

Mean Squared Error and Maximum Likelihood: Lecture XX

Charles B. Moss

October 12, 2026

Reading Question

- How would we justify $T = (x_1 + x_2) / 2$ as an estimator. First, do I need to justify it? Second, I could justify the estimator in two ways:
- Method of Moments estimator – I know that the expected value of the sum of Bernoulli variables is $n\theta$. Hence

$$\sum_{i=1}^n x_i = n\theta \Rightarrow \hat{\theta}_{MOM} = \frac{\sum_{i=1}^n x_i}{n} \quad (1)$$

or the method of moments estimate of the mean is simply the first moment.

- Alternatively, we could rely on maximum likelihood. The likelihood of the sample is simply a binomial function

$$L(\{x_i\}_{i=1}^n | \theta) = \binom{n}{r = \sum_{i=1}^n x_i} \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{(n - \sum_{i=1}^n x_i)} \quad (2)$$

- It is typically easier to do the calculus in log space

$$\begin{aligned} \ln(L) &= \ln(C_r^n) + \ln(\theta) \sum_{i=1}^n x_i + \left(n - \sum_{i=1}^n x_i \right) \ln(1 - \theta) \\ &\Rightarrow \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1 - \theta} = 0 \end{aligned} \quad (3)$$

$$\left(\sum_{i=1}^n x_i\right) \left(\frac{1}{\theta} + \frac{1}{1-\theta}\right) = \frac{n}{1-\theta}$$

$$\left(\sum_{i=1}^n x_i\right) \left(\frac{1-\theta}{\theta(1-\theta)} + \frac{\theta}{\theta(1-\theta)}\right) = \frac{n}{1-\theta} \quad (4)$$

$$\left(\sum_{i=1}^n x_i\right) \frac{1}{\theta(1-\theta)} = \frac{n}{1-\theta} \Rightarrow \hat{\theta}_{ML} = \frac{\sum_{i=1}^n x_i}{n}$$

- Same estimator - two justifications. What is the difference?

- 1 Mean Squared Error
- 2 Strategies for Choosing an Estimator
- 3 Best Linear Unbiased Estimator
- 4 Asymptotic Properties
- 5 Maximum Likelihood
- 6 Lomax Distribution

Mean Squared Error

- As stated in our discussion on closeness, one partial measure for the goodness of an estimator is

$$E \left[\left(\hat{\theta} - \theta \right)^2 \right] \quad (5)$$

- In the preceding example, the mean square error of the estimate can be written as

$$E \left[(T - \theta)^2 \right] \quad (6)$$

where θ is the true parameter value between zero and one.

- This expected value is conditioned on the probability of T at each level value of θ . For example, if $\theta = 0$ then the probability of each X becomes

$$P[X, \theta] = \theta^X (1 - \theta)^{1-X} \quad (7)$$

If the two events are independent

$$P[X_1, X_2, \theta] = \theta^{X_1+X_2} (1 - \theta)^{1-X_1-X_2} \quad (8)$$

The mean squared error at any theta can then be derived as

$$MSE(\theta) = P[0, 0, \theta] (0 - \theta)^2 + 2P[0, 1, \theta] (0.5 - \theta)^2 + P[1, 1, \theta] (1 - \theta)^2. \quad (9)$$

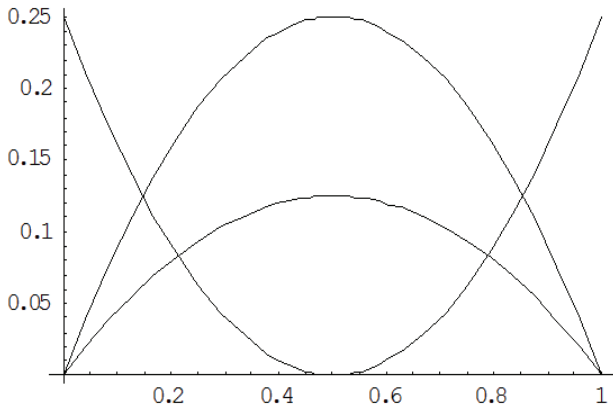
- The mean squared error for S can similarly be computed as

$$MSE(\theta) = P[0, \theta](0 - \theta)^2 + P[1, \theta](1 - \theta)^2 \quad (10)$$

- Finally, the mean square error of W can be written as

$$MSE(\theta) = (0.5 - \theta)^2 \quad (11)$$

- The mean squared error for each estimator is presented in Figure 1.



