CS 540 Introduction to Artificial Intelligence

Statistics Review

University of Wisconsin-Madison

Fall 2023

Review: Bayesian Inference

- Conditional Probability & Bayes Rule:

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

- Evidence *E*: what we can observe
- Hypothesis *H*: what we'd like to infer from evidence
  - Need to plug in prior, likelihood, etc.
- Usually do not know these probabilities. How to estimate?

Samples and Estimation

- Usually, we don't know the distribution P
  - Instead, we see a bunch of samples

- Typical statistics problem: **estimate distribution** from samples
  - Estimate probabilities $P(H)$, $P(E)$, $P(E|H)$
  - Estimate the mean $E[X]$
  - Estimate parameters $P_\theta(X)$

Samples and Estimation

- Estimate probability $P(H)$, $P(E)$, $P(E|H)$
- Estimate the mean $E[X]$
- Estimate parameters $P_\theta(X)$

- Example: Bernoulli with parameter *p* *(i.e., a weighted coin flip)*
  - $P(X = 1) = p$
  - Mean $E[X]$ is *p*

Examples: Sample Mean

- Bernoulli with parameter *p*
- See samples $x_1, x_2, \ldots, x_n$
  - Estimate mean with **sample mean**

$$\hat{\mathbb{E}}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - That is, counting heads

Break & Quiz

Q 2.1: You see samples of X given by $[0,1,1,2,2,0,1,2]$. Empirically estimate $\mathbb{E}[X^2]$

A. 9/8
B. 15/8
C. 1.5
D. There aren't enough samples to estimate $\mathbb{E}[X^2]$

Break & Quiz

Q 2.1: You see samples of X given by [0,1,1,2,2,0,1,2]. Empirically estimate $\mathbb{E}[X^2]$

A. 9/8
B. 15/8
C. 1.5
D. There aren't enough samples to estimate $\mathbb{E}[X^2]$

Break & Quiz

Q 2.1: You see samples of X given by [0,1,1,2,2,0,1,2]. Empirically estimate $\mathbb{E}[X^2]$

A. 9/8
B. 15/8
C. 1.5
D. There aren't enough samples to estimate $\mathbb{E}[X^2]$

$$E[X^2] \approx \frac{1}{n}\sum_i X_i^2$$
$$= \frac{1}{8}(0^2 + 1 + 1 + 4 + 4 + 0 + 1 + 4) = 15/8$$

Estimating Multinomial Parameters

- k-sized die (special case: k=2 coin)
- Face i has probability p_i, for $i=1...k$
- In n rolls, we observe face i showing up n_i times

$$\sum_{i=1}^{k} n_i = n$$

- Estimate $(p_{1,...,}\ p_k)$ from this data $(n_{1,...,}\ n_k)$

Maximum Likelihood Estimate (MLE)

- The MLE of multinomial parameters $(\widehat{p_1}, \ldots, \widehat{p_k})$

$$\widehat{p_i} = \frac{n_i}{n}$$

- Estimate using frequencies

Break & Quiz

Q 2.2: You are empirically estimating *P(X)* for some random variable *X* that takes on 100 values. You see 50 samples. How many of your *P(X=a)* estimates might be 0?

A. None.
B. Between 5 and 50, exclusive.
C. Between 50 and 100, inclusive.
D. Between 50 and 99, inclusive.

Break & Quiz

Q 2.2: You are empirically estimating *P(X)* for some random variable *X* that takes on 100 values. You see 50 samples. How many of your *P(X=a)* estimates might be 0?

For each a, your estimate is $\mathrm{P}(X = a) = \frac{\#\text{samples taking value } a}{50}$

A. None.
B. Between 5 and 50, exclusive.
C. Between 50 and 100, inclusive.
D. Between 50 and 99, inclusive.

Break & Quiz

Q 2.2: You are empirically estimating *P(X)* for some random variable *X* that takes on 100 values. You see 50 samples. How many of your *P(X=a)* estimates might be 0?

