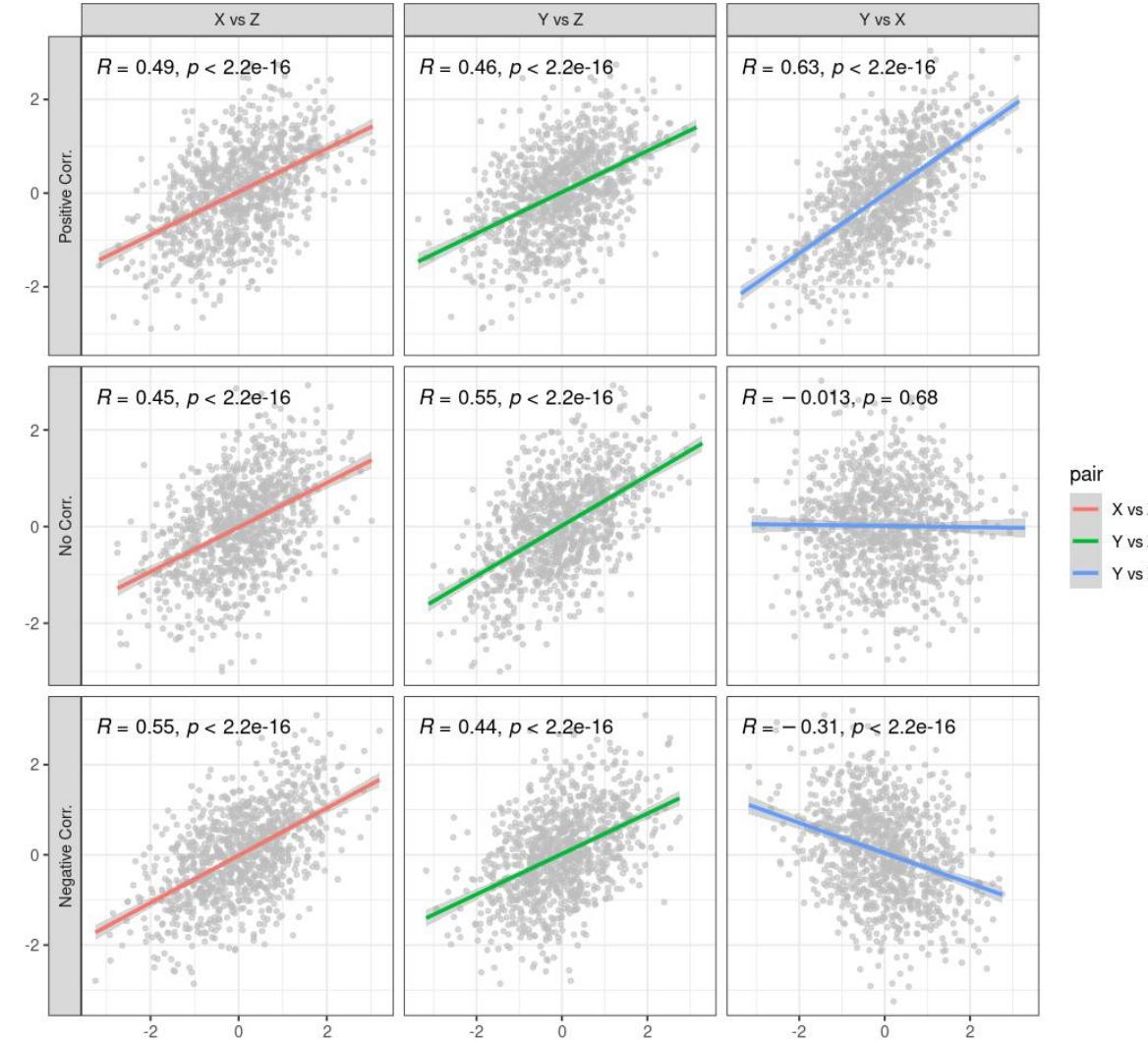
Question: How are X and Y correlated?

Well, everything is possible

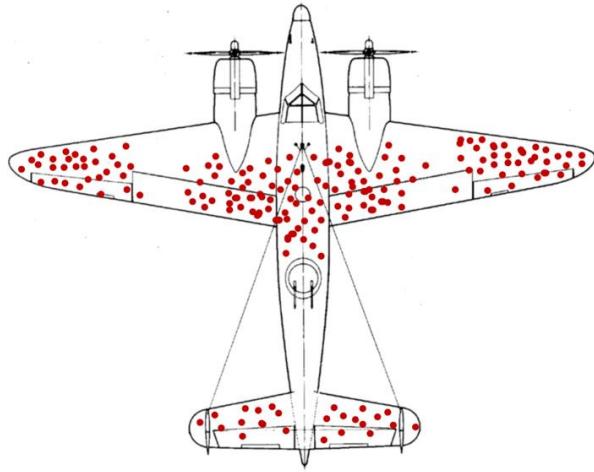
Simulation examples:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & ? & 0.5 \\ ? & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix} \right)$$

