

---

## 27. Maximum likelihood estimation

*Lehmann §7.1; Ferguson §17*

As I assume you already know, if  $X$  is a random variable (or vector) with density or mass function  $f_\theta(x)$  that depends on a parameter  $\theta$ , then the function  $f_\theta(x)$  viewed as a function of  $\theta$  is called the likelihood function of  $\theta$ . We often denote it this function by  $L(\theta)$ . Note that  $L(\theta) = f_\theta(x)$  is implicitly a function of  $x$ , but we suppress this fact in the notation. Let the set of possible values of  $\theta$  be the set  $\Omega$ .

If  $L(\theta)$  has a maximizer in  $\Omega$ , say  $\hat{\theta}$ , then of course  $\hat{\theta}$  is called the maximum likelihood estimator or MLE. Since the logarithm function is a strictly increasing function, any maximizer of  $L(\theta)$  also maximizes  $\ell(\theta) \stackrel{\text{def}}{=} \log L(\theta)$ . It is often much easier to maximize  $\ell(\theta)$  than  $L(\theta)$ .

**Example 27.1** Suppose  $\Omega = (0, \infty)$  and  $X \sim \text{binomial}(n, e^{-\theta})$ . Then

$$\ell(\theta) = \log \binom{n}{x} - x\theta + (n-x)\log(1 - e^{-\theta}),$$

so

$$\ell'(\theta) = -x + \frac{x-n}{1-e^\theta}.$$

Thus, setting  $\ell'(\theta) = 0$  yields  $\theta = -\log(x/n)$ . It isn't hard to verify that  $-\log(x/n)$  is in fact the maximizer of  $\ell(\theta)$ .

As the preceding example demonstrates, it is not always the case that a MLE exists — for if  $X = 0$  or  $X = n$ , then  $\log(-X/n)$  is not contained in  $\Omega$ . This is just one of the technical details that we will consider. Ultimately, we will show that the maximum likelihood estimator is, in many cases, asymptotically normal. However, this is not always the case; in fact, it is not even necessarily true that the MLE is consistent, as shown in Problem 27.1.

We begin the discussion of the consistency of the MLE by defining the so-called Kullback-Liebler information.

**Definition 27.1** If  $f_{\theta_0}(x)$  and  $f_{\theta_1}(x)$  are two densities, the Kullback-Leibler information number equals

$$K(f_{\theta_0}, f_{\theta_1}) = \mathbb{E}_{\theta_0} \log \frac{f_{\theta_0}(X)}{f_{\theta_1}(X)}.$$

If  $P_{\theta_0}(f_{\theta_0}(X) > 0 \text{ and } f_{\theta_1}(X) = 0) > 0$ , then  $K(f_{\theta_0}, f_{\theta_1})$  is defined to be  $\infty$ .

We may show that the Kullback-Leibler information must be nonnegative using a useful fact called Jensen's inequality.

**Theorem 27.1** *Jensen's inequality.* If  $g(t)$  is a convex function, then for any random variable  $X$ ,  $g(\mathbb{E} X) \leq \mathbb{E} g(X)$ . Furthermore, if  $g(t)$  is strictly convex, then  $\mathbb{E} g(X) = g(\mathbb{E} X)$  only if  $P(X = c) = 1$  for some constant  $c$ .

**Proof:** Let  $\mu = \mathbb{E} X$ . Then by the definition of convexity, there exists a linear function  $h(t)$  such that  $h(t) \leq g(t)$  for all  $t$  and  $h(\mu) = g(\mu)$ . Thus,  $h(X) \leq g(X)$ , which implies that  $\mathbb{E} h(X) \leq \mathbb{E} g(X)$ . But by the linearity of the expectation operator and  $h(t)$ ,  $\mathbb{E} h(X) = h(\mu) = g(\mu)$ . Furthermore, if  $g(t)$  is strictly convex, then  $h(t) < g(t)$  for all  $t \neq \mu$ , which means that  $\mathbb{E} h(X) < \mathbb{E} g(X)$  unless  $P(X = \mu) = 1$ . This proves the result. ■