For each a, your estimate is $\mathrm{P}(X = a) = \frac{\#\text{samples taking value } a}{50}$

A. None.
B. Between 5 and 50, exclusive.
C. Between 50 and 100, inclusive.
D. Between 50 and 99, inclusive.

Break & Quiz

Q 2.2: You are empirically estimating *P(X)* for some random variable *X* that takes on 100 values. You see 50 samples. How many of your *P(X=a)* estimates might be 0?

For each a, your estimate is $\mathrm{P}(X = a) = \frac{\#\text{samples taking value } a}{50}$

A. None.
B. Between 5 and 50, exclusive.
C. Between 50 and 100, inclusive.
D. **Between 50 and 99, inclusive.**

If you don't see a number at all in the 50 samples then the estimated probability of that number is 0.

You can see up to 50 different values in 50 samples. On the other hand, all 50 samples might have the same value in which case 99 values were never seen.

Regularized Estimate

- Hyperparameter $\epsilon > 0$

$$\widehat{p_i} = \frac{n_i + \epsilon}{n + k\epsilon}$$

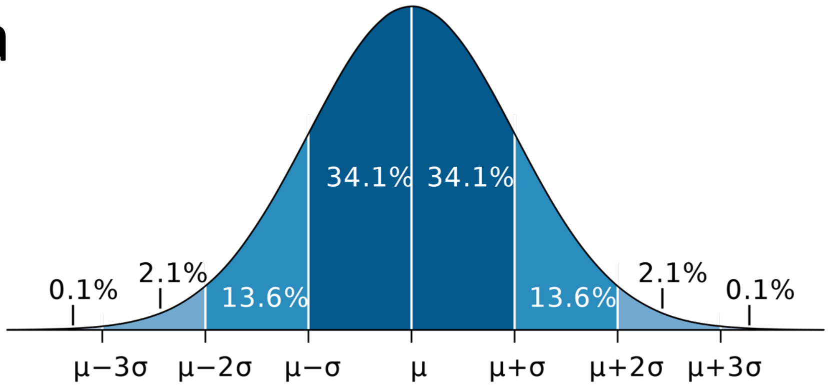
- Avoids zero when *n* is small
- Biased, but has smaller variance
- Equivalent to a specific Maximum A Posterori (MAP) estimate, or smoothing

Estimating 1D Gaussian Parameters

- Gaussian (aka Normal) distribution $N(\mu, \sigma^2)$
  - True mean μ, true variance σ^2
- Observe *n* data points from this distribution

$$x_1, \dots, x_n$$

- Estimate μ, σ^2 from this data



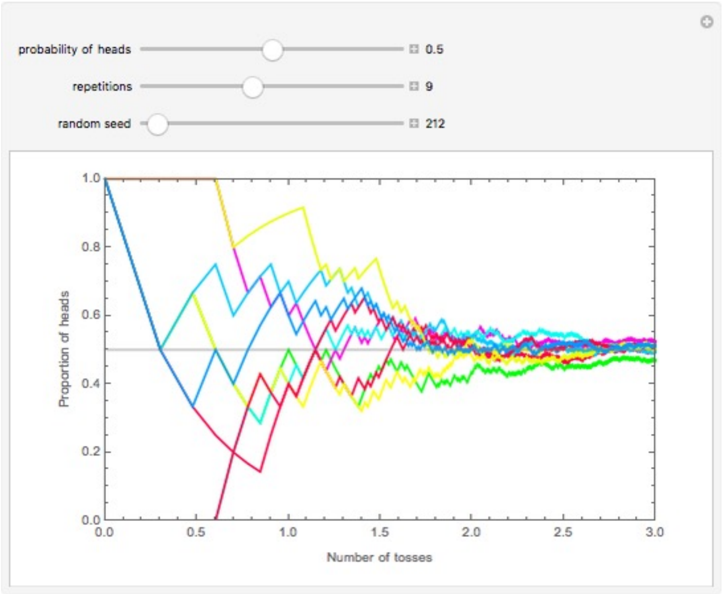
Wikipedia: Normal distribution

Estimating 1D Gaussian Parameters

- Mean estimate $\hat{\mu} = \dfrac{x_1 + \cdots + x_n}{n}$
- Variance estimates
  - Unbiased $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n-1}$
  - MLE $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n}$

