3 Maximum likelihood estimation

3.1 Overview

So far we have been constructing estimators by informal reasoning. Maximum likelihood estimation (MLE) gives us a **systematic** way to construct estimators for any parametric family. Recall that the MLE is defined as follows:

Definition 3.1. Given data $X_1, \dots, X_n$ drawn i.i.d. from a density f_θ, where θ is unknown, the **likelihood function** is

$$L(\theta;\, X_{1:n}) = \prod_{i=1}^{n} f_\theta(X_i).$$

The **log-likelihood function** is $\ell(\theta;\, X_{1:n}) = \ln L(\theta;\, X_{1:n})$. The **maximum likelihood estimator** (MLE) is the value of θ that maximizes L (equivalently, ℓ):

$$\hat{\theta}_n = \arg\max_\theta \ell(\theta;\, X_{1:n}).$$

To construct the MLE and then analyze its statistical properties, we take the following steps:

1. Compute the (log-)likelihood function;
2. Compute the maximum likelihood estimator by maximizing it;
3. Derive the *exact* sampling distribution of the estimator;
4. Derive the mean-squared error (MSE) of the estimator and identify its statistical properties (biased? consistent?).

Step 1 requires fluency with algebraic manipulations, especially log and exponent properties. Step 2 requires fluency in derivative rules. Step 3 requires fluency with change-of-variables and sums and averages of iid random variables. As such, these problems synthesize a lot of the technical skills needed in future statistical theory courses. Furthermore, the steps are cumulative: you cannot do Step 3 correctly if you muck up Steps 1 and 2.

3.2 Worked examples

Example 3.1 (Exponential distribution)**.** Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\theta)$, with density $f_\theta(x) = \theta e^{-\theta x}$ for $x > 0$.

Step 1: Log-likelihood.

$$L(\theta;\, X_{1:n}) = \prod_{i=1}^{n} \theta e^{-\theta X_i} = \theta^n e^{-\theta \sum_{i=1}^n X_i} = \theta^n e^{-\theta S_n},$$

where $S_n = X_1 + \cdots + X_n$. Taking the log:

$$\ell(\theta;\, X_{1:n}) = n \ln \theta - \theta S_n.$$

Step 2: MLE.
Setting the derivative to zero:

$$\ell'(\theta) = \frac{n}{\theta} - S_n = 0 \implies \hat{\theta}_n = \frac{n}{S_n} = \frac{1}{\bar{X}_n}.$$

Since $\ell''(\theta) = -n/\theta^2 < 0$, the log-likelihood is concave, confirming this is a maximum.

Step 3: Sampling distribution.
Since $\text{Exp}(\theta) = \text{Gamma}(1, \theta)$, by the MGF argument for sums of iid gammas,

$$\bar{X}_n \sim \text{Gamma}(n, n\theta).$$

Since $\hat{\theta}_n = 1/\bar{X}_n$, the sampling distribution of the MLE is the inverse gamma:

$$\hat{\theta}_n \sim \text{IG}(n, n\theta).$$

Read more about inverse gamma in the distribution familes part of the website.

Step 4: Bias, variance, and MSE.
To compute moments, let $Y = \bar{X}_n \sim \text{Gamma}(n, n\theta)$ and apply LOTUS with $g(Y) = Y^{-k}$:

$$\mathbb{E}[\hat{\theta}_n^k] = \int_0^\infty y^{-k} \cdot \frac{(n\theta)^n}{\Gamma(n)} y^{n-1} e^{-n\theta y}\, dy = \frac{(n\theta)^n}{\Gamma(n)} \cdot \frac{\Gamma(n-k)}{(n\theta)^{n-k}} = (n\theta)^k \frac{\Gamma(n-k)}{\Gamma(n)}.$$

Using $\Gamma(n) = (n-1)\Gamma(n-1)$:

$$\mathbb{E}[\hat{\theta}_n] = n\theta \cdot \frac{\Gamma(n-1)}{\Gamma(n)} = \frac{n}{n-1}\,\theta,$$

$$\mathbb{E}[\hat{\theta}_n^2] = (n\theta)^2 \cdot \frac{\Gamma(n-2)}{\Gamma(n)} = \frac{n^2}{(n-1)(n-2)}\,\theta^2.$$

