CHAPTER SEVEN

Point Estimation

Introduction

Given a parameter of interest, such as a population mean μ or population proportion p, the objective of point estimation is to use a sample to compute a number that represents in some sense a good guess for the true value of the parameter. The resulting number is called *a point estimate*. In Section 7.1, we present some general concepts of point estimation. In Section 7.2, we describe and illustrate two important methods for obtaining point estimates: the method of moments and the method of maximum likelihood.

Obtaining a point estimate entails calculating the value of a statistic such as the sample mean $\overline{X}$ or sample standard deviation S. We should therefore be concerned that the chosen statistic contains all the relevant information about the parameter of interest. The idea of no information loss is made precise by the concept of sufficiency, which is developed in Section 7.3. Finally, Section 7.4 further explores the meaning of efficient estimation and properties of maximum likelihood.

J.L. Devore and K.N. Berk, *Modern Mathematical Statistics with Applications*, Springer Texts in Statistics,
DOI 10.1007/978-1-4614-0391-3_7, © Springer Science+Business Media, LLC 2012

7.1 General Concepts and Criteria

Statistical inference is frequently directed toward drawing some type of conclusion about one or more parameters (population characteristics). To do so requires that an investigator obtain sample data from each of the populations under study. Conclusions can then be based on the computed values of various sample quantities. For example, let μ (a parameter) denote the average duration of anesthesia for a short-acting anesthetic. A random sample of $n = 10$ patients might be chosen, and the duration for each one determined, resulting in observed durations $x_1, x_2, \ldots, x_{10}$. The sample mean duration $\bar{x}$ could then be used to draw a conclusion about the value of μ. Similarly, if σ^2 is the variance of the duration distribution (population variance, another parameter), the value of the sample variance s^2 can be used to infer something about σ^2.

When discussing general concepts and methods of inference, it is convenient to have a generic symbol for the parameter of interest. We will use the Greek letter θ for this purpose. The objective of point estimation is to select a single number, based on sample data, that represents a sensible value for θ. Suppose, for example, that the parameter of interest is μ, the true average lifetime of batteries of a certain type. A random sample of $n = 3$ batteries might yield observed lifetimes (hours) $x_1 = 5.0$, $x_2 = 6.4$, $x_3 = 5.9$. The computed value of the sample mean lifetime is $\bar{x} = 5.77$, and it is reasonable to regard 5.77 as a very plausible value of μ, our "best guess" for the value of μ based on the available sample information.

Suppose we want to estimate a parameter of a single population (e.g., μ or σ) based on a random sample of size n. Recall from the previous chapter that before data is available, the sample observations must be considered random variables (rv's) $X_1, X_2, \ldots, X_n$. It follows that any function of the X_i's—that is, any statistic—such as the sample mean $\overline{X}$ or sample standard deviation S is also a random variable. The same is true if available data consists of more than one sample. For example, we can represent duration of anesthesia of m patients on anesthetic A and n patients on anesthetic B by $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, respectively. The difference between the two sample mean durations is $\overline{X} - \overline{Y}$, the natural statistic for making inferences about $\mu_1 - \mu_2$, the difference between the population mean durations.

DEFINITION

A point estimate of a parameter θ is a single number that can be regarded as a sensible value for θ. A point estimate is obtained by selecting a suitable statistic and computing its value from the given sample data. The selected statistic is called the **point estimator** of θ.

In the battery example just given, the estimator used to obtain the point estimate of μ was $\overline{X}$, and the point estimate of μ was 5.77. If the three observed lifetimes had instead been $x_1 = 5.6$, $x_2 = 4.5$, and $x_3 = 6.1$, use of the estimator $\overline{X}$ would have resulted in the estimate $\bar{x} = (5.6 + 4.5 + 6.1)/3 = 5.40$. The symbol $\hat{\theta}$ ("theta hat") is customarily used to denote both the estimator of θ and the point estimate resulting from a given sample.[1] Thus $\hat{\mu} = \overline{X}$ is read as "the point estimator of μ is the sample mean $\overline{X}$." The statement "the point estimate of μ is 5.77" can be written concisely as $\hat{\mu} = 5.77$. Notice that in writing $\hat{\theta} = 72.5$, there is no indication of how this point estimate was obtained (what statistic was used). It is recommended that both the estimator and the resulting estimate be reported.

