8.2.3 Maximum Likelihood Estimation

So far, we have discussed estimating the mean and variance of a distribution. Our methods have been somewhat ad hoc. More specifically, it is not clear how we can estimate other parameters. We now would like to talk about a systematic way of parameter estimation. Specifically, we would like to introduce an estimation method, called *maximum likelihood estimation* (MLE). To give you the idea behind MLE let us look at an example.

Example 8.7

I have a bag that contains 3 balls. Each ball is either red or blue, but I have no information in addition to this. Thus, the number of blue balls, call it θ, might be 0, 1, 2, or 3. I am allowed to choose 4 balls at random from the bag with replacement. We define the random variables X_1, X_2, X_3, and X_4 as follows

$$X_i = \begin{cases} 1 & \text{if the } i\text{th chosen ball is blue} \\ & \\ 0 & \text{if the } i\text{th chosen ball is red} \end{cases}$$

Note that X_i's are i.i.d. and $X_i \sim Bernoulli(\frac{\theta}{3})$. After doing my experiment, I observe the following values for X_i's.

$$x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1.$$

Thus, I observe 3 blue balls and 1 red balls.

1. For each possible value of θ, find the probability of the observed sample, $(x_1, x_2, x_3, x_4) = (1, 0, 1, 1)$.
2. For which value of θ is the probability of the observed sample is the largest?

Solution

Since $X_i \sim Bernoulli(\frac{\theta}{3})$, we have

$$P_{X_i}(x) = \begin{cases} \frac{\theta}{3} & \text{for } x = 1 \\ \\ 1 - \frac{\theta}{3} & \text{for } x = 0 \end{cases}$$

Since X_i's are independent, the joint PMF of X_1, X_2, X_3, and X_4 can be written as

$$P_{X_1X_2X_3X_4}(x_1, x_2, x_3, x_4) = P_{X_1}(x_1)P_{X_2}(x_2)P_{X_3}(x_3)P_{X_4}(x_4)$$

Therefore,

$$\begin{aligned} P_{X_1X_2X_3X_4}(1,0,1,1) &= \frac{\theta}{3} \cdot \left(1 - \frac{\theta}{3}\right) \cdot \frac{\theta}{3} \cdot \frac{\theta}{3} \\ &= \left(\frac{\theta}{3}\right)^3 \left(1 - \frac{\theta}{3}\right). \end{aligned}$$

Note that the joint PMF depends on θ, so we write it as $P_{X_1X_2X_3X_4}(x_1, x_2, x_3, x_4; \theta)$. We obtain the values given in Table 8.1 for the probability of $(1, 0, 1, 1)$.

θ	$P_{X_1X_2X_3X_4}(1,0,1,1;\theta)$
0	0
1	0.0247
2	0.0988
3	0

Table 8.1: Values of $P_{X_1X_2X_3X_4}(1,0,1,1;\theta)$ for Example 8.1

The probability of observed sample for $\theta = 0$ and $\theta = 3$ is zero. This makes sense because our sample included both red and blue balls. From the table we see that the probability of the observed data is maximized for $\theta = 2$. This means that the observed data is most likely to occur for $\theta = 2$. For this reason, we may choose $\hat{\theta} = 2$ as our estimate of θ. This is called the maximum likelihood estimate (MLE) of θ.

The above example gives us the idea behind the maximum likelihood estimation. Here, we introduce this method formally. To do so, we first define the **likelihood** function. Let X_1, X_2, X_3, ..., X_n be a random sample from a distribution with a parameter θ (In general, θ might be a vector, $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$.) Suppose that x_1, x_2,

x_3, ..., x_n are the observed values of X_1, X_2, X_3, ..., X_n. If X_i's are discrete random variables, we define the *likelihood* function as the probability of the observed sample sample as a function of θ:

$$\begin{aligned} L(x_1, x_2, \cdots, x_n; \theta) &= P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n; \theta) \\ &= P_{X_1 X_2 \cdots X_n}(x_1, x_2, \cdots, x_n; \theta). \end{aligned}$$

