# LECTURE 5: MAXIMUM LIKELIHOOD AND THE INFORMATION (CRAMÉR-RAO) BOUND

Michal Kolesár*

September 23, 2024

REFERENCE CB, Chapter 7.3; HMC, Chapter 6.2, or H22, Chapter 10.

In the previous lectures, we talked about evaluating estimators of $\theta$ based on data $X \sim F(\cdot \mid \theta)$, $\theta \in \Theta$, but so far we did not discuss general ways of finding them. We will now talk about one systematic way of coming up with estimators: using maximum likelihood, and we will discuss its properties. We will conclude in Section 5 by briefly discussing another method, the method of moments.

## 1. MAXIMUM LIKELIHOOD

Let $f_n(x \mid \theta)$ denote the joint probability density function (PDF) of the data $X$, consisting of $n$ observations. Viewed as a function of $\theta$, the joint PDF is called a likelihood function. To stress that the likelihood function views the $x$'s in the joint density as fixed, and is a function of $\theta$, it is usually denoted by

$$\mathcal{L}_n(\theta \mid x) = f_n(x \mid \theta).$$

By definition, the maximum likelihood estimator (MLE) $\hat{\theta}_{ML}$ of $\theta$ maximizes the likelihood function,

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta \mid x).$$

Thus, MLE is a parameter value such that it gives the greatest likelihood (not probability!) of observing $x$. Since $\log(x)$ is increasing in $x$, it is easy to see that $\hat{\theta}_{ML}$ also maximizes

$$\ell_n(\theta \mid x) := \log \mathcal{L}_n(\theta \mid x).$$

The function $\ell_n(\theta \mid x)$ is called the log-likelihood, and its derivative is called the *score*,

$$\mathcal{S}_n(\theta \mid x) := \frac{\partial}{\partial \theta} \ell_n(\theta \mid x).$$

*Email: mkolesar@princeton.edu.

If $\ell_n(\theta \mid x)$ is differentiable in $\theta$, and $\hat{\theta}_{ML}$ lies in the interior of $\Theta$, then $\hat{\theta}_{ML}$ satisfies $\mathcal{S}_n(\hat{\theta}_{ML} \mid X) = 0$.

If $X = (X_1, \ldots, X_n)$ is a random sample, then $f_n(x \mid \theta) = \prod_{i=1}^n f(x_i \mid \theta)$, and

$$\mathcal{S}_n(\theta \mid x_1, \ldots, x_n) = \sum_{i=1}^n \frac{\partial \log f(x_i \mid \theta)}{\partial \theta} = \sum_{i=1}^n \frac{\partial f(x_i \mid \theta)/\partial \theta}{f(x_i \mid \theta)}$$

Now you can see the reason why we took log of the likelihood function: it is easier to take the derivative of the sum than the derivative of the product.

*Remark 1.* To save on notation, we'll typically keep the conditioning on $x$ implicit and just write $\mathcal{L}_n(\theta)$, $\ell_n(\theta)$ and $\mathcal{S}_n(\theta)$.

*Remark 2.* It is a consequence of the factorization theorem that if $\hat{\theta}_{ML}$ is unique then it is a function of *any* sufficient statistic. That is, if $T$ is a sufficient statistic, then $\hat{\theta}_{ML}$ is a function of $T$ (HMC, Theorem 7.3.2).[1]

*Digression.* If the likelihood doesn't have a unique maximum, then one can always pick one that is a function of the sufficient statistic only, but there may be other estimates that maximize the likelihood which are not a function of a sufficient statistic. For example (see Romano and Siegel 1986, Example 8.13), suppose $X_i \sim \mathcal{U}[\theta - 1/2, \theta + 1/2]$. Then you can verify that $X_{(1)}, X_{(n)}$ are sufficient. But any $\hat{\theta}$ that satisfies $X_{(n)} - 1/2 < \hat{\theta} < X_{(1)} + 1/2$ maximizes the likelihood. This holds for $\hat{\theta} = (X_{(1)} + X_{(n)})/2$, which is a function of a sufficient statistic only. But $\hat{\theta} = (X_{(n)} - 1/2) + \cos(X_1)^2(X_{(1)} - X_{(n)} + 1)$ also maximizes the likelihood. ⊠

