Lecture 22

- Reading: Section 9.1 of the book (but not the sub-section on confidence intervals)

- Outline:
  - Parameter estimation: bias, mean squared error, asymptotic properties
  - Maximum-likelihood estimation

Parameter Estimation
- We have some underlying unknown parameter, $\theta$
- We have $n$ random variables, $X_1, X_2, \ldots, X_n$ (often, but not always, these are i.i.d.)
- Our estimator for $\theta$ is a random variable
  \[
  \hat{\Theta} = g(X_1, X_2, \ldots X_n)
  \]
  for some function $g$
- An example: $\theta$ is the probability of heads for some coin. I toss the coin $n$ times: $X_1 \ldots X_n$ are Bernoulli random variables, with $p_{X_i}(1) = \theta$, and $p_{X_i}(0) = 1 - \theta$. A “natural” estimator is
  \[
  \hat{\Theta} = \frac{X_1 + X_2 + \ldots + X_n}{n}
  \]

Examples
- The pollster problem: estimating the proportion of the population that prefers a particular political candidate.
- Medical drug trials: we’d like to estimate the “effectiveness” of a new drug for some disease. “Effectiveness” is defined as the proportion of people with the disease whose condition improves when given the drug.
- Survival analysis: Lightbulbs from a particular manufacturer all have a lifetime that is a random variable with an exponential distribution with the same parameter $\theta$. I take 1000 lightbulbs, and measure the length of time that each lightbulb lasts. How can I estimate $\theta$?
- I’d like to model the gain in price of a particular stock each day as a normal random variable with parameters $\mu$, $\sigma^2$. I observe the gain of the stock on each day for 100 days. How can I estimate $\mu$ and $\sigma^2$?

Two Topics Covered in This Lecture
- What makes a good estimator?
  - How can we compare different estimators?
    - Bias, mean squared error, consistency
- How can we derive estimators?
  - Maximum-likelihood estimation
Bias of an Estimator

- The bias of an estimator is
  \[ E[\hat{\Theta}] - \theta \]
  Note: in general \( E[\hat{\Theta}] \) is a function of \( \theta \), so the bias is a function of \( \theta \)

- An estimator is unbiased if for all \( \theta \),
  \[ E[\hat{\Theta}] - \theta = 0 \]

An example: \( X_1, \ldots, X_n \) are i.i.d. Bernoulli random variables with parameter \( \theta \). Define our estimator to be
  \[ \hat{\Theta} = \frac{X_1 + X_2 + \ldots + X_n}{n} \]
Then \( E[\hat{\Theta}] = n\theta/n = \theta \), hence this estimator is unbiased.

Mean Squared Error of an Estimator

- The mean squared error (MSE) for an estimator is
  \[ E[(\hat{\Theta} - \theta)^2] \]

- It is always the case that
  \[ E[(\hat{\Theta} - \theta)^2] = \left( E[\hat{\Theta}] - \theta \right)^2 + \text{Var}(\hat{\Theta}) \]
  \[ \text{BIAS}^2 + \text{VARIANCE} \]

Mean Squared Error (continued)

\( X_1, \ldots, X_n \) are i.i.d. Bernoulli random variables with parameter \( \theta \).
Define our estimator to be
  \[ \hat{\Theta} = \frac{X_1 + X_2 + \ldots + X_n}{n+1} \]
Then \( E[\hat{\Theta}] = n\theta/(n+1) \), hence the bias is \( -\theta/(n+1) \), and this estimator is biased.

Mean Squared Error (continued)

\( X_1, \ldots, X_n \) are i.i.d. Bernoulli random variables with parameter \( \theta \).
Define our estimator to be
  \[ \hat{\Theta} = \frac{X_1 + X_2 + \ldots + X_n}{n} \]
This has bias \( = 0 \), and \( \text{Var}(\hat{\Theta}) = \theta(1-\theta)/n \). Hence the MSE is
  \[ \frac{\theta(1-\theta)}{n} \]
Alternatively, consider \( \hat{\Theta} = \frac{X_1 + X_2 + \ldots + X_n}{n+1} \).
The bias is \( -\theta/(n+1) \). \( \text{Var}(\hat{\Theta}) = n\theta(1-\theta)/(n+1)^2 \). MSE is
  \[ \frac{\theta^2}{(n+1)^2} + \frac{n\theta(1-\theta)}{(n+1)^2} \]
e.g., for \( n = 10 \), \( \theta = 0.1 \), MSE for the first estimator is 0.009,
MSE for the second estimator is 0.0075
Asymptotic Properties

- Often we will consider a sequence of estimators \( \hat{\Theta}_n \) for \( n = 1, 2, \ldots \). For example, define
  \[
  \hat{\Theta}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}
  \]
  where each \( X_i \) is Bernoulli with parameter \( \theta \).
- An estimator is **asymptotically unbiased**, if for all \( \theta \),
  \[
  \lim_{n \to \infty} E[\hat{\Theta}_n] = \theta
  \]
- An estimator is **consistent** if for all \( \theta \), the sequence \( \hat{\Theta}_n \) converges to \( \theta \) in probability.

