MSSC 6010 / Comp. Probability

Instructor: Mehdi Maadooliat

Chapter 7

Department of Mathematics, Statistics and Computer Science

Special thanks to Prof. Ana Militino for providing the original slides of the book.

MARQUETTE UNIVERSITY
Be The Difference.

Chapter 7

Point Estimation

7.1 Introduction

- In general, the **pdf** of a random variable X is $f(x \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of parameters that characterize the **pdf**.

- The vector of parameters $\boldsymbol{\theta}$ is defined in a parameter space denoted Θ. For each value of $\boldsymbol{\theta} \in \Theta$, there is a different **pdf**.

1

- To obtain possible values for the vector of parameters, a random sample from the population of interest is taken and statistics called **estimators** are constructed.
- The values of the estimators are called **point estimates**.
- For example, $\overline{X}$ may be used as a point estimator for μ, in which case $\bar{x}$ is a point estimate of μ.
- Since estimators are statistics or functions of random variables, they themselves are random variables.
- Studying the sampling distributions of estimators as well as their statistical properties such as mean square error, bias or unbiasedness, efficiency, consistency, and robustness, all of which will be defined in this chapter, will give guidelines about which estimators to employ.

2

7.2 PROPERTIES OF POINT ESTIMATORS

7.2.1 Mean Square Error

- The goodness of an estimator is related to how close its estimates are to the true parameter.
- The difference between an estimator T for an unknown parameter θ, and the parameter θ itself is called the error.
- Since this quantity can be either positive or negative, it is common to square the error so that various estimators $T_1, T_2, \ldots,$ can be compared using a nonnegative measure of error.
- To that end, the **mean square error** of an estimator, denoted $MSE[T]$, is defined as $MSE[T] = E\left[(T-\theta)^2\right]$.
- Estimators with small MSE will have a distribution such that the values in the distribution will be close to the true parameter.

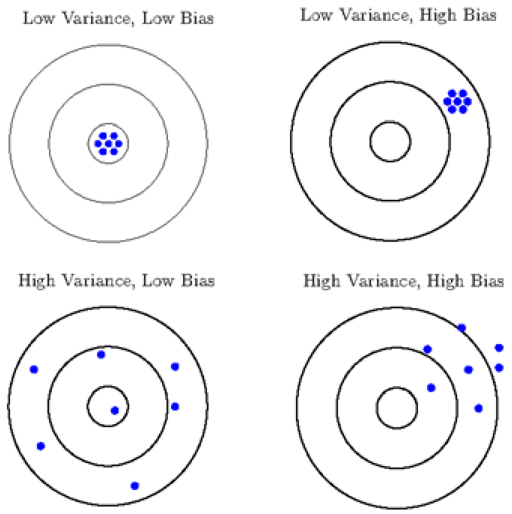
3

WHAT IS MSE?

- In fact, MSE consists of two nonnegative components, the variance of the estimator T and the squared bias of the estimator T, where bias is defined as $E[T] - \theta$ since

$$
\begin{aligned}
MSE[T] &= E\left[(T - E[T] + E[T] - \theta)^2\right] \\
&= E[T - E[T]]^2 + E\left[(E[T] - \theta)^2\right] \qquad (7.1) \\
&+ 2E\left[(T - E[T])(E[T] - \theta)\right] \\
&= \text{var}[T] + (E[T] - \theta)^2 + 2(E[T] - E[T])(E[T] - \theta) \\
&= \text{var}[T] + (E[T] - \theta)^2 \\
&= \text{var}[T] + (Bias[T])^2. \qquad (7.2)
\end{aligned}
$$

4



Estimators that minimize MSE for all possible values of θ do not always exist.
In other words, an estimator may have the minimum MSE for some values of θ and not others.

5

MARQUETTE UNIVERSITY Be The Difference.

7.2.2 Unbiased Estimators

When $E[T] = \theta$, T is an **unbiased** estimator of θ.
When an estimator is unbiased, its MSE is equal to its variance, that is, $MSE[T] = \text{var}[T]$. On the other hand, when $E[T] \neq \theta$, the estimator is biased.

Example 7.1 Show that the sample mean and the sample variance are unbiased estimators of the population mean and the population variance respectively.

$$E\left[\overline{X}\right] = E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = \frac{n\mu}{n} = \mu$$

6

MARQUETTE UNIVERSITY Be The Difference.

Solution: To show that S^2 is an unbiased estimator of σ^2, use the fact that

$$\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n} (X_i - \mu)^2 - n\left(\mu - \overline{X}\right)^2$$

$$E\left[S^2\right] = E\left[\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}\right] = E\left[\frac{\sum_{i=1}^{n}(X_i - \mu)^2 - n\left(\mu - \overline{X}\right)^2}{n-1}\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} E\left[(X_i - \mu)\right]^2 - nE\left[\left(\overline{X} - \mu\right)^2\right]\right]$$

$$= \frac{1}{n-1}\left[n\sigma^2 - n\frac{\sigma^2}{n}\right] = \sigma^2$$

■

7

MARQUETTE UNIVERSITY Be The Difference.

Example 7.2 Suppose $X \sim Pois(\lambda)$ where λ is unknown. Show
(a) $\overline{X}$ is an unbiased estimator of λ.
(b) $2\overline{X}$ is an unbiased estimator of 2λ.
(c) $\overline{X}^2$ is a biased estimator of λ^2.

Solution: To solve the problems, keep in mind that if $X \sim Pois(\lambda)$, $E[X] = \lambda$, and $\text{var}[X] = \lambda$.

(a) Since $E\left[\overline{X}\right] = E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = \sum_{i=1}^{n} \frac{E[X_i]}{n} = \frac{n\lambda}{n} = \lambda$, it follows that $\overline{X}$ is an unbiased estimator of λ.

(b) Since $E\left[2\overline{X}\right] = 2E\left[\overline{X}\right] = 2\lambda$, it follows that $2\overline{X}$ is an unbiased estimator of 2λ.

(c) Since $E\left[\overline{X}^2\right] = \text{var}\left[\overline{X}\right] + \mu_{\overline{X}}^2 = \frac{\lambda}{n} + \lambda^2$, it follows that $\overline{X}^2$ is a biased estimator of λ^2. However, $\overline{X}^2$ is an asymptotically unbiased estimator of λ^2. That is, as n tends to infinity, the estimator becomes unbiased. ■

8

MARQUETTE UNIVERSITY Be The Difference.

Example 7.3 Suppose $\{X_1, X_2, \ldots, X_n\}$ is a random sample from a $N(\mu, \sigma)$ distribution. Show that S is a biased estimator of σ.

Solution: Recall that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. Let $X = \frac{(n-1)S^2}{\sigma^2}$ and take the square root and the expected value of both sides:

$$E\left[\sqrt{X}\right] = E\left[\frac{\sqrt{n-1}}{\sigma} \cdot S\right].$$

Since $X \sim \chi^2_{n-1}$, the expected value of $\sqrt{X}$ is $\int_{-\infty}^{\infty} \sqrt{x} f(x) dx$, where $f(x)$ is the **pdf** of a chi-square random variable.

9

SOLUTION CONT'D

$$E\left[\sqrt{X}\right] = \int_0^\infty \sqrt{x}\frac{1}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{n-1}{2}}}x^{\frac{n-1}{2}-1}e^{-\frac{x}{2}}dx$$

$$= \frac{1}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{n-1}{2}}}\int_0^\infty x^{\frac{n-1}{2}-1+\frac{1}{2}}e^{-\frac{x}{2}}dx$$

$$= \frac{1}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{n-1}{2}}}\int_0^\infty x^{\frac{n}{2}-1}e^{-\frac{x}{2}}dx \qquad (7.3)$$

Next, use the change of variable $x/2 = t$ where $dx = 2dt$ in an attempt to force the right hand side of (7.3) to look like a gamma function. Specifically, recall that $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$ for $\alpha > 0$.

10

SOLUTION CONT'D

$$E\left[\sqrt{X}\right] = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{n-1}{2}}}\int_0^\infty (2t)^{\frac{n}{2}-1}\,e^{-t}\,2\,dt$$

$$= \frac{2^{\frac{n}{2}}}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{n-1}{2}}}\int_0^\infty t^{\frac{n}{2}-1}\,e^{-t}\,dt = \frac{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}$$

Since

$$E\left[\sqrt{X}\right] = E\left[\frac{\sqrt{n-1}}{\sigma}\,S\right] = \frac{\sqrt{2}\,\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})},$$

it follows that

$$E[S] = \sigma\,\frac{\sqrt{2}\Gamma\left(\frac{n}{2}\right)}{\sqrt{n-1}\Gamma\left(\frac{n-1}{2}\right)} \neq \sigma \qquad (7.4)$$

Therefore, S is a biased estimator of σ. ■

Note that the coefficient $\frac{\sqrt{2}\Gamma(\frac{n}{2})}{\sqrt{n-1}\Gamma\left(\frac{n-1}{2}\right)}$ is virtually 1 for values of $n \geq 20$.

11

7.2.3 EFFICIENCY

A desirable property of a good estimator is not only to be unbiased, but also to have a small variance; which translates into a small MSE for estimators, regardless of whether they are biased or unbiased. One way to compare the MSE of two estimators is by using **relative efficiency**. Given two estimators T_1 and T_2, the efficiency of T_1 relative to T_2, written $\mathit{eff}(T_1, T_2)$, is

$$\mathit{eff}(T_1, T_2) = \frac{MSE\,[T_2]}{MSE\,[T_1]}. \tag{7.5}$$

When the estimators in (7.5) are unbiased, the efficiency of T_1 relative to T_2 is simply the ratio of estimators variances written

$$\mathit{eff}(T_1, T_2) = \frac{\text{var}\,[T_2]}{\text{var}\,[T_1]}.$$

12

EFFICIENCY CONT'D

- The estimator T_1 is more efficient than the estimator T_2 if for any sample size, $MSE\,[T_1] \leq MSE\,[T_2]$, which then implies that $\mathit{eff}(T_1, T_2) \geq 1$.
- When the estimators are unbiased, the estimator T_1 is more efficient than the estimator T_2 if for any sample size, $\text{var}\,[T_1] \leq \text{var}\,[T_2]$, which also implies that $\mathit{eff}(T_1, T_2) \geq 1$.
- If a choice is to be made among a small number of unbiased estimators, simply compute the variance of all of the estimators and select the estimator with minimum variance.
- However, if the estimator that has the smallest variance among all possible unbiased estimators must be chosen, an infinite number of variances would need to be calculated. Clearly, this is not a viable solution.