*Example 1 (CB, Example 7.2.11, HMC Example 3.1.3).* Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Then

$$f(x_i \mid \mu, \sigma^2) = -\log\sqrt{2\pi} - (1/2)\log\sigma^2 - (x_i - \mu)^2/(2\sigma^2).$$

So

$$\ell_n(\mu, \sigma^2) = -n\log\sqrt{2\pi} - (n/2)\log\sigma^2 - \sum_{i=1}^n (x_i - \mu)^2/(2\sigma^2).$$

The first-order conditions (FOCs) are

$$\partial\ell_n/\partial\mu = \sum_{i=1}^n (x_i - \mu)/\sigma^2 = 0,$$
$$\partial\ell_n/\partial\sigma^2 = -n/(2\sigma^2) + \sum_{i=1}^n (x_i - \mu)^2/(2\sigma^4) = 0.$$

So $\hat{\mu}_{ML} = \overline{X}_n$ and $\hat{\sigma}^2_{ML} = \sum_{i=1}^n (X_i - \overline{X}_n)^2/n = (n-1)S_n^2/n$. ⊠

*Example 2.* Let $X_1, \ldots, X_n$ be a random sample from $U[0, \theta]$. Then $f(x_i \mid \theta) = \mathbb{1}\{0 \leq x_i \leq \theta\}/\theta$. So

$$\mathcal{L}_n(\theta) = \frac{1}{\theta^n}\mathbb{1}\{0 \leq x_{(1)}\}\mathbb{1}\{x_{(n)} \leq \theta\}.$$ ⊠

We conclude that $\hat{\theta}_{ML} = X_{(n)}$.

1. We emphasize *any* since the whole data is also trivially a sufficient statistic, so it's not just that MLE is a function of *a* sufficient statistic.

INVARIANCE A very useful property of MLE is that they are invariant to parametrization. In particular, suppose we're interested in some function $\tau(\theta)$, and we know that the MLE of $\theta$ is $\hat{\theta}_{ML}$. Then the MLE of $\tau$ is given by $\hat{\tau}_{ML} = \tau(\hat{\theta}_{ML})$ (HMC, Theorem 6.1.2, or CB, Theorem 7.2.10).

*Example 1 (continued).* The MLE of the coefficient of variation $cv = \sigma/\mu$ is given by

$$\widehat{cv}_{ML} = \frac{\hat{\sigma}_{ML}}{\overline{X}_n}. \qquad \boxtimes$$

## 2. FISHER INFORMATION

For a given $\theta \in \Theta$, let $\mathcal{X}_\theta = \{(x_1, \ldots, x_n) \colon f_n(x_1, \ldots, x_n \mid \theta) > 0\}$. Recall from lecture 1 that $\mathcal{X}_\theta$ is called the *support* of $f_n$. We make the following assumption:

*Assumption 1.* The support $\mathcal{X}_\theta$ of $X$ does not depend on $\theta$, so that we can denote it by $\mathcal{X}$. Also, the log-likelihood $\ell_n(\theta \mid x)$ is twice continuously differentiable in $\theta$ for all $x \in \mathcal{X}$.

We can think of $\mathcal{S}_n(\theta)$ as a random variable, since it depends on the data $X$. Its second moment is called Fisher information (if $\theta$ were a vector it'd be a matrix equal to the expected outer product of the score vector)

$$\mathcal{I}_n(\theta) = E[\mathcal{S}_n(\theta)^2] = E\left[\left(\frac{\partial \ell_n(\theta \mid X)}{\partial \theta}\right)^2\right].$$

Fisher information plays an important role in maximum likelihood estimation. The theorem below gives two information equalities:

*Theorem 3 (Information equalities). Under Assumption 1,*

$$E[\mathcal{S}_n(\theta)] = 0,$$

*and*

$$\mathcal{I}_n(\theta) = -E\left[\frac{\partial^2 \ell_n(\theta \mid X)}{\partial \theta^2}\right].$$

*Proof.* Since $\ell_n(\theta)$ is twice differentiable in $\theta$, $f_n(x \mid \theta)$ is twice differentiable in $\theta$ as well. Differentiating (under the integral sign) the identity $\int_{\mathcal{X}} f_n(x \mid \theta)\, \mathrm{d}x = 1$ with respect to $\theta$ yields

$$\int_{\mathcal{X}} \frac{\partial f_n(x \mid \theta)}{\partial \theta}\, \mathrm{d}x = 0. \tag{1}$$

for all $\theta \in \Theta$. Note that

$$\mathcal{S}_n(\theta) = \frac{\partial \log f_n(x \mid \theta)}{\partial \theta} = \frac{1}{f_n(x \mid \theta)} \frac{\partial f_n(x \mid \theta)}{\partial \theta}. \tag{2}$$