Therefore:

$$\text{bias}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta = \frac{\theta}{n-1} > 0.$$

The estimator is *biased upward*, but the bias vanishes as $n \to \infty$.

$$\text{var}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n^2] - \left(\mathbb{E}[\hat{\theta}_n]\right)^2 = \frac{n^2\theta^2}{(n-1)(n-2)} - \frac{n^2\theta^2}{(n-1)^2} = \frac{n^2\theta^2}{(n-1)^2(n-2)}.$$

Both $\text{bias}^2(\hat{\theta}_n)$ and $\text{var}(\hat{\theta}_n)$ go to zero as $n \to \infty$, so $\text{MSE}(\hat{\theta}_n) \to 0$: the estimator is **consistent**.

Example 3.2 (Uniform distribution)**.** Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$, with density

$$f_\theta(x) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta \\ 0, & \text{otherwise.} \end{cases}$$

Before doing any calculations, let us visualize $f_\theta(x)$ for two different values of θ; see Figure 30.

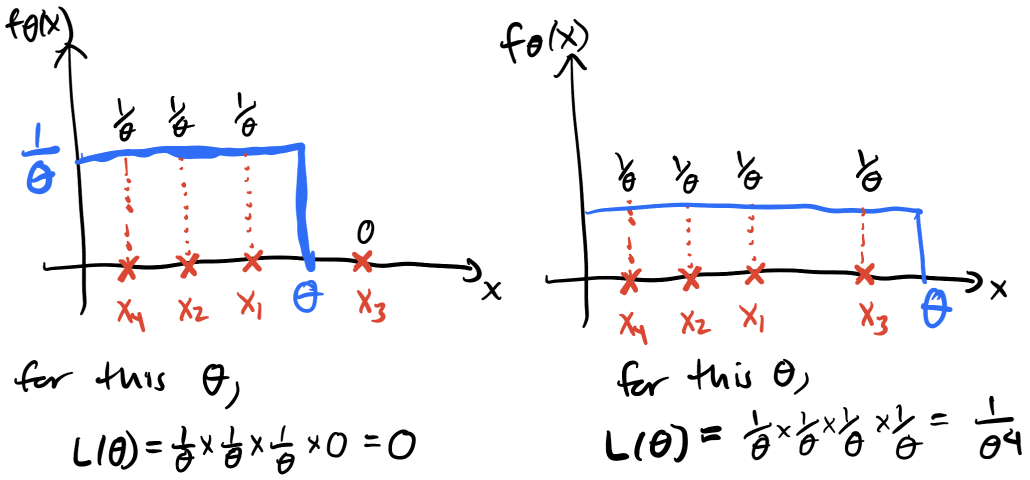

Figure 30: Density $f_\theta(x)$ for two different values of θ, with the data X_1, X_2, X_3, X_4 marked on the x-axis.

For both the left- and righthand plots, the data $X_1, \ldots, X_n$ stays the same: the red x's on the x-axis. Now, recall that the likelihood is the product $L(\theta;\, X_{1:n}) = \prod_{i=1}^n f_\theta(X_i)$. On the plots, this corresponds to the product of the heights of the vertical dotted lines. In the lefthand plot, $\theta < X_3$, so $f_\theta(X_3) = 0$, and the whole product is then zero, so $L(\theta) = 0$. This means the data could not have possibly been drawn from Unif$(0,\, \theta)$ for the particular value of θ in the lefthand plot. Indeed, if we *did* have $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0,\, \theta)$ for that θ, then all the X_i's would be forced to lie to the left of θ. In the righthand plot, the value of θ *does* contain all of the X_i, so the product of the $f_\theta(X_i)$'s is $L(\theta) = (1/\theta)^n$, with $n = 4$.