Example 7.1 An automobile manufacturer has developed a new type of bumper, which is supposed to absorb impacts with less damage than previous bumpers. The manufacturer has used this bumper in a sequence of 25 controlled crashes against a wall, each at 10 mph, using one of its compact car models. Let X = the number of crashes that result in no visible damage to the automobile. The parameter to be estimated is p = the proportion of all such crashes that result in no damage [alternatively, $p = P$(no damage in a single crash)]. If X is observed to be $x = 15$, the most reasonable estimator and estimate are

$$\text{estimator} \quad \hat{p} = \frac{X}{n} \qquad \text{estimate} = \frac{x}{n} = \frac{15}{25} = .60$$

■

If for each parameter of interest there were only one reasonable point estimator, there would not be much to point estimation. In most problems, though, there will be more than one reasonable estimator.

Example 7.2 Reconsider the accompanying 20 observations on dielectric breakdown voltage for pieces of epoxy resin introduced in Example 4.36 (Section 4.6).

24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88

The pattern in the normal probability plot given there is quite straight, so we now assume that the distribution of breakdown voltage is normal with mean value μ. Because normal distributions are symmetric, μ is also the median lifetime of the distribution. The given observations are then assumed to be the result of a random sample $X_1, X_2, \ldots, X_{20}$ from this normal distribution. Consider the following estimators and resulting estimates for μ:

a. Estimator $= \overline{X}$, estimate $= \bar{x} = \sum x_i/n = 555.86/20 = 27.793$

b. Estimator $= \widetilde{X}$, estimate $= \tilde{x} = (27.94 + 27.98)/2 = 27.960$

c. Estimator $= \overline{X}_e = [\min(X_i) + \max(X_i)]/2$ = the midrange, (average of the two extreme lifetimes), estimate $= [\min(x_i) + \max(x_i)]/2 = (24.46 + 30.88)/2 = 27.670$

d. Estimator $= \overline{X}_{\text{tr}(10)}$, the 10% trimmed mean (discard the smallest and largest 10% of the sample and then average),

$$\text{estimate} = \bar{x}_{\text{tr}(10)} = \frac{555.86 - 24.46 - 25.61 - 29.50 - 30.88}{16} = 27.838$$

[1] Following earlier notation, we could use $\hat{\Theta}$ (an uppercase theta) for the estimator, but this is cumbersome to write.

Each one of the estimators (a)–(d) uses a different measure of the center of the sample to estimate μ. Which of the estimates is closest to the true value? We cannot answer this without knowing the true value. A question that can be answered is, "Which estimator, when used on other samples of X_i's, will tend to produce estimates closest to the true value?" We will shortly consider this type of question. ■

Example 7.3 Studies have shown that a calorie-restricted diet can prolong life. Of course, controlled studies are much easier to do with lab animals. Here is a random sample of eight lifetimes (days) taken from a population of 106 rats that were fed a restricted diet (from "Tests and Confidence Sets for Comparing Two Mean Residual Life Functions," *Biometrics*, 1988: 103–115)

716 1144 1017 1138 389 1221 530 958

Let $X_1, \ldots, X_8$ denote the lifetimes as random variables, before the observed values are available. We want to estimate the population variance σ^2. A natural estimator is the sample variance:

$$\hat{\sigma}^2 = S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1} = \frac{\sum X_i^2 - (\sum X_i)^2/n}{n-1}$$

The corresponding estimate is

$$\hat{\sigma}^2 = s^2 = \frac{\sum x_i^2 - (\sum x_i)^2/8}{7} = \frac{6{,}991{,}551 - (7113)^2/8}{7} = \frac{667{,}205}{7} = 95{,}315$$

The estimate of σ would then be $\hat{\sigma} = s = \sqrt{95{,}315} = 309$

An alternative estimator would result from using divisor n instead of $n - 1$ (i.e., the average squared deviation):

$$\hat{\sigma}^2 = \frac{\sum (X_i - \overline{X})^2}{n} \qquad \text{estimate} = \frac{667{,}205}{8} = 83,401$$

We will indicate shortly why many statisticians prefer S^2 to the estimator with divisor n. ■

In the best of all possible worlds, we could find an estimator $\hat{\theta}$ for which $\hat{\theta} = \theta$ always. However, $\hat{\theta}$ is a function of the sample X_i's, so it is a random variable. For some samples, $\hat{\theta}$ will yield a value larger than θ, whereas for other samples $\hat{\theta}$ will underestimate θ. If we write

$$\hat{\theta} = \theta + \text{error of estimation}$$

then an accurate estimator would be one resulting in small estimation errors, so that estimated values will be near the true value.