Considering the Kullback-Leibler information once again, we first note that

$$\mathbb{E}_{\theta_0} \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} = \mathbb{E}_{\theta_1} I\{f_{\theta_0}(X) > 0\} \leq 1.$$

Therefore, by the strict convexity of the function  $-\log x$ ,

$$K(f_{\theta_0}, f_{\theta_1}) = E_{\theta_0} - \log \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} \geq -\log E_{\theta_0} \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} \geq 0, \quad (83)$$

with equality if and only if  $P_{\theta_0} \{f_{\theta_0}(X) = f_{\theta_1}(X)\} = 1$ . Inequality (83) is sometimes called the Shannon-Kolmogorov information inequality.

If  $X_1, \dots, X_n$  are iid with density  $f_{\theta_0}(x)$ , then  $\ell(\theta) = \sum_{i=1}^n \log f_{\theta_0}(x_i)$ . Thus, the Shannon-Kolmogorov information inequality may be used to prove the consistency of the maximum likelihood estimator in the case of a finite parameter space:

**Theorem 27.2** Suppose  $\Omega$  is finite and that  $X_1, \dots, X_n$  are iid with density  $f_{\theta_0}(x)$ . Furthermore, suppose that the model is identifiable, which is to say that different values of  $\theta$  lead to different distributions. Then if  $\hat{\theta}_n$  denotes the maximum likelihood estimator,  $\hat{\theta}_n \xrightarrow{P} \theta_0$ .

**Proof:** Notice that

$$\frac{1}{n} \sum_{i=1}^n \log \frac{f_{\theta}(X_i)}{f_{\theta_0}(X_i)} \xrightarrow{P} E_{\theta_0} \log \frac{f_{\theta}(X_i)}{f_{\theta_0}(X_i)} = -K(f_{\theta_0}, f_{\theta}). \quad (84)$$

The value of  $-K(f_{\theta_0}, f_{\theta})$  is strictly negative for  $\theta \neq \theta_0$  by the identifiability of the model. Therefore, since  $\hat{\theta}_n$  is the maximizer of the left hand side of Equation (84),

$$P(\hat{\theta}_n \neq \theta_0) = P\left(\max_{\theta \neq \theta_0} \frac{1}{n} \sum_{i=1}^n \log \frac{f_{\theta}(X_i)}{f_{\theta_0}(X_i)} > 0\right) \leq \sum_{\theta \neq \theta_0} P\left(\frac{1}{n} \sum_{i=1}^n \log \frac{f_{\theta}(X_i)}{f_{\theta_0}(X_i)} > 0\right) \rightarrow 0.$$

This implies that  $\hat{\theta}_n \rightarrow \theta_0$ . ■

The result of Theorem 27.2 may be extended in several ways; however, it is unfortunately *not* true in general that a maximum likelihood estimator is consistent, as seen in Problem 27.1. We will present the extension given in Lehmann, but we do so without proof.

If we return to the simple Example 27.1, we found that the MLE was found by solving the equation

$$\ell'(\theta) = 0. \quad (85)$$

Equation (85) is called the likelihood equation, and naturally a root of the likelihood equation is a good candidate for a maximum likelihood estimator. However, there may be no root and there may be more than one. It turns out the probability that at least one root exists goes to 1 as  $n \rightarrow \infty$ . Consider Example 27.1, in which no MLE exists whenever  $X = 0$  or  $X = n$ . In that case, both  $P(X = 0) = (1 - e^{-\theta})^n$  and  $P(X = n) = e^{-n\theta}$  go to zero as  $n \rightarrow \infty$ . In the case of multiple roots, one of these roots is typically consistent for  $\theta_0$ , as stated in the following theorem.

