6

Method of Maximum Likelihood

We now give a second method of finding estimators. While the method of moments estimators were very easy to find, it will be more complicated to determine maximum likelihood estimators. The trade off, as we will show, is that maximum likelihood estimators have some very nice statistical properties that the method of moments estimators do not possess.

Definition 6.1. If $Y_1, \ldots, Y_n$ is a random sample from a population whose density is $f(y|\theta)$, then the *likelihood function* is defined as

$$L(\theta) := \prod_{i=1}^{n} f(y_i|\theta).$$

Note that the likelihood function is the joint density function of the random sample $Y_1, \ldots, Y_n$ viewed as a function of the parameter θ for a fixed realization $y_1, \ldots, y_n$ of $Y_1, \ldots, Y_n$. The reason that we write $L(\theta)$ is to emphasize that the likelihood function is being viewed as a function only of θ.

Example 6.2. If $Y_1, \ldots, Y_n$ is a random sample from an $\mathrm{Exp}(\theta)$ population where $\theta > 0$ is a parameter so that their common density is

$$f(y|\theta) = \frac{1}{\theta} e^{-y/\theta}, \quad y > 0,$$

determine the likelihood function $L(\theta)$.

Solution. We find

$$\begin{aligned} L(\theta) = \prod_{i=1}^{n} f(y_i|\theta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-y_i/\theta} &= \frac{1}{\theta} e^{-y_1/\theta} \cdot \frac{1}{\theta} e^{-y_2/\theta} \cdots \frac{1}{\theta} e^{-y_n/\theta} \\ &= \frac{1}{\theta^n} \exp\left\{ -\frac{1}{\theta} \sum_{i=1}^{n} y_i \right\} \end{aligned}$$

for $\theta > 0$ provided that $y_1 > 0, \ldots, y_n > 0$.

Example 6.3. If $Y_1, \ldots, Y_n$ is a random sample from a $\mathcal{N}(\theta, \sigma^2)$ population where $\theta \in \mathbb{R}$ is a parameter so that their common density is

$$f(y|\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-\theta)^2}{2\sigma^2}\right\}, \quad -\infty < y < \infty,$$

determine the likelihood function $L(\theta)$.

Solution. We find

$$\begin{aligned} L(\theta) = \prod_{i=1}^{n} f(y_i|\theta) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y_i-\theta)^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} \exp\left\{-\frac{(y_i-\theta)^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\sum_{i=1}^{n} \frac{(y_i-\theta)^2}{2\sigma^2}\right\} \end{aligned}$$

for $-\infty < \theta < \infty$ provided $-\infty < y_1 < \infty, \ldots, -\infty < y_n < \infty$.

Definition 6.4. The *maximum likelihood estimator* of θ is that value of θ which maximizes $L(\theta)$. Call it $\hat{\theta}_{\text{MLE}}$. That is,

$$\hat{\theta}_{\text{MLE}} := \arg\max_{\theta} L(\theta).$$

In order to maximize $L(\theta)$ it is sometimes easier to work with $\log L(\theta)$ instead. In fact, the log-likelihood function is important enough to have its own notation.

Definition 6.5. If $Y_1, \ldots, Y_n$ is a random sample from a population whose density is $f(y|\theta)$, then the *log-likelihood function* $\ell(\theta)$ is defined as

$$\ell(\theta) := \log L(\theta)$$

where $L(\theta)$ is the likelihood function.

Remark. Since the logarithm function is monotonically increasing, it is clear that the value of θ where the maximum of $L(\theta)$ occurs is necessarily the value of θ where the maximum of $\ell(\theta)$ occurs, and that the converse is also true. That is,

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta) \quad \text{iff} \quad \hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \log L(\theta).$$

One technique for maximizing the log-likelihood function is to use the second derivative test from elementary calculus. That is, we find the critical points by solving $\ell'(\theta) = 0$ for θ, and then determine which critical point is the global maximum by considering $\ell''(\theta)$.