Taking expectation of (2) and plugging in (1) then yields

$$E[\mathcal{S}_n(\theta)] = \int_{\mathcal{X}} \frac{\partial f_n(x \mid \theta)}{\partial \theta}\, \mathrm{d}x = 0,$$

which proves the first assertion. Differentiating (1) again yields

$$\int_{\mathcal{X}} \frac{\partial^2 f_n(x \mid \theta)}{\partial \theta^2} dx = 0 \tag{3}$$

for all $\theta \in \Theta$. Note that

$$\frac{\partial^2 \ell_n(\theta \mid x)}{\partial \theta^2} = \frac{\partial \mathcal{S}_n(\theta)}{\partial \theta} = \frac{1}{f_n(x \mid \theta)} \frac{\partial^2 f_n(x \mid \theta)}{\partial \theta^2} - \frac{1}{f_n^2(x \mid \theta)} \left( \frac{\partial f_n(x \mid \theta)}{\partial \theta} \right)^2.$$

Taking an expectation and using (3) and (2) then yields

$$\begin{aligned} E\left[ \frac{\partial^2 \ell_n(\theta \mid x)}{\partial \theta^2} \right] &= \int_{\mathcal{X}} \frac{\partial^2 f_n(x \mid \theta)}{\partial \theta^2} \, \mathrm{d}x - \int_{\mathcal{X}} \frac{1}{f_n(x \mid \theta)} \left( \frac{\partial f_n(x \mid \theta)}{\partial \theta} \right)^2 \mathrm{d}x \\ &= - \int_{\mathcal{X}} \frac{1}{f_n(x \mid \theta)} \left( \frac{\partial f_n(x \mid \theta)}{\partial \theta} \right)^2 \mathrm{d}x = -E[\mathcal{S}_n(\theta)^2] = -\mathcal{I}(\theta), \end{aligned}$$

which is our second result. □

- If the support $\mathcal{X}$ depended on $\theta$, then we'd need to use Leibniz rule to differentiate under the integral sign, which would mess up the result.
- Why is the second information equality true? I have no intuition, it seems like a coincidence that the curvature of the likelihood equals the variance of the score. As we'll see later, the result will give us a number of options for doing inference based on MLE.

*Digression.* The key to Theorem 3 is differentiation under the integral sign. When are we allowed to differentiate like this? One set of sufficient conditions is as follows (Newey and McFadden (1994, Lemma 3.6) or Durrett (2019, Theorem A.5.3)). If (i) for almost all $z$, the derivative $\partial a(z, \theta)/\partial \theta$ exists and is continuous in the neighborhood $\mathcal{N}$ of $\theta_0$, (ii) $\int \sup_{\theta \in \mathcal{N}} \|\nabla_\theta a(z, \theta)\| \, \mathrm{d}z < \infty$, then $\int a(z, \theta) \, \mathrm{d}z$ is continuously differentiable with $\nabla_\theta \int a(z, \theta) \, \mathrm{d}z = \int \nabla_\theta a(z, \theta) \, \mathrm{d}z$. ⊠

*Remark 4 (Information for a random sample).* Let us now consider Fisher information for a random sample. Let $X = (X_1, \ldots, X_n)$ be a random sample from distribution $f_n(x \mid \theta)$. Then the joint log-likelihood is $\ell_n(\theta \mid x) = \sum_{i=1}^n \log f(x_i \mid \theta)$. By the second information equality, it therefore follows that

$$\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta),$$

where $\mathcal{I}_n(\theta)$ is the Fisher information based on $n$ observations. As a shorthand, we will write $\mathcal{I}(\theta)$ in place of $\mathcal{I}_1(\theta)$.

