Larry Wasserman

All of Statistics

A Concise Course in Statistical Inference

Springer

9.2 The Method of Moments

The first method for generating parametric estimators that we will study is called the method of moments. We will see that these estimators are not optimal but they are often easy to compute. They are are also useful as starting values for other methods that require iterative numerical routines.

Suppose that the parameter $\theta = (\theta_1, \ldots, \theta_k)$ has k components. For $1 \leq j \leq k$, define the j^{th} **moment**

$$\alpha_j \equiv \alpha_j(\theta) = \mathbb{E}_\theta(X^j) = \int x^j dF_\theta(x) \tag{9.2}$$

and the j^{th} **sample moment**

$$\widehat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j. \tag{9.3}$$

9.3 Definition. *The* **method of moments estimator** $\widehat{\theta}_n$ *is defined to be the value of* θ *such that*

$$\begin{aligned} \alpha_1(\widehat{\theta}_n) &= \widehat{\alpha}_1 \\ \alpha_2(\widehat{\theta}_n) &= \widehat{\alpha}_2 \\ &\vdots \quad \vdots \quad \vdots \\ \alpha_k(\widehat{\theta}_n) &= \widehat{\alpha}_k. \end{aligned} \tag{9.4}$$

Formula (9.4) defines a system of k equations with k unknowns.

9.4 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Then $\alpha_1 = \mathbb{E}_p(X) = p$ and $\widehat{\alpha}_1 = n^{-1}\sum_{i=1}^n X_i$. By equating these we get the estimator

$$\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \quad \blacksquare$$

9.5 Example. Let $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$. Then, $\alpha_1 = \mathbb{E}_\theta(X_1) = \mu$ and $\alpha_2 = \mathbb{E}_\theta(X_1^2) = \mathbb{V}_\theta(X_1) + (\mathbb{E}_\theta(X_1))^2 = \sigma^2 + \mu^2$. We need to solve the equations[1]

$$\begin{aligned} \widehat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \widehat{\sigma}^2 + \widehat{\mu}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2. \end{aligned}$$

This is a system of 2 equations with 2 unknowns. The solution is

$$\widehat{\mu} = \overline{X}_n$$

[1] Recall that $\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$. Hence, $\mathbb{E}(X^2) = \mathbb{V}(X) + (\mathbb{E}(X))^2$.

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2. \quad \blacksquare$$

9.6 Theorem. *Let $\widehat{\theta}_n$ denote the method of moments estimator. Under appropriate conditions on the model, the following statements hold:*

1. *The estimate $\widehat{\theta}_n$ exists with probability tending to 1.*
2. *The estimate is consistent: $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$.*
3. *The estimate is asymptotically Normal:*

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow N(0, \Sigma)$$

where

$$\Sigma = g\mathbb{E}_\theta(YY^T)g^T,$$

$Y = (X, X^2, \ldots, X^k)^T$, $g = (g_1, \ldots, g_k)$ *and* $g_j = \partial\alpha_j^{-1}(\theta)/\partial\theta$.

The last statement in the theorem above can be used to find standard errors and confidence intervals. However, there is an easier way: the bootstrap. We defer discussion of this until the end of the chapter.

9.3 Maximum Likelihood

The most common method for estimating parameters in a parametric model is the **maximum likelihood method**. Let $X_1, \ldots, X_n$ be IID with PDF $f(x;\theta)$.

9.7 Definition. *The* **likelihood function** *is defined by*

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta). \tag{9.5}$$

The **log-likelihood function** *is defined by* $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$.

The likelihood function is just the joint density of the data, except that we **treat it as a function of the parameter θ.** Thus, $\mathcal{L}_n : \Theta \to [0, \infty)$. The likelihood function is not a density function: in general, it is **not** true that $\mathcal{L}_n(\theta)$ integrates to 1 (with respect to θ).