Example 6.6. If $Y_1, \ldots, Y_n$ is a random sample from an $\operatorname{Exp}(\theta)$ population where $\theta > 0$ is a parameter so that their common density is

$$f(y|\theta) = \frac{1}{\theta} e^{-y/\theta}, \quad y > 0,$$

determine $\hat{\theta}_{\text{MLE}}$, the maximum likelihood estimator of θ.

Solution. We found in Example 6.2 that

$$L(\theta) = \frac{1}{\theta^n} \exp\left\{ -\frac{1}{\theta} \sum_{i=1}^{n} y_i \right\}$$

for $\theta > 0$ provided that $y_1 > 0, \ldots, y_n > 0$. In order to maximize the likelihood function $L(\theta)$, we will try to maximize the log-likelihood function $\ell(\theta)$ instead. Therefore,

$$\ell(\theta) = \log L(\theta) = -n \log \theta - \frac{1}{\theta} \sum_{i=1}^{n} y_i$$

and so

$$\ell'(\theta) = \frac{d}{d\theta} \ell(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} y_i.$$

Setting $\ell'(\theta) = 0$ implies that

$$\theta = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{y}$$

is the only critical point. Since

$$\ell''(\theta) = \frac{n}{\theta^2} - \frac{2}{\theta^3} \sum_{i=1}^{n} y_i$$

and

$$\ell''(\overline{y}) = \frac{n}{\overline{y}^2} - \frac{2}{\overline{y}^3} \sum_{i=1}^{n} y_i = \frac{n}{\overline{y}^2} - \frac{2}{\overline{y}^3} \cdot n\overline{y} = \frac{n}{\overline{y}^2} - \frac{2n}{\overline{y}^2} = -\frac{n}{\overline{y}^2} < 0$$

we deduce from the second derivative test that the critical point $\theta = \overline{y}$ is, in fact, where the global maximum occurs. Therefore,

$$\hat{\theta}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \overline{Y}.$$

Example 6.7. If $Y_1, \ldots, Y_n$ is a random sample from a $\mathcal{N}(\theta, \sigma^2)$ population where $\theta \in \mathbb{R}$ is a parameter so that their common density is

$$f(y|\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(y-\theta)^2}{2\sigma^2} \right\}, \quad -\infty < y < \infty,$$

determine $\hat{\theta}_{\text{MLE}}$, the maximum likelihood estimator of θ.

Solution. We found in Example 6.3 that

$$L(\theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\sum_{i=1}^n \frac{(y_i-\theta)^2}{2\sigma^2}\right\}$$

for $-\infty < \theta < \infty$ provided $-\infty < y_1 < \infty, \ldots, -\infty < y_n < \infty$. In order to maximize the likelihood function $L(\theta)$, we will try to maximize the log-likelihood function $\ell(\theta)$ instead. Therefore,

$$\ell(\theta) = \log L(\theta) = -n\log\sigma - \frac{n}{2}\log(2\pi) - \sum_{i=1}^n \frac{(y_i-\theta)^2}{2\sigma^2}$$

and so

$$\ell'(\theta) = \frac{d}{d\theta}\ell(\theta) = \sum_{i=1}^n \frac{(y_i-\theta)}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n y_i - \frac{n\theta}{\sigma^2} = \frac{n\overline{y}}{\sigma^2} - \frac{n\theta}{\sigma^2} = \frac{n(\overline{y}-\theta)}{\sigma^2}.$$

Setting $\ell'(\theta) = 0$ implies that

$$\theta = \frac{1}{n}\sum_{i=1}^n y_i = \overline{y}$$

is the only critical point. Since

$$\ell''(\theta) = -\frac{n}{\sigma^2} < 0$$

for all θ, we deduce from the second derivative test that the critical point $\theta = \overline{y}$ is, in fact, where the global maximum occurs. Therefore,

$$\hat{\theta}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^n Y_i = \overline{Y}.$$

Example 6.8. If $Y_1, \ldots, Y_n$ is a random sample from a population having density

$$f(y|\theta) = (\theta+1)y^\theta, \quad 0 < y < 1,$$

where $\theta > -1$ is a parameter, determine $\hat{\theta}_{\mathrm{MLE}}$, the maximum likelihood estimator of θ.