*Example 3.* Let us calculate Fisher information for a single sample $X \sim \mathcal{N}(\mu, \sigma^2)$, with $\sigma^2$ is known. Thus, our parameter is $\theta = \mu$. The log-likelihood is $\ell_1(\mu \mid x) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x - \mu)^2$. So

$$\frac{\partial \ell_1(\mu \mid x)}{\partial \mu} = \frac{x - \mu}{\sigma^2}, \tag{4}$$

and $\partial^2 \ell_1(\mu \mid x)/\partial \mu^2 = -1/\sigma^2$. So $-E[\partial^2 \ell_1(\mu \mid X)/\partial \mu^2] = 1/\sigma^2$. At the same time,

$$\mathcal{I}(\theta) = E[(\partial \ell_1(\mu \mid X)/\partial \mu)^2] = E[(X-\mu)^2/\sigma^4] = 1/\sigma^2,$$

which verifies the second information equality. The first information equality is also clear by taking expectation of (4). ⊠

*Example 4 (HMC, Example 6.2.1).* Let us calculate Fisher information for a Bernoulli$(\theta)$ distribution. Note that a Bernoulli distribution is discrete. So we use probability mass function instead of PDF. This yields the log-likelihood $\ell_1(\theta \mid x) = x \log \theta + (1-x)\log(1-\theta)$. So $\mathcal{S}_1(\theta) = X/\theta - (1-X)/(1-\theta)$ and $\partial^2 \ell_1(\theta \mid x)/\partial \theta^2 = -x/\theta^2 - (1-x)/(1-\theta)^2$. So

$$\begin{aligned} E[\mathcal{S}_1(\theta)^2] &= E[(X/\theta - (1-X)/(1-\theta))^2] \\ &= E[X^2/\theta^2] - 2E[X(1-X)/(\theta(1-\theta))] + E[(1-X)^2/(1-\theta)^2] \\ &= E[X/\theta^2] + E[(1-X)/(1-\theta)^2] = 1/\theta + 1/(1-\theta) \\ &= \frac{1}{\theta(1-\theta)}, \end{aligned}$$

where the third line follows since $x = x^2$, $x(1-x) = 0$, and $(1-x) = (1-x)^2$ if $x \in \{0,1\}$. At the same time,

$$\begin{aligned} -E[\partial^2 \ell_1(\theta \mid X)/\partial \theta^2] &= E[X/\theta^2 + (1-X)/(1-\theta)^2] = 1/\theta + 1/(1-\theta) \\ &= \frac{1}{\theta(1-\theta)}, \end{aligned}$$

as expected. ⊠

## 3. INFORMATION INEQUALITY

An important question in the theory of statistical estimation is whether there is a non-trivial bound such that no estimator can be more efficient (in the sense of having lower mean squared error (MSE)) than this bound. The theorem below is a result of this sort. The result is commonly known as the *Cramér-Rao bound*, but we shall follow the lead of Savage (1972) and call it the *Information inequality*, since the result is originally due to Fréchet (1943); Cramér (1946) and Rao (1945) just re-discovered it.

*Theorem 5 (CB, Theorem 7.3.9, HMC, Theorem 6.2.1). Let $X \sim f_n(x \mid \theta)$, and suppose that Assumption 1 holds. Suppose $\hat{\theta}(X)$ is an estimator with finite variance for all $\theta$ such that*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} E_\theta[\hat{\theta}(X)] = \int_{\mathcal{X}} \hat{\theta}(x) \frac{\partial}{\partial \theta} f_n(x \mid \theta)\, \mathrm{d}x$$

*(i.e. we can pass the derivative under the integral sign). Then*

$$\operatorname{var}_\theta(\hat{\theta}) \geq \frac{(dE_\theta[\hat{\theta}(X)]/d\theta)^2}{\mathcal{I}_n(\theta)}.$$

*In particular, if $\hat{\theta}$ is unbiased for $\theta$, then $\operatorname{var}_\theta(\hat{\theta}) \geq 1/\mathcal{I}_n(\theta)$.*

*Proof.* The first information equality gives $E_\theta[\mathcal{S}_n(\theta \mid X)] = 0$. So,

$$\operatorname{cov}_\theta(\hat{\theta}(X), \mathcal{S}_n(\theta \mid X)) = E_\theta[\hat{\theta}(X)\mathcal{S}_n(\theta \mid X)] = \int_{\mathcal{X}} \hat{\theta}(x) \frac{\partial f_n(x \mid \theta)}{\partial \theta} \, \mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\theta} E_\theta[\hat{\theta}(X)].$$

