Overview

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Philosophy

From 10 years teaching experience

Description of probability and statistics for data-science that

1. Is self contained

   $\Longrightarrow$ Comprehensive theoretical + intuitive explanations

2. Is useful in practice

   $\Longrightarrow$ Real-world data examples provided *for all topics*

3. Follows a logical flow

   $\Longrightarrow$ Probability/statistics/causal inference are *intertwined*

Materials

Book

115 videos with slides

200 exercises with solutions

102 Python notebooks using 23 real-world datasets

Everything (including free preprint) available at

https://www.ps4ds.net/

Contents

1. **Probability** Probability
2. **Discrete variables** Discrete variables
3. **Continuous variables** Continuous variables
4. **Multiple discrete variables** Multiple discrete variables
5. **Multiple continuous variables** Multiple continuous variables
6. **Discrete and continuous variables** Discrete and continuous variables
7. **Averaging** Averaging
8. **Correlation** Correlation
9. **Estimation of population parameters** Estimation of population parameters
10. **Hypothesis testing** Hypothesis testing
11. **Principal component analysis and low-rank models** Principal component analysis and low-rank models
12. **Regression and classification**

Probability

- Probability spaces encode *common-sense* properties
- Conditional probability
- Empirical probability estimation (*statistics*!)

Votes in United States Congress

Probability

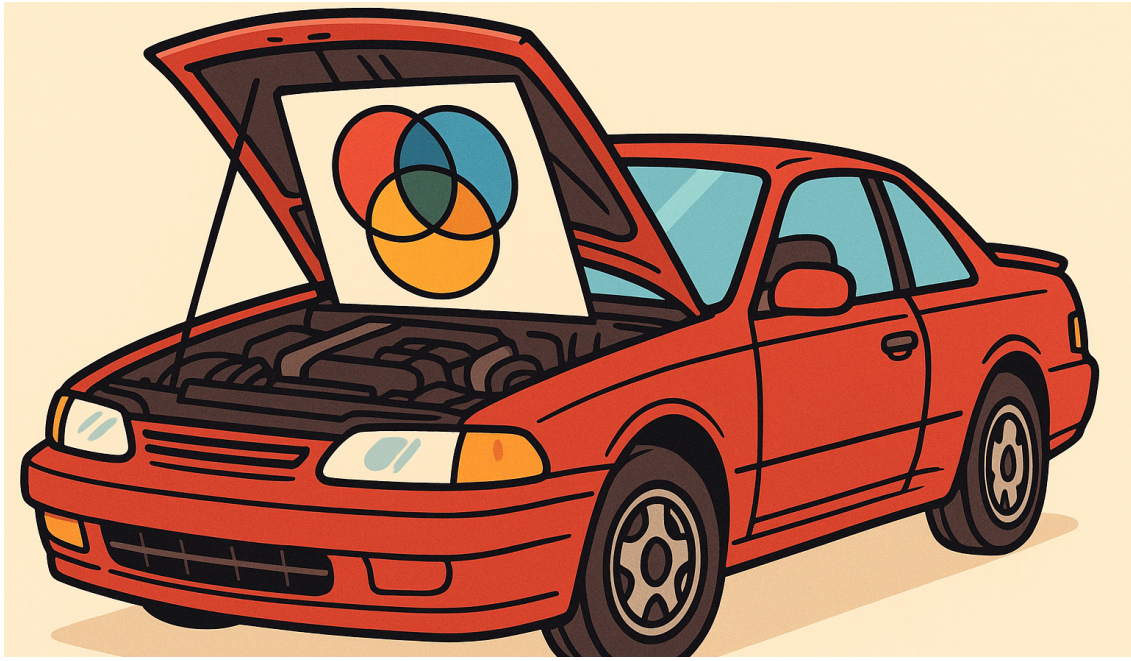
- Independence does not imply conditional independence and vice versa

- Probabilities often cannot be derived analytically, but can be approximated via Monte Carlo simulations

3x3 basketball tournament in Tokyo Olympics

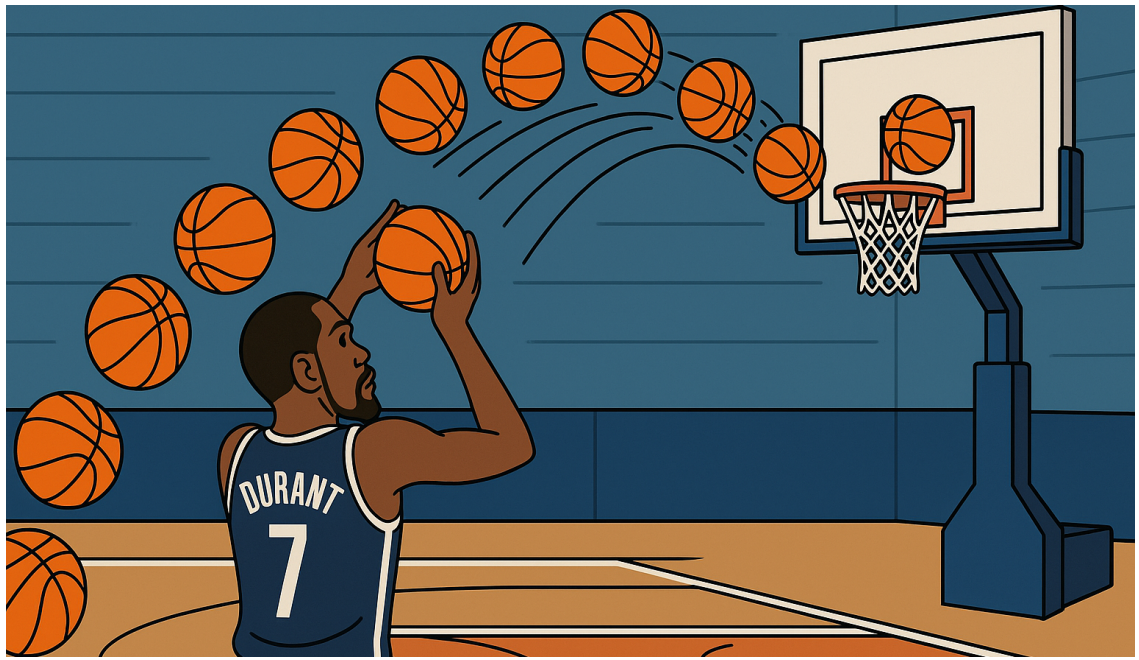
Discrete variables

- **Mathematician**: Functions in a probability space
- **Data scientist**: Variables described by probabilities