**Theorem 27.3** Suppose that  $X_1, \dots, X_n$  are iid with density  $f_{\theta_0}(x)$  for  $\theta_0$  in an open interval  $\Omega \subset R$ , where the model is identifiable (i.e., different values of  $\theta \in \Omega$  give different distributions). Furthermore, suppose that the loglikelihood function  $\ell(\theta)$  is differentiable and that the support  $\{x : f_{\theta}(x) > 0\}$  does not depend on  $\theta$ . Then with probability approaching 1 as  $n \rightarrow \infty$ , there exists  $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$  such that  $\ell'(\hat{\theta}_n) = 0$  and  $\hat{\theta}_n \xrightarrow{P} \theta_0$ .

Stated succinctly, Theorem 27.3 says that under certain regularity conditions, there is a consistent root of the likelihood equation. It is important to note that there is no guarantee that this consistent root is the MLE. However, if the likelihood equation only has a single root, we can be more precise:

**Corollary 27.1** Under the conditions of Theorem 27.3, if for every  $n$  there is a unique root of the likelihood equation, and this root is a local maximum, then this root is the MLE and the MLE is consistent.

**Proof:** The only thing that needs to be proved is the assertion that the unique root is the MLE. Denote the unique root by  $\hat{\theta}_n$  and suppose there is some other point  $\theta$  such that  $\ell(\theta) \geq \ell(\hat{\theta}_n)$ . Then there must be a local minimum between  $\hat{\theta}_n$  and  $\theta$ , which contradicts the assertion that  $\hat{\theta}_n$  is the unique root of the likelihood equation. ■

---

## Problems

**Problem 27.1** Suppose that for  $\theta \in (0, 1)$ ,  $X$  is a continuous random variable with density

$$f_{\theta}(x) = \frac{3(1-\theta)}{4\delta^3(\theta)} [\delta^2(\theta) - (x-\theta)^2] I\{|x-\theta| \leq \delta(\theta)\} + \frac{\theta}{2} I\{|x| \leq 1\}, \quad (86)$$

where  $\delta(\theta) > 0$  for all  $\theta$ .

- (a) Prove that  $f_{\theta}(x)$  is a legitimate density.
- (b) What condition on  $\delta(\theta)$  ensures that  $\{x : f_{\theta}(x) > 0\}$  does not depend on  $\theta$ ?
- (c) With  $\delta(\theta) = \exp\{-(1-\theta)^{-4}\}$ , let  $\theta = .125$ . Take samples of sizes  $n \in \{50, 250, 500\}$  from  $f_{\theta}(x)$ . In each case, graph the loglikelihood function and find the MLE. Also, try to identify the consistent root of the likelihood equation in each case.

**Hints:** To generate a sample from  $f_{\theta}(x)$ , note that  $f_{\theta}(x)$  is a mixture density, which means you can start by generating a standard uniform random variable. If it's less than  $\theta$ , generate a uniform variable on  $(-1, 1)$ . Otherwise, generate a variable with density  $3(\delta^2 - x^2)/4\delta^3$  on  $(-\delta, \delta)$  and then add  $\theta$ . You should be able to do this by inverting the cdf. Be very careful when graphing the loglikelihood and finding the MLE. In particular, make sure you evaluate the loglikelihood by hand specifically at each of the sample points in  $(0, 1)$ ; if you fail to do this, you'll miss the point of the problem and you'll get the MLE incorrect.

**Problem 27.2** In the situation of Problem 27.1, prove that the MLE is inconsistent.

**Problem 27.3** Suppose that  $X_1, \dots, X_n$  are iid with density  $f_{\theta}(x)$ , where  $\theta \in (0, \infty)$ . For each of the following forms of  $f_{\theta}(x)$ , prove that the likelihood equation has a unique solution and that this solution maximizes the likelihood function.