Mean Squared Error

A popular way to quantify the idea of $\hat{\theta}$ being close to θ is to consider the squared error $(\hat{\theta} - \theta)^2$. Another possibility is the absolute error $|\hat{\theta} - \theta|$, but this is more difficult to work with mathematically. For some samples, $\hat{\theta}$ will be quite close to θ and the resulting squared error will be very small, whereas the squared error will be quite large whenever a sample produces an estimate $\hat{\theta}$ that is far from the target. An omnibus measure of accuracy is the mean squared error (expected squared error), which entails averaging the squared error over all possible samples and resulting estimates.

DEFINITION

The **mean squared error** of an estimator $\hat{\theta}$ is $E[(\hat{\theta}-\theta)^2]$.

A useful result when evaluating mean squared error is a consequence of the following rearrangement of the shortcut for evaluating a variance $V(Y)$:

$$V(Y)=E\left(Y^2\right)-[E(Y)]^2 \quad \Rightarrow \quad E\left(Y^2\right)=V(Y)+[E(Y)]^2$$

That is, the expected value of the square of Y is the variance plus the square of the mean value. Letting $Y=\hat{\theta}-\theta$, the estimation error, the left-hand side is just the mean squared error. The first term on the right-hand side is $V(\hat{\theta}-\theta)=V(\hat{\theta})$ since θ is just a constant. The second term involves $E(\hat{\theta}-\theta)=E(\hat{\theta})-\theta$, the difference between the expected value of the estimator and the value of the parameter. This difference is called the **bias** of the estimator. Thus

$$\text{MSE}=V(\hat{\theta})+[E(\hat{\theta})-\theta]^2=\text{ variance of estimator }+(\text{bias})^2$$

Example 7.4

(Example 7.1 continued)

Consider once again estimating a population proportion of "successes" p. The natural estimator of p is the sample proportion of successes $\hat{p}=X/n$. The number of successes X in the sample has a binomial distribution with parameters n and p, so $E(X)=np$ and $V(X)=np(1-p)$. The expected value of the estimator is

$$E(\hat{p})=E\left(\frac{X}{n}\right)=\frac{1}{n}E(X)=\frac{1}{n}np=p$$

Thus the bias of $\hat{p}$ is $p-p=0$, giving the mean squared error as

$$E[(\hat{p}-p)^2]=V(\hat{p})+0^2=V\left(\frac{X}{n}\right)=\frac{1}{n^2}V(X)=\frac{p(1-p)}{n}$$

Now consider the alternative estimator $\hat{p}=(X+2)/(n+4)$. That is, add two successes and two failures to the sample and then calculate the sample proportion of successes. One intuitive justification for this estimator is that

$$\left|\frac{X}{n}-.5\right|=\left|\frac{X-.5n}{n}\right| \qquad \left|\frac{X+2}{n+4}-.5\right|=\left|\frac{X-.5n}{n+4}\right|$$

from which we see that the alternative estimator is always somewhat closer to .5 than is the usual estimator. It seems particularly reasonable to move the estimate toward .5 when the number of successes in the sample is close to 0 or n. For example, if there are no successes at all in the sample, is it sensible to estimate the population proportion of successes as zero, especially if n is small?