Discrete variables

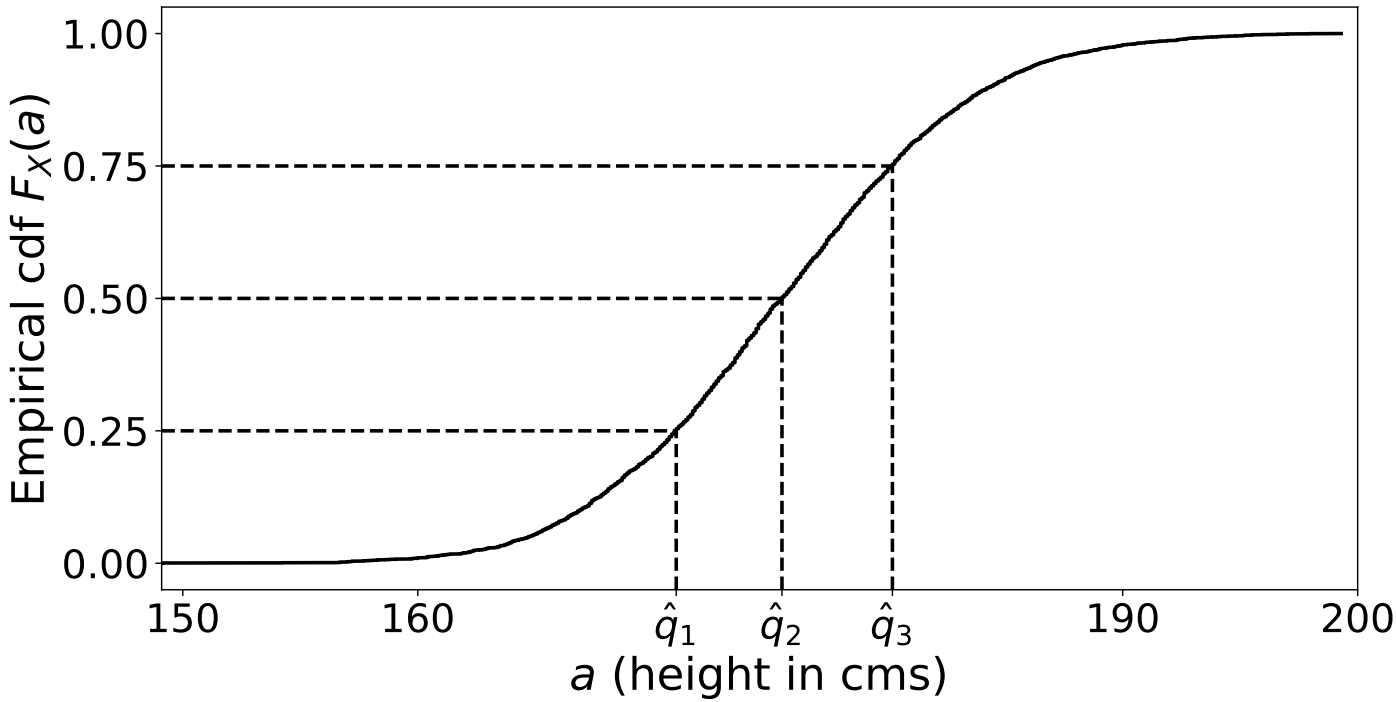
- **Nonparametric** modeling: empirical probabilities
- **Parametric** modeling: geometric, binomial, Poisson distributions fit via maximum likelihood



Kevin Durant's free throws

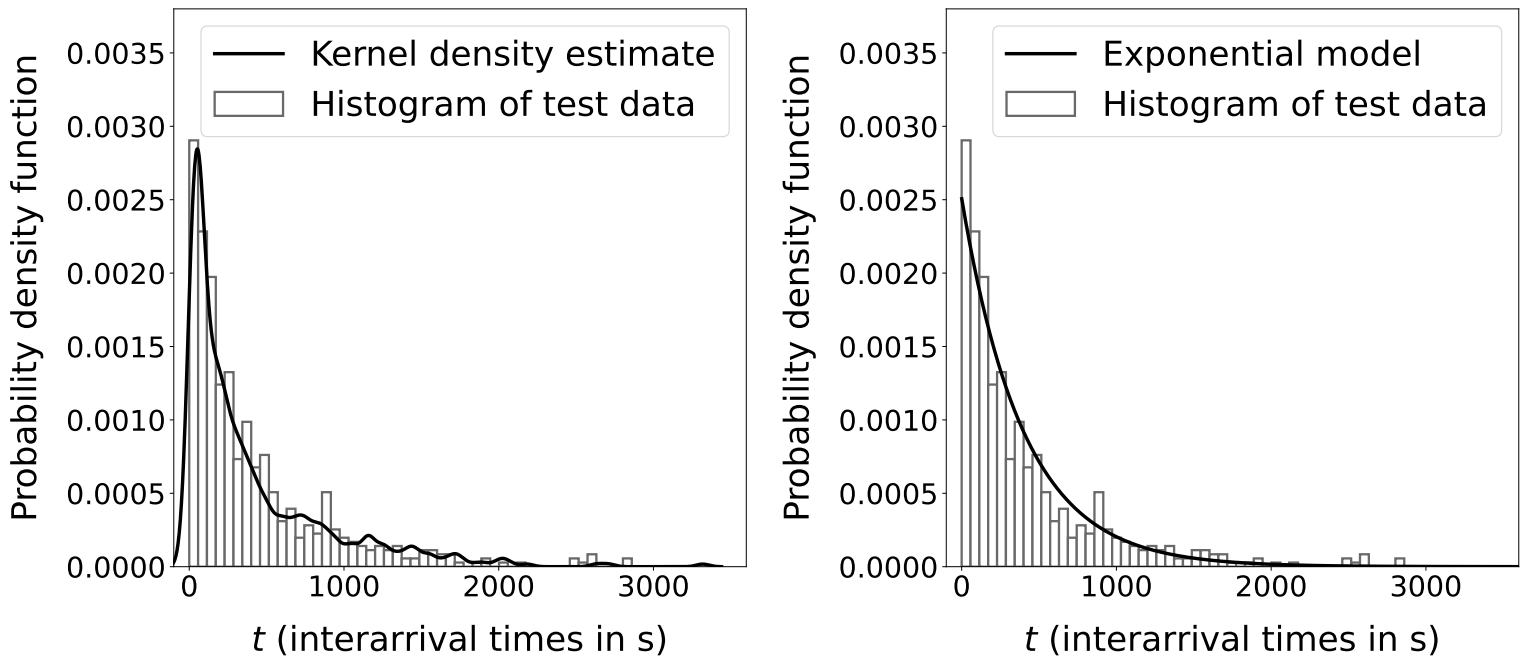
Continuous variables

- **Mathematician**: Functions in a probability space
- **Data scientist**: Variables described by probabilities
  - Cumulative distribution function (cdf) and quantiles
  - Probability density



Continuous variables

- **Nonparametric** modeling: empirical cdf / quantiles, histogram, kernel density estimation
- **Parametric** modeling: exponential / Gaussian distributions fit via maximum likelihood
- Simulation via inverse transform sampling

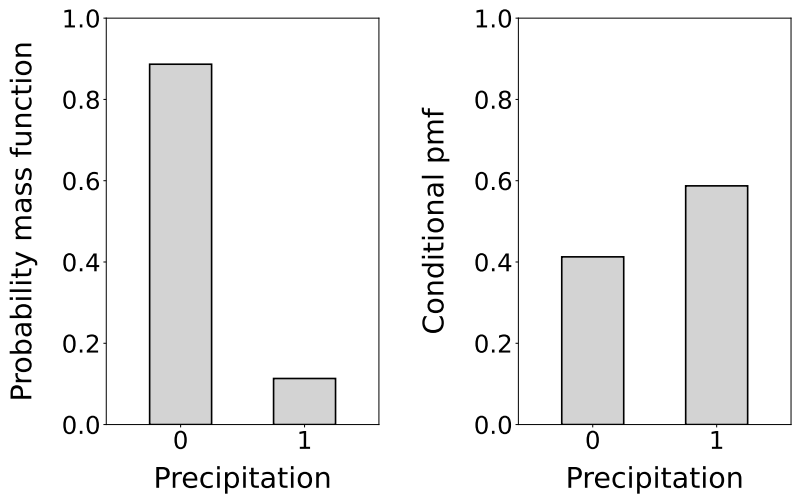


Interarrival times between telephone calls

Multiple discrete variables

- Joint, marginal, conditional distributions
- **Causal inference**: potential outcomes, randomization, confounders, Simpson's paradox

