Lecture 6 – Point Estimation

Last Modified: September 10, 2026

Outline:

6.1 Point Estimation

- **Definition:** Given a statistical model $\{f_\theta(x) \mid \theta \in \Theta\}$, any statistic $T = T(\underline{X})$ used for estimating $g(\theta)$ is called a **point estimator**.
  - Estimator – refers to the *function*, i.e., the random variable $T(\underline{X})$
  - Estimate – refers to a *particular value*, i.e., the realization $T(\underline{x})$
- In this lecture, we address two basic questions:
  1. How do we find estimators?
  2. How do we compare / evaluate different estimators?

6.1.1 Method of Moments

- For a sample $X_i \overset{\text{iid}}{\sim} f_\theta(x)$ for $\theta \in \Theta \subset \mathbb{R}^k$ we can define
  - **Theoretical moments:**
$$\mu_k = \mathbb{E}[X_i^k] \qquad \text{(depend on } \theta\text{)}$$
  - **Empirical moments:**
$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \qquad \text{(are statistics)}$$
- **Intuition:** A good estimation procedure should lead to a match between theoretical and empirical moments.
- In order to estimate θ, the method of moments requires to solve
$$\begin{aligned} \mu_1(\theta_1, \ldots, \theta_k) &= m_1 \\ \mu_2(\theta_1, \ldots, \theta_k) &= m_2 \\ &\vdots \\ \mu_k(\theta_1, \ldots, \theta_k) &= m_k \end{aligned}$$
  This is a system of k equations with k unknowns.
- **Example:** $X_i \overset{\text{iid}}{\sim} \mathsf{Exp}(\theta)$. Find the method of moments estimator of θ.

- In this case, $k = 1$ and theoretical first moment is

$$\mu_1 = \mathbb{E}[X_i] = \frac{1}{\theta}$$

- The empirical first moment is

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$$

- Equating these, we see that

$$\mu_1 = m_1 \quad \iff \quad \frac{1}{\theta} = \bar{X}_n \quad \iff \quad \theta = \frac{1}{\bar{X}_n}$$

- Hence, $\hat{\theta} = \frac{1}{\bar{X}_n}$ is the method of moments estimator.

- **Example:** $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$
  - $k = 2$ and the theoretical moments are

$$\mu_1 = \mathbb{E}[X_i] = \mu, \qquad \mu_2 = \mathbb{E}[X_i^2] = \mathbb{E}^2[X_i] + \mathsf{Var}(X_i) = \mu^2 + \sigma^2$$

  - Matching the theoretical and empirical moments gives:

$$\begin{array}{l} \mu_1(\mu, \sigma^2) = m_1 \\ \mu_2(\mu, \sigma^2) = m_2 \end{array} \quad \iff \quad \begin{array}{rl} \mu & = m_1 \\ \mu^2 + \sigma^2 & = m_2 \end{array} \quad \iff \quad \begin{array}{rl} \mu & = m_1 \\ \sigma^2 & = m_2 - m_1^2 \end{array}$$

  - Thus, the moment estimators are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n, \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \Big(\frac{1}{n}\sum_{i=1}^{n}\Big)^2$$

  - Letting $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n}$ denote the empirical mean, the estimator of the variance can also be expressed in terms of the empirical variance:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

    Hence, the moment estimators for the Normal model are exactly the empirical mean and variance.

- **Remark:** There is some arbitrariness in the method of moments. For example, why match the first k moments but not others? Why not match moments or some function g of the observe data?

- **Example:** $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, $\theta = \sigma^2$.
  - In this case, $\mu_1(\theta) = \mathbb{E}[X_1] = 0$ for all θ and thus "solving" the equation $\mu_1(\theta) = m_1$ is not possible in general.
  - As an alternative, one may consider the second moment:

$$\mu_2(\theta) = \mathbb{E}\big[X_i^2\big] = \sigma^2 \quad \iff \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

- But, one could also have considered matching the fourth moment:

$$\mu_4 = \mathbb{E}\big[X_1^4\big] = 3\sigma^4 \quad \iff \quad \hat{\sigma}^2 = \left(\frac{1}{3n}\sum_{i=1}^{n} X_i^4\right)^{1/2}$$

- **Example:** $X_i \overset{\text{iid}}{\sim} \mathsf{Unif}(0, \theta)$.
  - $k = 1$ and the first theoretical moment is

$$\mu_1 = \mathbb{E}[X_i] = \int_0^\theta x \frac{1}{\theta} dx = \frac{\theta}{2}$$

  - Solving for θ such that $\mu_1(\theta) = m_1$ gives the moment estimator:

$$\hat{\theta} = 2\bar{X}_n$$

    But does this estimator make sense?
  - Note that this estimator is not a function of the sufficient statistic $X_{(n)}$. So in a sense it is not using all the relevent information.
  - If $n = 3$ and $\underline{x} = (1, 24, 2)$ then $\hat{\theta} = 18$. However, we have observed 24 so we know that $\theta \geq 24$. .