The bias of the alternative estimator is

$$E\left(\frac{X+2}{n+4}\right) - p = \frac{1}{n+4}E(X+2) - p = \frac{np+2}{n+4} - p = \frac{2/n - 4p/n}{1+4/n}$$

This bias is not zero unless $p = .5$. However, as n increases the numerator approaches zero and the denominator approaches 1, so the bias approaches zero. The variance of the estimator is

$$V\left(\frac{X+2}{n+4}\right) = \frac{1}{(n+4)^2}V(X+2) = \frac{V(X)}{(n+4)^2} = \frac{np(1-p)}{(n+4)^2} = \frac{p(1-p)}{n+8+16/n}$$

This variance approaches zero as the sample size increases. The mean squared error of the alternative estimator is

$$\text{MSE} = \frac{p(1-p)}{n+8+16/n} + \left(\frac{2/n - 4p/n}{1+4/n}\right)^2$$

So how does the mean squared error of the usual estimator, the sample proportion, compare to that of the alternative estimator? If one MSE were smaller than the other for all values of p, then we could say that one estimator is always preferred to the other (using MSE as our criterion). But as Figure 7.1 shows, this is not the case at least for the sample sizes $n = 10$ and $n = 100$, and in fact is not true for any other sample size.

According to Figure 7.1, the two MSE's are quite different when n is small. In this case the alternative estimator is better for values of p near .5 (since it moves the sample proportion toward .5) but not for extreme values of p. For large n the two MSE's are quite similar, but again neither dominates the other.

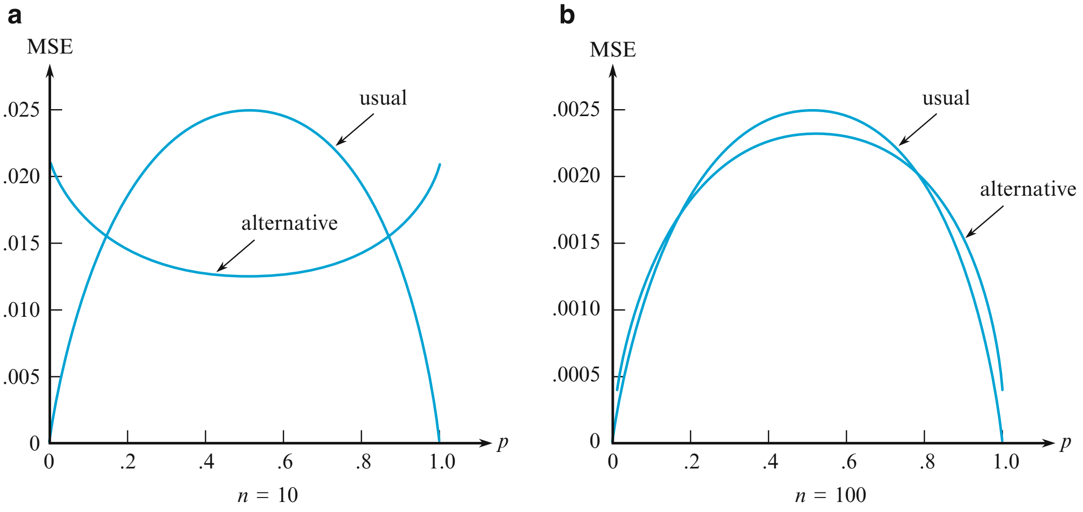


Figure 7.1 Graphs of MSE for the usual and alternative estimators of p ■

Seeking an estimator whose mean squared error is smaller than that of every other estimator for all values of the parameter is generally too ambitious a goal. One common approach is to restrict the class of estimators under consideration in some way, and then seek the estimator that is best in that restricted class. A very popular restriction is to impose the condition of unbiasedness.

Unbiased Estimators

Suppose we have two measuring instruments; one instrument has been accurately calibrated, but the other systematically gives readings smaller than the true value being measured. When each instrument is used repeatedly on the same object, because of measurement error, the observed measurements will not be identical. However, the measurements produced by the first instrument will be distributed about the true value in such a way that on average this instrument measures what it purports to measure, so it is called an unbiased instrument. The second instrument yields observations that have a systematic error component or bias.

DEFINITION A point estimator $\hat{\theta}$ is said to be an **unbiased estimator** of θ if $E(\hat{\theta}) = \theta$ for every possible value of θ. If $\hat{\theta}$ is not unbiased, the difference $E(\hat{\theta}) - \theta$ is called the **bias** of $\hat{\theta}$.