On the other hand, by the Cauchy-Schwarz inequality,

$$\operatorname{cov}_\theta(\hat{\theta}(X), \mathcal{S}_n(\theta \mid X))^2 \leq \operatorname{var}_\theta(\hat{\theta}(X)) \operatorname{var}_\theta(\mathcal{S}_n(\theta \mid X)) = \operatorname{var}_\theta(\hat{\theta}(X))\mathcal{I}_n(\theta),$$

which yields the first result. If $\hat{\theta}$ is unbiased, then $\frac{\mathrm{d}}{\mathrm{d}\theta} E_\theta[\hat{\theta}(X)] = \frac{\mathrm{d}}{\mathrm{d}\theta}\theta = 1$, which yields the second result. □

Some consequences:

1. If we find an unbiased estimator with variance equal to $1/\mathcal{I}_n(\theta)$, we have found the best unbiased estimator in the sense that it achieves the lowest MSE among all unbiased estimators, uniformly over $\theta \in \Theta$. Such an estimator is called uniformly minimum variance unbiased (UMVU)

2. If $\delta(X)$ is an estimator of $g(\theta)$, with bias $E[\delta(X) - g(\theta)] = b(\theta)$, then we can write the bound as
$$\operatorname{var}_\theta(\delta(X)) \geq \frac{(b'(\theta) + g'(\theta))^2}{\mathcal{I}_n(\theta)}.$$

*Remark 6.* If $\theta$ is a vector, and $\hat{\theta}$ is unbiased, then the bound can be written as $\operatorname{var}(\hat{\theta}) - \mathcal{I}_n(\theta)^{-1} \geq 0$ with the inequality meaning that the difference is positive semi-definite.

*Example 5.* Let us calculate the bound for random sample $X_1, \ldots, X_n$ from Bernoulli$(\theta)$ distribution. We have already seen that $\mathcal{I}_1(\theta) = 1/(\theta(1-\theta))$ in this case. So Fisher information for the sample is $\mathcal{I}_n(\theta) = n/(\theta(1-\theta))$. Thus, any unbiased estimator of $\theta$, under some regularity conditions, has variance no smaller than $\theta(1-\theta)/n$. On the other hand, let $\hat{\theta} = \overline{X}_n = \sum_{i=1}^n X_i/n$ be an estimator of $\theta$. Then $E_\theta[\hat{\theta}] = \theta$, i.e. $\hat{\theta}$ is unbiased, and $\operatorname{var}(\hat{\theta}) = \theta(1-\theta)/n$, which coincides with the information inequality Thus, $\overline{X}_n$ is the UMVU estimator of $\theta$. ⊠

*Example 6 (CB, Example 7.3.13).* Let us now consider a counterexample. Let $X \sim U[0, \theta]$. Then $f_1(x \mid \theta) = \mathbb{1}\{0 \leq x \leq \theta\}/\theta$, so that $\ell_1(\theta \mid x) = -\log(\theta) \cdot \mathbb{1}\{0 \leq x \leq \theta\}$, and $\mathcal{S}_1(\theta \mid x) = -1/\theta$. Thus, $-E[\partial \mathcal{S}_1(\theta \mid X)/\partial\theta] = -1/\theta^2 < 0$, while the Fisher information is positive, $E[\mathcal{S}_1^2(\theta \mid X)] = 1/\theta^2$. Thus, the second information equality does not hold in this example. The reason is that the support of the distribution depends on $\theta$. Moreover, consider an estimator $\hat{\theta} = ((n+1)/n)X_{(n)}$ of $\theta$ based on a random sample $X_1, \ldots, X_n$.

Then $E_\theta[X_{(n)}] = \theta$ and one can show that

$$\operatorname{var}(\hat{\theta}) = ((n+1)^2/n^2)\operatorname{var}(X_{(n)}) = \theta^2/(n(n+2)).$$

So $\hat{\theta}$ is unbiased, but its variance is smaller than $1/\mathcal{I}_n(\theta) = \theta^2/n^2$. Thus, the information bound does not hold in this example either, because it requires that the support $\mathcal{X}$ is independent of the parameter. ⊠

