CS 540 Introduction to Artificial Intelligence
Statistics

University of Wisconsin–Madison
Fall 2025, Section 3
September 8, 2025

Announcements

- HW 1 to be released Friday (9/12)
- Different classroom on 10/24: Psychology 113
- Class roadmap:

Probability
Statistics
Linear Algebra & PCA
Logic
NLP

Mostly Foundations

Last Class: Probability

- Outcomes
- Events
- Probability distribution
- Axioms of probability
- Random variables
- PDF & CDF
- Expectation & variance
- Joint probability
- Marginal probability
- Independence
- Conditional probability
- Conditional independence
- Chain rule

Mathematical tools for discussing and reasoning about randomness

Last Class: 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.

Today: Learning from Samples

- When our **evidence** consists of **data**

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

We observe many samples

First example: estimating the bias of a coin

Learning from Samples with Bayes

- Coin flip: **fair** or **biased**?
- I promise that either:
  - The coin is fair, or
  - The coin is heads 75% of the time.
- These happen with equal probability.
- We observe four samples: HHTH

Learning from Samples with Bayes

- We observe: HHTH
- Bayes' Rule:

$$P(\text{ fair} \mid HHTH\,) = \frac{P(\,HHTH \mid \text{fair}\,)P(\text{fair})}{P(HHTH)}$$

- Conditional independence

$$P(HHTH \mid \text{fair}) = P(\,H \mid \text{fair}\,) \cdot P(\,H \mid \text{fair}\,) \cdot P(\,T \mid \text{fair}\,) \cdot P(H \mid \text{fair})$$

$$= \quad \frac{1}{2} \quad \times \quad \frac{1}{2} \quad \times \quad \frac{1}{2} \quad \times \quad \frac{1}{2} \quad = \frac{1}{16}$$

Learning from Samples with Bayes

- Bayes' Rule:

$$P(\text{fair} \mid HHTH) = \frac{P(HHTH \mid \text{fair})P(\text{fair})}{P(HHTH)}$$

$$P(HHTH) = P(HHTH, \text{fair}) + P(HHTH, \text{biased})$$

$$= P(HHTH \mid \text{fair})P(\text{fair}) + P(HHTH \mid \text{biased})P(\text{biased})$$

$$= \frac{1}{16} \times \frac{1}{2} + \left(\frac{3}{4} \times \frac{3}{4} \times \frac{1}{4} \times \frac{3}{4}\right) \times \frac{1}{2}$$

$$= \frac{1}{32} + \frac{27}{512} \qquad = \frac{16}{512} + \frac{27}{512} \qquad = \frac{43}{512}$$

Learning from Samples with Bayes

- We observe: HHTH
- Bayes' Rule:

$$P(\text{ fair} \mid HHTH\,) = \frac{P(\,HHTH \mid \text{fair}\,)P(\text{fair})}{P(HHTH)}$$

$$= \frac{\frac{1}{16}\times\frac{1}{2}}{\frac{43}{512}}$$

$$\approx 0.372$$

Learning from Samples with Bayes

- Coin flip: **fair** or **biased**?
- I promise that either:
  - The coin is fair, or
  - The coin is heads 75% of the time.
- These happen with equal probability.
- We observe: HHTH

What if no professor is around to magically supply this information?

$$P(\text{fair} \mid HHTH) \approx 0.372$$

Samples and Estimation

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

- Typical statistics problem: **estimate distribution** from samples
  - Estimate probabilities
  - Estimate parameters
  - Estimate the mean

Technique: Empirical Estimation

- Goal: Estimate some underlying quantity
- Input: dataset of samples
- Output: what you saw in data

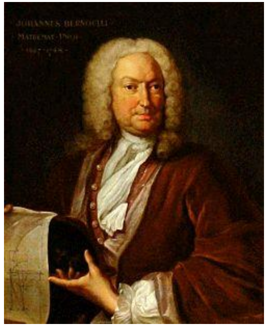
- Example:
  - Flip a coin 10 times, got 7 heads
  - Conclude: $P(\text{heads}) = 0.7$

Examples: Sample Mean

- Bernoulli distribution with parameter p
  - Models biased coin: $P(heads) = p$
  - Outcome set is $\{0,1\}$, not H/T

Named for Jacob Bernoulli…

not Johann Bernoulli,
his brother,



nor David Bernoulli,
Johann's son

Bernoulli's principle
in physics

Examples: Sample Mean

- Bernoulli distribution with parameter *p*
  - Models biased coin: $P(heads) = p$
  - Outcome set is $\{0,1\}$, not H/T

- If X is a Bernoulli random variable, then

$$\mathbb{E}[X] = 1 \cdot P(X = 1) + 0 \cdot P(X = 0)$$
$$= 1 \cdot p + 0 \cdot (1 - p)$$
$$= p$$

Examples: Sample Mean

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

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - That is, counting heads

Break & Quiz

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 $\mathbb{E}[X^2]$

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

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 $\mathrm{P}(X = a) = \dfrac{\#\text{samples taking value } a}{50}$

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.