- **Definition 7.2.1.** Let X and Y be two estimators of θ . We say that X is better (or more efficient) than Y if $E(X - \theta)^2 \leq E(Y - \theta)^2$ for all $\theta \in \Theta$ and strictly less than for at least one $\theta \in \Theta$.
 - When an estimator is dominated by another estimator, the dominated estimator is inadmissible.
 - **Definition 7.2.2.** Let $\hat{\theta}$ be an estimator of θ . We say that $\hat{\theta}$ is inadmissible if there is another estimator which is better in the sense that it produces a lower mean square error of the estimate. An estimator that is not inadmissible is admissible.

Strategies for Choosing an Estimator

- Subjective strategy: This strategy considers the likely outcome of θ and selects the estimator that is best in that likely neighborhood.
- Minimax Strategy: According to the minimax strategy, we choose the estimator for which the largest possible value of the mean squared error is the smallest.
 - **Definition 7.2.3:** Let $\hat{\theta}$ be an estimator of θ . It is a minimax estimator if for any other estimator of $\tilde{\theta}$, we have

$$\max_{\theta} E \left[\left(\hat{\theta} - \theta \right)^2 \right] \leq \max_{\theta} E \left[\left(\tilde{\theta} - \theta \right)^2 \right] \quad (12)$$

Best Linear Unbiased Estimator

- **Definition 7.2.4:** $\hat{\theta}$ is said to be an unbiased estimator of θ if $E[\hat{\theta}] = \theta$ for all $\theta \in \Theta$. We call $E[\hat{\theta} - \theta]$ bias.
- In our previous discussion T and S are unbiased estimators while W is biased.
- **Theorem 7.2.10:** The mean squared error is the sum of the variance and the bias squared. That is, for any estimator $\hat{\theta}$ of θ

$$E\left[(\hat{\theta} - \theta)^2\right] = V(\theta) + \left(E[\hat{\theta}] - \theta\right)^2 \quad (13)$$

- **Theorem 7.2.11** Let $\{X_i\}$, $i = 1, 2, \dots, n$ be independent and have a common mean μ and variance σ^2 . Consider the class of linear estimators of μ which can be written in the form

$$\hat{\mu} = \bar{X} = \sum_{i=1}^n a_i X_i \quad (14)$$

and impose the unbiasedness condition

$$E \left[\sum_{i=1}^n a_i X_i \right] = \mu \quad (15)$$

Then

$$V(\bar{X}) \leq V \left(\sum_{i=1}^n a_i X_i \right) \quad (16)$$

for all a_i satisfying the unbiasedness condition. Further, this condition holds with equality only for $a_i = 1/n$.

- To prove these points note that the a_i s must sum to one for unbiasedness

$$E \left[\sum_{i=1}^n a_i X_i \right] = \sum_{i=1}^n a_i E[X_i] = \sum_{i=1}^n a_i \mu = \mu \sum_{i=1}^n a_i \quad (17)$$

Therefore $\sum_{i=1}^n a_i = 1$ results in an unbiased estimator.