13

EFFICIENCY CONT'D

- Thankfully, it can be shown that if $T = \hat{\theta}$ is an unbiased estimator of θ and a random sample of size n, $X_1, X_2, \ldots, X_n$, has **pdf** $f(x|\theta)$, then the variance of the unbiased estimator, $\hat{\theta}$, must satisfy the inequality

$$\text{var}\big[\hat{\theta}\big] \geq \frac{1}{n \cdot E\left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta}\right)^2\right]}, \qquad (7.6)$$

  where $f(X|\theta)$ is the density function of the distribution of interest evaluated at the random variable X.

- In the discrete case, $p(X|\theta)$ is used instead of $f(X|\theta)$.
- In general, the probability distributions of both discrete and continuous distributions are referred to using the notation $f(x)$.
- The inequality in (7.6) is known as the **Cramér-Rao inequality**, and the quantity on the right hand side of the equation is known as the Cramér-Rao Lower Bound (CRLB).

14

CRAMER-RAO LOWER BOUND

DEFINITION 7.1: If $\hat{\theta}$ is an unbiased estimator of θ and

$$\text{var}\big[\hat{\theta}\big] = \frac{1}{n \cdot E\left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta}\right)^2\right]} \qquad (7.7)$$

then $\hat{\theta}$ is a **minimum variance unbiased** estimator of θ.

- Not all parameters have unbiased estimators whose variance equals the CRLB. However, when the variance of an unbiased estimator equals the CRLB, the estimator is **efficient** or **minimum variance**.
- The quantity in the denominator of (7.7) is known as the **Fisher information** about θ that is supplied by the sample.
- That is, the smaller the variance of the estimator, the greater the information.

15

MARQUETTE UNIVERSITY Be The Difference.

Example 7.5 Show that $\overline{X}$ is a minimum variance unbiased estimator of the mean λ of a Poisson population.

Solution: If $X \sim Pois(\lambda)$ then according to Box (**??**), $E[X] = \lambda$, $\text{var}[X] = \lambda$, and the pdf of X is

$$\mathbb{P}(X = x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}. \tag{7.8}$$

- Since $E\left[\overline{X}\right] = \sum_{i=1}^{n} \frac{E[X_i]}{n} = \frac{n\lambda}{n} = \lambda$,
- $\overline{X}$ is an unbiased estimator of λ, with variance $\frac{\lambda}{n}$ because the $\text{var}\left[\overline{X}\right] = \text{var}\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}\left[X_i\right] = \frac{n\lambda}{n^2} = \frac{\lambda}{n}$.
- Consequently, if the CRLB equals $\frac{\lambda}{n}$, $\overline{X}$ is a minimum variance unbiased estimator of λ according to Definition 7.1.
- By taking the natural logarithm of (7.8),

$$\ln \mathbb{P}(x|\lambda) = x\ln(\lambda) - \lambda - \ln(x!). \tag{7.9}$$

16

MARQUETTE UNIVERSITY Be The Difference.

- Taking the derivative of (7.9) with respect to λ gives

$$\frac{\partial \ln \mathbb{P}(x|\lambda)}{\partial \lambda} = \frac{x}{\lambda} - 1 = \frac{x-\lambda}{\lambda}.$$

- Hence

$$E\left[\left(\frac{\partial \ln \mathbb{P}(X|\lambda)}{\partial \lambda}\right)^2\right] = E\left[\left(\frac{X-\lambda}{\lambda}\right)^2\right] = \frac{E[(X-\lambda)^2]}{\lambda^2} = \frac{\text{var}[X]}{\lambda^2}.$$

- Therefore,

$$E\left[\left(\frac{\partial \ln \mathbb{P}(X|\lambda)}{\partial \lambda}\right)^2\right] = \frac{\text{var}[X]}{\lambda^2} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda},$$

  and the CRLB is

$$\frac{1}{n \cdot E\left[\left(\frac{\partial \ln f(X|\lambda)}{\partial \lambda}\right)^2\right]} = \frac{\lambda}{n}.$$

- Consequently, since $\overline{X}$ is unbiased and $\text{var}\left[\overline{X}\right] = \frac{\lambda}{n}$, it follows that $\overline{X}$ is a minimum variance unbiased estimator of λ.

17

MARQUETTE UNIVERSITY Be The Difference.

Example 7.6 Show that $\overline{X}$ is a minimum variance unbiased estimator of the mean θ of an exponential population.

Solution: If $X \sim Exp\left(\frac{1}{\theta}\right)$ then according to Box (**??**), when using the substitution $\theta = \frac{1}{\lambda}$, $E[X] = \theta$, $\text{var}[X] = \theta^2$, and the **pdf** of X is

$$f(x) = \begin{cases} \frac{1}{\theta}e^{-x/\theta} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}. \tag{7.10}$$

Since $E\left[\overline{X}\right] = \sum_{i=1}^{n} \frac{E[X_i]}{n} = \frac{n\theta}{n} = \theta$, it follows that $\overline{X}$ is an unbiased estimator of θ, with variance $\frac{\theta^2}{n}$ since $\text{var}\left[\overline{X}\right] = \text{var}\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}\left[X_i\right] = \frac{n\theta^2}{n^2} = \frac{\theta^2}{n}$. Consequently, if the CRLB equals $\frac{\theta^2}{n}$, $\overline{X}$ is a minimum variance unbiased estimator of θ according to Definition 7.1. By taking the natural logarithm of (7.10),

$$\ln f(x|\theta) = -\ln(\theta) - \frac{x}{\theta}. \tag{7.11}$$

MARQUETTE UNIVERSITY Be The Difference.

Taking the derivative of (7.11) with respect to θ gives

$$\frac{\partial \ln f(x|\theta)}{\partial \theta} = -\frac{1}{\theta} + \frac{x}{\theta^2} = \frac{x-\theta}{\theta^2}.$$

Hence

$$E\left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta}\right)^2\right] = E\left[\left(\frac{X-\theta}{\theta^2}\right)^2\right] = \frac{E[(X-\theta)^2]}{\theta^4} = \frac{\text{var}[X]}{\theta^4}.$$

Therefore,

$$E\left[\left(\frac{\partial \ln f(X|\lambda)}{\partial \theta}\right)^2\right] = \frac{\text{var}[X]}{\theta^4} = \frac{\theta^2}{\theta^4} = \frac{1}{\theta^2},$$

and the CRLB is

$$\frac{1}{n \cdot E\left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta}\right)^2\right]} = \frac{\theta^2}{n}.$$

Consequently, since $\overline{X}$ is unbiased and $\text{var}\left[\overline{X}\right] = \frac{\theta^2}{n}$, it follows that $\overline{X}$ is a minimum variance unbiased estimator of θ. ■

MARQUETTE UNIVERSITY Be The Difference.

EXAMPLE

Example 7.7 ▷ ***Comparing Estimators: Blue Jean Length*** ◁ Suppose the true manufactured length of new 32L blue jeans follows a normal distribution with unknown μ and $\sigma = 0.5$ inches. It is known that 32L blue jeans sold in stores have a length of at least 31 inches. If a random sample of size $n = 3$ of 32L blue jeans is taken to estimate μ, which of the following estimators $\hat{\mu}_1$ or $\hat{\mu}_2$ is better in terms of bias, variance, and relative efficiency where $\hat{\mu}_1 = 0.33 \cdot (X_1 + X_2 + X_3)$ and $\hat{\mu}_2 = 0.50 \cdot (X_1 + X_2)$?

20

MARQUETTE UNIVERSITY Be The Difference.

SOLUTION

$$E[\hat{\mu}_1] = 0.33 \cdot E[X_1 + X_2 + X_3] = 0.33 \cdot (E[X_1] + E[X_2] + E[X_3])$$
$$= 0.33(\mu + \mu + \mu) = 0.99\mu,$$

it follows that $\hat{\mu}_1$ is a biased estimator of μ with bias $0.99\mu - \mu = -0.01\mu$. On the other hand

$$E[\hat{\mu}_2] = 0.50 \cdot E[X_1 + X_2] = 0.50 \cdot \big(E[X_1] + E[X_2]\big) = 0.50 \cdot (\mu + \mu) = \mu$$

which makes $\hat{\mu}_2$ an unbiased estimator of μ. The variances of $\hat{\mu}_1$ and $\hat{\mu}_2$ are

$$\text{var}[\hat{\mu}_1] = \text{var}\big[0.33 \cdot \big(X_1 + X_2 + X_3\big)\big]$$
$$= 0.33^2 \cdot (\text{var}[X_1] + \text{var}[X_2] + \text{var}[X_3])$$
$$= 0.33^2 \cdot (0.25 + 0.25 + 0.25) = 0.081675, \text{ and}$$

$$\text{var}[\hat{\mu}_2] = \text{var}\big[0.50 \cdot \big(X_1 + X_2\big)\big] = 0.50^2 \cdot (\text{var}[X_1] + \text{var}[X_2])$$
$$= 0.25 \cdot (0.25 + 0.25) = 0.125, \text{ respectively.}$$

21

SOLUTION CONT'D

Before looking at the relative efficiency of $\hat{\mu}_1$ to $\hat{\mu}_2$, compute the *MSE* for each estimator using the fact that $MSE = \text{Variance} + \text{Bias}^2$.

$$MSE[\hat{\mu}_1] = 0.081675 + (0.01\mu)^2 = 0.081675 + 0.0001\mu^2$$
$$MSE[\hat{\mu}_2] = 0.125 + 0^2 = 0.125$$

Since

$$\textit{eff}(\hat{\mu}_1, \hat{\mu}_2) = \frac{MSE\,(\hat{\mu}_2)}{MSE\,(\hat{\mu}_1)} = \frac{0.125}{0.081675 + 0.0001\mu^2} < 1 \text{ for all } |\mu| > 20.82,$$

conclude that $\hat{\mu}_2$ is both more efficient and has smaller *MSE* than does $\hat{\mu}_1$, since it is known that $\mu \geq 31$ inches according to the problem.

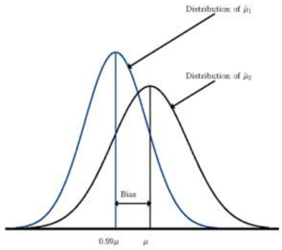


Figure 7.2: Graphical representations for the sampling distributions of $\hat{\mu}_1$ and $\hat{\mu}_2$.

