

## 3. Maximum Likelihood

### Basic Theory

Suppose again that we have an observable **random variable**  $X$  for an **experiment**, that takes values in a set  $S$ . Suppose also that distribution of  $X$  depends on an unknown parameter  $\theta$ , taking values in a parameter space  $\Theta$ . Specifically, we will denote the **probability density function** of  $X$  on  $S$  by  $f_\theta$  for  $\theta \in \Theta$ . Of course, our data variable  $X$  will almost always be vector-valued. The parameter  $\theta$  may also be vector-valued.

The **likelihood function**  $L$  is the function obtained by reversing the roles of  $x$  and  $\theta$  in the probability density function; that is, we view  $\theta$  as the variable and  $x$  as the given information (which is precisely the point of view in estimation):

$$L_x(\theta) = f_\theta(x), \quad \theta \in \Theta, x \in S$$

In the method of maximum likelihood, we try to find a value  $u(x)$  of the parameter  $\theta$  that maximizes  $L_x(\theta)$  for each  $x \in S$ . If we can do this, then the statistic  $u(X)$  is called a **maximum likelihood estimator** of  $\theta$ . The method is intuitively appealing--we try to find the values of the parameters that would have most likely produced the data we in fact observed.

Since the natural logarithm function is strictly increasing, the maximum value of  $L_x(\theta)$ , if it exists, will occur at the same points as the maximum value of  $\ln(L_x(\theta))$ . This latter function is called the **log likelihood function** and in many cases is easier to work with than the likelihood function (typically because the probability density function  $f_\theta(x)$  has a product structure).

### Vector of Parameters

An important special case is when  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$  is a vector of  $k$  real parameters, so that  $\Theta \subseteq \mathbb{R}^k$ . In this case, the maximum likelihood problem is to maximize a function of several variables. If  $\Theta$  is a continuous set, the methods of calculus can be used. If the maximum value of  $L_x$  occurs at a point  $\theta$  in the interior of  $\Theta$ , then  $L_x$  has a local maximum at  $\theta$ . Therefore, assuming that the likelihood function is differentiable, we can find this point by solving

$$\frac{\partial^1 L_x(\theta)}{\partial \theta_i} = 0, \quad i \in \{1, 2, \dots, k\}$$

or equivalently

$$\frac{\partial^1 \ln(L_x(\theta))}{\partial \theta_i} = 0, \quad i \in \{1, 2, \dots, k\}$$

On the other hand, the maximum value may occur at a boundary point of  $\Theta$ , or may not exist at all.

## Random Sample

Consider next the case where our outcome variable  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is a random sample of size  $n$  from the distribution with of a random variable  $X$  taking values in  $R$ , with probability density function  $g_\theta$ ,  $\theta \in \Theta$ . Then  $\mathbf{X}$  takes values in  $S = R^n$ , and the joint probability density function of  $\mathbf{X}$  is the product of the marginal probability density functions. Thus, the likelihood function in this special case becomes

$$L_x(\theta) = \prod_{i=1}^n g_\theta(x_i), \quad \mathbf{x} = (x_1, x_2, \dots, x_n) \in S, \theta \in \Theta$$

and hence the log likelihood function becomes

$$\ln(L_x(\theta)) = \sum_{i=1}^n \ln(g_\theta(x_i)), \quad \mathbf{x} = (x_1, x_2, \dots, x_n) \in S, \theta \in \Theta$$

## Examples and Special Cases

In the following subsections, we will study maximum likelihood estimation in a number of classical cases.