- (a) *Weibull:* For some constant  $a > 0$ ,

$$f_{\theta}(x) = a\theta^a x^{a-1} \exp\{-(\theta x)^a\} I\{x > 0\}$$

- (b) *Cauchy:*

$$f_{\theta}(x) = \frac{\theta}{\pi} \frac{1}{\pi^2 + \theta^2}$$

- (c)

$$f_{\theta}(x) = \frac{3\theta^2\sqrt{3}}{2\pi(x^3 + \theta^3)} I\{x > 0\}$$

**Problem 27.4** Find the MLE and its asymptotic distribution given a random sample of size  $n$  from  $f_{\theta}(x) = (1-\theta)\theta^x$ ,  $x = 0, 1, 2, \dots$ ,  $\theta \in (0, 1)$ .

---

## 28. Asymptotic normality of the MLE

*Lehmann §7.2 and 7.3; Ferguson §18*

As seen in the preceding topic, the MLE is not necessarily even consistent, so the title of this topic is slightly misleading — however, “Asymptotic normality of the consistent root of the likelihood equation” is a bit too long! It will be necessary to review a few facts regarding Fisher information before we proceed.

As you are probably already aware, for a density (or mass) function  $f_\theta(x)$ , we define the Fisher information function to be

$$I(\theta) = \mathbb{E}_\theta \left\{ \frac{d}{d\theta} \log f_\theta(X) \right\}^2. \quad (87)$$

Loosely speaking,  $I(\theta)$  is the amount of information about  $\theta$  contained in a single observation from the density  $f_\theta(x)$ . However, this interpretation doesn't always make sense — for example, it is possible to have  $I(\theta) = 0$  for a very informative observation (see Example 7.2.1 on page 462 of Lehmann).

Suppose that  $f_\theta(x)$  is twice differentiable with respect to  $\theta$  and that

$$0 = \frac{d}{d\theta} \mathbb{E}_\theta \frac{f_\theta(X)}{f_\theta(X)} = \mathbb{E}_\theta \frac{\frac{d}{d\theta} f_\theta(X)}{f_\theta(X)} \quad (88)$$

and

$$0 = \frac{d^2}{d\theta^2} \mathbb{E}_\theta \frac{f_\theta(X)}{f_\theta(X)} = \mathbb{E}_\theta \frac{\frac{d^2}{d\theta^2} f_\theta(X)}{f_\theta(X)}. \quad (89)$$

Equations (88) and (89) may appear a bit cryptic; this is one of the rare occasions in which our insistence on steering clear of measure theory is a hindrance. If we consider the example in which  $f_\theta(x)$  is the density for a continuous random variable in the usual sense (i.e.,  $f_\theta(x) = F'_\theta(x)$  for some continuous  $F_\theta(x)$ ), then Equations (88) and (89) are merely the statements

$$\frac{d}{d\theta} \int f_\theta(x) dx = \int \frac{d}{d\theta} f_\theta(x) dx \quad \text{and} \quad \frac{d^2}{d\theta^2} \int f_\theta(x) dx = \int \frac{d^2}{d\theta^2} f_\theta(x) dx.$$

Thus, Equations (88) and (89) are merely statements that we may interchange the order of differentiation and integration.

Equations (88) and (89) give two additional expressions for  $I(\theta)$ . From Equation (88) follows

$$I(\theta) = \text{Var}_\theta \left\{ \frac{d}{d\theta} \log f_\theta(X) \right\}, \quad (90)$$

and Equation (89) implies

$$I(\theta) = -\mathbb{E}_\theta \left\{ \frac{d^2}{d\theta^2} \log f_\theta(X) \right\}. \quad (91)$$

In many cases, Equation (91) is the easiest form of the information to work with.

Equations (90) and (91) make clear a helpful property of the information, namely that for independent random variables, the information about  $\theta$  contained in the joint sample is simply the sum of the individual information components. In particular, if we have an iid sample from  $f_\theta(x)$ , then the information about  $\theta$  equals  $nI(\theta)$ .

If  $\eta = g(\theta)$  for some invertible and differentiable function  $g(\cdot)$ , then since

$$\frac{d}{d\eta} = \frac{d\theta}{d\eta} \frac{d}{d\theta} = \frac{1}{g'(\theta)} \frac{d}{d\theta},$$

by the chain rule, we conclude that

$$I(\eta) = \frac{I(\theta)}{\{g'(\theta)\}^2}. \quad (92)$$

The reason that we need the Fisher information is that we will show that under certain regularity conditions,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N\left\{0, \frac{1}{I(\theta_0)}\right\}, \quad (93)$$

where  $\hat{\theta}_n$  is the consistent root of the likelihood equation.