9.8 Definition. *The* **maximum likelihood estimator** MLE, *denoted by* $\widehat{\theta}_n$, *is the value of θ that maximizes $\mathcal{L}_n(\theta)$.*

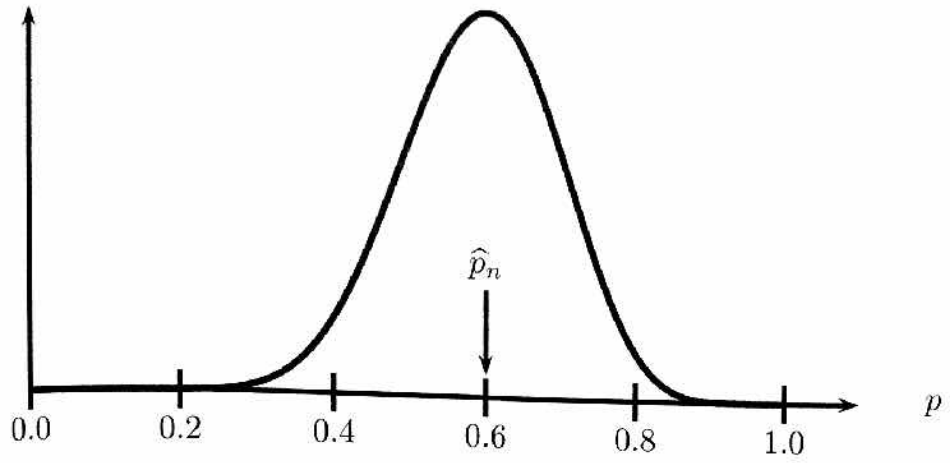

FIGURE 9.1. Likelihood function for Bernoulli with $n = 20$ and $\sum_{i=1}^n X_i = 12$. The MLE is $\widehat{p}_n = 12/20 = 0.6$.

The maximum of $\ell_n(\theta)$ occurs at the same place as the maximum of $\mathcal{L}_n(\theta)$, so maximizing the log-likelihood leads to the same answer as maximizing the likelihood. Often, it is easier to work with the log-likelihood.

9.9 Remark. If we multiply $\mathcal{L}_n(\theta)$ by any positive constant c (not depending on θ) then this will not change the MLE. Hence, we shall often drop constants in the likelihood function.

9.10 Example. Suppose that $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. The probability function is $f(x; p) = p^x(1-p)^{1-x}$ for $x = 0, 1$. The unknown parameter is p. Then,

$$\mathcal{L}_n(p) = \prod_{i=1}^n f(X_i; p) = \prod_{i=1}^n p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$$

where $S = \sum_i X_i$. Hence,

$$\ell_n(p) = S \log p + (n - S) \log(1 - p).$$

Take the derivative of $\ell_n(p)$, set it equal to 0 to find that the MLE is $\widehat{p}_n = S/n$. See Figure 9.1. ■

9.11 Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. The parameter is $\theta = (\mu, \sigma)$ and the likelihood function (ignoring some constants) is:

$$\begin{aligned} \mathcal{L}_n(\mu, \sigma) &= \prod_i \frac{1}{\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(X_i - \mu)^2 \right\} \\ &= \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2 \right\} \end{aligned}$$

$$= \sigma^{-n} \exp\left\{-\frac{nS^2}{2\sigma^2}\right\} \exp\left\{-\frac{n(\overline{X}-\mu)^2}{2\sigma^2}\right\}$$

where $\overline{X} = n^{-1}\sum_i X_i$ is the sample mean and $S^2 = n^{-1}\sum_i (X_i - \overline{X})^2$. The last equality above follows from the fact that $\sum_i (X_i - \mu)^2 = nS^2 + n(\overline{X}-\mu)^2$ which can be verified by writing $\sum_i (X_i - \mu)^2 = \sum_i (X_i - \overline{X} + \overline{X} - \mu)^2$ and then expanding the square. The log-likelihood is

$$\ell(\mu, \sigma) = -n \log \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\overline{X}-\mu)^2}{2\sigma^2}.$$

Solving the equations

$$\frac{\partial \ell(\mu,\sigma)}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial \ell(\mu,\sigma)}{\partial \sigma} = 0,$$

we conclude that $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma} = S$. It can be verified that these are indeed global maxima of the likelihood. ∎