- The final condition can be demonstrated through the identity

$$\sum_{i=1}^n \left(a_i - \frac{1}{n} \right)^2 = \sum_{i=1}^n a_i^2 - \frac{2}{n} \sum_{i=1}^n a_i + \frac{1}{n} \quad (18)$$

If $a_i = 1/n$ then $\sum_{i=1}^n (a_i - 1/n)^2 = 0$. Thus, any other a_i must yield a higher variance $\left(\sum_{i=1}^n (a_i - 1/n)^2 \geq 0 \right)$

$$\sum_{i=1}^n \left(a_i - \frac{1}{n} \right)^2 \sigma^2 \geq 0 \Rightarrow \sum_{i=1}^n a_i^2 \sigma^2 \geq \left(\frac{1}{n} \right) \sigma^2 \quad (19)$$

- **Theorem 7.2.12 Amemiya p. 129:** Consider the problem of minimizing

$$\sum_{i=1}^n a_i^2 \quad (20)$$

with respect to $\{a_i\}$ subject to the condition

$$\sum_{i=1}^n a_i b_i = 1. \quad (21)$$

- To derive this solution Amemiya starts by imposing the result

$$\sum_{i=1}^n \left(a_i - \frac{b_i}{\sum_{i=1}^n b_i^2} \right)^2 = \sum_{i=1}^n a_i^2 - 2 \frac{\sum_{i=1}^n a_i b_i}{\sum_{i=1}^n b_i^2} + \frac{1}{\sum_{i=1}^n b_i^2} \quad (22)$$

- The trick here is focusing on second and third term

$$\sum_{i=1}^n a_i b_i = 1 \Rightarrow -2 \frac{\sum_{i=1}^n a_i b_i}{\sum_{i=1}^n b_i^2} = -2 \frac{1}{\sum_{i=1}^n b_i^2} \quad (23)$$

- Therefore Equation 22 becomes

$$\sum_{i=1}^N a_i^2 - \frac{1}{\sum_{i=1}^N b_i^2} \quad (24)$$

Asymptotic Properties

- **Definition 7.2.5.** We say that $\hat{\theta}$ is a consistent estimator of θ if

$$\text{plim}_{n \rightarrow \infty} \hat{\theta} = \theta \quad (25)$$

Maximum Likelihood

- The basic concept behind maximum likelihood estimation is to choose that set of parameters that maximize the likelihood of drawing a particular sample.
 - For example, suppose that we know that a sample of random variables have a variance of one, but an unknown mean.
 - Let the sample be $X = \{ 5, 6, 7, 8, 10 \}$. The probability of each of these points based on the unknown mean(μ) can be written as

$$\begin{aligned}
 f(5|\mu) &= \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{(5-\mu)^2}{2} \right] \\
 f(6|\mu) &= \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{(6-\mu)^2}{2} \right] \\
 &\vdots \\
 f(10|\mu) &= \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{(10-\mu)^2}{2} \right]
 \end{aligned} \tag{26}$$

- Continued

- Continued

- Assuming that the sample is independent so that the joint distribution function can be written as the product of the marginal distribution functions, the probability of drawing the entire sample based on a given mean can then be written as

$$L(X|\mu) = \frac{1}{(2\pi)^{-5/2}} \exp \left[-\frac{(5-\mu)^2}{2} - \frac{(6-\mu)^2}{2} - \dots - \frac{(10-\mu)^2}{2} \right] \quad (27)$$

- The function $L(X|\mu)$ is typically referred to as the likelihood function.

- Continued

- Continued

- The value of μ that maximize the likelihood function of the sample can then be defined by

$$\max_{\mu} L(X|\mu) \quad (28)$$

Under the current scenario, we find it easier, however, to maximize the natural logarithm of the likelihood function

$$\begin{aligned} \max_{\mu} \ln(L(X|\mu)) &\Rightarrow \frac{\partial}{\partial \mu} \left[K - \frac{(5-\mu)^2}{2} - \frac{(6-\mu)^2}{2} - \dots - \frac{(10-\mu)^2}{2} \right] \\ &= -(5-\mu) - (6-\mu) - \dots - (10-\mu) = 0 \\ \hat{\mu}_{MLE} &= \frac{5+6+7+8+9+10}{6} \end{aligned} \quad (29)$$