### The Bernoulli Distribution

Suppose that we have a coin with unknown probability  $p$  of heads. We toss the coin  $n$  times and record the sequence of heads and tails. Thus, the data  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is a random sample of size  $n$  from the [Bernoulli distribution](#) with success parameter  $p$ . Let

$$Y = \sum_{i=1}^n X_i$$

denote the number of heads, so that the proportion of heads (the sample mean) is

$$M = \frac{Y}{n}$$

1. Suppose that  $p$  varies in the interval  $(0, 1)$ . Show that  $M$  is the maximum likelihood estimator of  $p$ . Recall that  $M$  is also the [method of moments](#) estimator of  $p$ .
2. Suppose that the coin is either fair or two-headed, so  $p$  takes values in  $\{\frac{1}{2}, 1\}$ . Show that the maximum likelihood estimator of  $p$  is the statistic given below, and interpret the result:

$$U = \begin{cases} 1, & Y = n \\ \frac{1}{2}, & Y < n \end{cases}$$

Exercises 1 and 2 show that the maximum likelihood estimator of a parameter, like the solution to any maximization problem, depends critically on the domain.

3. Show that

$$\text{a. } \mathbb{E}(U) = \begin{cases} 1, & p = 1 \\ \frac{1}{2} + \left(\frac{1}{2}\right)^{n+1}, & p = \frac{1}{2} \end{cases}$$

b.  $U$  is biased, but is asymptotically unbiased.

4. Show that

$$\text{a. } \text{MSE}(U) = \begin{cases} 0, & p = 1 \\ \left(\frac{1}{2}\right)^{n+2}, & p = \frac{1}{2} \end{cases}$$

b.  $U$  is consistent.

5. Show that  $U$  is uniformly better than  $M$  on the parameter space  $\{\frac{1}{2}, 1\}$

## Other Basic Distributions

In the following exercises, recall that if  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is a random sample from a distribution with mean  $\mu$  and variance  $\sigma^2$ , then the [method of moments](#) estimators of  $\mu$  and  $\sigma^2$  are, respectively,

$$M = \frac{1}{n} \sum_{i=1}^n X_i, \quad T^2 = \frac{1}{n} \sum_{i=1}^n (X_i - M)^2$$

Of course,  $M$  is the [sample mean](#), and  $T^2 = \frac{n-1}{n} S^2$  where  $S^2$  is the [sample variance](#). In the exercises that follow, we will compute the maximum likelihood estimators for these parameters for several families of distributions.

6. Suppose that  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is a random sample from the [Poisson distribution](#) with unknown parameter  $a \in (0, \infty)$ . Show that the maximum likelihood estimator of  $a$  is the sample mean  $M$ . Recall that for the Poisson distribution, the parameter  $a$  is both the mean and the variance.

7. Suppose that  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is a random sample from the [normal distribution](#) with unknown mean  $\mu \in \mathbb{R}$  and variance  $\sigma^2 \in (0, \infty)$ . Show that the maximum likelihood estimators of  $\mu$  and  $\sigma^2$  are  $M$

and  $T^2$ , respectively.

8. Suppose that  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is a random sample from the [gamma distribution](#) with known shape parameter  $k$  and unknown scale parameter  $b \in (0, \infty)$ .

- Show that the method of moments estimator of  $b$  is  $W = \frac{M}{k}$ .
- Show that  $W$  is also the maximum likelihood estimator of  $b$ .

9. Run the [gamma estimation experiment](#) 1000 times, updating every 10 runs, for several values of the sample size  $n$ , shape parameter  $k$ , and scale parameter  $b$ . In each case, compare the [method of moments estimator](#)  $V$  of  $b$  when  $k$  is unknown with the method of moments and maximum likelihood estimator  $W$  of  $b$  when  $k$  is known. Which estimator seems to work better in terms of mean square error?

10. Suppose that  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is a random sample from the [beta distribution](#) with left parameter  $a \in (0, \infty)$  and right parameter  $b = 1$ . Show that the maximum likelihood estimator of  $a$  is

$$V = \frac{-n}{\sum_{i=1}^n \ln(X_i)}$$

11. Run the [beta estimation experiment](#) 1000 times, updating every 10 runs, for several values of the sample size  $n$  and the parameter  $a$ . In each case, compare the [method of moments estimator](#)  $U$  with the maximum likelihood estimator  $V$ . Which estimator seems to work better in terms of mean square error?