**Example 28.1** Suppose that  $X_1, \dots, X_n$  are iid Poisson( $\theta_0$ ) random variables. Then the likelihood equation has a unique root, namely  $\hat{\theta}_n = \bar{X}_n$ , and we know that by the central limit theorem  $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, \theta_0)$ . However, the Fisher information for a single observation in this case is

$$-E_{\theta} \left\{ \frac{d}{d\theta} f_{\theta}(X) \right\} = E_{\theta} \frac{X}{\theta^2} = \frac{1}{\theta}.$$

Thus, in this example, equation (93) holds.

Rather than stating all of the regularity conditions necessary to prove Equation (93), we work backwards, figuring out the conditions as we go through the proof. The first step is to expand  $\ell'(\hat{\theta}_n)$  in a power series around  $\theta_0$ :

$$\ell'(\hat{\theta}_n) = \ell'(\theta_0) + (\hat{\theta}_n - \theta_0)\ell''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2\ell'''(\theta_n^*) \quad (94)$$

for some  $\theta_n^*$  between  $\hat{\theta}_n$  and  $\theta_0$ . Clearly, the validity of Equation (94) hinges on the existence of a continuous third derivative of  $\ell(\theta)$ . Rewriting equation (94) gives

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\sqrt{n}\{\ell'(\hat{\theta}_n) - \ell'(\theta_0)\}}{\ell''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)\ell'''(\theta_n^*)} = \frac{\frac{1}{\sqrt{n}}\{\ell'(\theta_0) - \ell'(\hat{\theta}_n)\}}{-\frac{1}{n}\ell''(\theta_0) - \frac{1}{2n}(\hat{\theta}_n - \theta_0)\ell'''(\theta_n^*)}. \quad (95)$$

Let's consider the pieces of Equation (95) individually. If the conditions of Theorem 27.3 are met, then  $\ell'(\hat{\theta}_n) \xrightarrow{P} 0$ . If Equation (88) holds and  $I(\theta_0) < \infty$ , then

$$\frac{1}{\sqrt{n}}\ell'(\theta_0) = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n \frac{d}{d\theta} \log f_{\theta_0}(X_i) \right) \xrightarrow{\mathcal{L}} N\{0, I(\theta_0)\}$$

by the central limit theorem and Equation (90). If Equation (89) holds, then

$$\frac{1}{n}\ell'(\theta_0) = \frac{1}{n} \sum_{i=1}^n \frac{d^2}{d\theta^2} \log f_{\theta_0}(X_i) \xrightarrow{P} -I(\theta_0)$$

by the weak law of large numbers and Equation (91). Finally, we would like to have the term involving  $\ell'''(\theta_n^*)$  disappear, so clearly it is enough to show that  $\frac{1}{n}\ell'''(\theta)$  is bounded in probability in a neighborhood of  $\theta_0$ .

Putting all of these facts together gives a theorem.

**Theorem 28.1** Suppose that the conditions of Theorem 27.3 are satisfied, and let  $\hat{\theta}_n$  denote a consistent root of the likelihood equation. Assume also that  $\ell'''(\theta)$  exists and is continuous, that equations (88) and (89) hold, and that  $\frac{1}{n}\ell'''(\theta)$  is bounded in probability in a neighborhood of  $\theta_0$ . Then if  $0 < I(\theta_0) < \infty$ ,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N\left\{0, \frac{1}{I(\theta_0)}\right\},$$

where  $\hat{\theta}_n$  is the consistent root of the likelihood equation.