That is, $\hat{\theta}$ is unbiased if its probability (i.e., sampling) distribution is always "centered" at the true value of the parameter. Suppose $\hat{\theta}$ is an unbiased estimator; then if $\theta = 100$, the $\hat{\theta}$ sampling distribution is centered at 100; if $\theta = 27.5$, then the $\hat{\theta}$ sampling distribution is centered at 27.5, and so on. Figure 7.2 pictures the distributions of several biased and unbiased estimators. Note that "centered" here means that the expected value, not the median, of the distribution of $\hat{\theta}$ is equal to θ.

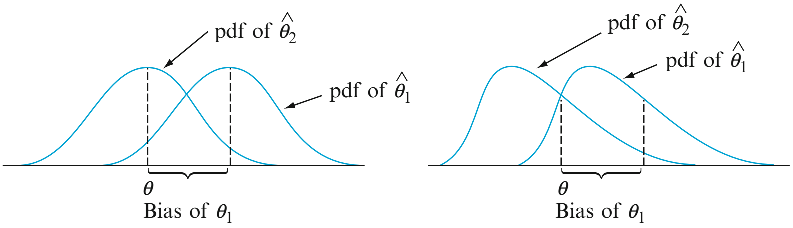


Figure 7.2 The pdf's of a biased estimator $\hat{\theta}_1$ and an unbiased estimator $\hat{\theta}_2$ for a parameter θ

It may seem as though it is necessary to know the value of θ (in which case estimation is unnecessary) to see whether $\hat{\theta}$ is unbiased. This is usually not the case, however, because unbiasedness is a general property of the estimator's sampling distribution—where it is centered—which is typically not dependent on any particular parameter value. For example, in Example 7.4 we showed that $E(\hat{p}) = p$ when $\hat{p}$ is the sample proportion of successes. Thus if $p = .25$, the sampling distribution of $\hat{p}$ is centered at .25 (centered in the sense of mean value), when $p = .9$ the sampling distribution is centered at .9, and so on. It is not necessary to know the value of p to know that $\hat{p}$ is unbiased.

PROPOSITION When X is a binomial rv with parameters n and p, the sample proportion $\hat{p} = X/n$ is an unbiased estimator of p.

Example 7.5 Suppose that X, the reaction time to a stimulus, has a uniform distribution on the interval from 0 to an unknown upper limit θ (so the density function of X is rectangular in shape with height $1/\theta$ for $0 \le x \le \theta$). An investigator wants to estimate θ on the basis of a random sample $X_1, X_2, \ldots, X_n$ of reaction times. Since θ is the largest possible time in the entire population of reaction times, consider as a first estimator the largest sample reaction time: $\hat{\theta}_b = \max(X_1, \ldots, X_n)$. If $n = 5$ and $x_1 = 4.2$, $x_2 = 1.7$, $x_3 = 2.4$, $x_4 = 3.9$, $x_5 = 1.3$, the point estimate of θ is $\hat{\theta}_b = \max(4.2,\ 1.7,\ 2.4,\ 3.9,\ 1.3) = 4.2$.

Unbiasedness implies that some samples will yield estimates that exceed θ and other samples will yield estimates smaller than θ — otherwise θ could not possibly be the center (balance point) of $\hat{\theta}_b$'s distribution. However, our proposed estimator will never overestimate θ (the largest sample value cannot exceed the largest population value) and will underestimate θ unless the largest sample value equals θ. This intuitive argument shows that $\hat{\theta}_b$ is a biased estimator. More precisely, using our earlier results on order statistics, it can be shown (see Exercise 50) that

$$E(\hat{\theta}_b) = \frac{n}{n+1} \cdot \theta < \theta \qquad \left(\text{since} \frac{n}{n+1} < 1\right)$$

The bias of $\hat{\theta}_b$ is given by $n\theta/(n + 1) - \theta = -\theta/(n + 1)$, which approaches 0 as n gets large.