12. Suppose that  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is a random sample from the [Pareto distribution](#) with shape parameter  $a \in (0, \infty)$ . Show that the maximum likelihood estimator of  $a$  is

$$V = \frac{n}{\sum_{i=1}^n \ln(X_i)}$$

13. Run the [Pareto estimation experiment](#) 1000 times, updating every 10 runs, for several values of the sample size  $n$  and the parameter  $a$ . In each case, compare the [method of moments estimator](#)  $U$  with the maximum likelihood estimator  $V$ . Which estimator seems to work better in terms of mean square error?

## Uniform Distributions

In this section we will study two estimation problems that are a good source of insight and counterexamples. In a sense, our first estimation problem is the continuous analogue of an estimation problem studied in the section on [Order Statistics](#) in the chapter [Finite Sampling Models](#). Suppose that  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is a random sample from the [uniform distribution](#) on the interval  $[0, a]$ , where  $a \in (0, \infty)$  is an unknown parameter.

14. Show that the method of moments estimator of  $a$  is  $U = 2M$ .

15. Show that

- a.  $U$  is unbiased.
- b.  $\text{var}(U) = \frac{a^2}{3n}$  so  $U$  is consistent.

16. Show that the maximum likelihood estimator of  $a$  is  $X_{n,n}$ , the  $n^{\text{th}}$  order statistic.

17. Show that

- a.  $\mathbb{E}(X_{n,n}) = \frac{n}{n+1} a$
- b.  $\text{bias}(X_{n,n}) = -\frac{a}{n+1}$  so that  $X_{n,n}$  is negatively biased but asymptotically unbiased.

18. Show that

- a.  $\text{var}(X_{n,n}) = \frac{n}{(n+2)(n+1)^2} a^2$
- b.  $\text{MSE}(X_{n,n}) = \frac{2}{(n+1)(n+2)} a^2$  so that  $X_{n,n}$  is consistent.

Now let  $V = \frac{n+1}{n} X_{n,n}$ .

19. Show that

- a.  $V$  is unbiased.
- b.  $\text{var}(V) = \frac{a^2}{n(n+2)}$  so that  $V$  is consistent.

20. Show that the asymptotic relative efficiency of  $V$  to  $U$  is infinite.

The last exercise shows that  $V$  is a much better estimator than  $U$ ; in fact, an estimator such as  $V$ , whose mean square error decreases on the order of  $\frac{1}{n^2}$ , is called **super efficient**. Now, having found a really good estimator, let's see if we can find a really bad one. A natural candidate is an estimator based on  $X_{n,1}$ , the *first* order statistic.

21. Show that

- a. If  $X$  is uniformly distributed on  $[0, a]$  then so is  $a - X$
- b.  $(a - X_1, a - X_2, \dots, a - X_n)$  is also a random sample from the uniform distribution on  $[0, a]$
- c.  $X_{n,1}$  has the same distribution as  $a - X_{n,n}$ .

22. Show that  $\mathbb{E}(X_{n,1}) = \frac{a}{n+1}$ , and hence  $W = (n+1)X_{n,1}$  is unbiased.

23. Show that  $\text{var}(W) = \frac{n}{n+2} a^2$ , so  $W$  is not even consistent.

24. Run the **uniform estimation experiment** 1000 times, updating every 10 runs, for several values of the sample size  $n$  and the parameter  $a$ . In each case, compare the empirical bias and mean square error of the estimators with their theoretical values. Rank the estimators in terms of empirical mean square error.

Our next series of exercises will show that the maximum likelihood estimator is not necessarily unique. Suppose that  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is a random sample from the **uniform distribution** on the interval  $[a, a + 1]$ , where  $a \in \mathbb{R}$  is an unknown parameter.