6.1.2 Method of Maximum Likelihood

- **Definition:** Given a statistical model $\{f_\theta(x) \mid \theta \in \Theta\}$, assuming $\underline{X} = \underline{x}$ is observed, the **likelihood function** is given by

$$L(\theta \mid \underline{x}) = f_{\underline{X}}(\underline{x}), \quad \text{for } \theta \in \Theta$$

- The likelhood function looks at the joint distribution of the sample in a different way:
  - $f_{\underline{X}}(\underline{x}) = \prod_{i=1}^n f_\theta(x_i)$ is a function of $\underline{x}$ for fixed θ:
  - $L(\theta \mid \underline{x})$ is a function is a function of θ for fixed $\underline{x}$
- In the discrete case:
  - The likelihood is the probability of observing $\underline{x}$ when θ is the true parameter, i.e,

$$L(\theta \mid \underline{x}) = f_{\underline{X}}(\underline{x}) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$

  - Comparing probabilities, we see that

$$\frac{L(\theta_1 \mid \underline{x})}{L(\theta_2 \mid \underline{x})} > 1 \quad \implies \quad \begin{array}{l}\text{the probability of observing } \underline{x} \text{ is} \\ \text{larger if } \theta = \theta_1 \text{ than if } \theta = \theta_2\end{array}$$

    Sometime we say that θ_1 is *more likely* than θ_2.
- **Remark:** The statement *more likely* is not the same as the statement *more probable*. Note that θ is not a random variable so there is no probability to assign to these events. See https://en.wikipedia.org/wiki/Likelihood_function#Historical_remarks for more information on the history of this terminology.

- **Intuition:** The likelihood function $L(\theta \mid \underline{x})$ induces an **ordering** among the θ. We prefer values of θ which make the realization $\underline{x}$ more probable. Thus, we only care about the ordering, not the precise value of L.

- **Likelihood principle:** If $\underline{x}$ and $\underline{x}^*$ are such that $L(\theta \mid \underline{x})$ is proportional to $L(\theta \mid \underline{x}^*)$, i.e. there exists a constant $c(\underline{x}, \underline{x}^*)$ such that

$$L(\theta \mid \underline{x}) = c(\underline{x}, \underline{x}^*) L(\theta \mid \underline{x}^*), \quad \text{for all } \theta \in \Theta$$

  the inference based on $\underline{x}$ and $\underline{x}^*$ should coincide. Note that under this conditions, these likelihood functions leads to the same ordering on θ.

- **Example.** Consider $X_i \overset{i.i.d.}{\sim} \mathsf{Po}(\theta)$. Suppose $\underline{x} = (1, 2)$ and $\underline{x}^* = (0, 3)$. Then

$$L(\theta \mid \underline{x}) = e^{-2\theta} \frac{\theta^3}{1!2!}, \quad \text{and} \quad L(\theta \mid \underline{x}^*) = e^{-2\theta} \frac{\theta^3}{0!3!}.$$

  So $L(\theta \mid \underline{x}) = 3L(\theta \mid \underline{x}^*)$, and thus the likelihood principle implies that the inference based on $\underline{x}$ and $\underline{x}^*$ should coincide.

- **Definition:** Given a statistical model $\{f_\theta(x) \mid \theta \in \Theta\}$ and random sample $\underline{X}$ with realization $\underline{x}$ the **maximum likelihood** estimate is $\hat{\theta} = \hat{\theta}(\underline{x})$ such that

$$L(\hat{\theta} \mid \underline{x}) = \max_{\theta \in \Theta} L(\theta \mid \underline{x})$$

  Then, the **Maximum Likelihood Estimator** (MLE) is $\hat{\theta} = \hat{\theta}(\underline{X})$. Note that this is a random variable because it a function of the sample.

- **Finding the MLE:** Assume that $\Theta \subset \mathbb{R}$ and the likelihood function is smooth.