Solution. We begin by noting that the likelihood function is

$$L(\theta) = \prod_{i=1}^n f(y_i|\theta) = \prod_{i=1}^n (\theta+1)y_i^\theta = (\theta+1)^n \left(\prod_{i=1}^n y_i\right)^\theta$$

for $\theta > 0$ provided that $0 < y_1 < 1, \ldots, 0 < y_n < 1$. In order to maximize the likelihood function $L(\theta)$, we will try to maximize the log-likelihood function $\ell(\theta)$ instead.

Therefore,

$$\ell(\theta) = \log L(\theta) = n \log(\theta + 1) + \theta \sum_{i=1}^{n} \log y_i$$

and so

$$\ell'(\theta) = \frac{d}{d\theta}\ell(\theta) = \frac{n}{\theta + 1} + \sum_{i=1}^{n} \log y_i.$$

Setting $\ell'(\theta) = 0$ implies that

$$\theta = -\frac{n}{\sum_{i=1}^{n} \log y_i} - 1$$

is the only critical point. Since

$$\ell''(\theta) = -\frac{n}{(\theta + 1)^2} < 0$$

for all θ, we deduce from the second derivative test that the critical point

$$\theta = -\frac{n}{\sum_{i=1}^{n} \log y_i} - 1$$

is, in fact, where the global maximum occurs. Therefore,

$$\hat{\theta}_{\text{MLE}} = -\frac{n}{\sum_{i=1}^{n} \log Y_i} - 1.$$

It is perhaps worth noting that $0 < \prod_{i=1}^{n} y_i < 1$ since $0 < y_i < 1$ for all i. This implies that

$$\sum_{i=1}^{n} \log y_i < 0.$$

Hence,

$$-\frac{n}{\sum_{i=1}^{n} \log Y_i} > 0$$

which in turn implies that $\hat{\theta}_{\text{MLE}} > -1$ as required.

We end our study of maximum likelihood estimation with an example that shows that determining the MLE is not always as simple as setting $\ell'(\theta) = 0$ and solving.

Example 6.9. If $Y_1, \ldots, Y_n$ is a random sample from a Uniform$(0, \theta)$ population where $\theta > 0$ is a parameter so that their common density is

$$f(y|\theta) = \frac{1}{\theta}, \quad 0 \le y \le \theta,$$

then

$$L(\theta) = \prod_{i=1}^{n} f(y_i|\theta) = \theta^{-n} \quad \text{and} \quad \ell(\theta) = \log L(\theta) = -n \log \theta.$$

We see, however, that

$$\ell'(\theta) = -\frac{n}{\theta}$$

and so setting $\ell'(\theta) = 0$ gives nonsense. Thus, we must be more careful. In fact, we need to be more careful with our definition of $L(\theta)$. That is,

$$L(\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

where

$$\begin{aligned} f(y_1|\theta) &= \theta^{-1}, \quad 0 \le y_1 \le \theta, \\ f(y_2|\theta) &= \theta^{-1}, \quad 0 \le y_2 \le \theta, \\ &\vdots \\ f(y_n|\theta) &= \theta^{-1}, \quad 0 \le y_n \le \theta, \end{aligned}$$

and so

$$L(\theta) = \theta^{-n}$$

for $\theta > 0$ provided that $0 \le y_1 \le \theta, 0 \le y_2 \le \theta, \ldots, 0 \le y_n \le \theta$. Note that another way to write the constraint is as follows:

$$0 \le \min\{y_1, \ldots, y_n\} \le \max\{y_1, \ldots, y_n\} \le \theta.$$

Therefore, the likelihood function is

$$L(\theta) = \theta^{-n} \quad \text{for} \quad \theta \ge \max\{y_1, \ldots, y_n\}$$

provided $\min\{y_1, \ldots, y_n\} \ge 0$. Recall that $\hat{\theta}_{\text{MLE}}$ is that value of θ where the maximum of $L(\theta)$ occurs. Since θ^{-n} is monotonically decreasing, its maximum value necessarily occurs where θ is smallest; that is at $\theta = \max\{y_1, \ldots, y_n\}$. Thus,

$$\hat{\theta}_{\text{MLE}} = \max\{Y_1, \ldots, Y_n\}.$$