Lomax Distribution

- The Lomax distribution is a variant of the Pareto distribution family

$$f(x; \alpha, \lambda) = \frac{\alpha}{\lambda} \left[1 + \frac{x}{\lambda} \right]^{-(\alpha+1)} \quad (30)$$

- This yields the log-likelihood function for the sample

$$\text{Log}(L) = n \ln(\alpha) - n \ln(\lambda) - (\alpha + 1) \sum_{i=1}^n \ln \left(1 + \frac{x_i}{\lambda} \right) \quad (31)$$

- It is a little difficult to analytically solve for both variables simultaneously, so assume initially assume that we know λ and simply wish to solve for α .
- Taking the first-order condition of the log-likelihood function with respect to α

$$\begin{aligned}\frac{\partial \text{Log}(L)}{\partial \alpha} &= \frac{n}{\alpha} - \sum_{i=1}^n \ln \left(1 + \frac{x_i}{\lambda} \right) = 0 \\ \Rightarrow \hat{\alpha}_{ML} &= \frac{n}{\sum_{i=1}^n \ln \left(1 + \frac{x_i}{\lambda} \right)}\end{aligned}\tag{32}$$

- After the simple mechanics of finding the estimate value, I would like you to contemplate the second-order Taylor series expansion of the log-Likelihood function

$$\ln(L) = f(\beta_0) + \nabla_x f(\beta) |_{\beta=\beta_0} (\beta - \beta_0) + \frac{1}{2} (\beta - \beta_0)' \nabla_{xx}^2 f(\beta) |_{\beta=\beta_0} (\beta - \beta_0) + O(\beta^3) \quad (33)$$

- Consider some substitutions in Equation 33. First, assume that I want to substitute $z = \beta - \beta_0$ which I want to be a mean zero variable. To do this, I need to change the variance term. Specifically, I want to substitute

$$\Sigma = - [\nabla_{xx}^2 f(\beta) |_{\beta=\beta_0}]^{-1} \quad (34)$$

- Making these substitutions, I have

$$\ln(L) = f(\beta_0) + \nabla_x f(\beta) |_{\beta=\beta_0} z - \frac{1}{2} z' \Sigma z + O(z^3) \quad (35)$$

- Basically, we have two sets of results:
 - Consistency $\lim_{n \rightarrow \infty} \hat{\beta} = \beta$.
 - The variance term referred to as the Cramer-Rao lower bound

$$\begin{aligned} \hat{\Sigma}_{ML} &\approx \left(E \left[-\nabla_{xx}^2 f(\beta) |_{\beta=\hat{\beta}} \right] \right)^{-1} \\ &\approx \left[-\frac{1}{n} \nabla_{xx}^2 f(\beta) |_{\beta=\hat{\beta}} \right]^{-1} \end{aligned} \quad (36)$$

A Little More of the Lomax Mechanics

- The First-Order Conditions

$$\nabla_{\theta} \ln(L) = \begin{bmatrix} \frac{n}{\alpha} - \sum_{i=1}^n \ln\left(1 + \frac{x_i}{\lambda}\right) \\ -\frac{n}{\lambda} + (\alpha + 1) \sum_{i=1}^n \frac{x_i}{\lambda^2 \left[1 + \frac{x_i}{\lambda}\right]} \end{bmatrix}' \quad (37)$$

where $\theta = (\alpha \quad \lambda)'$.

$$\nabla_{\theta\theta}^2 \ln(L) = \begin{bmatrix} -\frac{n}{\alpha^2} & \sum_{i=1}^n \frac{x_i}{\lambda^2 \left(1 + \frac{x_i}{\lambda}\right)} \\ \sum_{i=1}^n \frac{x_i}{\lambda^2 \left(1 + \frac{x_i}{\lambda}\right)} & \frac{n}{\lambda^2} + (1 + \alpha) \sum_{i=1}^n \left[\frac{x_i^2}{\lambda^4 \left(1 + \frac{x_i}{\lambda}\right)^2} - \frac{2x_i}{\lambda^3 \left(1 + \frac{x_i}{\lambda}\right)} \right] \end{bmatrix} \quad (38)$$