It is easy to modify $\hat{\theta}_b$ to obtain an unbiased estimator of θ. Consider the estimator

$$\hat{\theta}_u = \frac{n+1}{n} \cdot \hat{\theta}_b = \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n)$$

Using this estimator on the data gives the estimate (6/5)(4.2) = 5.04. The fact that $(n + 1)/n > 1$ implies that $\hat{\theta}_u$ will overestimate θ for some samples and underestimate it for others. The mean value of this estimator is

$$\begin{aligned} E(\hat{\theta}_u) &= E\left[\frac{n+1}{n} \cdot \max(X_1, \ldots, X_n)\right] = \frac{n+1}{n} \cdot E[\max(X_1, \ldots, X_n)] \\ &= \frac{n+1}{n} \cdot \frac{n}{n+1}\theta = \theta \end{aligned}$$

If $\hat{\theta}_u$ is used repeatedly on different samples to estimate θ, some estimates will be too large and others will be too small, but in the long run there will be no systematic tendency to underestimate or overestimate θ. ■

Statistical practitioners who buy into the **Principle of Unbiased Estimation** would employ an unbiased estimator in preference to a biased estimator. On this basis, the sample proportion of successes should be preferred to the alternative estimator of p, and the unbiased estimator $\hat{\theta}_u$ should be preferred to the biased estimator $\hat{\theta}_b$ in the uniform distribution scenario of the previous example.

Example 7.6 Let's turn now to the problem of estimating σ^2 based on a random sample $X_1, \ldots, X_n$. First consider the estimator $S^2 = \sum (X_i - \overline{X}^2)/(n-1)$, the sample variance as we have defined it. Applying the result $E(Y^2) = V(Y) + [E(Y)]^2$ to

$$S^2 = \frac{1}{n-1}\left[\sum X_i^2 - \frac{(\sum X_i)^2}{n}\right]$$

from Section 1.4 gives

$$\begin{aligned} E(S^2) &= \frac{1}{n-1}\left\{\sum E(X_i^2) - \frac{1}{n}E\left[\left(\sum X_i\right)^2\right]\right\} \\ &= \frac{1}{n-1}\left\{\sum (\sigma^2+\mu^2) - \frac{1}{n}\left\{V\left(\sum X_i\right) + \left[E\left(\sum X_i\right)\right]^2\right\}\right\} \\ &= \frac{1}{n-1}\left\{n\sigma^2 + n\mu^2 - \frac{1}{n}n\sigma^2 - \frac{1}{n}(n\mu)^2\right\} \\ &= \frac{1}{n-1}\left\{n\sigma^2 - \sigma^2\right\} = \sigma^2 \end{aligned}$$

Thus we have shown that the **sample variance S^2 is an unbiased estimator of σ^2**.

The estimator that uses divisor n can be expressed as $(n-1)S^2/n$, so

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

This estimator is therefore biased. The bias is $(n-1)\sigma^2/n - \sigma^2 = -\sigma^2/n$. Because the bias is negative, the estimator with divisor n tends to underestimate σ^2, and this is why the divisor $n-1$ is preferred by many statisticians (although when n is large, the bias is small and there is little difference between the two).

This is not quite the whole story, however. Suppose the random sample has come from a normal distribution. Then from Section 6.4 , we know that the rv $(n-1)S^2/\sigma^2$ has a chi-squared distribution with $n-1$ degree of freedom. The mean and variance of a chi-squared variable are df and 2 df, respectively. Let's now consider estimators of the form

$$\hat{\sigma}^2 = c\sum (X_i - \overline{X})^2$$

The expected value of the estimator is

$$E\left[c\sum (X_i - \overline{X})^2\right] = c(n-1)E(S^2) = c(n-1)\sigma^2$$

so the bias is $c(n-1)\sigma^2 - \sigma^2$. The only unbiased estimator of this type is the sample variance, with $c = 1/(n-1)$.

Similarly, the variance of the estimator is

$$V\left[c\sum(X_i - \overline{X})^2\right] \quad = \quad V\left[c\sigma^2\,\frac{(n-1)S^2}{\sigma^2}\right] \quad = \quad c^2\sigma^4[2(n-1)]$$

Substituting these expressions into the relationship MSE = variance + (bias)^2, the value of c for which MSE is minimized can be found by taking the derivative with respect to c, equating the resulting expression to zero, and solving for c. The result is $c = 1/(n + 1)$. So in this situation, the principle of unbiasedness and the principle of minimum MSE are at loggerheads.