7.2.4 Consistent Estimators

The next property of estimators which is considered is **consistency**. Consistency is a property of a sequence of estimators rather than a single estimator. However, it is rather common to refer to an estimator as being consistent. Sequence of estimators means that the same estimation procedure is carried out for each sample of size n. If T is an estimator of θ and $X_1, X_2, \ldots$ are observed according to a distribution $f(x|\theta)$, a sequence of estimators $T_1, T_2, \ldots, T_n$ can be constructed by performing the same estimation procedure for samples of sizes $1, 2, \ldots, n$ respectively. In other words, the sequence is

$$T_1 = t(X_1), T_2 = t(X_1, X_2), \ldots, T_n = t(X_1, X_2, \ldots, X_n).$$

A sequence of estimators T_n (defined for all n) is a **consistent** estimator of the parameter θ for every $\theta \in \Theta$ if

$$\lim_{n\to\infty} \mathbb{P}(|T_n - \theta| \geq \epsilon) = 0, \text{ for all } \epsilon > 0. \qquad (7.12)$$

An equivalent statement of (7.12) is that a sequence of estimators T_n (defined for all n) is a **consistent** estimator of the parameter θ for every $\theta \in \Theta$ if

$$\lim_{n\to\infty} \mathbb{P}(|T_n - \theta| < \epsilon) = 1, \text{ for all } \epsilon > 0. \qquad (7.13)$$

Both definitions (7.12) and (7.13) state that a consistent sequence of estimators **converges in probability** to the parameter θ, where θ is the parameter the consistent sequence of estimators is estimating. In practical terms, this implies that the variance of a consistent estimator decreases as n increases and that the expected value of T_n tends to θ as n increases. Further, given a consistent sequence of estimators, say T_n, Chebyshev's inequality guarantees that

$$\mathbb{P}(|T_n - \theta| \geq \epsilon) = \mathbb{P}(|T_n - \theta|^2 \geq \epsilon^2) \leq \frac{E[(T_n - \theta)^2]}{\epsilon^2},$$

for every $\theta \in \Theta$. Since $E_\theta\left[(T_n - \theta)^2\right]$ can be expressed as

$$E_\theta\left[(T_n - \theta)^2\right] = \text{var}[T_n] + (Bias[T_n])^2,$$

MARQUETTE UNIVERSITY
Be The Difference.

REMINDER

A sequence of estimators T_n (defined for all n) is a **consistent** estimator of the parameter θ for every $\theta \in \Theta$ if

$$\lim_{n\to\infty} \mathbb{P}(|T_n - \theta| \geq \epsilon) = 0, \text{ for all } \epsilon > 0. \qquad (7.12)$$

Chebyshev's inequality guarantees that

$$\mathbb{P}(|T_n - \theta| \geq \epsilon) = \mathbb{P}(|T_n - \theta|^2 \geq \epsilon^2) \leq \frac{E[(T_n - \theta)^2]}{\epsilon^2},$$

for every $\theta \in \Theta$. Since $E_\theta\left[(T_n - \theta)^2\right]$ can be expressed as

$$E_\theta\left[(T_n - \theta)^2\right] = \text{var}[T_n] + (Bias[T_n])^2,$$

if

$$\lim_{n\to\infty} \text{var}[T_n] = 0 \quad \text{and} \quad \lim_{n\to\infty} (Bias[T_n])^2 = 0, \qquad (7.14)$$

then T_n is a consistent sequence of estimators of θ. Whenever the

MARQUETTE UNIVERSITY Be The Difference.

Example 7.8 Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample of size n from a distribution with mean μ and variance σ^2. Show that $\overline{X}_n$ is a consistent estimator of μ.

Solution: For $\overline{X}_n$ to be a consistent estimator of μ, it must be shown that

$$\lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) = 0 \text{ for all } \epsilon > 0.$$

Using Chebyshev's inequality and the fact that $E\left[\overline{X}_n\right] = \mu$ and $\text{var}\left[\overline{X}_n\right] = \sigma^2/n$,

$$\mathbb{P}(|\overline{X}_n - \mu| \geq k\sigma/\sqrt{n}) \leq \frac{1}{k^2}.$$

By setting $\epsilon = k\sigma/\sqrt{n}$, $k = \sqrt{n}\epsilon/\sigma$ so that

$$\frac{1}{k^2} = \frac{\sigma^2}{n\epsilon^2},$$

MARQUETTE UNIVERSITY Be The Difference.

SOLUTION CONT'D

from which it follows that

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}. \quad (7.15)$$

Given that $\sigma^2 < \infty$ (finite), by taking the limit as $n \to \infty$ on both sides of the $\leq$ sign of (7.15) gives

$$\lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) = 0 \text{ for all } \epsilon.$$

Consequently, $\overline{X}_n$ is a consistent estimator of μ. This is essentially the **weak law of large numbers**. ■

MARQUETTE UNIVERSITY
Be The Difference.

7.3 Point Estimation Techniques

- Two methods are considered: the method of moments and the method of maximum likelihood.
- the information in a random sample $X_1, X_2, \ldots, X_n$ is used to make inferences about the unknown θ.
- The observed values of the random sample are denoted $x_1, x_2, \ldots, x_n$.
- Further, a random sample $X_1, X_2, \ldots, X_n$ is referred to with the boldface $\mathbf{X}$ the observed values in a random sample $x_1, x_2, \ldots, x_n$ with the boldface $\mathbf{x}$.
- The joint pdf of $X_1, X_2, \ldots, X_n$ is given by

$$\begin{aligned} f(\mathbf{x}|\theta) &= f(x_1, x_2, \ldots, x_n|\theta) \\ &= f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_\mathbf{n}|\theta) = \prod_{i=1}^{n} f(x_i|\theta). \end{aligned} \quad (7.17)$$

28

MARQUETTE UNIVERSITY
Be The Difference.

7.2.3.1 Method of Moments Estimators

- The idea behind the **method of moments** is to equate population moments about the origin to their corresponding sample moments, where the $\boldsymbol{r}^{\textbf{th}}$ **sample moment about the origin**, denoted m_r, is defined as

$$m_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r, \quad (7.18)$$

and subsequently to solve for estimators of the unknown parameters.

- Recall that the r^{th} population moment about the origin of a random variable X denoted α_r, was defined as $E[X^r]$.
- It follows that $\alpha_r = E[X^r] = \sum_{i=1}^{\infty} x_i^r \mathbb{P}(X = x_i)$ for discrete X, and that $\alpha_r = E[X^r] = \int_{-\infty}^{\infty} x^r f(x)\,dx$ for continuous X.

29

MARQUETTE UNIVERSITY
Be The Difference.

- Specifically, given a random sample $X_1, X_2, \ldots, X_n$ from a population with **pdf** $f(x|\theta_1, \theta_2, \ldots, \theta_k)$, the method of moments estimators, denoted $\tilde{\theta}_i$ for $i = 1, \ldots, k$ are found by equating the first k population moments about the origin to their corresponding sample moments and solving the resulting system of simultaneous equations given in Equation (7.19).

$$\begin{cases} \alpha_1(\theta_1, \ldots, \theta_k) &= m_1 \\ \alpha_2(\theta_1, \ldots, \theta_k) &= m_2 \\ \vdots & \vdots \\ \alpha_k(\theta_1, \ldots, \theta_k) &= m_k \end{cases} \tag{7.19}$$

MARQUETTE UNIVERSITY
Be The Difference.

Example 7.10 Given a random sample of size n from a $Bin(1, \pi)$ population, find the method of moments estimator of π.

Solution: The first sample moment m_1 is $\overline{X}$ and the first population moment about zero for the binomial random variable is $\alpha_1 = E\left[X^1\right] = 1 \cdot \pi$. By equating the first population moment to the first sample moment,

$$\alpha_1(\pi) = \pi \stackrel{\text{set}}{=} \overline{X} = m_1,$$

which implies that the method of moments estimator for π, is $\tilde{\pi} = \overline{X}$. ■

Example 7.11 Given a random sample of size m from a $Bin(n,\pi)$ population, find the method of moments estimator of π.

Solution: The first sample moment m_1 is $\overline{X}$ and the first population moment about zero for the binomial random variable is $\alpha_1 = E\left[X^1\right] = n \cdot \pi$. By equating the first population moment to the first sample moment,

$$\alpha_1(\pi) = n\pi \stackrel{\text{set}}{=} \overline{X} = m_1,$$

which implies that the method of moments estimator for π, is $\tilde{\pi} = \frac{\overline{X}}{n}$. ■

Example 7.12 Given a random sample of size n from a $Pois(\lambda)$ population, find the method of moments estimator of λ.

Solution: The first sample moment m_1 is $\overline{X}$ and the first population moment about zero for a Poisson random variable is $\alpha_1 = E\left[X^1\right] = \lambda$. By equating the first population moment to the first sample moment,

$$\alpha_1(\pi) = \lambda \stackrel{\text{set}}{=} \overline{X} = m_1,$$

which implies that the method of moments estimator for λ, is $\tilde{\lambda} = \overline{X}$. ■

Example 7.13 Given a random sample of size n from a $N(\mu, \sigma)$ population, find the method of moments estimators of μ and σ^2.

Solution: The first and second sample moments m_1 and m_2 are $\overline{X}$ and $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ respectively. The first and second population moments about zero for a normal random variable are $\alpha_1 = E\left[X^1\right] = \mu$ and $\alpha_2 = E\left[X^2\right] = \sigma^2 + \mu^2$. By equating the first two population moments to the first two sample moments,

$$\begin{cases} \alpha_1(\mu, \sigma^2) = \mu \overset{\text{set}}{=} \overline{X} = m_1 \\ \alpha_2(\mu, \sigma^2) = \sigma^2 + \mu^2 \overset{\text{set}}{=} \frac{1}{n}\sum_{i=1}^{n} X_i^2 = m_2. \end{cases} \tag{7.20}$$

Solving the system of equations in (7.20) yields $\tilde{\mu} = \overline{X}$ and $\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = S_u^2$ as the method of moments estimators for μ and σ^2 respectively. ■

Example 7.14 Given a random sample of size n from a $Gamma(\alpha, \lambda)$ population, find the method of moments estimators of α and λ.

Solution: $E[X] = \frac{\alpha}{\lambda}$, and the var $[X] = \frac{\alpha}{\lambda^2}$ for a random variable X that follows a gamma distribution. The first and second sample moments m_1 and m_2 are $\overline{X}$ and $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ respectively. The first and second population moments for a gamma random variable are

$$\alpha_1 = E\left[X^1\right] = \frac{\alpha}{\lambda},$$

and

$$\alpha_2 = E\left[X^2\right] = \sigma^2 + E[X]^2 = \frac{\alpha}{\lambda^2} + \frac{\alpha^2}{\lambda^2} = \frac{\alpha(1+\alpha)}{\lambda^2}$$

EXAMPLE 7.14 CONT'D

By equating the first two population moments to the first two sample moments,

$$\begin{cases} \alpha_1(\alpha,\lambda) = \dfrac{\alpha}{\lambda} \overset{\text{set}}{=} \overline{X} = m_1 \\ \alpha_2(\alpha,\lambda) = \dfrac{\alpha(1+\alpha)}{\lambda^2} \overset{\text{set}}{=} \dfrac{1}{n}\sum_{i=1}^{n} X_i^2 = m_2. \end{cases} \quad (7.21)$$

When it is recalled that $S_u^2 = \frac{\sum_i^n (X_i - \overline{X})^2}{n}$, the system of equations in (7.21) can be solved to obtain $\tilde{\alpha} = \frac{\overline{X}^2}{S_u^2}$ and $\tilde{\lambda} = \frac{\overline{X}}{S_u^2}$ as the method of moments estimators for α and λ respectively. ■

7.3.2 Likelihood and Maximum Likelihood Estimators

When sampling from a population described by a **pdf** $f(x|\theta)$, knowledge of θ provides knowledge of the entire population. The idea behind maximum likelihood is to select the value for θ that makes the observed data most likely under the assumed probability model.