- Using the R-Code 'Lecture21.R' with 250 observations, the optimal vector is

$$\hat{\theta} = \begin{bmatrix} \hat{\alpha} \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} 2.2983 \\ 0.5831 \end{bmatrix} \quad (39)$$

- Putting these values into Equation 37 yields

$$\begin{aligned} \nabla_{\theta} \ln(L) &= \begin{bmatrix} 2.4200e - 06 \\ 1.9156e - 05 \end{bmatrix} \\ \text{grad}(\text{loglike}, \text{bML}\$pars) &= \begin{bmatrix} -2.4219e - 06 \\ -1.9173e - 06 \end{bmatrix} \end{aligned} \quad (40)$$

where `grad()` is the gradient routine from R.

- Note that the difference between signs in Equation 40 is due to the fact that R_{solnp} minimizes - so we multiply the original objective function by negative 1. Basically, the solutions are identical.

- Next, we examine the result for our Hessian in Equation 38 and the numerical approximation from R

$$\nabla_{\theta\theta} \ln(L) = \begin{bmatrix} -47.3278 & 129.9951 \\ 129.995 & -389.9551 \end{bmatrix}$$

$$\text{hess}(\text{loglike}, \text{bML\$pars}) = \begin{bmatrix} 47.3278 & -129.9951 \\ -192.9951 & 389.9551 \end{bmatrix} \quad (41)$$

Again the results are identical except for the signs.

- In addition, the Rsolnp routine provides an estimate of the Hessian

$$\hat{H} = \begin{bmatrix} 48.6647 & -132.5905 \\ -132.5905 & 394.9814 \end{bmatrix} \quad (42)$$

- As a final point, note that likelihood function is a product function

$$\prod_{i=1}^n \frac{\alpha}{\lambda} \left[1 + \frac{x_i}{\lambda} \right]^{-(\alpha+1)} \quad (43)$$

If we wanted an average - across all events, we would take the n^{th} root of this result. Continuing this thought process through, we would divide the log-likelihood function by n . Hence, Equation 31 becomes

$$\frac{1}{n} \text{Log} (L) = \ln (\alpha) - \ln (\lambda) - \frac{(\alpha + 1)}{n} \sum_{i=1}^n \ln \left(1 + \frac{x_i}{\lambda} \right) \quad (44)$$

- Since we are only worried about the gradient going to zero, dividing by n is only really important for the Hessian. In addition, we correct for the negative term in Equation 34 so our estimated variance term from the Cramer-Rao lower bound becomes

$$\hat{\Sigma} = \begin{bmatrix} \frac{1}{\alpha^2} & -\frac{1}{n} \sum_{i=1}^n \frac{x_i}{\lambda^2 \left(1 + \frac{x_i}{\lambda}\right)} \\ -\frac{1}{n} \sum_{i=1}^n \frac{x_i}{\lambda^2 \left(1 + \frac{x_i}{\lambda}\right)} & -\frac{a}{\lambda^2} - \frac{(1+\alpha)}{n} \sum_{i=1}^n \left[\frac{x_i^2}{\lambda^4 \left(1 + \frac{x_i}{\lambda}\right)^2} - \frac{2x_i}{\lambda^3 \left(1 + \frac{x_i}{\lambda}\right)} \right] \end{bmatrix} \quad (45)$$

- With these changes, the Cramer-Rao lower bound of the variance becomes

$$\begin{aligned}\Sigma_{CR} &= \begin{bmatrix} 0.1893 & -0.5160 \\ -0.5160 & 1.5598 \end{bmatrix}^{-1} \\ &= \begin{bmatrix} 62.6135 & 20.8728 \\ 20.8728 & 7.5992 \end{bmatrix}\end{aligned}\tag{46}$$

Readings

- Moss 7.4
- David Salsburg. 2002. *A Lady Tasting Tea* – Amazon eTextbook \$12.99.