  - If the likelihood has unique maximizer $\hat{\theta}$, then it satisfies the derivative conditions:

$$\frac{\partial}{\partial \theta} L(\theta \mid \underline{x})\Big|_{\theta=\hat{\theta}} = 0, \quad \text{and} \quad \frac{\partial^2}{\partial \theta^2} L(\theta \mid \underline{x})\Big|_{\theta=\hat{\theta}} < 0$$

  - It is often convenient to work with the **log-likelihood function:**

$$\ell(\theta \mid \underline{x}) = \log L(\theta \mid \underline{x})$$

  - Because the logarithm is strictly increasing, it does not affect the location of the maximum, and the derivative conditions can be stated equivalently as

$$\frac{\partial}{\partial \theta} \ell(\theta \mid \underline{x})\Big|_{\theta=\hat{\theta}} = 0, \quad \text{and} \quad \frac{\partial^2}{\partial \theta^2} \ell(\theta \mid \underline{x})\Big|_{\theta=\hat{\theta}} < 0$$

- **Example:** $X_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, $\Theta = [0, 1]$

  - The likelihood is

$$L(\theta \mid \underline{x}) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}.$$

  - First we consider the case $\sum_{i=1}^n x_i = n$ (all 1's). The likelihood is

$$L(1, \ldots, 1 \mid \underline{x}) = \theta^n$$

    which is increasing in θ. Thus, the maximum is achieved at $\hat{\theta} = 1$.

- Next, we consider the case $\sum_{i=1}^n x_i = 0$ (all 0's). The likelihood is
$$L(0, \ldots, 0 \mid \underline{x}) = (1-\theta)^n$$
which is decreasing in θ. Thus, the maximum is achieved at $\hat{\theta} = 0$.
- Finally, we consider the case $0 < \sum_{i=1}^n x_i < n$. Noting that
$$L(0 \mid \underline{x}) = L(1 \mid \underline{x}) = 0 \quad \text{and} \quad L(\theta \mid \underline{x}) > 0 \quad \text{for } 0 < \theta < 1$$
we conclude that the maximum does no occur at the boundary points and we can restrict our attention to the interval $\theta \in (0, 1)$.
- To find the maximizer, it is convenient to work the with logarithm of likelihood function (a.k.a. the log likelihood), which is given by
$$\ell(\theta \mid \underline{x}) = \sum_{i=1}^n \log f_\theta(x_i) = \sum_{i=1}^n x_i \log(\theta) + \left(n - \sum_{i=1}^n x_i\right) \log(1-\theta).$$
- This function is differentiable in $\theta \in (0, 1)$ with
$$\begin{aligned}\frac{\partial}{\partial\theta}\ell(\theta \mid \underline{x}) &= \frac{1}{\theta}\sum_{i=1}^n x_i - \frac{1}{1-\theta}\Big(n - \sum_{i=1}^n x_i\Big) \\ &= \frac{1}{\theta(1-\theta)}\left[(1-\theta)\sum_{i=1}^n x_i - \theta\Big(n - \sum_{i=1}^n x_i\Big)\right] \\ &= \frac{n}{\theta(1-\theta)}\left[\frac{1}{n}\sum_{i=1}^n x_i - \theta\right]\end{aligned}$$
- The derivative has a unique root at $\theta = \bar{x}_n$. Since the derivative is strictly negative $\theta < \bar{x}_n$ and strictly positive for $\theta > \bar{x}_n$ we conclude that this root is the unique global maximum, and thus the maximum likelihood estimate is $\hat{\theta} = \bar{x}_n$.
- Note that an alternative approach to verifying that the root of the derivative corresponds to a maximum is to the consider the second derivative;
$$\frac{\partial^2}{\partial\theta^2}\ell(\theta \mid \underline{x}) = -\frac{1}{\theta^2}\sum_{i=1}^n x_i - \frac{1}{(1-\theta)^2}\left(n - \sum_{i=1}^n x_i\right) < 0, \quad \text{for all } \theta,$$
Since the second derivative is strictly negative for all θ, we conclude that the root corresponds to the unique global maximum.
- Combining all three cases, we see that the MLE is given by the sample mean:
$$\hat{\theta} = \bar{X}_n \quad \text{is the MLE}$$

- **Example:** $X_i \overset{\text{iid}}{\sim} \mathsf{Unif}(0, \theta)$, $\Theta = \mathbb{R}_+$
  - The likelihood is
  $$L(\theta \mid \underline{x}) = \frac{1}{\theta^n}\mathbf{1}_{x_{(n)} \le \theta}$$
  - This function is not differentiable in θ because there is a discontinuity $\theta = x_{(n)}$
  - However the maximum can be determined by inspection:
    1. $L(\theta \mid \underline{x})$ is zero for $\theta < x_{(n)}$
    2. $L(\theta \mid \underline{x})$ is strictly positive an decreasing in θ for $\theta \ge x_{(n)}$

    Thus, the unique maximum $\hat{\theta} = X_{(n)}$ is the MLE.