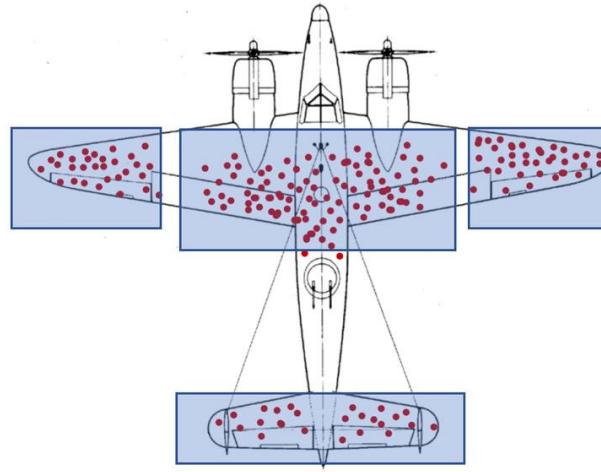
When is correlation transitive?



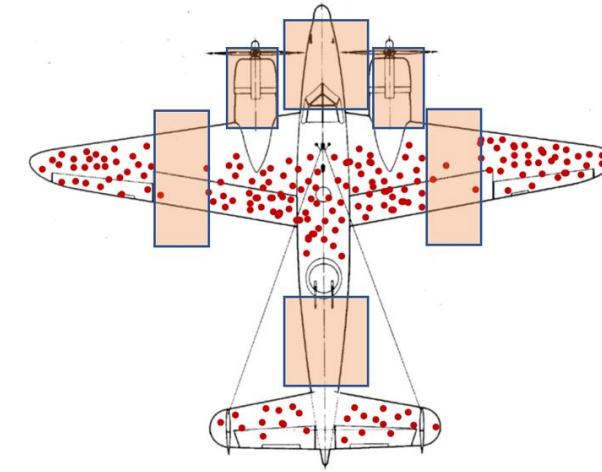
Survival (selection) bias



Our data is only from returning flights. Here we is a visualization of the places that bullet holes were observed.



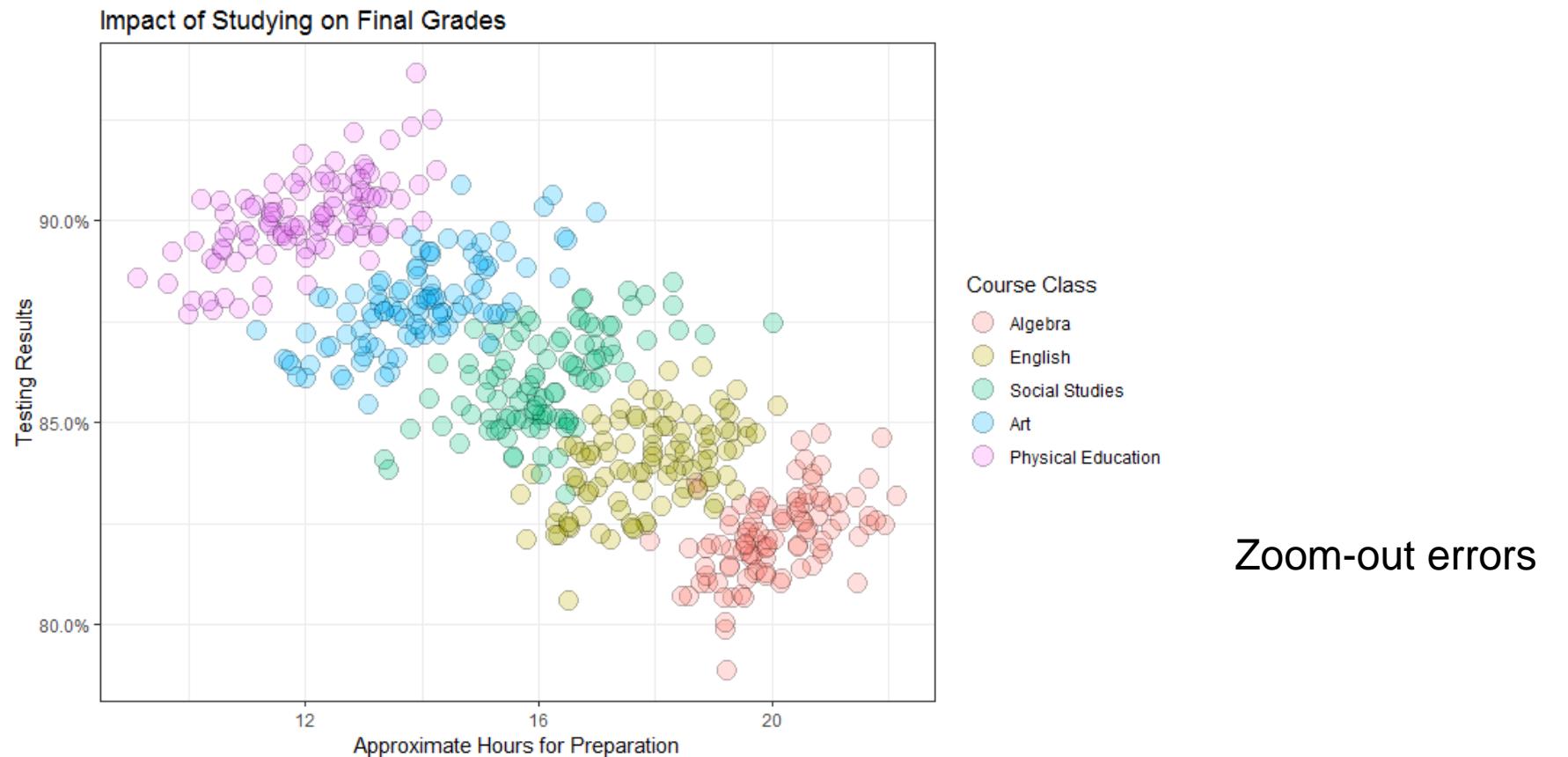
And initial guess at how to fix this might be to apply additional armor plating to the parts of the plane with the most holes...



