CS 540 Introduction to Artificial Intelligence
Statistics & Math Review
University of Wisconsin-Madison
Spring 2022
Samples and Estimation

• Usually, we don’t know the distribution $P$
  – Instead, we see a bunch of samples

• Typical statistics problem: **estimate parameters** from samples
  – Estimate probability $P(H)$
  – Estimate the mean $E[X]$
  – Estimate parameters $P_\theta(X)$
Samples and Estimation

• Typical statistics problem: estimate **parameters** from samples
  – Estimate probability $P(H)$
  – Estimate the mean $E[X]$
  – Estimate parameters $P_\theta(X)$

• Example: Bernoulli with parameter $p$
  – 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
    $$\hat{\mathbb{E}}[X] = \frac{1}{n} \sum_{i=1}^{n} x_i$$
  - No different from counting heads
Break & Quiz

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

A. 9/8
B. 15/8
C. 1.5
D. There aren’t enough samples to estimate $E[X^2]$
Break & Quiz

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

A. 9/8
B. 15/8
C. 1.5
D. There aren’t enough samples to estimate $E[X^2]$
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.
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.
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 \((p_1, \ldots, p_k)\) from this data \((n_1, \ldots, n_k)\)
Maximum Likelihood Estimate (MLE)

- The MLE of multinomial parameters \((\hat{p}_1, ..., \hat{p}_k)\)
  \[
  \hat{p}_i = \frac{n_i}{n}
  \]
- “frequency estimate”
Regularized Estimate

- Equivalent to a specific Maximum A Posteriori (MAP) estimate, or smoothing
- Hyperparameter $\epsilon > 0$
  \[
  \hat{p}_i = \frac{n_i + \epsilon}{n + k\epsilon}
  \]
- Avoids zero when $n$ is small
- Biased, but has smaller variance
Estimating 1D Gaussian Parameters

- Gaussian distribution $N(\mu, \sigma^2)$
- Observe $n$ data points from this distribution $x_1, \ldots, x_n$
- Estimate $\mu, \sigma^2$ from this data
Estimating 1D Gaussian Parameters

- Mean estimate
  \[ \hat{\mu} = \frac{x_1 + \cdots + x_n}{n} \]

- Variance estimates
  - Unbiased
    \[ s^2 = \frac{\sum_{i=1}^{n} (x_i - \hat{\mu})^2}{n - 1} \]
  - MLE
    \[ \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \hat{\mu})^2}{n} \]
Estimation Theory

• How do we know that the sample mean is a good estimate of the true mean?
  – Concentration inequalities
    \[ P(|\bar{X} - \hat{E}[X]| \geq t) \leq \exp(-2nt^2) \]
  – Law of large numbers
  – Central limit theorems, etc.