Estimation Theory

- Is the sample mean a good estimate of the true mean?
  - Law of large numbers
  - Central limit theorems



Wolfram Demo

Estimation Errors

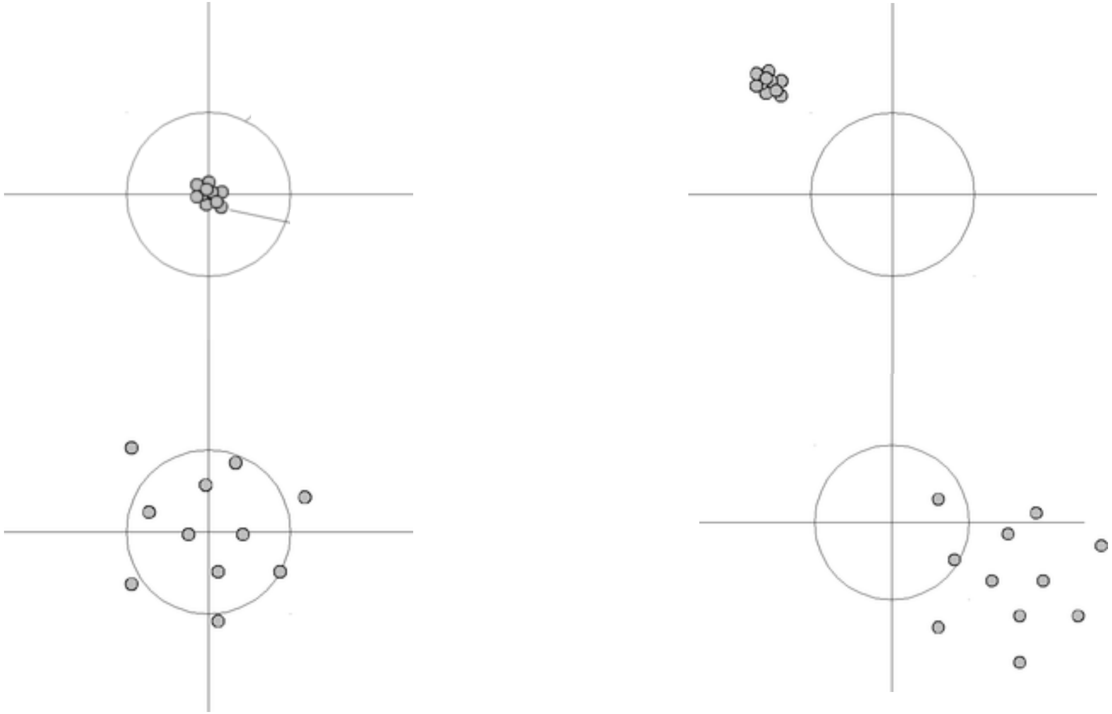
- With finite samples, likely error in the estimate.
- Mean squared error
  - $\mathrm{MSE}[\hat{\theta}] = \mathbb{E}\,[(\hat{\theta} - \theta)^2]$
- Bias / Variance Decomposition
  - $\mathrm{MSE}[\hat{\theta}] = \underbrace{\mathbb{E}\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right]}_{\text{Variance}} + \underbrace{\left(\mathbb{E}[\hat{\theta}] - \theta\right)^2}_{\text{Bias}}$