.... However this is where planes that *returned* had bullet holes. The planes we want to protect are the ones that did *not* return, so we should place armor there.

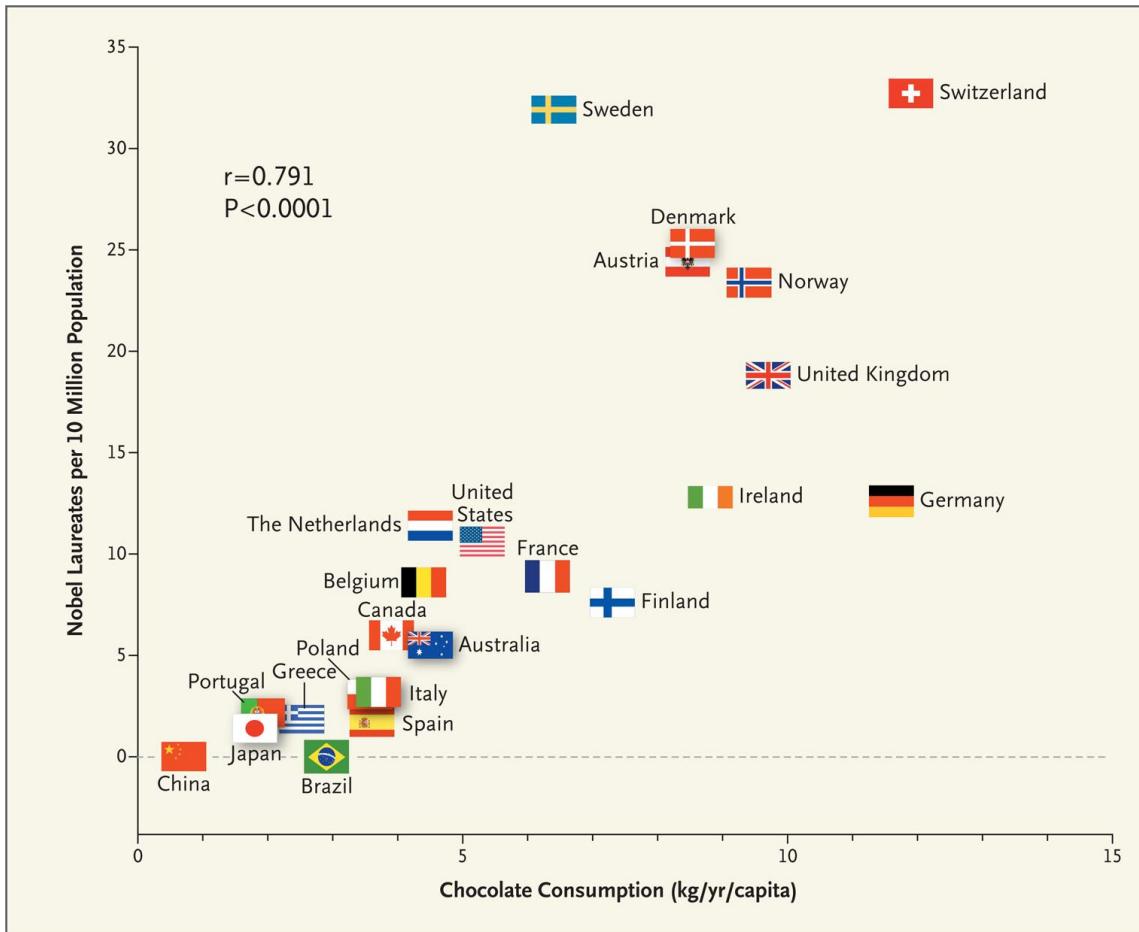
Simpson's paradox

The trend reverses when groups of data are combined



Ecological fallacy

Making false claims at individual level based on group averages



Zoom-in errors

Thank you!

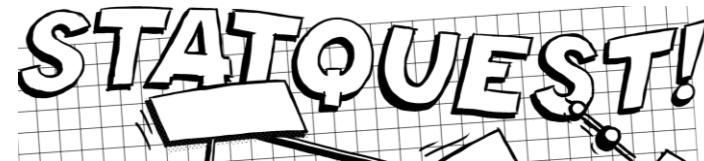