Coos Bay \ Corvallis	0	1
0	84.70	4.02
1	5.62	5.72



Precipitation in Oregon

3-point shooting

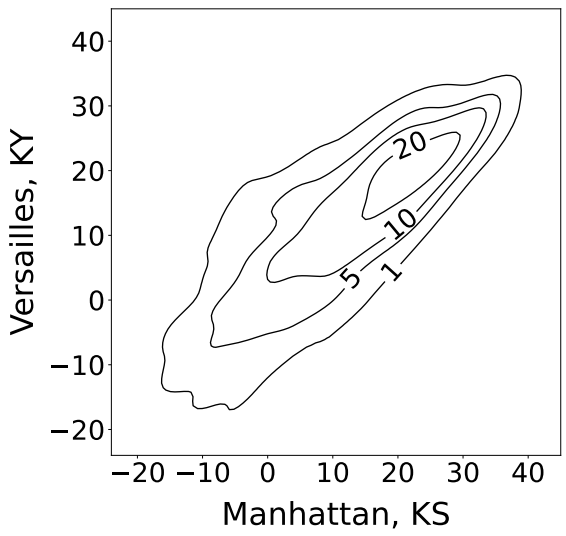
Multiple discrete variables

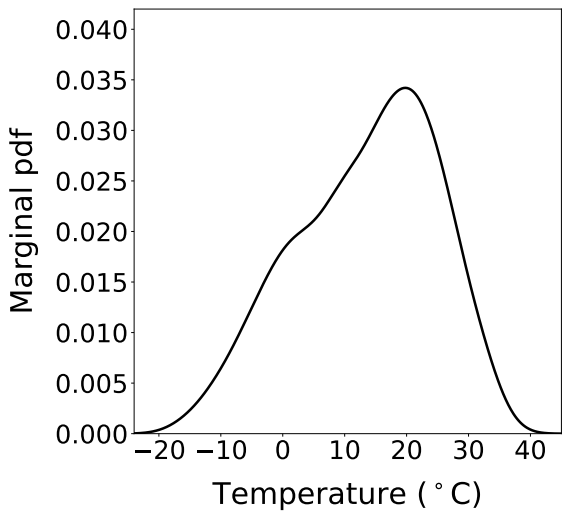
- Unless variables are very few, independence assumptions are essential for tractable estimation (curse of dimensionality)
- Naive Bayes / Markov chains

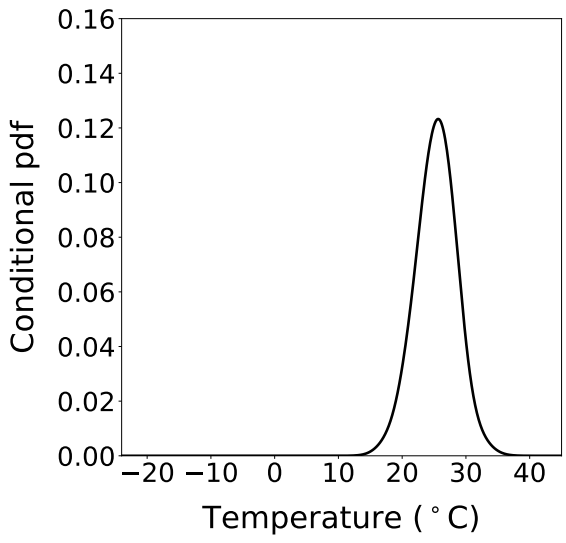
Predicting political affiliation from votes

Multiple continuous variables

- Joint, marginal, conditional densities
- **Nonparametric** modeling: kernel density estimation



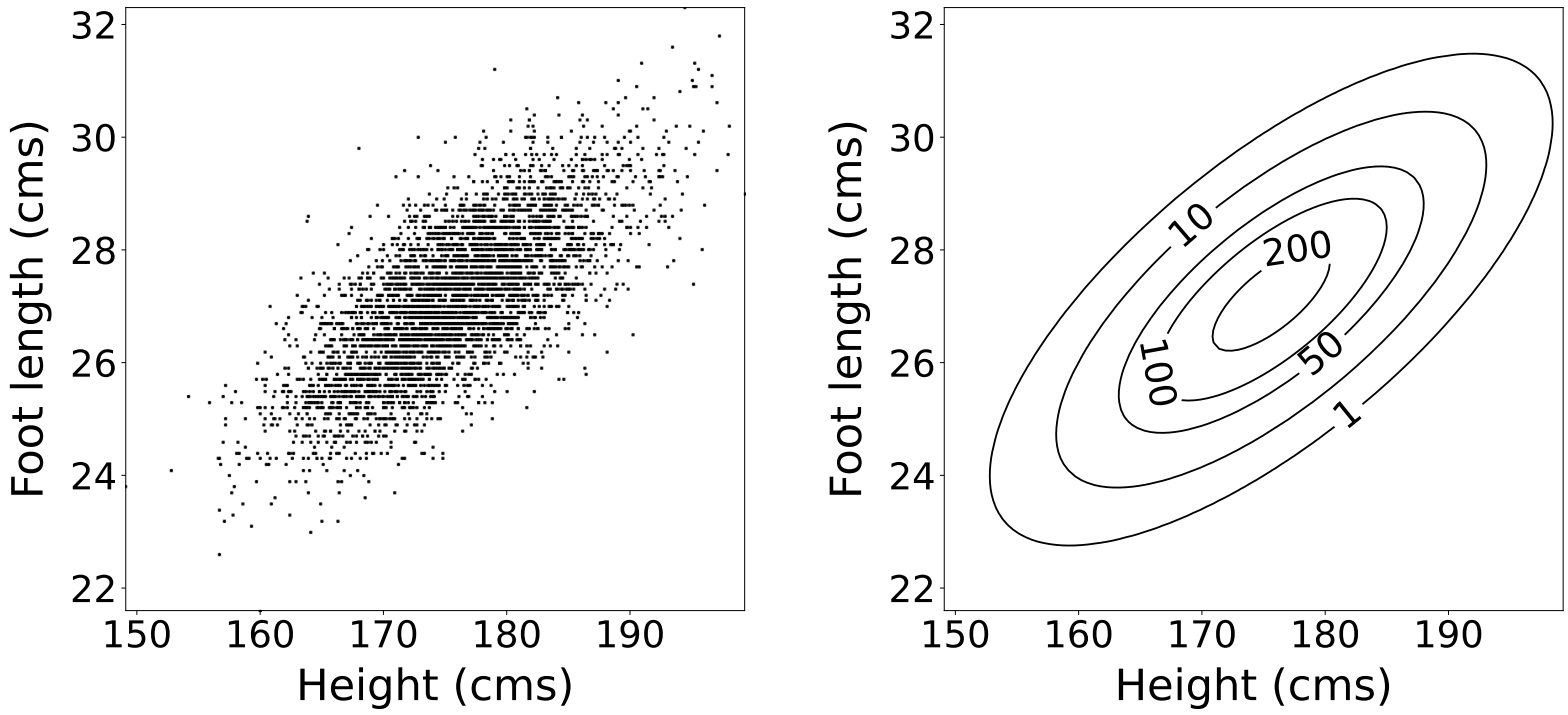




Temperature in Versailles and Manhattan

Multiple continuous variables

- **Parametric** modeling: Gaussian random vectors
- Joint simulation must use conditional distributions!