As a final blow, even though S^2 is unbiased for estimating σ^2, *it is not true* that the sample standard deviation S is unbiased for estimating σ. This is because the square root function is not linear, so the expected value of the square root is not the square root of the expected value. Well, if S is biased, why not find an unbiased estimator for σ and use it rather than S? Unfortunately there is no estimator of σ that is unbiased irrespective of the nature of the population distribution (although in special cases, e.g., a normal distribution, an unbiased estimator does exist). Fortunately the bias of S is not serious unless n is quite small. So we shall generally employ it as an estimator. ■

In Example 7.2, we proposed several different estimators for the mean μ of a normal distribution. If there were a unique unbiased estimator for μ, the estimation dilemma could be resolved by using that estimator. Unfortunately, this is not the case.

PROPOSITION

If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with mean μ, then $\overline{X}$ is an unbiased estimator of μ. If in addition the distribution is continuous and symmetric, then $\widetilde{X}$ and any trimmed mean are also unbiased estimators of μ.

The fact that $\overline{X}$ is unbiased is just a restatement of one of our rules of expected value: $E(\overline{X}) = \mu$ for every possible value of μ (for discrete as well as continuous distributions). The unbiasedness of the other estimators is more difficult to verify; the argument requires invoking results on distributions of order statistics from Section 5.5.

According to this proposition, the principle of unbiasedness by itself does not always allow us to select a single estimator. When the underlying population is normal, even the third estimator in Example 7.2 is unbiased, and there are many other unbiased estimators. What we now need is a way of selecting among unbiased estimators.

Estimators with Minimum Variance

Suppose $\hat{\theta}_1$ and $\hat{\theta}_2$ are two estimators of θ that are both unbiased. Then, although the distribution of each estimator is centered at the true value of θ, the spreads of the distributions about the true value may be different.

PRINCIPLE OF MINIMUM VARIANCE UNBIASED ESTIMATION

Among all estimators of θ that are unbiased, choose the one that has minimum variance. The resulting $\hat{\theta}$ Is called the **minimum variance unbiased estimator (MVUE)** of θ. Since MSE = variance + (bias)^2, seeking an unbiased estimator with minimum variance is the same as seeking an unbiased estimator that has minimum mean squared error.

Figure 7.3 pictures the pdf's of two unbiased estimators, with the first $\hat{\theta}$ having smaller variance than the second estimator. Then the first $\hat{\theta}$ is more likely than the second one to produce an estimate close to the true θ. The MVUE is, in a certain sense, the most likely among all unbiased estimators to produce an estimate close to the true θ.

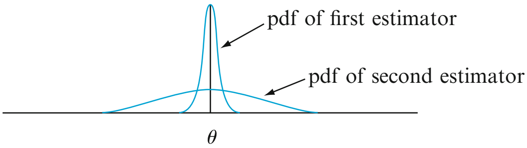


Figure 7.3 Graphs of the pdf's of two different unbiased estimators

Example 7.7 We argued in Example 7.5 that when $X_1, \ldots, X_n$ is a random sample from a uniform distribution on $[0, \theta]$, the estimator

$$\hat{\theta}_1 = \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n)$$

is unbiased for θ (we previously denoted this estimator by $\hat{\theta}_u$). This is not the only unbiased estimator of θ. The expected value of a uniformly distributed rv is just the midpoint of the interval of positive density, so $E(X_i) = \theta/2$. This implies that $E(\overline{X}) = \theta/2$, from which $E(2\overline{X}) = \theta$. That is, the estimator $\hat{\theta}_2 = 2\overline{X}$ is unbiased for θ.