Multinomial Distribution

- k-sided die (special case: $k = 2$ is a coin)
- Face i has probability p_i, for $i = 1, \ldots, k$
- Over n rolls, we observe face i showing up n_i times

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

Maximum Likelihood Estimate (MLE)

- Bayes' Rule:

$$P(H \mid E) = \frac{\overbrace{P(E \mid H)}^{\text{likelihood}}\overbrace{P(H)}^{\text{prior}}}{\underbrace{P(E)}_{\text{evidence}}}$$

- What if we don't have a prior?
- One solution: solve

$$\max_H P(E \mid H)$$

"maximum likelihood"

MLE for Multinomial

- Estimate $(p_1, \ldots, p_k)$ from this data $(n_1, \ldots, n_k)$

- Solve $\max_H P(E \mid H)$

- Solution (using calculus) sets $\hat{p}_i = \frac{n_i}{n}$

Regularized Estimate

- Hyperparameter $\epsilon > 0$

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

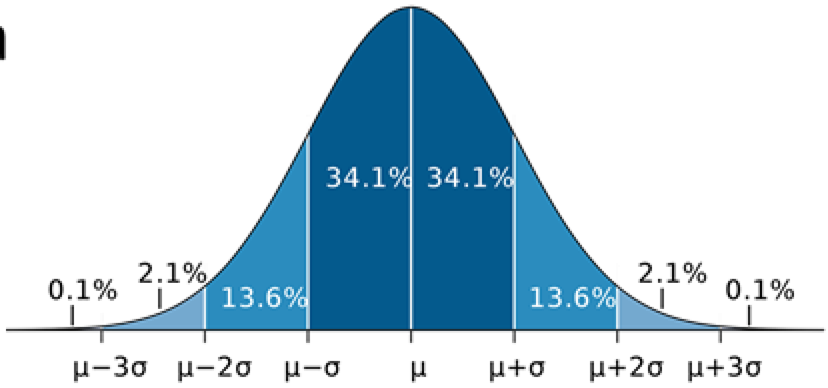
- Avoids zero when *n* is small
- Biased, but has smaller variance
- Equivalent to a specific Maximum A Posteriori (MAP) estimate, or smoothing