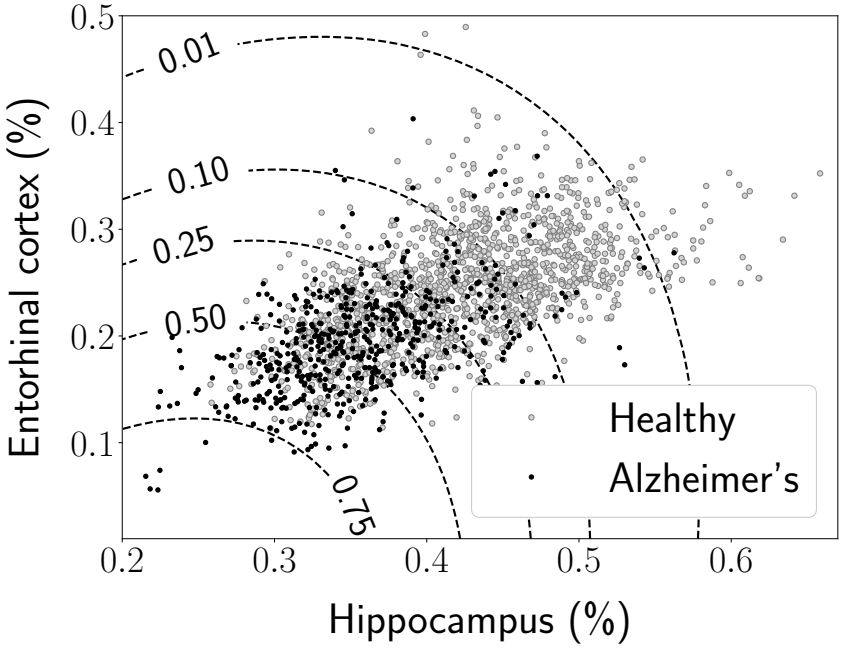


Anthropometric data

Discrete and continuous variables

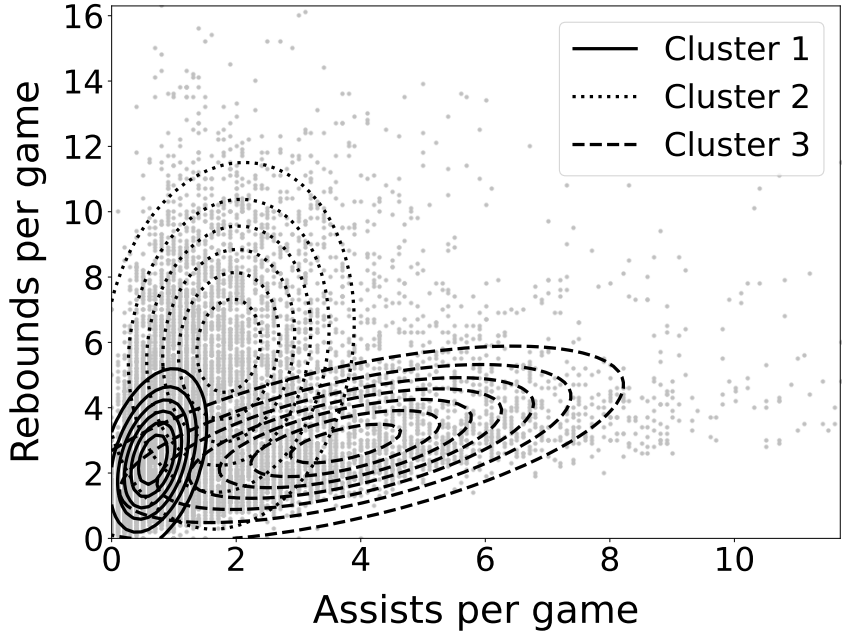
- Conditional probabilities / densities
- Parametric mixture models

Gaussian discriminant analysis



Alzheimer's diagnosis

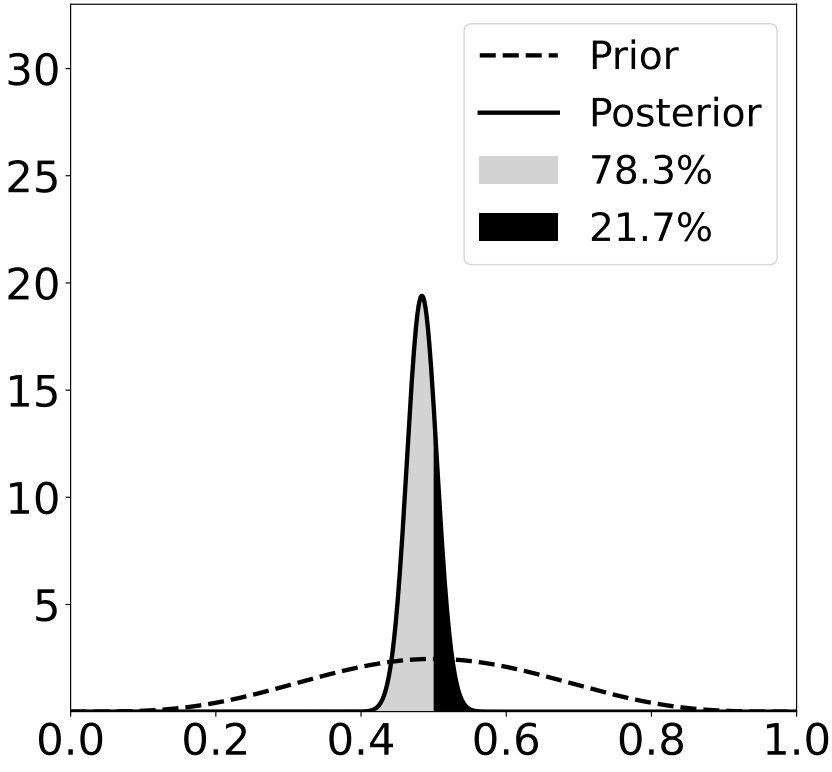
Gaussian mixture models



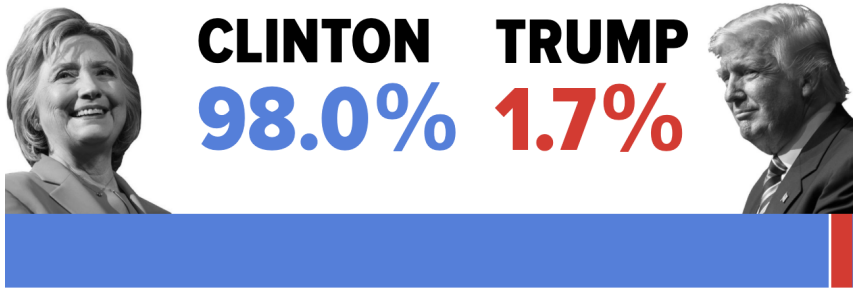
NBA stats

Discrete and continuous variables

- Bayesian models
- How not to predict an election!