When $x_1, x_2, \ldots, x_n$

are the observed values of a random variable X from a population with parameter θ, the notation $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$ will be used to indicate that the distribution depends on the parameter θ, and $\mathbf{x}$ to indicate the distribution is dependent on the observed values from the sample. Once the sample values are observed, $L(\theta|\mathbf{x})$ can still be evaluated in a formal sense, although it no longer has a probability interpretation (in the discrete case) as does (7.17).

$L(\theta|\mathbf{x})$ is the **likelihood function** of θ for $\mathbf{x}$ and is denoted by

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta). \tag{7.22}$$

The key difference between (7.17) and (7.22) is that the joint **pdf** given in (7.17) is a function of $\mathbf{x}$ for a given θ and the likelihood function given in (7.22) is a function of θ for given $\mathbf{x}$.

The value of θ that maximizes $L(\theta|\mathbf{x})$ is called the **maximum likelihood estimate** (mle) of θ. Another way to think of the mle is the mode of the likelihood function. The maximum likelihood estimate is denoted as $\hat{\theta}(\mathbf{x})$, and the maximum likelihood estimator (MLE), a statistic, as $\hat{\theta}(\mathbf{X})$.

38

LOG-LIKELIHOOD VS LIKELIHOOD

In general, the likelihood function may be difficult to manipulate, and it is usually more convenient to work with the natural logarithm of $L(\theta|\mathbf{x})$, called the **log-likelihood function**, since it converts products into sums. Finding the value θ that maximizes the log-likelihood function $\big(\ln L(\theta|\mathbf{x})\big)$ is equivalent to finding the value of θ that maximizes $L(\theta|\mathbf{x})$ since the natural logarithm is a monotonically increasing function. If $L(\theta|\mathbf{x})$ is differentiable with respect to θ, a possible mle is the solution to

$$\frac{\partial\big(\ln L(\theta|\mathbf{x})\big)}{\partial\theta} = 0. \tag{7.23}$$

```
L <- logL <- NULL; par(mfrow=c(1,2));
n <- 10; mus <- seq(0,10,length=100); mu <- 5; x <- rnorm(n,mean=mu);
Like <- function(mu, data=x) {prod(dnorm(x,mean=mu))}
logLike <- function(mu, data=x) {sum(dnorm(x,mean=mu,log=TRUE))}
max.L <- optim(1, Like, data=x, control=list(fnscale=-1))
# max.L <- optim(1, Like, data=x, control=list(fnscale=-1), method="Brent",lower=0,upper=10)
max.logL <- optim(1, logLike, data=x, control=list(fnscale=-1))
for (i in 1:length(mus)) {L[i] <- Like(mus[i], x); logL[i] <- logLike(mus[i], x)}
plot(mus,L,type="l"); abline(v=max.L$par, col=2)
plot(mus,logL,type="l"); abline(v=max.logL$par, col=2)
```

39

NOTE:

Note that a possible mle is the solution to (7.23). A possible solution is used since a solution to (7.23) is a necessary but not sufficient condition for the solution to be a maximum. Since the solution to (7.23) could be a local or global minimum, a local or global maximum, or a point of inflection. Recall that stationary points where,

$$\frac{\partial^2 (\ln L(\theta|\mathbf{x}))}{\partial \theta^2}\Big|_{\theta=\hat{\theta}(\mathbf{x})} < 0, \tag{7.24}$$

indicate some type of maximum either local or global. Further, the solution to (7.23) does not include points on the boundaries of the parameter space. Consequently, when evaluating the maximum of $L(\theta|\mathbf{x})$, the boundaries of the parameter space Θ as well as solutions to (7.23) must be evaluated.

Example 7.15 Given a random sample of size n taken from a *Bernoulli*(π) distribution, compute the maximum likelihood estimate and maximum likelihood estimator of the parameter π.

Solution: According to Box (**??**), the **pdf** for $X \sim \textit{Bernoulli}(\pi)$ is

$$P(X = x|\pi) = \pi^x (1-\pi)^{1-x},$$

where x takes on the value 1 with probability π and 0 with probability $1-\pi$. The likelihood function for the n observed values is

$$L(\pi|\mathbf{x}) = \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i}.$$

Taking the natural logarithm of the likelihood function gives

$$\ln L(\pi|\mathbf{x}) = \ln\left[\prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i}\right] = \sum_{i=1}^{n} \ln\left[\pi^{x_i}(1-\pi)^{1-x_i}\right]$$
$$= \sum_{i=1}^{n} \left[x_i \ln \pi + (1-x_i)\ln(1-\pi)\right]. \qquad (7.25)$$

To find the value that maximizes (7.25), take the first-order partial derivative of $\ln L(\pi|\mathbf{x})$ with respect to π and set the answer equal to zero.

$$\frac{\partial \ln L(\pi|\mathbf{x})}{\partial \pi} = \frac{\sum_{i=1}^{n} x_i}{\pi} - \frac{n - \sum_{i=1}^{n} x_i}{1-\pi} \overset{set}{=} 0. \qquad (7.26)$$

The solution to (7.26) is $\pi = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$. For $\pi = \bar{x}$ to be a maximum, the second-order partial derivative of the log-likelihood function must be negative at $\pi = \bar{x}$. The second-order partial derivative is

$$\frac{\partial^2 \ln L(\pi|\mathbf{x})}{\partial \pi^2} = \frac{-\sum_{i=1}^{n} x_i}{\pi^2} - \frac{n - \sum_{i=1}^{n} x_i}{(1-\pi)^2}.$$

Evaluating the second-order partial derivative at $\pi = \bar{x}$ yields

$$\frac{\partial^2 \ln L(\pi|\mathbf{x})}{\partial \pi^2} = \frac{-n\bar{x}}{\bar{x}^2} - \frac{(n - n\bar{x})}{(1-\bar{x})^2} = -\frac{n}{\bar{x}} - \frac{n}{1-\bar{x}},$$

which is less than zero since $0 \le \bar{x} \le 1$ and $n > 0$. Finally, since the values of the likelihood function at the boundaries of the parameter space, $\pi = 0$ and $\pi = 1$, are 0, it follows that $\pi = \bar{x}$ is the value that maximizes the likelihood function. The maximum likelihood estimate $\hat{\pi}(\mathbf{x}) = \bar{x}$ and the maximum likelihood estimator $\hat{\pi}(\mathbf{X}) = \overline{X}$. ∎

MARQUETTE UNIVERSITY
Be The Difference.

Example 7.16 ▷ ***MLEs with S: Oriental Cockroaches*** ◁ A laboratory is interested in testing a new child friendly pesticide on *Blatta orientalis* (oriental cockroaches). The scientists from the lab apply the new pesticide to 81 randomly selected *Blatta orientalis* oothecae (eggs). The results from the experiment are stored in the data frame `Roacheggs` in the variable `eggs`. A zero in the variable `eggs` indicates that nothing hatched from the egg while a 1 indicates the birth of a cockroach. Assuming the selected *Blatta orientalis* eggs are representative of the population of *Blatta orientalis* eggs, estimate the proportion of *Blatta orientalis* eggs that result in a birth after being sprayed with the child friendly pesticide. Use either `nlm()` in R or `nlmin()` in S-PLUS to solve the problem iteratively and to produce a graph of the log-likelihood function.

44

MARQUETTE UNIVERSITY
Be The Difference.

Solution: Note that whether or not a *Blatta orientalis* egg hatches is a Bernoulli trial with unknown parameter π. Using the maximum likelihood estimate from Example 7.15 on page 59, $\hat{\pi}(\mathbf{x}) = \bar{x} = 0.21$.

```
> library(PASWR)
> attach(Roacheggs)
> mean(eggs)
[1] 0.2098765
```

Both R and S-PLUS have iterative procedures that will minimize a given function. The minimization function in R is `nlm()`, while the minimization function in S-PLUS is `nlmin()`. The required arguments for both functions are `f()` and `p` where `f()` is the function to be minimized and `p` is a vector of initial values for the parameter(s).

Since both `nlm()` and `nlmin()` are minimization procedures and finding a maximum likelihood estimate is a maximization procedure, the functions `nlm()` and `nlmin()` on the negative of the log-likelihood function are used.

45

R CODE:

MARQUETTE UNIVERSITY
Be The Difference.

```
➢ p <- seq(0.1, 0.9, 0.001)
➢ negloglike <- function(p){-(sum(eggs)*log(p)+sum(1-eggs)*log(1-p))}
➢ nlm(negloglike, 0.2)
$minimum
[1] 41.61724
$estimate
[1] 0.209876
➢ par(pty = "s")
➢ p <- seq(0.1, 0.9, 0.001)
➢ plot(p, - negloglike(p), type = "l", ylab = "L",col=6,lwd=3)
➢ abline(v = mean(eggs), col = 13, lwd = 3)
```

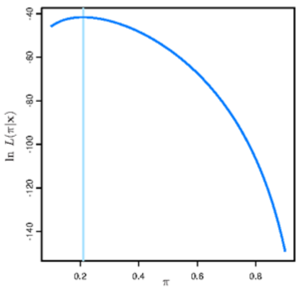


Figure 7.3: Illustration of the ln $L(\pi|x)$ function for Example 7.16

46

MORE ON OPTIMIZATION IN R

MARQUETTE UNIVERSITY
Be The Difference.

The function `optimize()`, available in both R and S-PLUS, approximate a local optimum of a continuous univariate function (`f`) within a given interval. The function searches the user provided interval for either a minimum (default) or maximum of the function `f`. To solve Example

```
➢ loglike <- function(p){(sum(eggs)*log(p)+sum(1-eggs)*log(1-p))}
➢ optimize(f=loglike,interval=c(0,1),maximum=TRUE)
$maximum
[1] 0.2098906
$objective
[1] -41.61724

➢ optim(0.5,negloglike)
$par
[1] 0.2098633
$value
[1] 41.61724

➢ optim(0.5,loglike,control=list(fnscale=-1))
$par
[1] 0.2098633
$value
[1] -41.61724
```

47

MARQUETTE UNIVERSITY Be The Difference.

Example 7.17 Let $X_1, X_2, \ldots, X_m$ be a random sample from a $Bin(n, \pi)$ population. Compute the maximum likelihood estimator and the maximum likelihood estimate for the parameter π. Verify your answer with simulation by generating 1,000 random values from a $Bin(n = 3, \pi = 0.5)$ population.