Maximum-Likelihood Estimators

- We have observations \( X_1, X_2, \ldots, X_n \). We'd like to define an estimator for a parameter \( \theta \). How can we do this?
- A common case: \( X_i \)'s are i.i.d., they have a common PMF \( p_{X_i}(x_i; \theta) \) that depends on \( \theta \). We have observations \( x_1, x_2, \ldots, x_n \). The **maximum-likelihood** estimate is then
  \[
  \hat{\theta} = \text{argmax}_\theta \left( \prod_{i=1}^n p_{X_i}(x_i; \theta) \right)
  \]
- The continuous case: \( X_i \)'s are i.i.d. with PDF \( f_{X_i}(x_i; \theta) \). The **maximum-likelihood** estimate is then
  \[
  \hat{\theta} = \text{argmax}_\theta \left( \prod_{i=1}^n f_{X_i}(x_i; \theta) \right)
  \]

Example 1: Maximum-Likelihood Estimation for Bernoulli R.V.s

- \( X_1 \ldots X_n \) are Bernoulli, \( p_{X_i}(1; \theta) = \theta \), \( p_{X_i}(0; \theta) = 1 - \theta \)
- Observed values are \( x_1, x_2, \ldots, x_n \), the likelihood function is
  \[
  \prod_{i=1}^n p_{X_i}(x_i; \theta) = \theta^s (1 - \theta)^{n-s}
  \]
  where \( s = \sum_{i=1}^n x_i \)
- Maximizing this function with respect to \( \theta \) gives
  \[
  \hat{\theta} = \frac{s}{n} = \frac{\sum_{i=1}^n x_i}{n}
  \]
  Our final estimator is
  \[
  \hat{\Theta} = \frac{\sum_{i=1}^n X_i}{n}
  \]

The Log-Likelihood Function

- The maximum-likelihood estimate is
  \[
  \hat{\theta} = \text{argmax}_\theta \left( \prod_{i=1}^n p_{X_i}(x_i; \theta) \right)
  \]
- Equivalently, we can choose \( \theta \) to maximize the **log-likelihood**:
  \[
  \hat{\theta} = \text{argmax}_\theta \left( \log \prod_{i=1}^n p_{X_i}(x_i; \theta) \right) = \text{argmax}_\theta \left( \sum_{i=1}^n \log p_{X_i}(x_i; \theta) \right)
  \]
  (This can be more convenient in many cases.)
Example 2: the Mean of a Normal Distribution

- $X_1 \ldots X_n$ are i.i.d. with PDF
  \[ f_{X_i}(x_i; \theta) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma^2}\right) \]
- Given observations $x_1, x_2, \ldots, x_n$,
  \[ \sum_{i=1}^{n} \log f_{X_i}(x_i; \theta) = -n \log \left(\frac{1}{\sqrt{2\pi} \sigma} \right) - \frac{n}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2 \]
- Maximizing this function w.r.t. $\theta$ gives $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i$, hence our estimator is $\hat{\Theta} = \frac{1}{n} \sum_{i=1}^{n} X_i$

This estimator is unbiased, consistent, and has MSE equal to $\sigma^2/n$.

Example 3: Uniform Distribution

- $X_1, X_2, \ldots, X_n$ are i.i.d., where $f_{X_i}(x_i; \theta) = \frac{1}{\theta}$ for $0 \leq x_i \leq \theta$, $f_{X_i}(x_i; \theta) = 0$ otherwise.
- Given observations $x_1, x_2, \ldots, x_n$,
  \[ \prod_{i=1}^{n} f_{X_i}(x_i; \theta) = \frac{1}{\theta^n} \]
  if $0 \leq x_i \leq \theta$ for all $x_i$, otherwise $\prod_{i=1}^{n} f_{X_i}(x_i; \theta) = 0$.
- This function is maximized at $\hat{\theta} = \max_i x_i$, hence the maximum likelihood estimator is $\hat{\Theta} = \max_i X_i$

Example 4: Mean and Variance of a Normal

- $X_1 \ldots X_n$ are i.i.d. with PDF
  \[ f_{X_i}(x_i; \theta, \nu) = \frac{1}{\sqrt{2\pi\nu}} \exp\left(-\frac{(x_i - \theta)^2}{2\nu}\right) \]
- Given observations $x_1, x_2, \ldots, x_n$,
  \[ \sum_{i=1}^{n} \log f_{X_i}(x_i; \theta, \nu) = -n \log \left(\frac{1}{\sqrt{2\pi\nu}} \right) - \frac{n}{2\nu} \sum_{i=1}^{n} (x_i - \theta)^2 \]
- Maximizing this function w.r.t. $\theta, \nu$ gives
  \[ \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \hat{\nu} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\theta})^2 \]

General Properties of Maximum-Likelihood Estimators

- If $X_1 \ldots X_n$ are i.i.d., then under some mild additional assumptions, the maximum-likelihood estimator is consistent.
- Moreover, again under mild conditions, the maximum-likelihood estimator is asymptotically normal, i.e.,
  \[ \frac{\hat{\Theta}_n - \theta}{\sigma_n} \]
  approaches a normal distribution with $\mu = 0$, $\sigma^2 = 1$, where $\sigma^2_n = \text{Var}(\hat{\Theta}_n)$.
Maximum-Likelihood Estimators (the General Case)

- We have observations $X_1, X_2, \ldots, X_n$. We’d like to define an estimator for a parameter $\theta$. How can we do this?

- The general case: assume that the PMF for $X = (X_1, X_2, \ldots, X)$ is $p_X(x_1, \ldots, x_n; \theta)$. Assume that we have observations $x_1 \ldots x_n$. Then the **maximum-likelihood** estimate is

$$\hat{\theta} = \arg\max_{\theta} p_X(x_1, \ldots, x_n; \theta)$$

- The continuous case:

$$\hat{\theta} = \arg\max_{\theta} f_X(x_1, \ldots, x_n; \theta)$$

where $f_X(\cdot; \theta)$ is the PDF parameterized by $\theta$. 