Biden vs Trump 2020

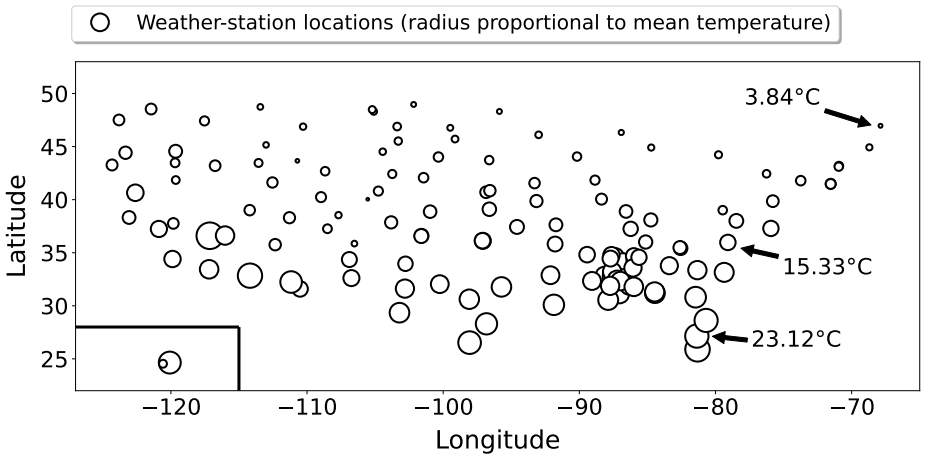


Clinton vs Trump 2016

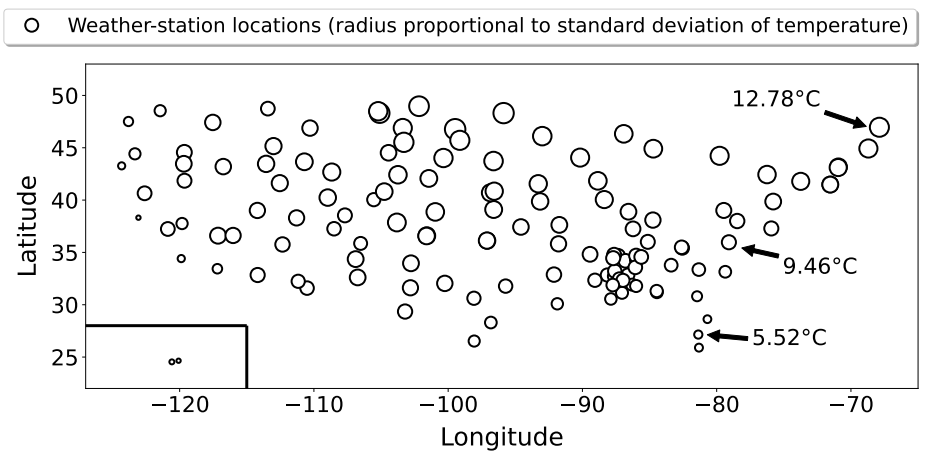
Averaging

- The mean and the sample mean
- Linearity of expectation, sensitivity to outliers
- The variance and the sample variance

Mean temperatures

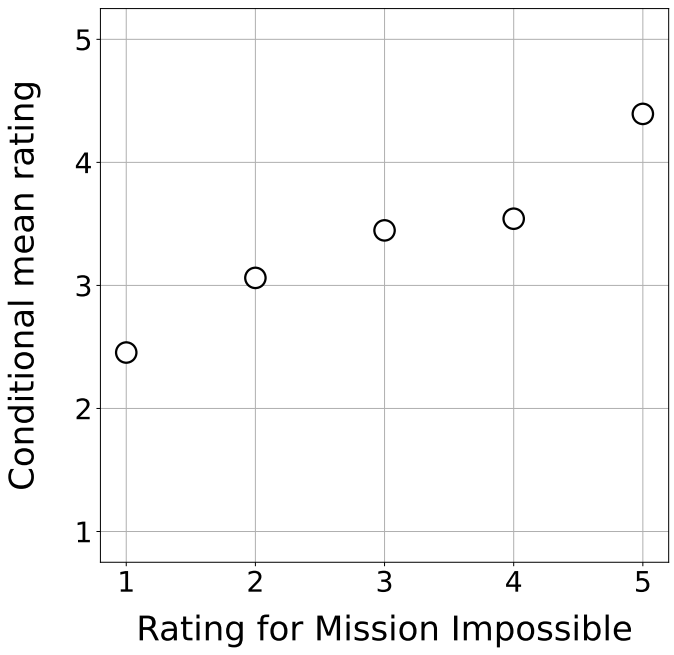


Standard deviation

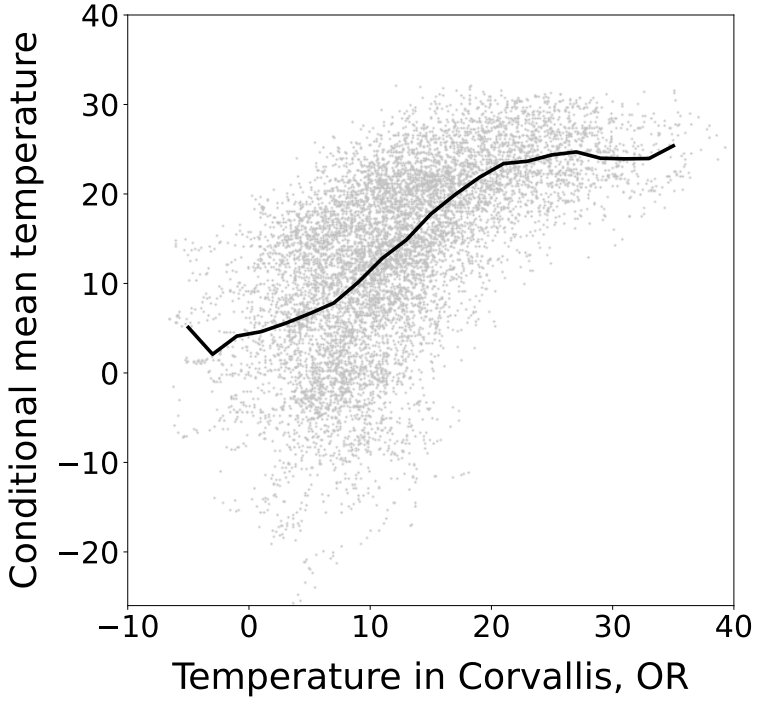


Averaging

- Iterated expectation and the conditional mean
- **Causal inference**: Average treatment effect