To get a more compact formula, we may use the vector notation, $\mathbf{X} = (X_1, X_2, \cdots, X_n)$. Thus, we may write

$$L(\mathbf{x}; \theta) = P_{\mathbf{X}}(\mathbf{x}; \theta).$$

If X_1, X_2, X_3, ..., X_n are jointly continuous, we use the joint PDF instead of the joint PMF. Thus, the likelihood is defined by

$$L(x_1, x_2, \cdots, x_n; \theta) = f_{X_1 X_2 \cdots X_n}(x_1, x_2, \cdots, x_n; \theta).$$

Let X_1, X_2, X_3, ..., X_n be a random sample from a distribution with a parameter θ. Suppose that we have observed $X_1 = x_1$, $X_2 = x_2$, $\cdots$, $X_n = x_n$.

1. If X_i's are discrete, then the **likelihood function** is defined as

$$L(x_1, x_2, \cdots, x_n; \theta) = P_{X_1 X_2 \cdots X_n}(x_1, x_2, \cdots, x_n; \theta).$$

2. If X_i's are jointly continuous, then the likelihood function is defined as

$$L(x_1, x_2, \cdots, x_n; \theta) = f_{X_1 X_2 \cdots X_n}(x_1, x_2, \cdots, x_n; \theta).$$

In some problems, it is easier to work with the **log likelihood function** given by

$$\ln L(x_1, x_2, \cdots, x_n; \theta).$$

Example 8.8

For the following random samples, find the likelihood function:

1. $X_i \sim Binomial(3, \theta)$, and we have observed $(x_1, x_2, x_3, x_4) = (1, 3, 2, 2)$.

2. $X_i \sim Exponential(\theta)$ and we have observed $(x_1, x_2, x_3, x_4) = (1.23, 3.32, 1.98, 2.12)$.

Solution

Remember that when we have a random sample, X_i's are i.i.d., so we can obtain the joint PMF and PDF by multiplying the marginal (individual) PMFs and PDFs.

1. If $X_i \sim Binomial(3, \theta)$, then

$$P_{X_i}(x;\theta) = \binom{3}{x}\theta^x(1-\theta)^{3-x}$$

Thus,

$$\begin{aligned} L(x_1, x_2, x_3, x_4; \theta) &= P_{X_1 X_2 X_3 X_4}(x_1, x_2, x_3, x_4; \theta) \\ &= P_{X_1}(x_1;\theta)P_{X_2}(x_2;\theta)P_{X_3}(x_3;\theta)P_{X_4}(x_4;\theta) \\ &= \binom{3}{x_1}\binom{3}{x_2}\binom{3}{x_3}\binom{3}{x_4}\theta^{x_1+x_2+x_3+x_4}(1-\theta)^{12-(x_1+x_2+x_3+x_4)}. \end{aligned}$$

Since we have observed $(x_1, x_2, x_3, x_4) = (1, 3, 2, 2)$, we have

$$\begin{aligned} L(1, 3, 2, 2; \theta) &= \binom{3}{1}\binom{3}{3}\binom{3}{2}\binom{3}{2}\theta^8(1-\theta)^4 \\ &= 27 \qquad \theta^8(1-\theta)^4. \end{aligned}$$

2. If $X_i \sim Exponential(\theta)$, then

$$f_{X_i}(x;\theta) = \theta e^{-\theta x}u(x),$$

where $u(x)$ is the unit step function, i.e., $u(x) = 1$ for $x \geq 0$ and $u(x) = 0$ for $x < 0$. Thus, for $x_i \geq 0$, we can write

$$\begin{aligned} L(x_1, x_2, x_3, x_4; \theta) &= f_{X_1 X_2 X_3 X_4}(x_1, x_2, x_3, x_4; \theta) \\ &= f_{X_1}(x_1;\theta)f_{X_2}(x_2;\theta)f_{X_3}(x_3;\theta)f_{X_4}(x_4;\theta) \\ &= \theta^4 e^{-(x_1+x_2+x_3+x_4)\theta}. \end{aligned}$$