Step 1: Likelihood Let us now generalize the situation in the figure to arbitrary n. If θ is smaller than any one of the X_i's, then $L(\theta) = 0$. If θ is larger than all of the X_i's, then $L(\theta) = 1/\theta^n$. We can equivalently write this in the following form:

$$L(\theta;\, X_{1:n}) = \begin{cases} 0, & \theta < \max(X_1, \ldots, X_n), \\ \dfrac{1}{\theta^n}, & \theta > \max(X_1, \ldots, X_n). \end{cases} \tag{29}$$

Step 2: MLE. The function $L(\theta)$ is plotted in Figure 31. We see by inspection that the largest value of $L(\theta)$ is achieved at $\theta = \max(X_1, \ldots, X_n)$. This must be the MLE. So

$$\hat{\theta}_n = \max(X_1, \ldots, X_n).$$

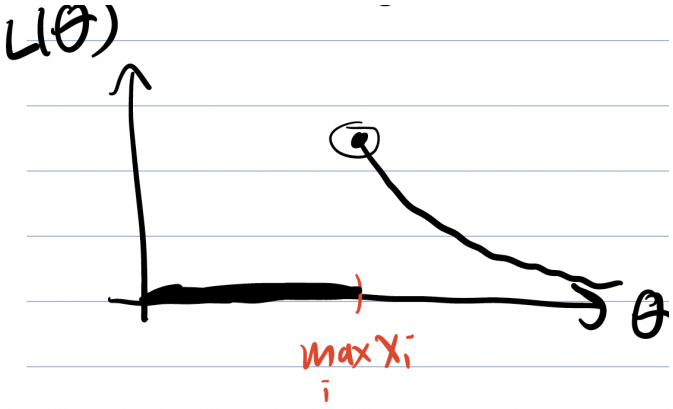

Figure 31: The likelihood $L(\theta)$ from (29), for the uniform example.

Step 3: Sampling distribution.

The range of $\hat{\theta}_n$ is $[0, \theta]$. For $0 < t < \theta$:

$$F_{\hat{\theta}_n}(t) = P\big(\max(X_1, \dots, X_n) \le t\big) = P(\{X_1 \le t\} \cup \{X_2 \le t\} \cup \cdots \cup \{X_n \le t\})$$
$$= P(X_1 \le t)^n = \left(\frac{t}{\theta}\right)^n.$$

Differentiating, the pdf of $\hat{\theta}_n$ is:

$$f_{\hat{\theta}_n}(t) = \frac{n\, t^{n-1}}{\theta^n}, \quad 0 < t < \theta.$$

Remark 3.1. When $\theta = 1$, this pdf is a special case of the Beta distribution family , which we will see next lecture when talking about Bayesian inference.

Step 4: Bias, variance, and MSE.

$$\mathbb{E}[\hat{\theta}_n] = \int_0^\theta t \cdot \frac{n\, t^{n-1}}{\theta^n}\, dt = \frac{n}{\theta^n} \cdot \frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\,\theta.$$

So $\text{bias}(\hat{\theta}_n) = \frac{n}{n+1}\,\theta - \theta = -\frac{\theta}{n+1} < 0$. The estimator is *biased downward* (which makes sense intuitively, since $\max(X_i) \le \theta$ always).

$$\mathbb{E}[\hat{\theta}_n^2] = \int_0^\theta t^2 \cdot \frac{n\, t^{n-1}}{\theta^n}\, dt = \frac{n}{\theta^n} \cdot \frac{\theta^{n+2}}{n+2} = \frac{n}{n+2}\,\theta^2.$$

$$\text{var}(\hat{\theta}_n) = \frac{n}{n+2}\,\theta^2 - \left(\frac{n}{n+1}\right)^2 \theta^2 = \frac{n\,\theta^2}{(n+1)^2(n+2)}.$$

Since both $\text{bias}^2(\hat{\theta}_n) \to 0$ and $\text{var}(\hat{\theta}_n) \to 0$ as $n \to \infty$, we have $\text{MSE}(\hat{\theta}_n) \to 0$: the estimator is **consistent**.