Bias / Variance

Low Bias | High Bias

Low Variance

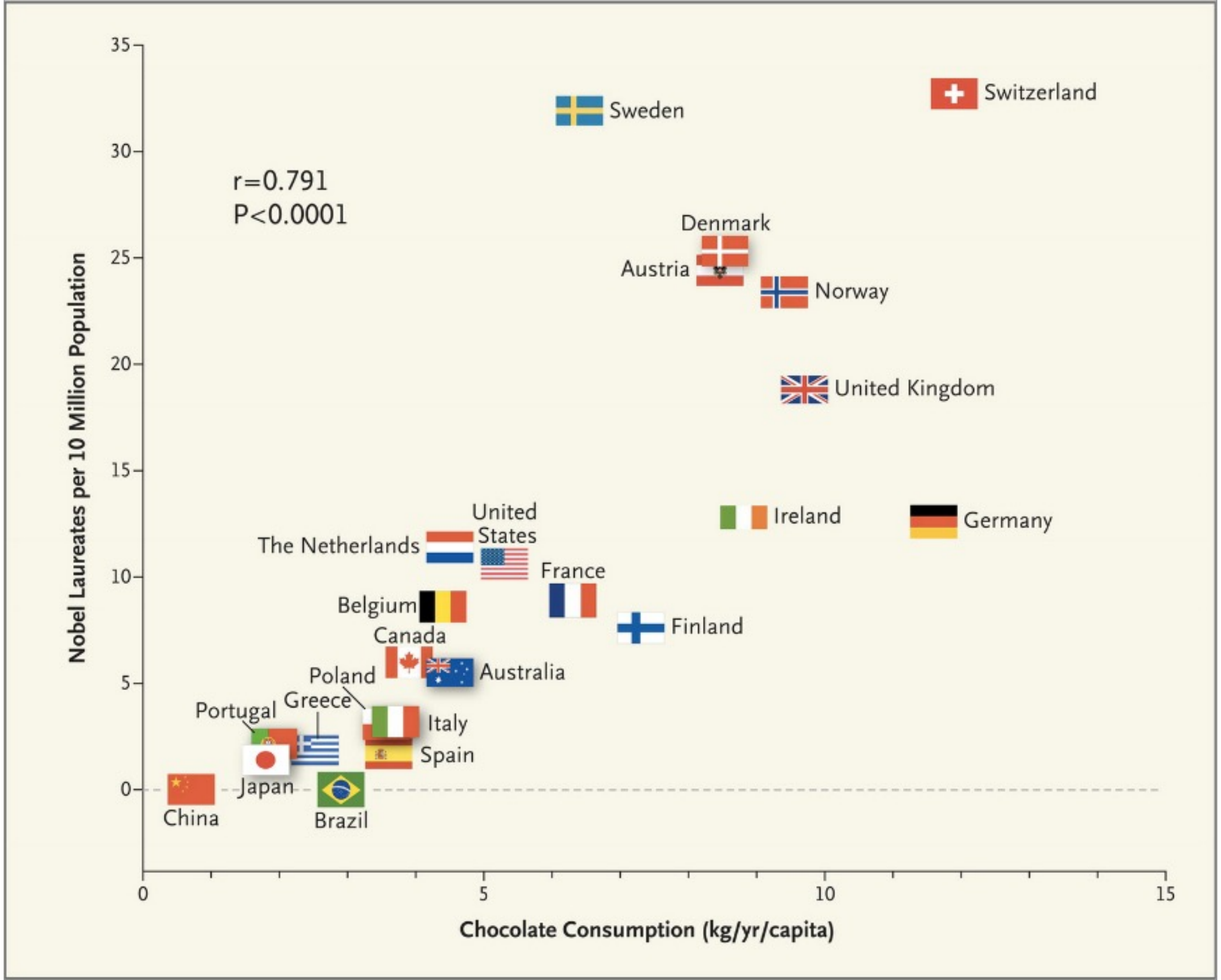
High Variance



Wikipedia: Bias-variance tradeoff

Correlation vs. Causation

- Conditional probabilities only define correlation (aka association)
- $P(Y|X)$ “large” does not mean X causes Y
- Example: X=yellow finger, Y=lung cancer
- Common cause: smoking



https://www.nejm.org/doi/full/10.1056/NEJMon1211064