Since we have observed $(x_1, x_2, x_3, x_4) = (1.23, 3.32, 1.98, 2.12)$, we have

$$L(1.23, 3.32, 1.98, 2.12; \theta) = \theta^4 e^{-8.65\theta}.$$

Now that we have defined the likelihood function, we are ready to define maximum likelihood estimation. Let X_1, X_2, X_3, $\ldots$, X_n be a random sample from a distribution with a parameter θ. Suppose that we have observed $X_1 = x_1$, $X_2 = x_2$, $\cdots$, $X_n = x_n$. The maximum likelihood estimate of θ, shown by $\hat{\theta}_{ML}$ is the value that maximizes the likelihood function

$$L(x_1, x_2, \cdots, x_n; \theta).$$

Figure 8.1 illustrates finding the maximum likelihood estimate as the maximizing value of θ for the likelihood function. There are two cases shown in the figure: In the first graph, θ is a discrete-valued parameter, such as the one in Example 8.7 . In the second one, θ is a continuous-valued parameter, such as the ones in Example 8.8. In both cases, the maximum likelihood estimate of θ is the value that maximizes the likelihood function.

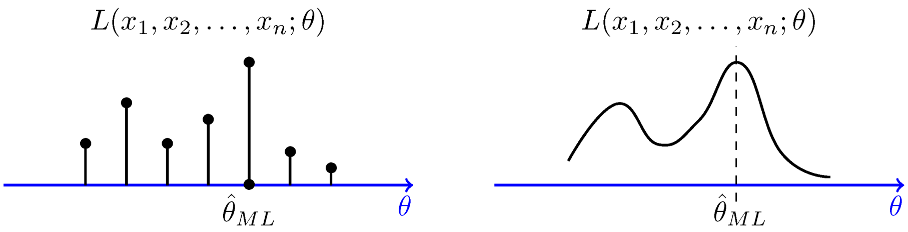

Figure 8.1 - The maximum likelihood estimate for θ.

Let us find the maximum likelihood estimates for the observations of Example 8.8.

Example 8.9

For the following random samples, find the maximum likelihood estimate of θ:

1. $X_i \sim Binomial(3, \theta)$, and we have observed $(x_1, x_2, x_3, x_4) = (1, 3, 2, 2)$.
2. $X_i \sim Exponential(\theta)$ and we have observed $(x_1, x_2, x_3, x_4) = (1.23, 3.32, 1.98, 2.12)$.

Solution

1. In Example 8.8., we found the likelihood function as

$$L(1, 3, 2, 2; \theta) = 27 \qquad \theta^8(1-\theta)^4.$$

To find the value of θ that maximizes the likelihood function, we can take the derivative and set it to zero. We have

$$\frac{dL(1, 3, 2, 2; \theta)}{d\theta} = 27\Big[\qquad 8\theta^7(1-\theta)^4 - 4\theta^8(1-\theta)^3\Big].$$

Thus, we obtain

$$\hat{\theta}_{ML} = \frac{2}{3}.$$

2. In Example 8.8., we found the likelihood function as

   $$L(1.23, 3.32, 1.98, 2.12; \theta) = \theta^4 e^{-8.65\theta}.$$

   Here, it is easier to work with the log likelihood function, $\ln L(1.23, 3.32, 1.98, 2.12; \theta)$. Specifically,

   $$\ln L(1.23, 3.32, 1.98, 2.12; \theta) = 4 \ln \theta - 8.65\theta.$$

   By differentiating, we obtain

   $$\frac{4}{\theta} - 8.65 = 0,$$

   which results in

   $$\hat{\theta}_{ML} = 0.46$$

It is worth noting that technically, we need to look at the second derivatives and endpoints to make sure that the values that we obtained above are the maximizing values. For this example, it turns out that the obtained values are indeed the maximizing values.