25. Show that the method of moments estimator of  $a$  is  $U = M - \frac{1}{2}$ .

26. Show that

- $U$  is unbiased.
- $\text{var}(U) = \frac{1}{12n}$  so  $U$  is consistent.

27. Show that any statistic  $V \in [X_{n,n} - 1, X_{n,1}]$  is a maximum likelihood estimator of  $a$ .

## The Invariance Property

Returning to the general setting, suppose now that  $h$  is a one-to-one function from the parameter space  $\Theta$  onto a set  $\Lambda$ . We can view  $\lambda = h(\theta)$  as a new parameter taking values in the space  $\Lambda$ , and it is easy to re-parameterize the probability density function with the new parameter. Thus, let

$$\bar{f}_\lambda(\mathbf{x}) = f_{h^{-1}(\lambda)}(\mathbf{x}), \quad \mathbf{x} \in S, \lambda \in \Lambda$$

The corresponding likelihood function is

$$\bar{L}_\mathbf{x}(\lambda) = L_\mathbf{x}(h^{-1}(\lambda)), \quad \lambda \in \Lambda, \mathbf{x} \in S$$

28. Suppose that  $u(\mathbf{x}) \in \Theta$  maximizes  $L_\mathbf{x}$  for  $\mathbf{x} \in S$ . Show that  $h(u(\mathbf{x})) \in \Lambda$  maximizes  $\bar{L}_\mathbf{x}$  for  $\mathbf{x} \in S$ .

It follows from Exercise 28 that if  $U$  is a maximum likelihood estimator for  $\theta$ , then  $V = h(U)$  is a maximum likelihood estimator for  $\lambda = h(\theta)$ . This result is known as the **invariance property**.

29. Suppose that  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is a random sample from the Poisson distribution with parameter  $a > 0$ , and let  $p = \mathbb{P}(X_i = 0) = e^{-a}$ . Find the maximum likelihood estimator of  $p$  in two ways:

- Directly, by finding the likelihood function corresponding to the parameter  $p$ .
- By using the result of [Exercise 6](#) and the invariance property.

If the function  $h$  is not one-to-one, the maximum likelihood problem for the new parameter vector  $\lambda = h(\theta)$  is not well-defined, because we cannot parameterize the probability density function in terms of  $\lambda$ . However,

there is a natural generalization of the maximum likelihood problem in this case. Define

$$\bar{L}_x(\lambda) = \max \{L_x(\theta) : (\theta \in \Theta) \text{ and } (h(\theta) = \lambda)\}, \quad \lambda \in \Lambda, \mathbf{x} \in S$$

30. Suppose again that  $u(x) \in \Theta$  maximizes  $L_x$  for  $x \in S$ . Show that  $h(u(x)) \in \Lambda$  maximizes  $\bar{L}_x$  for  $x \in S$ .

The result in the last exercise extends the invariance property to many-to-one transformations of the parameter: if  $U$  is a maximum likelihood estimator for  $\theta$ , then  $V = h(U)$  is a maximum likelihood estimator for  $\lambda = h(\theta)$ .

31. Suppose that  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is a random sample of size  $n$  from the Bernoulli distribution with unknown success parameter  $p \in (0, 1)$ . Find the maximum likelihood estimator of  $p(1 - p)$ , which is the variance of the sampling distribution.

32. Suppose that  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is a random sample from the normal distribution with unknown mean  $\mu \in \mathbb{R}$  and variance  $\sigma^2 \in (0, \infty)$ . Find the maximum likelihood estimator of  $\mu^2 + \sigma^2$ , which is the second moment about 0 for the sampling distribution.

---

[Virtual Laboratories](#) > [7. Point Estimation](#) > [1](#) [2](#) **[3](#)** [4](#) [5](#) [6](#)

[Contents](#) | [Applets](#) | [Data Sets](#) | [Biographies](#) | [External Resources](#) | [Key words](#) | [Feedback](#) | ©