Solution: The likelihood function is

$$L(\pi|\mathbf{x}) = \prod_{i=1}^{m} \binom{n}{x_i} \pi^{x_i}(1-\pi)^{n-x_i}$$
$$= \binom{n}{x_1} \pi^{x_1}(1-\pi)^{n-x_1} \times \cdots \times \binom{n}{x_m} \pi^{x_m}(1-\pi)^{n-x_m}, \tag{7.27}$$

and the log-likelihood function is

$$\ln L(\pi|\mathbf{x}) = \ln \left[\prod_{i=1}^{m} \binom{n}{x_i} \pi^{x_i}(1-\pi)^{n-x_i} \right]$$

48

MARQUETTE UNIVERSITY Be The Difference.

and the log-likelihood function is

$$\ln L(\pi|\mathbf{x}) = \ln \left[\prod_{i=1}^{m} \binom{n}{x_i} \pi^{x_i}(1-\pi)^{n-x_i} \right]$$

$$\ln L(\pi|\mathbf{x}) = \sum_{i=1}^{m} \ln \left[\binom{n}{x_i} \pi^{x_i}(1-\pi)^{n-x_i} \right]$$
$$= \sum_{i=1}^{m} \left[\ln \binom{n}{x_i} + x_i \ln \pi + (n - x_i) \ln(1-\pi) \right]. \tag{7.28}$$

Next, look for the value that maximizes the log-likelihood function by taking the first-order partial derivative of (7.28) and setting the answer to zero.

$$\frac{\partial \ln L(\pi|\mathbf{x})}{\partial \pi} = \frac{\sum_{i=1}^{m} x_i}{\pi} - \frac{mn - \sum_{i=1}^{m} x_i}{1-\pi} \overset{\text{set}}{=} 0. \tag{7.29}$$

49

MARQUETTE UNIVERSITY
Be The Difference.

The solution to (7.29) is $\pi = \frac{\sum_{i=1}^m x_i}{mn} = \frac{\bar{x}}{n}$. For $\pi = \frac{\bar{x}}{n}$ to be a maximum, the second-order partial derivative of the log-likelihood function must be negative at $\pi = \frac{\bar{x}}{n}$. The second-order partial derivative is

$$\frac{\partial^2 \ln L(\pi|\mathbf{x})}{\partial \pi^2} = \frac{-\sum_{i=1}^m x_i}{\pi^2} - \frac{mn - \sum_{i=1}^m x_i}{(1-\pi)^2}.$$

Evaluating the second-order partial derivative at $\pi = \frac{\bar{x}}{n}$ and using the substitution $\sum_{i=1}^m x_i = m\bar{x}$ yields

$$\begin{aligned}\frac{\partial^2 \ln L(\pi|\mathbf{x})}{\partial \pi^2} &= -\frac{m\bar{x}}{(\frac{\bar{x}}{n})^2} - \frac{mn - m\bar{x}}{(1-\frac{\bar{x}}{n})^2} \\ &= -\frac{mn^2}{\bar{x}} - \frac{m(n-\bar{x})}{\frac{(n-\bar{x})^2}{n^2}} = -\frac{mn^2}{\bar{x}} - \frac{mn^2}{n-\bar{x}} < 0.\end{aligned}$$

Finally, since the values of the likelihood function at the boundaries of the parameter space, $\pi = 0$ and $\pi = 1$, are 0, it follows that $\pi = \frac{\bar{x}}{n}$ is the value that maximizes the likelihood function. The maximum likelihood estimate $\hat{\pi}(\mathbf{x}) = \frac{\bar{x}}{n}$ and the maximum likelihood estimator $\hat{\pi}(\mathbf{X}) = \frac{\bar{X}}{n}$.

50

MARQUETTE UNIVERSITY
Be The Difference.

Example 7.18 Let $X_1, X_2, \ldots, X_m$ be a random sample from a $Pois(\lambda)$ population. Compute the maximum likelihood estimator and the maximum likelihood estimate for the parameter λ. Verify your answer with simulation by generating 1,000 random values from a $Pois(\lambda = 5)$ population.

Solution: The likelihood function is

$$L(\lambda|\mathbf{x}) = \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = e^{-n\lambda} \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!}, \qquad (7.30)$$

and the log-likelihood function is

$$\ln L(\lambda|\mathbf{x}) = \ln \left[e^{-n\lambda} \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} \right] = -n\lambda + \sum_{i=1}^n x_i \ln \lambda - \sum_{i=1}^n \ln(x_i!). \qquad (7.31)$$

51

Next, look for the value that maximizes the log-likelihood function by taking the first-order partial derivative of (7.31) and setting the answer to zero.

$$\frac{\partial \ln L(\lambda|\mathbf{x})}{\partial \lambda} = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} \overset{\text{set}}{=} 0. \tag{7.32}$$

The solution to (7.32) is $\lambda = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$. For $\lambda = \bar{x}$ to be a maximum, the second-order partial derivative of the log-likelihood function must be negative at $\lambda = \bar{x}$. The second-order partial derivative is

$$\frac{\partial^2 \ln L(\lambda|\mathbf{x})}{\partial \lambda^2} = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2}.$$

Evaluating the second-order partial derivative at $\lambda = \bar{x}$ yields

$$\frac{\partial^2 \ln L(\lambda|\mathbf{x})}{\partial \lambda^2} = -\frac{n\bar{x}}{\bar{x}^2} = -\frac{n}{\bar{x}} < 0.$$

Finally, since the values of the likelihood function at the boundaries of the parameter space, $\lambda = 0$ and $\lambda = \infty$, are 0, it follows that $\lambda = \bar{x}$ is the value that maximizes the likelihood function. The maximum likelihood estimate $\hat{\lambda}(\mathbf{x}) = \bar{x}$ and the maximum likelihood estimator $\hat{\lambda}(\mathbf{X}) = \overline{X}$.

To simulate $\hat{\lambda}(\mathbf{x}) = \bar{x}$, generate 1,000 random values from a $Pois(\lambda = 5)$ population.

```
> set.seed(99)
> mean(rpois(1000, 5))
[1] 5.073
```

Example 7.19 A box contains five pieces of candy. Some of the candies are alcoholic, and some are not. In an attempt to estimate the proportion of alcoholic candies, a sample of size $n = 3$ is taken with replacement which results in (a, a, n) (two alcoholic candies and one nonalcoholic candy). Write out the maximum likelihood function and use it to select the maximum likelihood estimate of π, the true proportion of alcoholic candies.

Solution: The possible values for π are $\frac{0}{5}$, $\frac{1}{5}$, $\frac{2}{5}$, $\frac{3}{5}$, $\frac{4}{5}$, and $\frac{5}{5}$. Since there is at least one alcoholic candy and there is at least one nonalcoholic candy, the values $\pi = 0$ and $\pi = 1$ must be ruled out. In this case,

the observed sample values are $\mathbf{x}$=(a, a, n). The likelihood function is

$$\begin{aligned} L(\pi|\mathbf{x}) &= f(\mathbf{x}|\pi) \\ &= f(\mathrm{a}|\pi) \times f(\mathrm{a}|\pi) \times f(\mathrm{n}|\pi). \end{aligned}$$

54

Box	π	$L(\pi\|\mathrm{a, a, n})$
aaaan	$\frac{4}{5}$	$\frac{4}{5} \cdot \frac{4}{5} \cdot \frac{1}{5} = \frac{16}{125}$
aaann	$\frac{3}{5}$	$\frac{3}{5} \cdot \frac{3}{5} \cdot \frac{2}{5} = \frac{18}{125}$
aannn	$\frac{2}{5}$	$\frac{2}{5} \cdot \frac{2}{5} \cdot \frac{3}{5} = \frac{12}{125}$
annnn	$\frac{1}{5}$	$\frac{1}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} = \frac{4}{125}$

Since the value $\pi = \frac{3}{5}$ maximizes the likelihood function, consider $\hat{\pi}(\mathbf{x}) = \frac{3}{5}$ to be the maximum likelihood estimate for the proportion of candies that are alcoholic. ■

55

Example 7.20 ▷ ***General MLE*** ◁ The random variable X can take on the values 0, 1, 2, and 3 with probabilities $\mathbb{P}(X = 0) = p^3$, $\mathbb{P}(X = 1) = (1-p)p^2$, $\mathbb{P}(X = 2) = (1-p)^2$, and $\mathbb{P}(X = 3) = 2p(1-p)$ where $0 < p < 1$.

(a) Do the given probabilities for the random variable X satisfy the conditions for a probability distribution of X?

(b) Find the maximum likelihood estimate for p if a random sample of size $n = 150$ resulted in a 0 twenty-four times, a 1 fifty-four times, a 2 thirty-two times, and a 3 forty times.

(c) Graph the log-likelihood function and determine its maximum using either the function `nlmin()` or the function `nlm()`.

Solution: The answers are:

SOLUTION:

(a) For the distribution of X to be a valid **pdf**, it must satisfy the following two conditions.

(1) $p(x) \geq 0$ for all x.

(2) $\sum_x p(x) = 1$.

Condition (1) is satisfied since $0 < p < 1$. Condition (2) is also satisfied since

$$\sum_x p(x) = p^3 + (1-p)p^2 + (1-p)^2 + 2p(1-p)$$
$$= p^3 + p^2 - p^3 + 1 + p^2 - 2p + 2p - 2p^2 = 1.$$

(b) The likelihood function is

$$L(p|\mathbf{x}) = \left[(p^3)\right]^{24}\left[(1-p)p^2\right]^{54}\left[(1-p)^2\right]^{32}[2p(1-p)]^{40}$$
$$= 2^{40}p^{220}(1-p)^{158},$$

SOLUTION CONT'D:

and the log-likelihood function is

$$\ln[L(p|\mathbf{x})] = 40\ln 2 + 220\ln p + 158\ln(1-p). \qquad (7.33)$$

Next, look for the value that maximizes the log-likelihood function by taking the first-order partial derivative of (7.33) with respect to p and setting the answer equal to zero.

$$\frac{\partial \ln[L(p|\mathbf{x})]}{\partial p} = \frac{220}{p} - \frac{158}{1-p} \overset{\text{set}}{=} 0. \qquad (7.34)$$

The solution to (7.34) is $p = 0.58$. In order for $p = 0.58$ to be a maximum, the second-order partial derivative of (7.33) with respect to p must be negative. Since,

$$\frac{\partial^2 \ln[L(\pi|\mathbf{x})]}{\partial p^2} = -\frac{220}{p^2} - \frac{158}{(1-p)^2} < 0 \text{ for all } p,$$

this value is a global maximum. Therefore, the maximum likelihood estimate of p, $\hat{p}(\mathbf{x}) = 0.58$.