Note that the value of the maximum likelihood estimate is a function of the observed data. Thus, as any other estimator, the maximum likelihood estimator (MLE), shown by $\hat{\Theta}_{ML}$ is indeed a random variable. The MLE estimates $\hat{\theta}_{ML}$ that we found above were the values of the random variable $\hat{\Theta}_{ML}$ for the specified observed d

The Maximum Likelihood Estimator (MLE)

Let X_1, X_2, X_3, ..., X_n be a random sample from a distribution with a parameter θ. Given that we have observed $X_1 = x_1$, $X_2 = x_2$, $\cdots$, $X_n = x_n$, a maximum likelihood estimate of θ, shown by $\hat{\theta}_{ML}$ is a value of θ that maximizes the likelihood function

$$L(x_1, x_2, \cdots, x_n; \theta).$$

A maximum likelihood estimator (MLE) of the parameter θ, shown by $\hat{\Theta}_{ML}$ is a random variable $\hat{\Theta}_{ML} = \hat{\Theta}_{ML}(X_1, X_2, \cdots, X_n)$ whose value when $X_1 = x_1$, $X_2 = x_2$, $\cdots$, $X_n = x_n$ is given by $\hat{\theta}_{ML}$.

Example 8.10

For the following examples, find the maximum likelihood estimator (MLE) of θ:

1. $X_i \sim Binomial(m, \theta)$, and we have observed X_1, X_2, X_3, ..., X_n.
2. $X_i \sim Exponential(\theta)$ and we have observed X_1, X_2, X_3, ..., X_n.

Solution

1. Similar to our calculation in Example 8.8., for the observed values of $X_1 = x_1$, $X_2 = x_2$, $\cdots$, $X_n = x_n$, the likelihood function is given by

$$\begin{aligned} L(x_1, x_2, \cdots, x_n; \theta) &= f_{X_1 X_2 \cdots X_n}(x_1, x_2, \cdots, x_n; \theta) \\ &= \prod_{i=1}^{n} f_{X_i}(x_i; \theta) \\ &= \prod_{i=1}^{n} \binom{m}{x_i} \theta^{x_i} (1-\theta)^{m-x_i} \\ &= \left[\prod_{i=1}^{n} \binom{m}{x_i}\right] \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{mn - \sum_{i=1}^{n} x_i}. \end{aligned}$$

Note that the first term does not depend on θ, so we may write $L(x_1, x_2, \cdots, x_n; \theta)$ as

$$L(x_1, x_2, \cdots, x_n; \theta) = c \qquad \theta^s (1-\theta)^{mn-s},$$

where c does not depend on θ, and $s = \sum_{k=1}^{n} x_i$. By differentiating and setting the derivative to 0 we obtain

$$\hat{\theta}_{ML} = \frac{1}{mn} \sum_{k=1}^{n} x_i.$$

This suggests that the MLE can be written as

$$\hat{\Theta}_{ML} = \frac{1}{mn} \sum_{k=1}^{n} X_i.$$

2. Similar to our calculation in Example 8.8., for the observed values of $X_1 = x_1$, $X_2 = x_2$, $\cdots$, $X_n = x_n$, the likelihood function is given by

$$\begin{aligned} L(x_1, x_2, \cdots, x_n; \theta) &= \prod_{i=1}^{n} f_{X_i}(x_i; \theta) \\ &= \prod_{i=1}^{n} \theta e^{-\theta x_i} \\ &= \theta^n e^{-\theta \sum_{k=1}^{n} x_i}. \end{aligned}$$

Therefore,

$$\ln L(x_1, x_2, \cdots, x_n; \theta) = n \ln \theta - \sum_{k=1}^{n} x_i \theta.$$