9.12 Example (A Hard Example). Here is an example that many people find confusing. Let $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$. Recall that

$$f(x;\theta) = \begin{cases} 1/\theta & 0 \leq x \leq \theta \\ 0 & \text{otherwise.} \end{cases}$$

Consider a fixed value of θ. Suppose $\theta < X_i$ for some i. Then, $f(X_i;\theta) = 0$ and hence $\mathcal{L}_n(\theta) = \prod_i f(X_i;\theta) = 0$. It follows that $\mathcal{L}_n(\theta) = 0$ if any $X_i > \theta$. Therefore, $\mathcal{L}_n(\theta) = 0$ if $\theta < X_{(n)}$ where $X_{(n)} = \max\{X_1, \ldots, X_n\}$. Now consider any $\theta \geq X_{(n)}$. For every X_i we then have that $f(X_i;\theta) = 1/\theta$ so that $\mathcal{L}_n(\theta) = \prod_i f(X_i;\theta) = \theta^{-n}$. In conclusion,

$$\mathcal{L}_n(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \theta \geq X_{(n)} \\ 0 & \theta < X_{(n)}. \end{cases}$$

See Figure 9.2. Now $\mathcal{L}_n(\theta)$ is strictly decreasing over the interval $[X_{(n)}, \infty)$. Hence, $\widehat{\theta}_n = X_{(n)}$. ∎

The maximum likelihood estimators for the multivariate Normal and the multinomial can be found in Theorems 14.5 and 14.3.

9.4 Properties of Maximum Likelihood Estimators

Under certain conditions on the model, the maximum likelihood estimator $\widehat{\theta}_n$ possesses many properties that make it an appealing choice of estimator. The main properties of the MLE are:

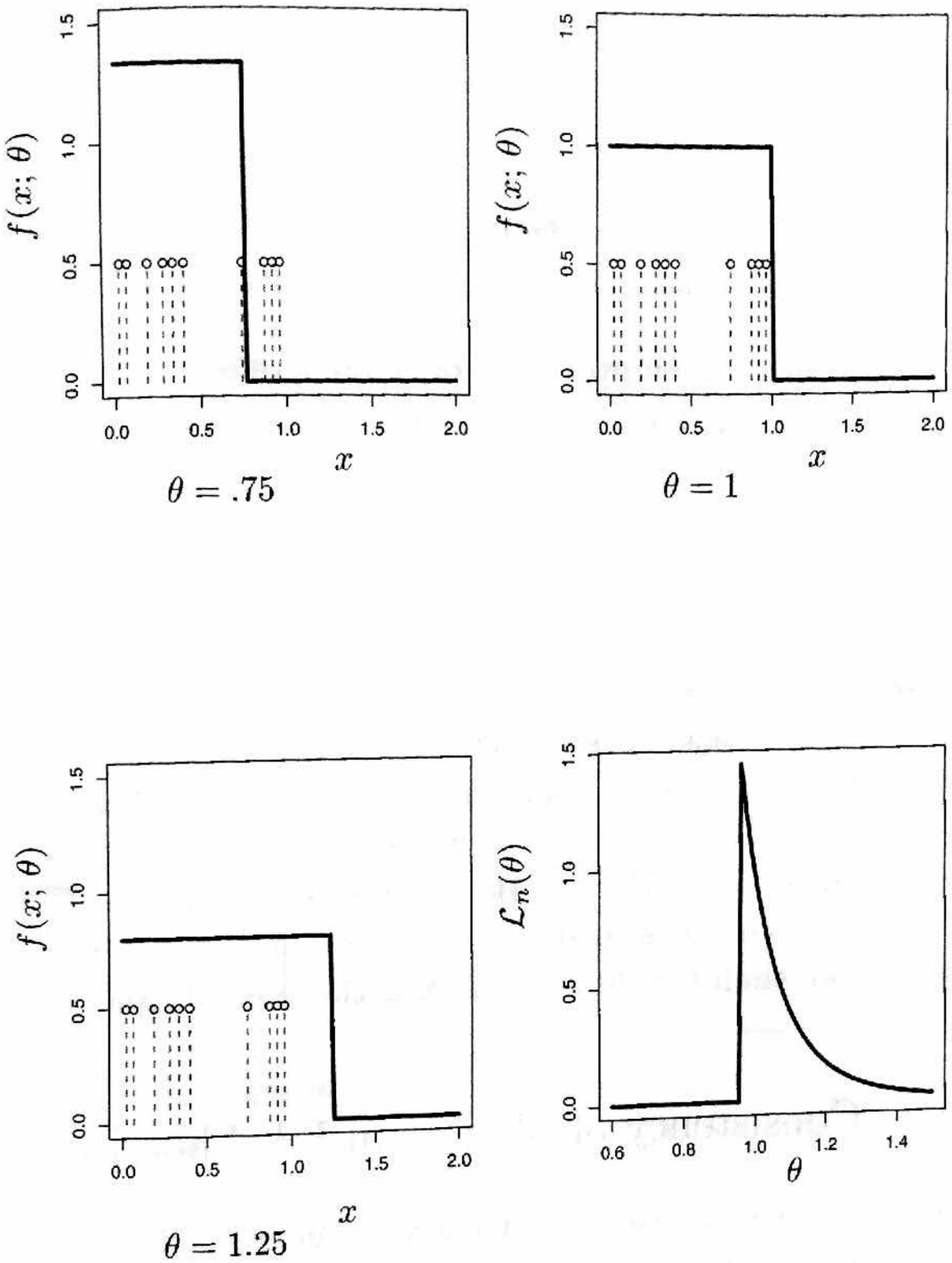

FIGURE 9.2. Likelihood function for Uniform $(0, \theta)$. The vertical lines show the observed data. The first three plots show $f(x;\theta)$ for three different values of θ. When $\theta < X_{(n)} = \max\{X_1, \ldots, X_n\}$, as in the first plot, $f(X_{(n)};\theta) = 0$ and hence $\mathcal{L}_n(\theta) = \prod_{i=1}^n f(X_i;\theta) = 0$. Otherwise $f(X_i;\theta) = 1/\theta$ for each i and hence $\mathcal{L}_n(\theta) = \prod_{i=1}^n f(X_i;\theta) = (1/\theta)^n$. The last plot shows the likelihood function.

1. The MLE is **consistent**: $\widehat{\theta}_n \xrightarrow{P} \theta_\star$ where $\theta_\star$ denotes the true value of the parameter θ;

2. The MLE is **equivariant**: if $\widehat{\theta}_n$ is the MLE of θ then $g(\widehat{\theta}_n)$ is the MLE of $g(\theta)$;

3. The MLE is **asymptotically Normal**: $(\widehat{\theta} - \theta_\star)/\widehat{\mathsf{se}} \rightsquigarrow N(0,1)$; also, the estimated standard error $\widehat{\mathsf{se}}$ can often be computed analytically;

4. The MLE is **asymptotically optimal** or **efficient**: roughly, this means that among all well-behaved estimators, the MLE has the smallest variance, at least for large samples;

5. The MLE is approximately the Bayes estimator. (This point will be explained later.)

We will spend some time explaining what these properties mean and why they are good things. In sufficiently complicated problems, these properties will no longer hold and the MLE will no longer be a good estimator. For now we focus on the simpler situations where the MLE works well. The properties we discuss only hold if the model satisfies certain **regularity conditions.** These are essentially smoothness conditions on $f(x;\theta)$. **Unless otherwise stated, we shall tacitly assume that these conditions hold.**

9.5 Consistency of Maximum Likelihood Estimators

Consistency means that the MLE converges in probability to the true value. To proceed, we need a definition. If f and g are PDF's, define the **Kullback-Leibler distance** [2] between f and g to be

$$D(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx. \tag{9.6}$$

It can be shown that $D(f,g) \geq 0$ and $D(f,f) = 0$. For any $\theta, \psi \in \Theta$ write $D(\theta,\psi)$ to mean $D(f(x;\theta), f(x;\psi))$.