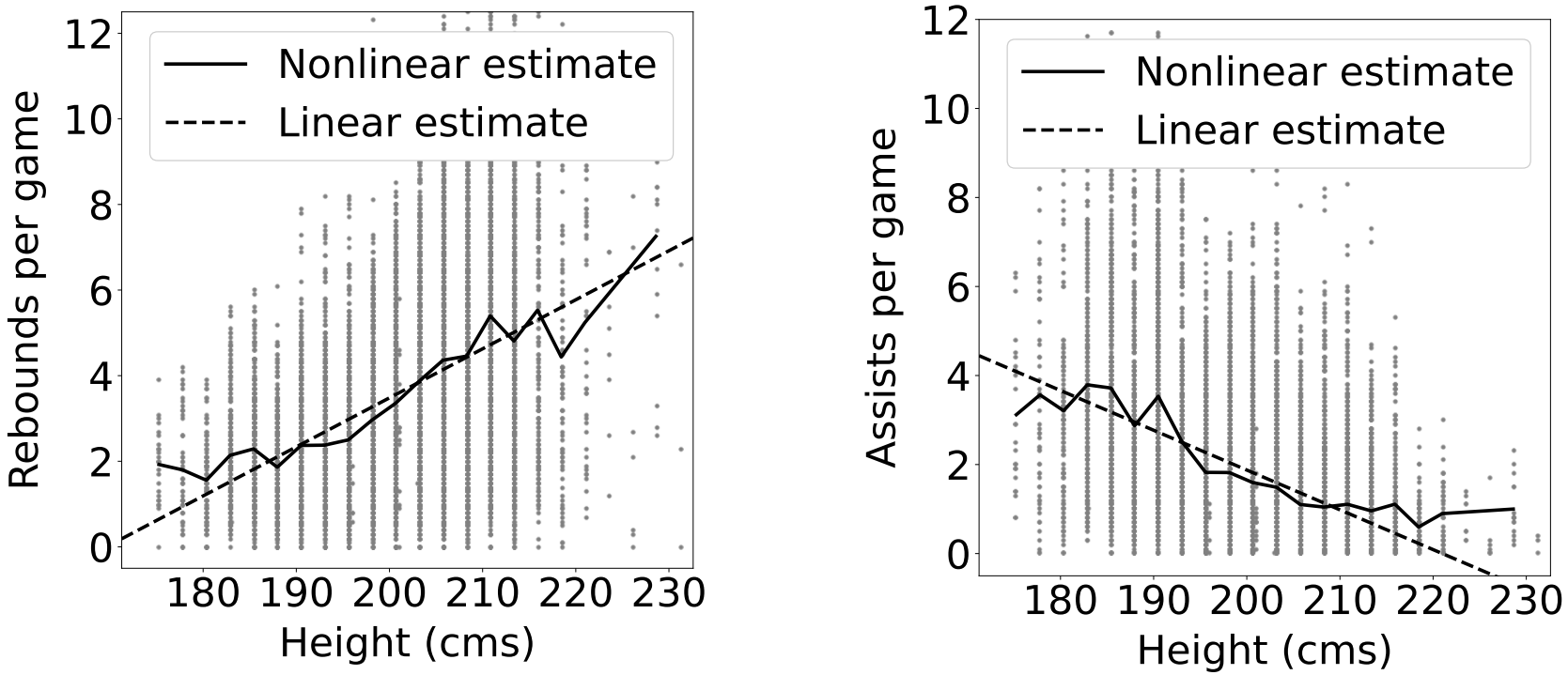
Movie ratings



Temperature

Correlation

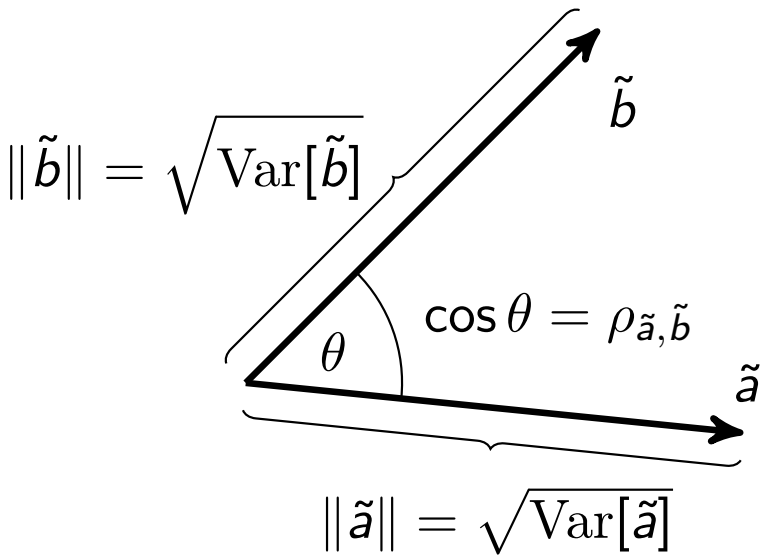
- Covariance and correlation coefficient
- Sample correlation
- Simple linear regression



Basketball stats vs height

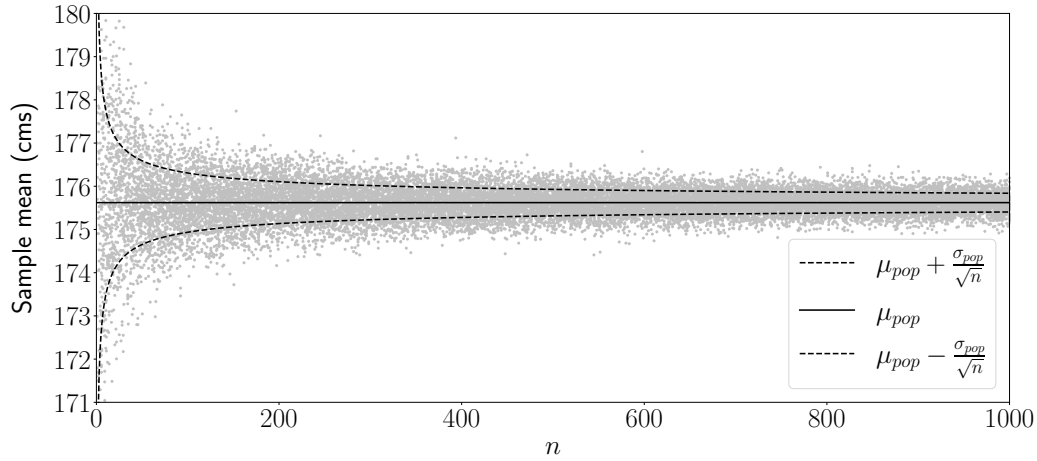
Correlation

- Geometric analysis
- Uncorrelation does not imply independence
- Correlation does not imply **causation**!



Estimation of population parameters

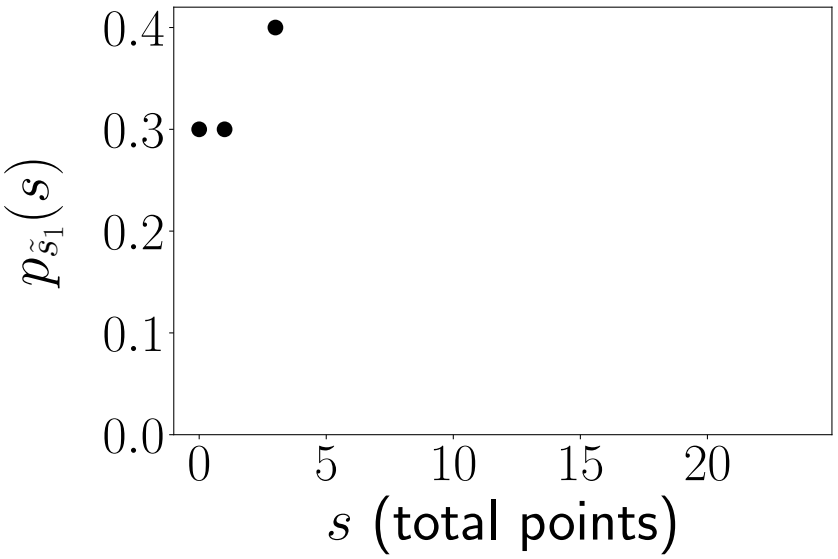
- Random sampling
- Bias and standard error
- Law of large numbers (and exceptions!)