58

SOLUTION CONT'D:

(c) Generic S code to graph the log-likelihood function depicted in Figure 7.4 on the following page follows.

```
➢ par(pty = "s")
➢ p <- seq(0.01, 0.99, 0.001)
➢ negloglike <-
  function(p){-40 * log(2) - 220 * log(p) - 158 * log(1 - p)}
➢ plot(p, -negloglike(p), type = "l", col = 6, lwd = 3)
➢ abline(v = 0.58, col = 13, lwd = 3)
```

To compute the maximum of the log-likelihood function, use the command `nlmin(loglike, 0.001)` with S-PLUS and the command `nlm(negloglike, 0.001)` with R.

```
➢ nlm(negloglike, 0.001) # R
$minimum
[1] 229.1760
$estimate
[1] 0.58201
```

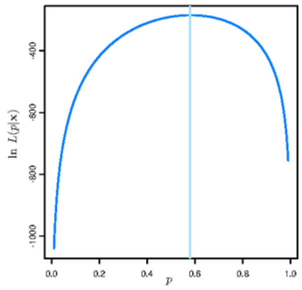


Figure 7.4: Illustration of the ln $L(p|\mathbf{x})$ function for Example 7.20.

59

MARQUETTE UNIVERSITY
Be The Difference.

Example 7.21 A farmer cans and sells mild and hot peppers at the local market. The farmer recently hired an assistant to label his products. The assistant is new to working with peppers and has mislabelled some of the hot peppers as mild peppers. The farmer performs a random check of 100 of the mild pepper cans labelled by the assistant to assess his work. Out of the 100 cans labelled mild peppers, it turns out that 8 are actually hot peppers.

(a) Which of the following proportions, 0.05, 0.08, or 0.10, maximizes the likelihood function?

(b) What is the maximum likelihood estimate for the proportion of cans the assistant has mislabelled?

Solution: The answers are:

60

MARQUETTE UNIVERSITY
Be The Difference.

SOLUTION:

(a) First define the random variable X as the number of mislabeled cans. In this definition of the random variable X, it follows that $n = 100$ and $m = 1$ since $X \sim Bin(100, \theta)$. The likelihood function for a random sample of size m from a $Bin(n, \pi)$ population was computed in (7.27) as

$$L(\pi|\mathbf{x}) = \prod_{i=1}^{m} \binom{n}{x_i} \pi^{x_i}(1-\pi)^{n-x_i}.$$

Since $m = 1$ here, it follows that the likelihood function is

$$L(\pi|\mathbf{x}) = \binom{n}{x} \pi^{x}(1-\pi)^{n-x}.$$

Consequently, the value for π that maximizes

$$\mathbb{P}(X = 8|\pi) = \binom{100}{8} \pi^{8} \cdot (1-\pi)^{92}$$

is the solution to the problem. The likelihoods for the three values of π are

61

SOLUTION CONT'D:

$$\mathbb{P}(X = 8|0.05) = \binom{100}{8} 0.05^8 \cdot (1-0.05)^{92} = 0.0648709,$$

$$\mathbb{P}(X = 8|0.08) = \binom{100}{8} 0.08^8 \cdot (1-0.08)^{92} = 0.1455185,$$

and

$$\mathbb{P}(X = 8|0.10) = \binom{100}{8} 0.10^8 \cdot (1-0.10)^{92} = 0.1148230.$$

Conclude that the value $\pi = 0.08$ is the value that maximizes the likelihood function among the three values of π provided.

(b) Recall that the maximum likelihood estimator for a binomial distribution was computed in Example as $\hat{\pi}(\mathbf{X}) = \frac{\sum_{i=1}^{m} x_i}{mn}$.

Therefore, the maximum likelihood estimate for the proportion of mislabeled cans is $\hat{\pi}(\mathbf{x}) = \frac{8}{1 \cdot 100} = 0.08$. ■

Example 7.22 ▷ ***I.I.D. Uniform Random Variables*** ◁
Suppose $\{X_1, X_2, \ldots, X_n\}$ is a random sample from a $Unif(0, \theta)$ distribution. Find the maximum likelihood estimator of θ. Find the maximum likelihood estimate for a randomly generated sample of 1,000 $Unif(0, 3)$ random variables.

Solution: According to Box (**??**), the **pdf** of a random variable $X \sim Unif(0, \theta)$ is

$$f(x|\theta) = \frac{1}{\theta}, \quad 0 \le x \le \theta.$$

The likelihood function is

$$L(\theta|\mathbf{x}) = \begin{cases} \frac{1}{\theta^n} & \text{for } 0 \le x_1 \le \theta, 0 \le x_2 \le \theta, \ldots, 0 \le x_n \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

In this problem, the standard calculus approach fails since the maximum of the likelihood function occurs at a point of discontinuity. Consider the graph in Figure 7.5 on the following page.

MARQUETTE UNIVERSITY
Be The Difference.

Clearly $\frac{1}{\theta^n}$ is maximized for small values of θ.

However, the likelihood function is only defined for $\theta \geq \max(x_i)$.

Specifically, if $\theta < \max(x_i)$, $L(\theta|\mathbf{x}) = 0$. It follows

then that the maximum likelihood estimator is $\hat{\theta}(\mathbf{X}) = \max(X_i)$. The following code finds the maximum likelihood estimate of 1,000 randomly generated *Unif* $(0, 3)$ random variables.

```
> set.seed(2)
> max(runif(1000, 0, 3))
[1] 2.998667
```

Thus, even though a standard calculus approach could not be used, the mle 2.998667 is quite good for $\theta = 3$. ■

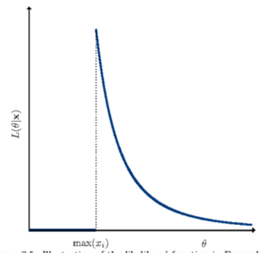


Figure 7.5: Illustration of the likelihood function in Example 7.22

MARQUETTE UNIVERSITY
Be The Difference.

Example 7.23 Suppose $\{X_1, X_2, \ldots, X_n\}$ is a random sample from a $N(\mu, \sigma)$ distribution, where σ is assumed known. Find the maximum likelihood estimator of μ.

Solution: According to Box ?? on page ??, the **pdf** of a random variable $X \sim N(\mu, \sigma)$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

The likelihood function is

$$L(\mu|\mathbf{x}) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}, \tag{7.35}$$

and the log-likelihood function is

$$\ln L(\mu|\mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}. \tag{7.36}$$

SOLUTION CONT'D:

To find the value of μ that maximizes $\ln L(\mu|\mathbf{x})$, take the first-order partial derivative of (7.36) with respect to μ, set the answer equal to zero, and solve. The first-order partial derivative of $\ln L(\mu|\mathbf{x})$ with respect to μ is

$$\frac{\partial \ln L(\mu, \sigma^2|\mathbf{x})}{\partial \mu} = \frac{\sum_{i=1}^{n}(x_i - \mu)}{\sigma^2} \overset{\text{set}}{=} 0. \quad (7.37)$$

The solution to (7.37) is $\mu = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$. For $\mu = \bar{x}$ to be a maximum, the second-order partial derivative of the log-likelihood function with respect to μ must be negative at $\mu = \bar{x}$. The second-order partial derivative of (7.36) is

$$\frac{\partial^2 \ln L(\mu|\mathbf{x})}{\partial \mu^2} = -\frac{n}{\sigma^2} < 0. \quad (7.38)$$

Since (7.35), goes to zero at $\pm\infty$, the boundary values, it follows that $\mu = \bar{x}$ is a global maximum. Consequently, the maximum likelihood estimator of μ is $\hat{\mu}(\mathbf{X}) = \overline{X}$, and the maximum likelihood estimate of μ is $\hat{\mu}(\mathbf{x}) = \bar{x}$. ■

66

Example 7.24 Suppose $\{X_1, X_2, \ldots, X_n\}$ is a random sample from a $N(\mu, \sigma)$ distribution, where μ is assumed known. Find the maximum likelihood estimator of σ^2.

Solution: According to Box ?? on page ??, the pdf of a random variable $X \sim N(\mu, \sigma)$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

The likelihood function is

$$L(\sigma^2|\mathbf{x}) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}, \quad (7.39)$$

and the log-likelihood function is

$$\ln L(\sigma^2|\mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}. \quad (7.40)$$

67

SOLUTION CONT'D:

To find the value of σ^2 that maximizes $\ln L(\sigma^2|\mathbf{x})$, take the first-order partial derivative of (7.40) with respect to σ^2, set the answer equal to zero, and solve. The first-order partial derivative of $\ln L(\sigma^2|\mathbf{x})$ with respect to σ^2 is

$$\frac{\partial \ln L(\mu,\sigma^2|\mathbf{x})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^4} \overset{\text{set}}{=} 0. \qquad (7.41)$$

The solution to (7.41) is $\sigma^2 = \frac{\sum_{i=1}^n (x_i-\mu)^2}{n}$. For $\sigma^2 = \frac{\sum_{i=1}^n (x_i-\mu)^2}{n}$ to be a maximum, the second-order partial derivative of the log-likelihood function with respect to σ^2 must be negative at $\sigma^2 = s_u^2$. For notational ease, let $r = \sigma^2$ in (7.40) so that

$$\ln L(r|\mathbf{x}) = \ln L(\sigma^2|\mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(r) - \frac{\sum_{i=1}^n (x_i-\mu)^2}{2r}. \qquad (7.42)$$

SOLUTION CONT'D:

The second-order partial derivative of (7.42) is

$$\frac{\partial^2 \ln L(r|\mathbf{x})}{\partial r^2} = \frac{n}{2}r^{-2} - \sum_{i=1}^n (x_i-\mu)^2 r^{-3} \overset{?}{<} 0. \qquad (7.43)$$

Multiplying the left hand side of (7.43) by r^3 gives

$$\frac{n}{2}r - \sum_{i=1}^n (x_i-\mu)^2 \overset{?}{<} 0. \qquad (7.44)$$

By substituting the value for the mle, $r = \frac{\sum_{i=1}^n (x_i-\mu)^2}{n}$, the ? above the $<$ can be removed since

$$\frac{r}{2} < \frac{\sum_{i=1}^n (x_i-\mu)^2}{n} = \sigma^2 = r.$$

Since (7.39), goes to zero at $\pm\infty$, the boundary values, it follows that $\sigma^2 = \frac{\sum_{i=1}^n (x_i-\mu)^2}{n}$ is a global maximum. Consequently, the maximum likelihood estimator of σ^2 is $\widehat{\sigma^2}(\mathbf{X}) = \frac{\sum_{i=1}^n (X_i-\mu)^2}{n}$, and the maximum likelihood estimate of σ^2 is $\widehat{\sigma^2}(\mathbf{x}) = \frac{\sum_{i=1}^n (x_i-\mu)^2}{n}$. ■

Example 7.25 Use `random.seed(33)` to generate 1,000 $N(4,1)$ random variables. Write log-likelihood functions for the simulated random variables and verify that the simulated maximum likelihood estimates for μ and σ^2 are reasonably close to the true parameters. Produce side by side graphs of $\ln L(\mu|\mathbf{x})$ and $\ln L(\sigma^2|\mathbf{x})$ indicating where the simulated maximum occurs in each graph.