The theorem is proved by noting that under the stated regularity conditions,  $\ell'(\hat{\theta}_n) \xrightarrow{P} 0$  so that the numerator in (95) converges in distribution to  $N\{0, I(\theta_0)\}$  by Slutsky's theorem. Furthermore, the denominator in (95) converges to  $I(\theta_0)$ , so Slutsky's theorem gives the desired result.

Sometimes, it is not possible to find an exact zero of  $\ell'(\theta)$ . One way to get a numerical approximation to a zero of  $\ell'(\theta)$  is to use Newton's method, in which we start at a point  $\theta_0$  and then set

$$\theta_1 = \theta_0 - \frac{\ell'(\theta_0)}{\ell''(\theta_0)}. \quad (96)$$

Ordinarily, after finding  $\theta_1$  we would set  $\theta_0$  equal to  $\theta_1$  and apply Equation (96) iteratively.

However, we may show that by using a *single step* of Newton's method, starting from a  $\sqrt{n}$ -consistent estimator of  $\theta_0$ , we may obtain an estimator with the same asymptotic distribution as  $\hat{\theta}_n$ . The proof of the following theorem is left as an exercise:

**Theorem 28.2** Suppose that  $\tilde{\theta}_n$  is any  $\sqrt{n}$ -consistent estimator of  $\theta_0$  (i.e.,  $\sqrt{n}(\tilde{\theta}_n - \theta_0)$  is bounded in probability). Then under the conditions of Theorem 28.1, if we set

$$\delta_n = \tilde{\theta}_n - \frac{\ell'(\tilde{\theta}_n)}{\ell''(\tilde{\theta}_n)}, \quad (97)$$

then

$$\sqrt{n}(\delta_n - \theta_0) \xrightarrow{\mathcal{L}} N\left(0, \frac{1}{I(\theta_0)}\right).$$

## Problems

**Problem 28.1** Do Problems 2.1 on p. 553 and 2.12 on p. 555.

**Problem 28.2 (a)** Show that under the conditions of Theorem 28.1, including  $0 < I(\theta_0) < \infty$ , then if  $\hat{\theta}_n$  is a consistent root of the likelihood equation,  $P_{\theta_0}(\hat{\theta}_n \text{ is a local maximum}) \rightarrow 1$ .

**(b)** Using the result of part (a), show that for any two sequences  $\hat{\theta}_{1n}$  and  $\hat{\theta}_{2n}$  of consistent roots of the likelihood equation,  $P_{\theta_0}(\hat{\theta}_{1n} = \hat{\theta}_{2n}) \rightarrow 1$ .

**Problem 28.3** Prove Theorem 28.2.

**Hint:** Start with  $\sqrt{n}(\delta_n - \theta_0) = \sqrt{n}(\delta_n - \tilde{\theta}_n) + \sqrt{n}(\tilde{\theta}_n - \theta_0)$ , then expand  $\ell'(\tilde{\theta}_n)$  in a Taylor series about  $\theta_0$  and rewrite  $\sqrt{n}(\tilde{\theta}_n - \theta_0)$  using this expansion.

**Problem 28.4** Suppose that the following is a random sample from a logistic density with cdf  $F_\theta(x) = (1 + \exp\{\theta - x\})^{-1}$  (I'll cheat and tell you that I used  $\theta = 2$ .)

1.0944 6.4723 3.1180 3.8318 4.1262  
1.2853 1.0439 1.7472 4.9483 1.7001  
1.0422 0.1690 3.6111 0.9970 2.9438

(a) Evaluate the unique root of the likelihood equation numerically. Then, taking the sample median as our known  $\sqrt{n}$ -consistent estimator  $\tilde{\theta}_n$  of  $\theta$ , evaluate the estimator  $\delta_n$  in equation (97) numerically.

(b) Find the asymptotic distributions of  $\sqrt{n}(\tilde{\theta}_n - 2)$  and  $\sqrt{n}(\delta_n - 2)$ . Then, simulate 200 samples of size  $n = 15$  from the logistic distribution with  $\theta = 2$ . Find the sample variances of the resulting sample medians and  $\delta_n$ -estimators. How well does the asymptotic theory match reality?