By differentiating and setting the derivative to 0 we obtain

$$\hat{\theta}_{ML} = \frac{n}{\sum_{k=1}^{n} x_i}.$$

This suggests that the MLE can be written as

$$\hat{\Theta}_{ML} = \frac{n}{\sum_{k=1}^{n} X_i}.$$

The examples that we have discussed had only one unknown parameter θ. In general, θ could be a vector of parameters, and we can apply the same methodology to obtain the MLE. More specifically, if we have k unknown parameters θ_1, θ_2, $\cdots$, θ_k, then we need to maximize the likelihood function

$$L(x_1, x_2, \cdots, x_n; \theta_1, \theta_2, \cdots, \theta_k)$$

to obtain the maximum likelihood estimators $\hat{\Theta}_1$, $\hat{\Theta}_2$, $\cdots$, $\hat{\Theta}_k$. Let's look at an example.

Example 8.11

Suppose that we have observed the random sample X_1, X_2, X_3, $\ldots$, X_n, where $X_i \sim N(\theta_1, \theta_2)$, so

$$f_{X_i}(x_i; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{(x_i-\theta_1)^2}{2\theta_2}}.$$

Find the maximum likelihood estimators for θ_1 and θ_2.

Solution

The likelihood function is given by

$$L(x_1, x_2, \cdots, x_n; \theta_1, \theta_2) = \frac{1}{(2\pi)^{\frac{n}{2}} {\theta_2}^{\frac{n}{2}}} \exp\left(-\frac{1}{2\theta_2} \sum_{i=1}^{n} (x_i - \theta_1)^2 \right).$$

Here again, it is easier to work with the log likelihood function

$$\ln L(x_1, x_2, \cdots, x_n; \theta_1, \theta_2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \theta_2 - \frac{1}{2\theta_2} \sum_{i=1}^{n} (x_i - \theta_1)^2.$$

We take the derivatives with respect to θ_1 and θ_2 and set them to zero:

$$\begin{aligned} \frac{\partial}{\partial \theta_1} \ln L(x_1, x_2, \cdots, x_n; \theta_1, \theta_2) &= \frac{1}{\theta_2} \sum_{i=1}^{n} (x_i - \theta_1) = 0 \\ \frac{\partial}{\partial \theta_2} \ln L(x_1, x_2, \cdots, x_n; \theta_1, \theta_2) &= -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2} \sum_{i=1}^{n} (x_i - \theta_1)^2 = 0. \end{aligned}$$

By solving the above equations, we obtain the following maximum likelihood estimates for θ_1 and θ_2:

$$\begin{aligned} \hat{\theta}_1 &= \frac{1}{n} \sum_{i=1}^{n} x_i, \\ \hat{\theta}_2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \theta_1)^2. \end{aligned}$$

We can write the MLE of θ_1 and θ_2 as random variables $\hat{\Theta}_1$ and $\hat{\Theta}_1$:

$$\hat{\Theta}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i,$$
$$\hat{\Theta}_2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \Theta_1)^2.$$

Note that $\hat{\Theta}_1$ is the sample mean, $\overline{X}$, and therefore it is an unbiased estimator of the mean. Here, $\hat{\Theta}_2$ is very close to the sample variance which we defined as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2.$$

In fact,

$$\hat{\Theta}_2 = \frac{n-1}{n} S^2.$$

Since we already know that the sample variance of unbiased estimator of the variance, we conclude that $\hat{\Theta}_2$ is a biased estimator of the variance:

$$E\hat{\Theta}_2 = \frac{n-1}{n}\theta_2.$$

Nevertheless, the bias is very small here and it goes to zero as n gets large.

Note: Here, we caution that we cannot always find the maximum likelihood estimator by setting the derivative to zero. For example, if θ is an integer-valued parameter (such as the number of blue balls in Example 8.9.), then we cannot use differentiation and we need to find the maximizing value in another way. Even if θ is a real-valued parameter, we cannot always find the MLE by setting the derivative to zero. For example, the maximum might be obtained at the endpoints of the acceptable ranges. We will see an example of such scenarios in the Solved Problems section (Section 8.2.5).