Solution: The code provided is for R. To have the given code function in S-PLUS, replace the function `nlm()` with `nlmin()`.

```
➢ par(mfrow = c(1, 2))
➢ n <- 1000; sigma <- 1; set.seed(33); x <- rnorm(n, 4, sigma)
➢ mu <- seq(2, 6, length = n)
➢ negloglikemu <- function(mu) {
  n/2 * log(2*pi) + n/2 * log(sigma^2) +
  (sum(x^2) - 2*mu*sum(x) + n*mu^2)/(2*sigma^2)
  }
➢ EM <- nlm(negloglikemu, 2)$estimate
➢ EM
➢ [1] 4.019708
```


R CODE CONT'D:

```
➢ mu1 <- 4
➢ negloglike <- function(sigma2) {
  n/2 * log(2*pi) + n/2 * log(sigma2) +
  (sum((x - mu1)^2))/(2 * sigma2)
  }
➢ ES <- nlm(negloglike, 0.5)$estimate
➢ ES
[1] 1.000426
```

Note that the maximum likelihood estimates for μ and σ^2 from the simulation are 4.019708 and 1.000426 respectively which are reasonably close to the parameters $\mu = 4$ and $\sigma^2 = 1$.

Code for graph of $\ln L(\mu|\mathbf{x})$ versus μ

```
➢ plot(mu, -negloglikemu(mu), type="n")
➢ lines(mu, -negloglikemu(mu), lwd=2)
➢ abline(v = EM, lty = 2)
```

MARQUETTE UNIVERSITY
Be The Difference.

R CODE CONT'D:

Code for graph of $\ln L(\sigma^2|\mathbf{x})$ versus σ^2

```
➢ sigma2 <- seq(0.5, 1.5, length = 1000)
➢ plot(sigma2, -negloglike(sigma2), type="n")
➢ lines(sigma2, -negloglike(sigma2), lwd=2)
➢ abline(v = ES, lty=2)
```

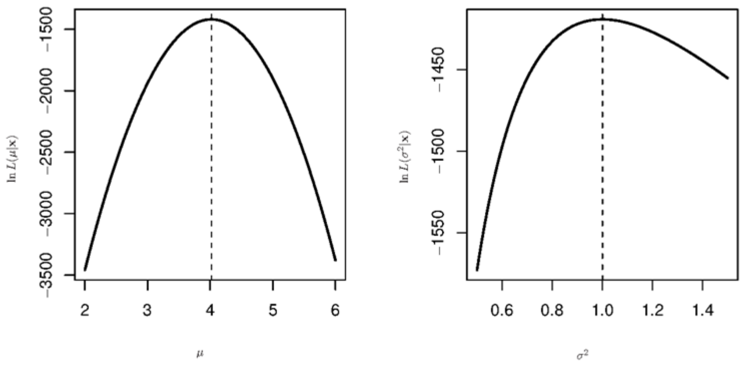


Figure 7.6: Illustration of $\ln L(\mu|\mathbf{x})$ and $\ln L(\sigma^2|\mathbf{x})$

72

MARQUETTE UNIVERSITY
Be The Difference.

7.3.2.3 Properties of Maximum Likelihood Estimators

Now that the Fisher information has been examined and several problems have been worked with maximum likelihood estimation, the properties of maximum likelihood estimators are formally enumerated:

1. MLEs are not necessarily unbiased. For example, when sampling from a $N(\mu, \sigma)$ population, the MLE of σ^2 is $\widehat{\sigma^2}(\mathbf{X}) = \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{n}$ which is a biased estimator of σ^2. However, although some MLEs may be biased, all MLEs are consistent which makes them asymptoticall unbiased. Symbolically, MLEs $\not\Rightarrow$ unbiased estimators; however, MLEs $\Rightarrow$ asymptotically unbiased estimators since MLEs $\Rightarrow$ consistent

2. If T is a MLE of θ and g is any function, then $g(T)$ is the MLE of $g(\theta)$. This is known as the **invariance** property of MLEs. For example, if $\overline{X}$ is the MLE of θ, then $\overline{X}^2$ is the MLE of θ^2.

73

MARQUETTE UNIVERSITY Be The Difference.

PROPERTIES OF MLE'S CONT'D:

3. When certain regularity conditions on $f(x|\theta)$ are satisfied, and an efficient estimator exists for the estimated parameter, the efficient estimator is the MLE of the estimated parameter. Be careful, not all MLEs are efficient! However, if an efficient estimator exists, the efficient estimator is also the MLE. That is, efficiency $\Rightarrow$ MLE, but MLE $\not\Rightarrow$ efficiency necessarily.

4. Under certain regularity conditions on $f(x|\theta)$, the MLE $\hat{\theta}(\mathbf{X})$ of θ based on a sample of size n from $f(x|\theta)$ is asymptotically normally distributed with mean θ, and variance $I_n(\theta)^{-1}$. That is as $n \longrightarrow \infty$,

$$\hat{\theta}(\mathbf{X}) \sim N\left(\theta, \sqrt{I_n(\theta)^{-1}}\right). \qquad (7.51)$$

The statement in (7.51) is the basis for large sample hypothesis tests (covered in Chapter **??**) and confidence intervals (covered in Chapter **??**).

MARQUETTE UNIVERSITY Be The Difference.

Note that the asymptotic variance of MLEs equals the Cramér-Rao lower bound since they are asymptotically efficient. That is, MLEs $\Rightarrow$ asymptotic efficiency. Consequently, a reasonable approximation to the distribution of $\hat{\theta}(\mathbf{X})$ for large sample sizes can be obtained. However, a normal distribution for $\hat{\theta}(\mathbf{X})$ cannot be guaranteed when the sample size is small.

Example 7.28 In Example 7.17 on page 68, it was found that the sample proportion of successes for a random sample of size m from a $Bin(n, \pi)$ distribution had $\hat{\pi} = \frac{\sum_{i=1}^{m} x_i}{mn}$ for its mle. That is, the MLE for the binomial proportion π is $\hat{\pi}(\mathbf{X}) = \frac{\sum_{i=1}^{m} X_i}{mn}$. What is the MLE for the variance of the sample proportion of successes where the random variable $\hat{\pi}$ is defined as $\frac{\sum_{i=1}^{m} X_i}{mn}$?

Solution: Given that $X \sim Bin(n, \pi)$, the variance of X is $n\pi(1-\pi)$. Therefore,

$$\text{var}[\hat{\pi}] = \text{var}\left[\frac{\sum_{i=1}^{m} X_i}{mn}\right] = \frac{\sum_{i=1}^{m} \text{var}[X_i]}{m^2 n^2} = \frac{mn\pi(1-\pi)}{m^2 n^2} = \frac{\pi(1-\pi)}{mn}.$$

Since $\text{var}[\hat{\pi}]$ is a function of the MLE $\hat{\pi}(\mathbf{X})$, it follows using the invariance property of MLEs that the MLE of the variance of $\hat{\pi}$ is

$$\widehat{\text{var}}\left[\hat{\pi}(\mathbf{X})\right] = \frac{\hat{\pi}(1-\hat{\pi})}{mn}.$$

Note: Many texts will list the MLE of the variance of the sample proportion of successes in a binomial distribution as $\frac{\hat{\pi}(1-\hat{\pi})}{n}$ because they use $m = 1$ in their definition of $\hat{\pi}$. ■

Example 7.29 ▷ ***MOM and MLE for a Gamma*** ◁ Given a random sample of size n from a population with pdf

$$f(x|\theta) = \frac{x}{\theta^2} e^{-\frac{x}{\theta}}, \quad x \geq 0, \quad \theta > 0,$$

(a) Find an estimator of θ using the method of moments.

(b) Find an estimator of θ using the method of maximum likelihood.

(c) Are the method of moments and maximum likelihood estimators of θ unbiased?

(d) Compute the variance of the MLE of θ.

(e) Is the MLE of θ efficient?

Solution: Since $X \sim Gamma\left(\alpha = 2, \lambda = \frac{1}{\theta}\right)$, according to Box (**??**) $E[X] = \frac{\alpha}{\lambda} = 2\theta$ and $\text{var}[X] = \frac{\alpha}{\lambda^2} = 2\theta^2$.

(a) Equating the first population moment about the origin to the first sample moment about the origin gives

$$\alpha_1(\theta) = 2\theta \overset{\text{set}}{=} \overline{X} = m_1,$$

which implies that the method of moments estimator for θ is $\tilde{\theta} = \frac{\overline{X}}{2}$.

(b) The likelihood equation is given as

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i) = \frac{\prod_{i=1}^{n} x_i}{\theta^{2n}} e^{\frac{-\sum_{i=1}^{n} x_i}{\theta}}, \qquad (7.52)$$

and the log-likelihood function is

$$\ln L(\theta|\mathbf{x}) = -2n\ln(\theta) + \sum_{i=1}^{n} \ln(x_i) - \frac{\sum_{i=1}^{n} x_i}{\theta}. \qquad (7.53)$$

To find the value of θ that maximizes $\ln L(\theta|\mathbf{x})$, take the first-order partial derivative of (7.53) with respect to θ, set the answer equal to zero, and solve. The first-order partial derivative of $\ln L(\theta|\mathbf{x})$ with respect to θ is

$$\frac{\partial \ln L(\theta|\mathbf{x})}{\partial\theta} = -\frac{2n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} \overset{\text{set}}{=} 0. \qquad (7.54)$$

The solution to (7.54) is $\theta = \frac{\overline{X}}{2}$ which agrees with the method of moments estimator. However, to ensure that $\theta = \frac{\overline{X}}{2}$ is a maximum, the second-order partial derivative with respect to θ must be negative. The second-order partial derivative of (7.53) is

$$\frac{\partial^2 \ln L(\theta|\mathbf{x})}{\partial\theta^2} = \frac{2n}{\theta^2} - \frac{2\sum_{i=1}^{n} x_i}{\theta^3} \overset{?}{<} 0. \qquad (7.55)$$

MARQUETTE UNIVERSITY
Be The Difference.

By using $\theta = \frac{\overline{X}}{2}$ in (7.55) arrive at the expression

$$-\frac{8n}{\overline{X}^2} \overset{?}{<} 0. \tag{7.56}$$

The ? above the < in (7.56) can be removed since $\int_0^0 f(x)\,dx = 0 \Rightarrow \overline{X} > 0$. Finally since (7.52) goes to zero as $\theta \to \infty$, it can be concluded that $\theta = \frac{\overline{X}}{2}$ is a global maximum. Consequently, the maximum likelihood estimator of θ is $\hat{\theta}(\mathbf{X}) = \frac{\overline{X}}{2}$.