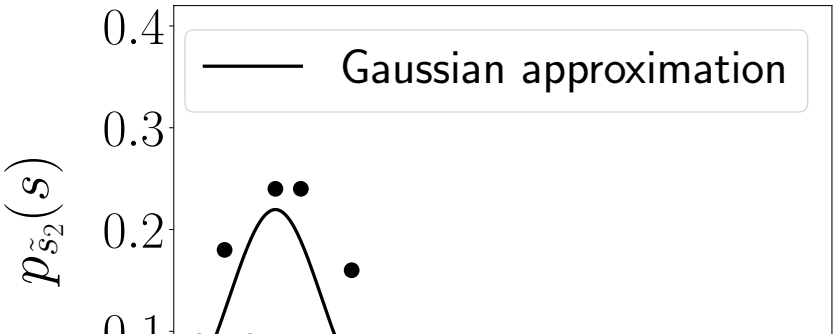


Height estimation

Estimation of population parameters

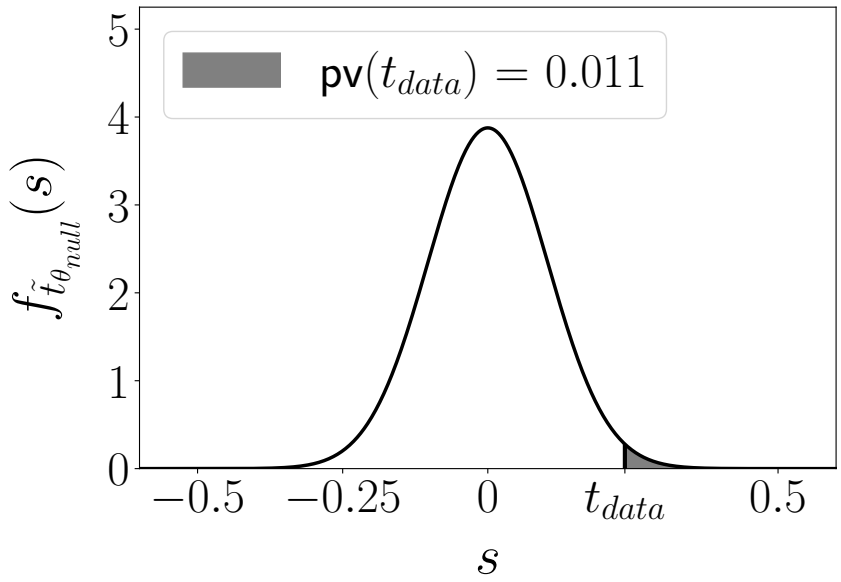
- Central limit theorem
- Confidence intervals
- Bootstrapping





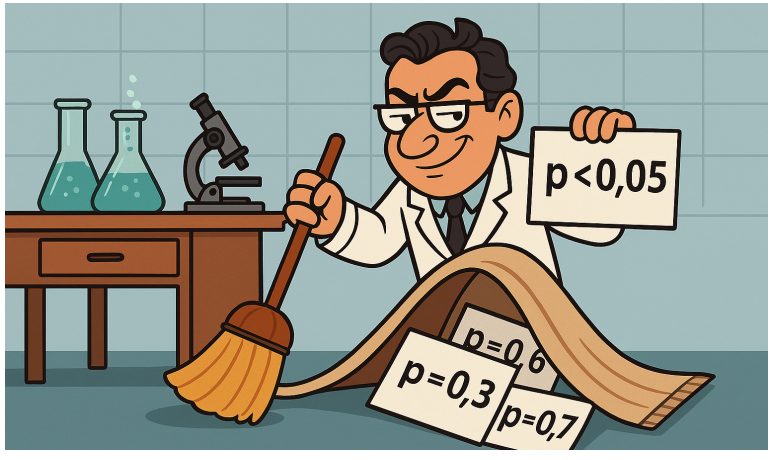
Hypothesis testing

- Null hypothesis and p value
- **Parametric** testing: Statistical significance and power
- **Nonparametric** testing: Permutation test



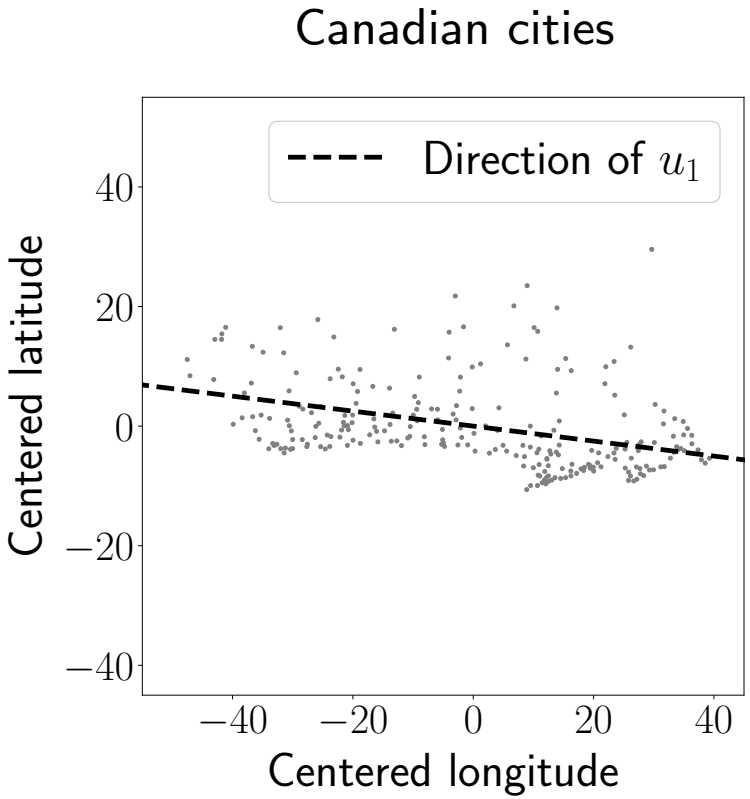
Hypothesis testing

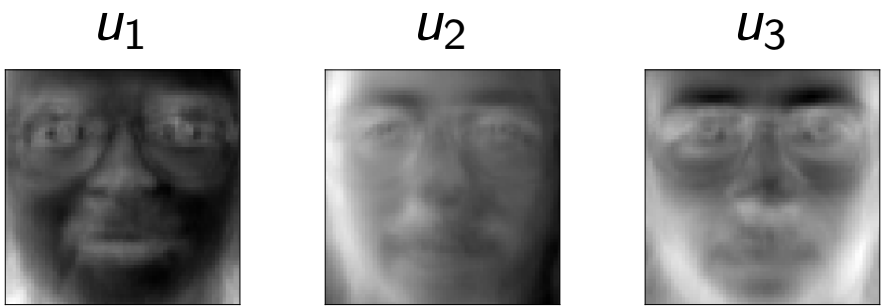
- Multiple testing
- Hypothesis testing and **causal inference**
- Practical significance, publication bias, p-hacking



Principal component analysis

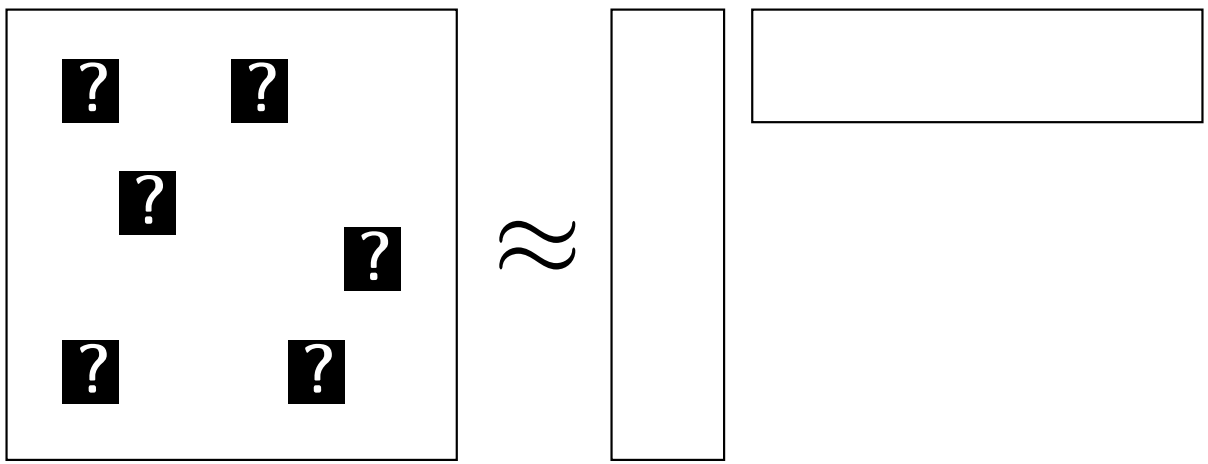
- Covariance matrix and sample covariance matrix
- Principal directions and principal components
- Dimensionality reduction