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

$$x_1, \ldots, x_n$$

- Estimate μ, σ^2 from this data



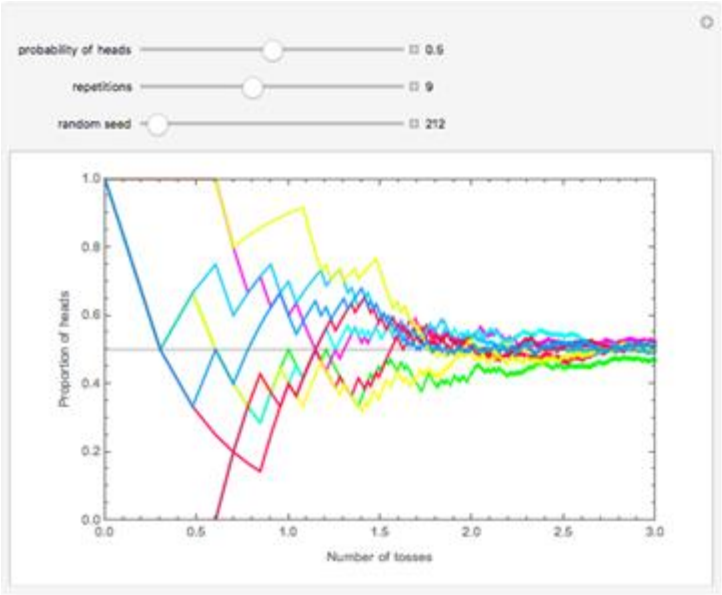
Wikipedia: Normal distribution

Estimating 1D Gaussian Parameters

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

Estimation Theory

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



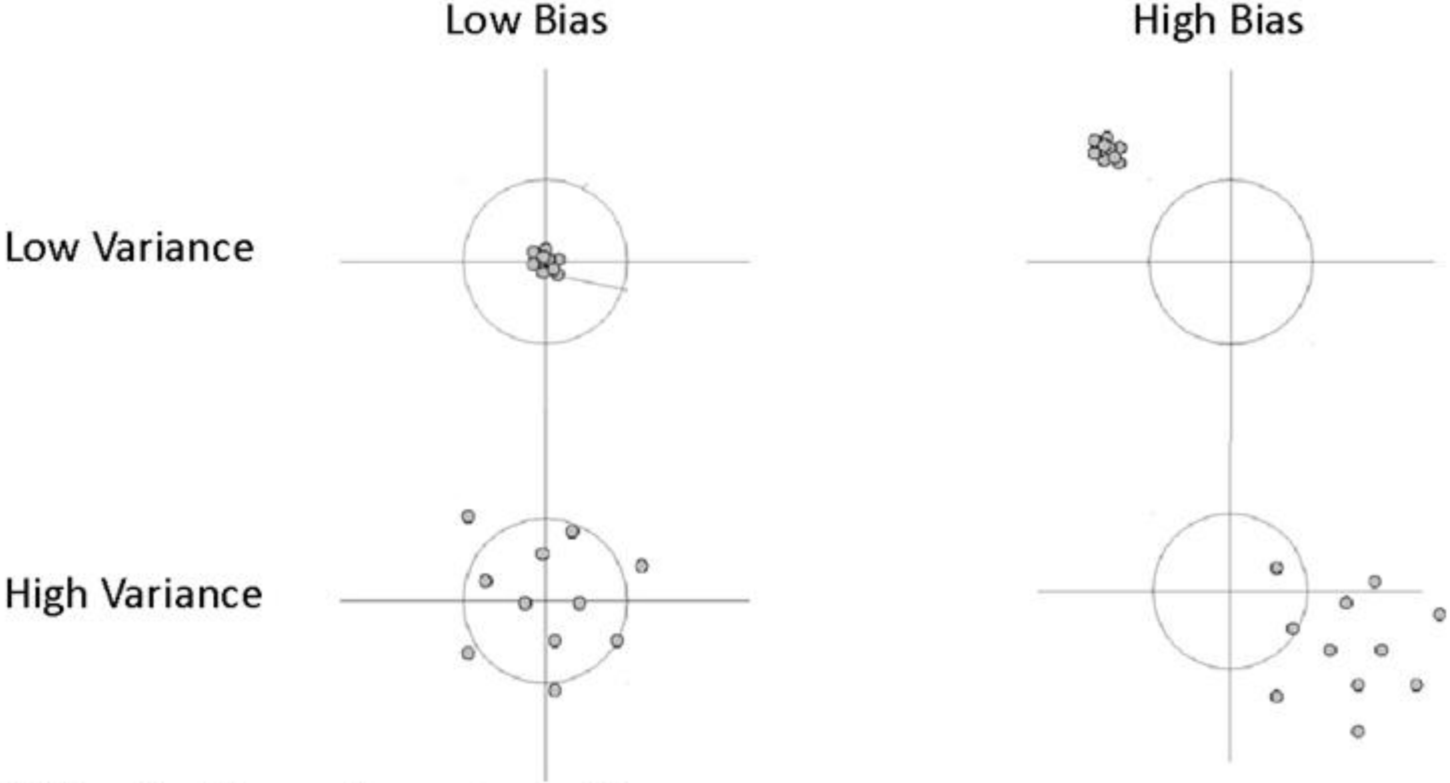
Wolfram Demo

Estimation Errors

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

Variance Bias

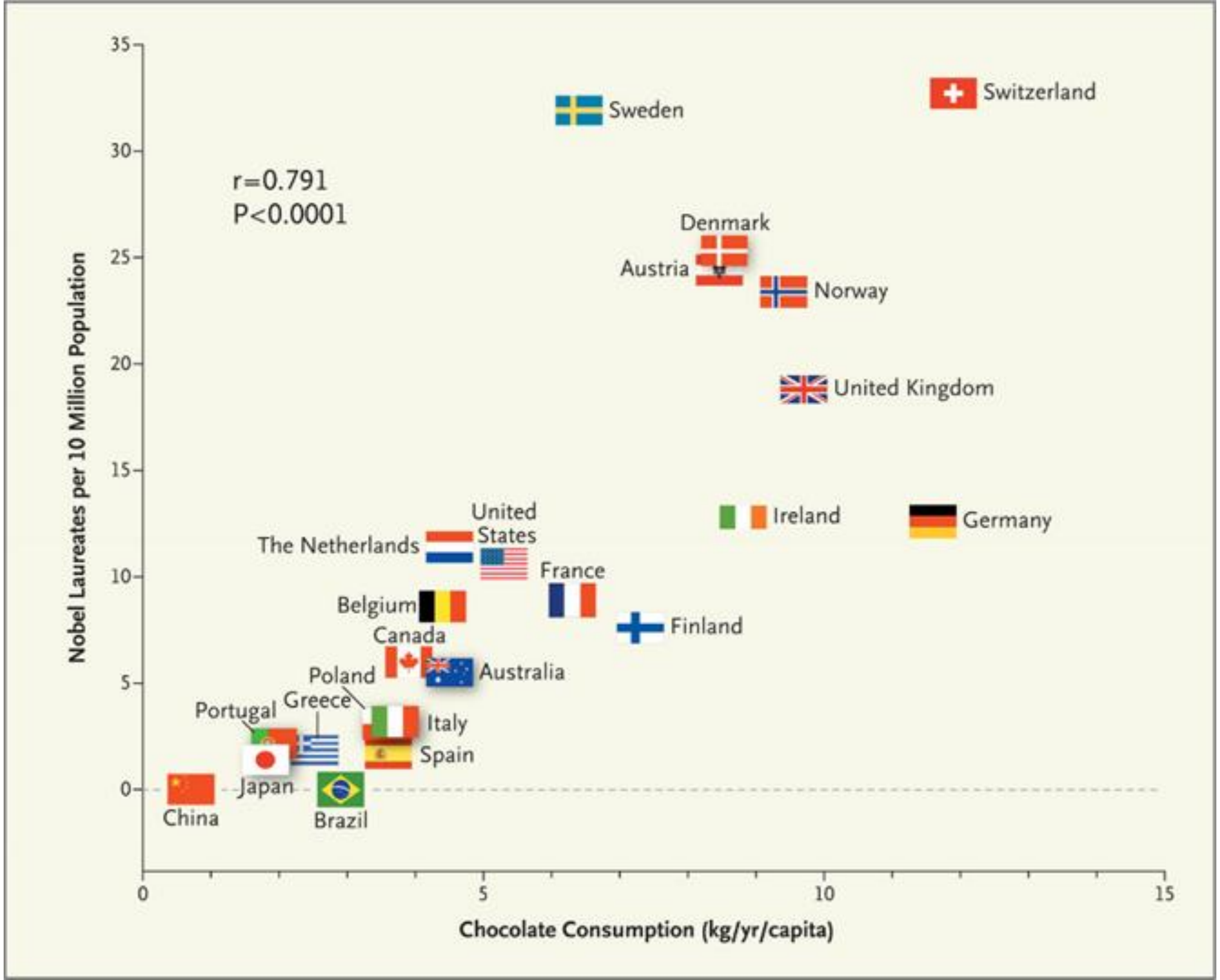
Bias / Variance



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

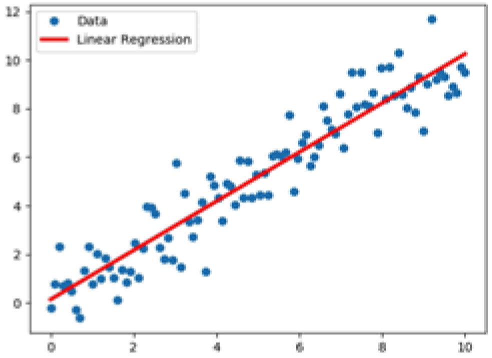