## 4. ATTAINABILITY OF THE INFORMATION BOUND

Can we always attain the bound in the sense that there exists an unbiased estimator with variance equal to $1/\mathcal{I}_n(\theta)$? To answer this question note that the crucial step in the derivation of the bound was the Cauchy-Schwarz inequality $\operatorname{cov}_\theta(\hat{\theta}(X), \mathcal{S}_n(\theta \mid X))^2 \leq \operatorname{var}(\hat{\theta}(X))\mathcal{I}_n(\theta)$. Thus, an unbiased estimator $\hat{\theta}(X)$ will attain the information inequality if and only if the inequality above holds as an equality which, in turn, happens if and only if $\hat{\theta}(X)$ and $\mathcal{S}_n(\theta \mid X)$ are linearly dependent. In this case there exist functions $a(\theta)$ and $b(\theta)$ such that $\mathcal{S}_n(\theta \mid X) = a(\theta)(\hat{\theta}(X) - b(\theta))$. By the first information equality, $E[\mathcal{S}_n(\theta \mid X)] = 0$. Since $\hat{\theta}(X)$ is unbiased, this implies that $b(\theta) = \theta$. Thus, there exists an unbiased estimator which attains the information bound if and only if there exists some function $a(\theta)$ such that

$$\mathcal{S}_n(\theta \mid x)/a(\theta) + \theta$$

does not depend on $\theta$ (Corollary 7.3.15 in CB), which also gives a direct recipe for constructing such an estimator.

*Digression.* One can show that this is possible in only a limited range of cases: only if the underlying family of distributions constitutes what's called an exponential family (see Chapter 2 in Lehmann and Casella (1998) for references). ⊠

*Example 7 (CB, Example 7.3.14, 7.3.16).* As an example, let $X_1, \ldots, X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, and put $\theta = (\mu, \sigma^2)'$. Then the log-likelihood equals

$$\ell_n(\theta \mid X) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2,$$

which implies

$$\frac{\partial \ell_n(\theta \mid X)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^4} = \frac{n}{2\sigma^4}\left(\sum_{i=1}^n (X_i - \mu)^2/n - \sigma^2\right),$$

and

$$\frac{\partial \ell_n(\theta \mid X)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu) = \frac{n}{\sigma^2}(\overline{X}_n - \mu).$$

Thus, whether $\sigma^2$ is known, we can always make $\frac{\partial \ell_n(\theta|X)}{\partial \mu}/a(\theta) + \mu$ not depend on $\theta$ by setting $a(\theta) = n/\sigma^2$, which gives $\hat{\theta} = \overline{X}_n$ as the best unbiased estimator of $\mu$.

What about $\sigma^2$? Suppose $\mu$ is known. We need to find $a(\sigma^2)$ such that $\frac{\partial \ell_n(\theta|X)}{\partial \sigma^2}/a(\sigma^2) + \sigma^2$ doesn't depend on $\sigma^2$. This is possible by letting $a(\sigma^2) = 2\sigma^4/n$, which yields the estimator $\hat{\sigma}^2 = \sum_i (X_i - \mu)^2/n$ (which is the maximum likelihood estimator). Its variance is given by the inverse of

$$\mathcal{I}_n(\sigma) = -\frac{\partial}{\partial \sigma^2}\left(\frac{n}{2\sigma^4}\right) E\left[\left(\sum_{i=1}^{n}(X_i-\mu)^2/n - \sigma^2\right)\right] + \frac{n}{2\sigma^4}\frac{\partial}{\partial \sigma^2}\sigma^2 = \frac{n}{2\sigma^4}.$$

If $\mu$ is unknown, then by analogous reasoning, we need to find $a(\theta)$ such that $\frac{n}{2\sigma^4 a(\theta)} \cdot (\sum_{i=1}^{n}(X_i - \mu)^2/n - \sigma^2) + \sigma^2$ does not depend on unknown parameters. This is not possible when $\mu$ is unknown. Therefore, there is no unbiased estimator of $\sigma^2$ with variance given by the (2,2) element of the inverse of

$$\mathcal{I}_n(\mu, \sigma) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

So the information bound for the variance of unbiased estimators of $\sigma^2$ is given by $2\sigma^4/n$ whether $\mu$ is known, but it is achievable only when $\mu$ is known. ⊠

## 5. METHOD OF MOMENTS

This is another general method for constructing estimators. Let $X_1, \ldots, X_n$ be a random sample from some distribution. There is a $k$-dimensional parameter of interest $\theta$ that satisfies the system of equations

$$E[g(X_i, \theta)] = 0,$$

where $g$ is a vector-valued function (with dimension $k$). Then the method of moments estimator $\hat{\theta}_{MM}$ of $\theta$ is the solution of the above system of equations if we replace the population expectation by sample average

$$\frac{1}{n}\sum_i g(X_i, \theta) = 0.$$

It is implicitly assumed here that the solution exists and is unique. In the simplest case, the moments have the separable form $g_j(X_i, \theta) = X_i^j - f_j(\theta)$.