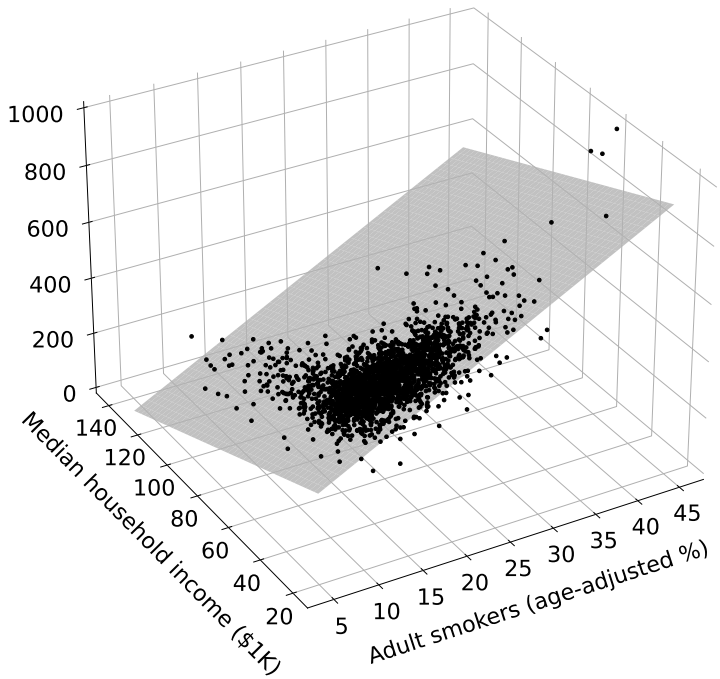
Low-rank models

- Singular-value decomposition and truncation
- Matrix completion for collaborative filtering

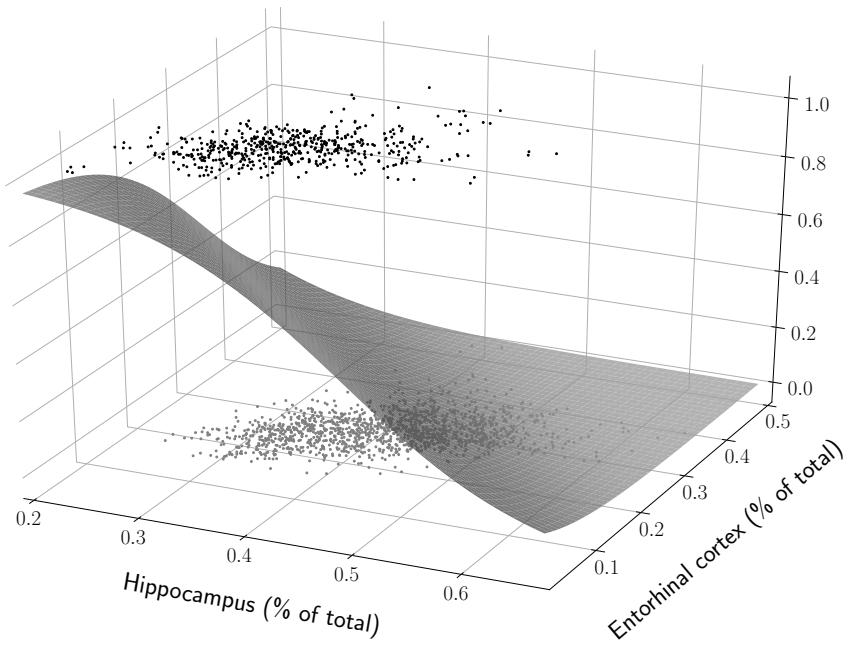


Linear regression and classification

- Linear minimum MSE estimation, ordinary least squares, causal inference
- Overfitting and regularization (ridge regression, sparse regression)
- Logistic and softmax regression



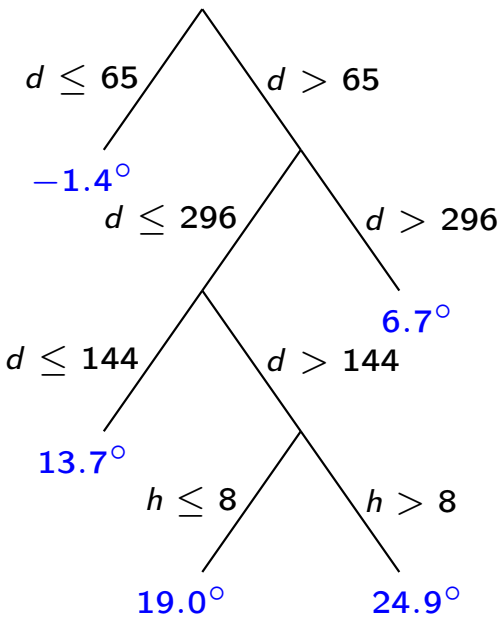
Mortality prediction

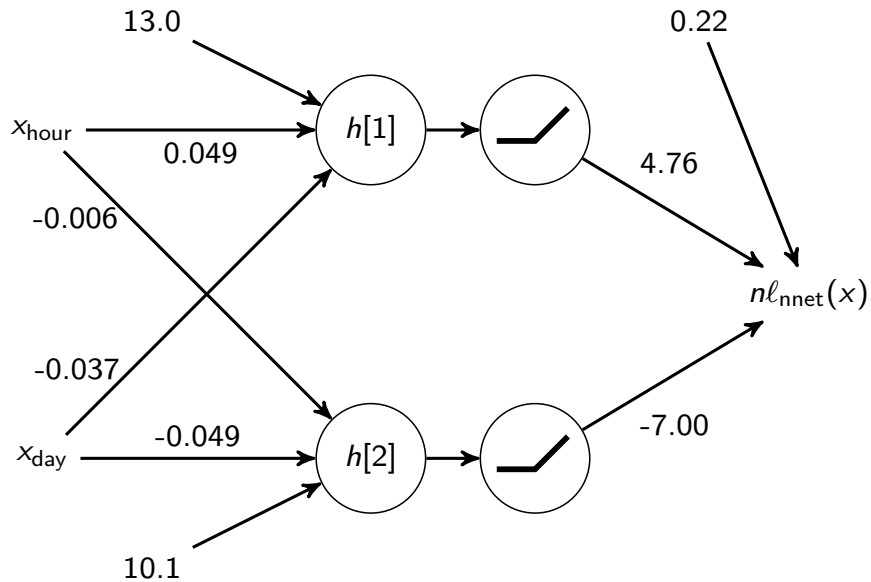


Alzheimer's diagnosis

Nonlinear regression and classification

- Regression and classification trees
- Ensembles: bagging / random forests / boosting
- Neural networks and deep learning





Philosophy

From 10 years teaching experience

Description of probability and statistics for data-science that

1. Is self contained
   $\Longrightarrow$ Comprehensive theoretical + intuitive explanations

2. Is useful in practice
   $\Longrightarrow$ Real-world data examples provided *for all topics*

3. Follows a logical flow
   $\Longrightarrow$ Probability/statistics/causal inference are *intertwined*

https://www.ps4ds.net/