Q&A

Where to get help?

- <https://chat.openai.com/>
- <https://stats.stackexchange.com/>
- <https://www.google.com>
- <https://www.3blue1brown.com>
- <https://statquest.org/>

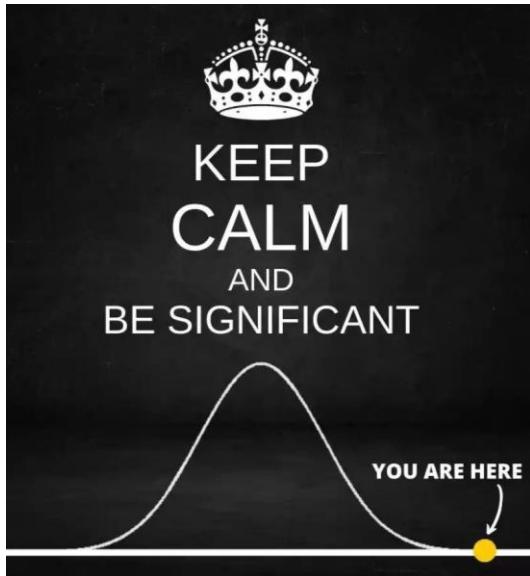


Cross Validated

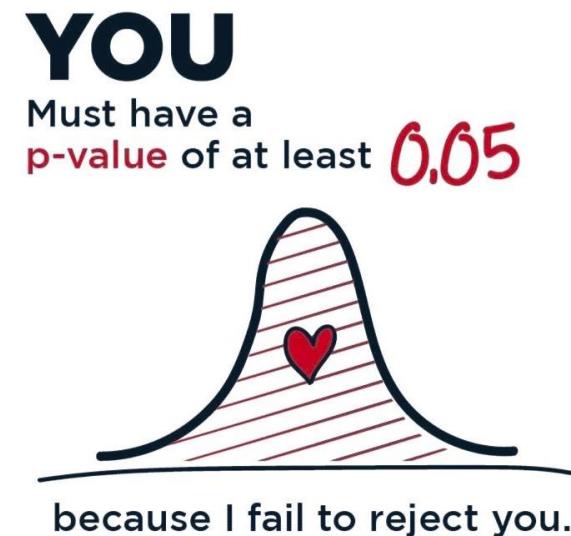
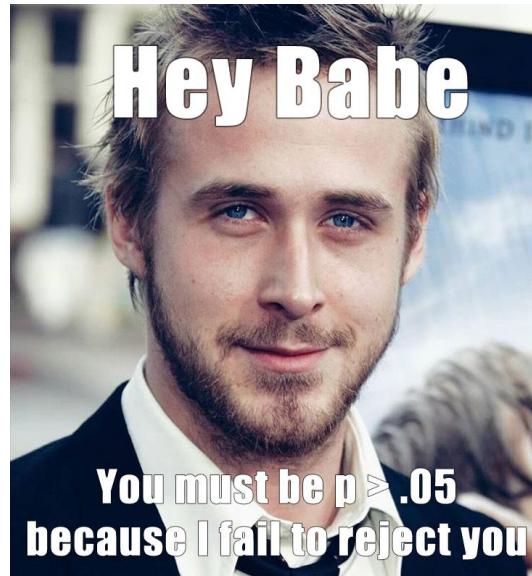


Valentine's special!

Probability/distribution



Hypothesis testing



Model fitting