https://www.nejm.org/doi/full/10.1056/NEJMon1211064

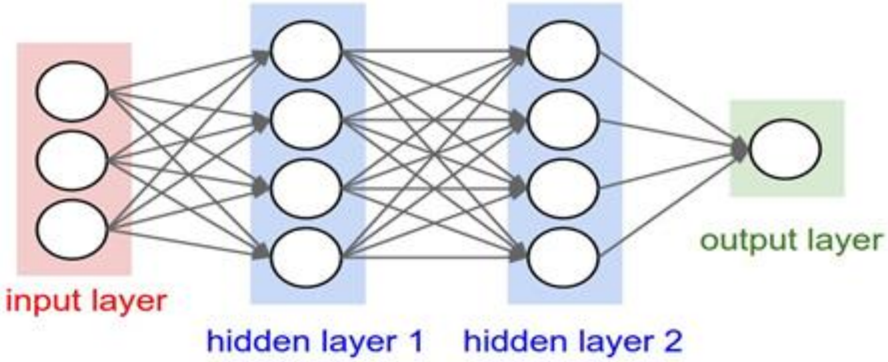
Looking Forward: Linear Algebra

Linear Algebra: What is it good for?

- Study of **linear** functions: simple, tractable
- In AI/ML: building blocks for all models
  - e.g., linear regression; part of neural networks



Hieu Tran



Stanford CS231n

Linear Algebra: What is it?

- Study of **linear** functions: simple, tractable
  - Basic linear function: $f(x) = ax + b$
  - High dimensions: $f(\vec{x}) = A\vec{x} + \vec{b}$

- Perform algebra on these systems
  – Given $\vec{y} = A\vec{x} + \vec{b}$, solve for $\vec{x}$