*Example 8 (CB, Example 7.2.1).* Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Then $E[X_i] = \mu$ and $E[X_i^2] = \mu^2 + \sigma^2$. Thus, $\hat{\mu}_{MM} = \sum_{i=1}^{n} X_i/n$ and $\hat{\mu}_{MM}^2 + \hat{\sigma}_{MM}^2 = \sum_{i=1}^{n} X_i^2/n$. So $\hat{\sigma}_{MM}^2 = \sum_{i=1}^{n} X_i^2/n - (\sum_{i=1}^{n} X_i/n)^2$. ⊠

*Example 9 (CB, Example 7.2.2).* Let $X_1, \ldots, X_n$ be a random sample from Binomial$(k, p)$,

with both $k, p$ unknown. Now,

$$P(X_i = j) = \binom{k}{j} p^j (1-p)^{k-j}.$$

So $E[X_i - kp] = 0$ and $E[X_i^2] = kp(1-p) + k^2p^2$. The first equation implies $\hat{p}_{MM}\hat{k}_{MM} = \overline{X}_n$. Plugging this into the second equation, we get

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \overline{X}_n(1-\hat{p}_{MM}) + \overline{X}_n^2 \implies \hat{p}_{MM} = 1 - \frac{n-1}{n}\frac{S_n^2}{\overline{X}_n},$$

and thus $\hat{k}_{MM} = \overline{X}_n^2/(\overline{X}_n - (n-1)S_n^2/n)$. Note that it's possible to get negative estimates of $k$ and $p$. ⊠

The idea of the method of moments is old, dating back to at least Pearson in the 19th century. There is a generalization of it which allows for more moments than the dimensionality of the parameter. It is called generalized method of moments (GMM) and will be studied extensively later on, as the main workhorse of econometrics.

*Remark 7.* MLE is always inside $\Theta$ (Why?). In contrast, method of moment estimates may not be.

*Remark 8.* An immediate consequence of Theorem 3 is that the maximum likelihood estimator based on a random sample $X_1, \ldots, X_n$ can be viewed as a method of moments estimator based on the equality $E[\mathcal{S}_1(\theta \mid X_i)] = 0$.

## REFERENCES


Cramér, Harald. 1946. "A Contribution to the Theory of Statistical Estimation." *Scandinavian Actuarial Journal* 1946, no. 1 (January): 85–94. https://doi.org/10.1080/03461238.1946.10419631.

Durrett, Rick. 2019. *Probability: Theory and Examples.* 5th ed. New York, NY: Cambridge University Press. https://doi.org/doi.org/10.1017/9781108591034.

Fréchet, Maurice. 1943. "Sur l'extension de Certaines Evaluations Statistiques Au Cas de Petits Echantillons." *Revue de l'Institut International de Statistique* 11 (3/4): 182. https://doi.org/10.2307/1401114.

Lehmann, Erich L., and George Casella. 1998. *Theory of Point Estimation.* 2nd ed. New York, NY: Springer. https://doi.org/doi.org/10.1007/b98854.

Newey, Whitney K., and Daniel L. McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." Chap. 36 in *Handbook of Econometrics,* edited by Robert F. Engle and Daniel L. McFadden, 4:2111–2245. New York, NY: Elsevier. https://doi.org/10.1016/S1573-4412(05)80005-4.

Rao, C. Radhakrishna. 1945. "Information and the Accuracy Attainable in the Estimation of Statistical Parameters." *Bulletin of Calcutta Mathematical Society* 37 (3): 81–91. https://doi.org/10.1007/978-1-4612-0919-5_16.

Romano, Joseph P., and Andrew F. Siegel. 1986. *Counterexamples in Probability and Statistics.* New York, NY: Routledge. https://doi.org/10.1201/9781315140421.

Savage, Leonard J. 1972. *The Foundations of Statistics.* 2nd ed. New York, NY: Dover Publications.