(c) Since both the method of moments and the method of maximum likelihood returned the same estimator for θ, that is $\hat{\theta}(\mathbf{X}) = \tilde{\theta} = \frac{\overline{X}}{2}$, the question is

$$E\left[\hat{\theta}(\mathbf{X})\right] = E\left[\tilde{\theta}\right] \overset{?}{=} \theta.$$

MARQUETTE UNIVERSITY
Be The Difference.

Both $\tilde{\theta}$ and $\hat{\theta}(\mathbf{X})$ are therefore unbiased estimators since

$$E\left[\hat{\theta}(\mathbf{X})\right] = E\left[\tilde{\theta}\right] = E\left[\frac{\overline{X}}{2}\right] = \frac{\sum_{i=1}^{n} E\left[X_i\right]}{2n} = \frac{n \cdot 2\theta}{2n} = \theta.$$

(d) The variance of the MLE of θ is

$$\operatorname{var}\left[\hat{\theta}(\mathbf{X})\right] = \operatorname{var}\left[\frac{\overline{X}}{2}\right] = \operatorname{var}\left[\frac{\sum_{i=1}^{n} X_i}{2n}\right] = \frac{n\operatorname{var}\left[X\right]}{4n^2} = \frac{n2\theta^2}{4n^2} = \frac{\theta^2}{2n}.$$

(e) For $\hat{\theta}(\mathbf{X}) = \frac{\overline{X}}{2}$ to be considered an efficient or minimum variance estimator of θ, the variance of $\frac{\overline{X}}{2}$ must equal the CRLB. That is, does

$$\operatorname{var}\left[\hat{\theta}(\mathbf{X})\right] = \frac{\theta^2}{2n} \overset{?}{=} \frac{1}{n \cdot E\left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta}\right)^2\right]}?$$

Since $f(x|\theta) = \frac{x}{\theta^2}e^{-\frac{x}{\theta}}$ for $x \geq 0$, and $\theta > 0$, it follows that $\ln f(x|\theta) = \ln x - 2\ln\theta - \frac{x}{\theta}$, and that $\frac{\partial \ln f(x|\theta)}{\partial\theta} = \frac{x-2\theta}{\theta^2}$. Consequently,

$$\frac{1}{n \cdot E\left[\left(\frac{\partial \ln f(X|\theta)}{\partial\theta}\right)^2\right]} = \frac{1}{n \cdot E\left[\left(\frac{X-2\theta}{\theta^2}\right)^2\right]} = \frac{1}{\frac{n\cdot\text{var}[X]}{\theta^4}} = \frac{1}{\frac{n\cdot 2\theta^2}{\theta^4}} = \frac{\theta^2}{2n},$$

and conclude that $\frac{X}{2}$ is an efficient estimator of θ. ■

Example 7.30 ▷ ***MLEs for Exponentials*** ◁ Given a random sample of size n from an exponential distribution with pdf

$$f(x) = \frac{1}{\theta}e^{-\frac{x}{\theta}} \quad x \geq 0, \quad \theta > 0, \tag{7.57}$$

(a) Find the MLE of θ^2.

(b) Show that the MLE of θ^2 is a biased estimator of θ^2.

(c) Provide an unbiased estimator of θ^2.

(d) Find the variance of your MLE of θ^2.

(e) Find the variance of your unbiased estimator of θ^2.

MARQUETTE UNIVERSITY
Be The Difference.

Solution: To find the MLE of θ^2, there are two possibilities. First, the MLE of θ could be found and the invariance property could be used to say that this estimate squared is the MLE of θ^2. (See problem **??** of this chapter.) Second, and this is the current approach, the MLE of θ^2 can be found directly.

(a) For notational ease, use the change of variable $\theta^2 = p$, and $\theta = \sqrt{p}$ in (7.57). The resulting **pdf** using the change of variable is

$$f(x) = \frac{1}{\sqrt{p}} e^{-\frac{x}{\sqrt{p}}} \quad x \geq 0, \quad p > 0.$$

The likelihood function is

$$L(p|\mathbf{x}) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{p}} e^{-\frac{x_i}{\sqrt{p}}} = \frac{1}{(\sqrt{p})^n} e^{-\frac{\sum_{i=1}^{n} x_i}{\sqrt{p}}}, \qquad (7.58)$$

MARQUETTE UNIVERSITY
Be The Difference.

and the log-likelihood function is

$$\ln L(p|\mathbf{x}) = -\frac{n}{2} \ln p - \frac{\sum_{i=1}^{n} x_i}{\sqrt{p}}. \qquad (7.59)$$

To find the value of p that maximizes $\ln L(p|\mathbf{x})$, take the first-order partial derivative of (7.59) with respect to p, set the answer equal to zero, and solve. The first-order partial derivative of $\ln L(p|\mathbf{x})$ with respect to p is

$$\frac{\partial \ln L(p|\mathbf{x})}{\partial p} = -\frac{n}{2p} + \frac{\sum_{i=1}^{n} x_i}{2p^{\frac{3}{2}}} \stackrel{\text{set}}{=} 0. \qquad (7.60)$$

The solution to (7.60) is $p = \bar{x}^2$. For $p = \bar{x}^2$ to be a maximum, the second-order partial derivative of the log-likelihood function with respect to p must be negative at $p = \bar{x}^2$. The second-order partial

MARQUETTE UNIVERSITY
Be The Difference.

derivative of (7.59) is

$$\frac{\partial^2 \ln L(p|\mathbf{x})}{\partial p^2} = \frac{n}{2p^2} - \frac{3\sum_{i=1}^n x_i}{4p^{\frac{5}{2}}} \overset{?}{<} 0. \tag{7.61}$$

By substituting $p = \bar{x}^2$ in the right hand side of (7.61), the ? above the $<$ can be removed since $\bar{x} < \frac{3\bar{x}}{2}$ because $\bar{x} > 0$ for any sample due to the fact that $\mathbb{P}(X = 0) = 0$ for any continuous distribution. Finally, since as $p \to \infty$, $L(p|\mathbf{x}) \to 0$, it can be concluded that the MLE of $p = \theta^2$ is $\hat{p}(\mathbf{X}) = \widehat{\theta^2}(\mathbf{X}) = \overline{X}^2$.

(b) Next, show that $\overline{X}^2$ is a biased estimator of θ^2. The easiest way to determine the mean and variance of $\overline{X}^2$ is with moment generating functions. It is known that the moment generating function of an exponential random variable, X, is $M_X(t) = (1-\theta t)^{-1}$. Furthermore,

MARQUETTE UNIVERSITY
Be The Difference.

if $Y = \sum_{i=1}^n c_i X_i$ and each X_i has a moment generating function $M_{X_i}(t)$, then the moment generating function of Y is $M_Y(t) = \prod_{i=1}^n M_{X_i}(c_i t)$. In the case where $Y = \overline{X} = \frac{\sum_{i=1}^n X_i}{n}$, each $c_i = \frac{1}{n}$. For the special case of the exponential, the moment generating function for $\overline{X}$ is

$$M_{\overline{X}}(t) = M_Y(t) = \prod_{i=1}^n \left(1 - \theta \cdot \frac{t}{n}\right)^{-1} = \left(1 - \frac{\theta t}{n}\right)^{-n}.$$

Thus, to calculate the mean and variance of $\overline{X}^2$, take the first through fourth derivatives of $M_{\overline{X}}(t)$ and evaluate them when $t = 0$ to find $E\left[\overline{X}^i\right]$ for $i = 1, 2, 3$ and 4. The first, second, third, and fourth

MARQUETTE UNIVERSITY Be The Difference.

derivatives of $M_{\overline{X}}(t)$ respectively are

$$M'_{\overline{X}}(t) = -n\left(1-\frac{\theta}{n}t\right)^{-n-1}\left(-\frac{\theta}{n}\right)$$
$$M''_{\overline{X}}(t) = \theta(-n-1)\left(1-\frac{\theta}{n}t\right)^{-n-2}\left(-\frac{\theta}{n}\right)$$
$$M'''_{\overline{X}}(t) = \frac{\theta^2(n+1)}{n}(-n-2)\left(1-\frac{\theta}{n}t\right)^{-n-3}\left(-\frac{\theta}{n}\right)$$
$$M^{(4)}_{\overline{X}}(t) = \frac{\theta^3(n+1)(n+2)}{n^2}(-n-3)\left(1-\frac{\theta}{n}t\right)^{-n-4}\left(-\frac{\theta}{n}\right)$$

Evaluating these derivatives at $t = 0$ gives the expected values of $\overline{X}$ to the first, second, third, and fourth powers.

MARQUETTE UNIVERSITY Be The Difference.

$$M'_{\overline{X}}(0) = \theta = E\left[\overline{X}\right]$$
$$M''_{\overline{X}}(0) = \frac{\theta^2(n+1)}{n} = E[\overline{X}^2]$$
$$M'''_{\overline{X}}(0) = \frac{\theta^3(n+1)(n+2)}{n^2} = E[\overline{X}^3]$$
$$M^{(4)}_{\overline{X}}(0) = \frac{\theta^4(n+1)(n+2)(n+3)}{n^3} = E[\overline{X}^4]$$

Since $E[\overline{X}^2] = \dfrac{\theta^2(n+1)}{n} \neq \theta^2$, $\overline{X}^2$ is a biased estimator of θ^2.

(c) An unbiased estimator of θ^2 would be to use the quantity $\frac{n\overline{X}^2}{n+1}$.

(d) The variance of $\overline{X}^2$ can be computed as $E[\overline{X}^4] - \left(E[\overline{X}^2]\right)^2$.

$$\begin{aligned}\text{var}[\overline{X}^2] &= \frac{\theta^4(n+1)(n+2)(n+3)}{n^3} - \left(\frac{\theta^2(n+1)}{n}\right)^2 \\ &= \frac{2\theta^4(2n^2+5n+3)}{n^3} \\ &= \frac{2\theta^4\big((2n+3)(n+1)\big)}{n^3} \qquad (7.62)\end{aligned}$$

MARQUETTE UNIVERSITY
Be The Difference.

(e) The variance of the unbiased estimator of θ^2 is

$$\begin{aligned}\text{var}\left[\frac{n\overline{X}^2}{n+1}\right] &= \frac{n^2}{(n+1)^2}\text{var}[\overline{X}^2] \\ &= \frac{n^2}{(n+1)^2} \cdot \frac{2\theta^4\big((2n+3)(n+1)\big)}{n^3} \\ &= \frac{2\theta^4(2n+3)}{n(n+1)}.\end{aligned}$$

MARQUETTE UNIVERSITY
Be The Difference.

QUESTIONS?

- ANY QUESTION?

92