We will say that the model $\mathfrak{F}$ is **identifiable** if $\theta \neq \psi$ implies that $D(\theta,\psi) > 0$. This means that different values of the parameter correspond to different distributions. We will assume from now on the the model is identifiable.

[2] This is not a distance in the formal sense because $D(f,g)$ is not symmetric.

Let $\theta_\star$ denote the true value of θ. Maximizing $\ell_n(\theta)$ is equivalent to maximizing

$$M_n(\theta) = \frac{1}{n}\sum_i \log \frac{f(X_i;\,\theta)}{f(X_i;\,\theta_\star)}.$$

This follows since $M_n(\theta) = n^{-1}(\ell_n(\theta) - \ell_n(\theta_\star))$ and $\ell_n(\theta_\star)$ is a constant (with respect to θ). By the law of large numbers, $M_n(\theta)$ converges to

$$\begin{aligned}
\mathbb{E}_{\theta_\star}\left(\log \frac{f(X_i;\,\theta)}{f(X_i;\,\theta_\star)}\right) &= \int \log\left(\frac{f(x;\,\theta)}{f(x;\,\theta_\star)}\right) f(x;\,\theta_\star)dx \\
&= -\int \log\left(\frac{f(x;\,\theta_\star)}{f(x;\,\theta)}\right) f(x;\,\theta_\star)dx \\
&= -D(\theta_\star,\theta).
\end{aligned}$$

Hence, $M_n(\theta) \approx -D(\theta_\star,\theta)$ which is maximized at $\theta_\star$ since $-D(\theta_\star,\theta_\star) = 0$ and $-D(\theta_\star,\theta) < 0$ for $\theta \neq \theta_\star$. Therefore, we expect that the maximizer will tend to $\theta_\star$. To prove this formally, we need more than $M_n(\theta) \xrightarrow{\text{P}} -D(\theta_\star,\theta)$. We need this convergence to be uniform over θ. We also have to make sure that the function $D(\theta_\star,\theta)$ is well behaved. Here are the formal details.

9.13 Theorem. *Let $\theta_\star$ denote the true value of θ. Define*

$$M_n(\theta) = \frac{1}{n}\sum_i \log \frac{f(X_i;\,\theta)}{f(X_i;\,\theta_\star)}$$

and $M(\theta) = -D(\theta_\star,\theta)$. Suppose that

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\text{P}} 0 \tag{9.7}$$

and that, for every $\epsilon > 0$,

$$\sup_{\theta:|\theta-\theta_\star|\geq\epsilon} M(\theta) < M(\theta_\star). \tag{9.8}$$

Let $\widehat{\theta}_n$ denote the MLE. *Then $\widehat{\theta}_n \xrightarrow{\text{P}} \theta_\star$.*

The proof is in the appendix.

9.6 Equivariance of the MLE

9.14 Theorem. *Let $\tau = g(\theta)$ be a function of θ. Let $\widehat{\theta}_n$ be the* MLE *of θ. Then $\widehat{\tau}_n = g(\widehat{\theta}_n)$ is the* MLE *of τ.*

PROOF. Let $h = g^{-1}$ denote the inverse of g. Then $\widehat{\theta}_n = h(\widehat{\tau}_n)$. For any τ, $\mathcal{L}(\tau) = \prod_i f(x_i; h(\tau)) = \prod_i f(x_i; \theta) = \mathcal{L}(\theta)$ where $\theta = h(\tau)$. Hence, for any τ, $\mathcal{L}_n(\tau) = \mathcal{L}(\theta) \leq \mathcal{L}(\widehat{\theta}) = \mathcal{L}_n(\widehat{\tau})$. ■

9.15 Example. Let $X_1, \ldots, X_n \sim N(\theta, 1)$. The MLE for θ is $\widehat{\theta}_n = \overline{X}_n$. Let $\tau = e^\theta$. Then, the MLE for τ is $\widehat{\tau} = e^{\widehat{\theta}} = e^{\overline{X}}$. ■