If X is uniformly distributed on the interval $[A, B]$, then $V(X) = \sigma^2 = (B-A)^2/12$ (Exercise 23 in Chapter 4). Thus, in our situation, $V(X_i) = \theta^2/12$, $V(\overline{X}) = \sigma^2/n = \theta^2/(12n)$, and $V(\hat{\theta}_2) = V(2\overline{X}) = 4V(\overline{X}) = \theta^2/(3n)$. The results of Exercise 50 can be used to show that $V(\hat{\theta}_1) = \theta^2/[n(n+2)]$. The estimator $\hat{\theta}_1$ has smaller variance than does $\hat{\theta}_2$ if $3n < n(n+2)$—that is, if $0 < n^2 - n = n(n-1)$. As long as $n > 1$, $V(\hat{\theta}_1) < V(\hat{\theta}_2)$, so $\hat{\theta}_1$ is a better estimator than $\hat{\theta}_2$. More advanced methods can be used to show that $\hat{\theta}_1$ is the MVUE of θ—every other unbiased estimator of θ has variance that exceeds $\theta^2/[n(n+2)]$. ■

One of the triumphs of mathematical statistics has been the development of methodology for identifying the MVUE in a wide variety of situations. The most important result of this type for our purposes concerns estimating the mean μ of a normal distribution. For a proof in the special case that σ is known, see Exercise 45.

THEOREM

Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with parameters μ and σ. Then the estimator $\hat{\mu} = \overline{X}$ is the MVUE for μ.

Whenever we are convinced that the population being sampled is normal, the result says that $\overline{X}$ should be used to estimate μ. In Example 7.2, then, our estimate would be $\bar{x} = 27.793$.

Once again, in some situations such as the one in Example 7.6, it is possible to obtain an estimator with small bias that would be preferred to the best unbiased estimator. This is illustrated in Figure 7.4. However, MVUEs are often easier to obtain than the type of biased estimator whose distribution is pictured.

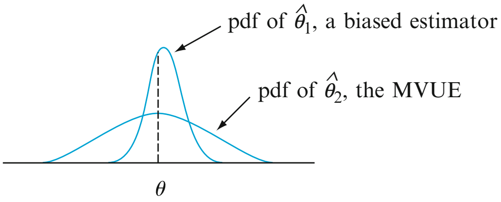

pdf of $\hat{\theta}_1$, a biased estimator
pdf of $\hat{\theta}_2$, the MVUE
θ

Figure 7.4 A biased estimator that is preferable to the MVUE

More Complications

The last theorem does not say that in estimating a population mean μ, the estimator $\overline{X}$ should be used irrespective of the distribution being sampled.

Example 7.8 Suppose we wish to estimate the number of calories θ in a certain food. Using standard measurement techniques, we will obtain a random sample $X_1, \ldots, X_n$ of n calorie measurements. Let's assume that the population distribution is a member of one of the following three families:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\theta)^2/(2\sigma^2)} \qquad -\infty < x < \infty \tag{7.1}$$

$$f(x) = \frac{1}{\pi[1 + (x-\theta)^2]} \qquad -\infty < x < \infty \tag{7.2}$$

$$f(x) = \begin{cases} \dfrac{1}{2c} & -c \le x - \theta \le c \\ 0 & \text{otherwise} \end{cases} \tag{7.3}$$

The pdf (7.1) is the normal distribution, (7.2) is called the Cauchy distribution, and (7.3) is a uniform distribution. All three distributions are symmetric about θ, which is therefore the median of each distribution. The value θ is also the mean for the normal and uniform distributions, but the mean of the Cauchy distribution fails to exist. This happens because, even though the Cauchy distribution is bell-shaped like the normal distribution, it has much heavier tails (more probability far out) than the normal curve. The uniform distribution has no tails. The four estimators for μ considered earlier are $\overline{X}, \widetilde{X}, \overline{X}_e$ (the average of the two extreme observations), and $\overline{X}_{\text{tr}(10)}$, a trimmed mean.

The very important moral here is that the best estimator for μ depends crucially on which distribution is being sampled. In particular,

1. If the random sample comes from a normal distribution, then $\overline{X}$ is the best of the four estimators, since it has minimum variance among all unbiased estimators.

2. If the random sample comes from a Cauchy distribution, then $\overline{X}$ and $\overline{X}_e$ are terrible estimators for μ, whereas $\widetilde{X}$ is quite good (the MVUE is not known); $\overline{X}$ is bad because it is very sensitive to outlying observations, and the heavy tails of the Cauchy distribution make a few such observations likely to appear in any sample.
3. If the underlying distribution is the particular uniform distribution in (7.3), then the best estimator is $\overline{X}_e$; in general, this estimator is greatly influenced by outlying observations, but here the lack of tails makes such observations impossible.
4. *The trimmed mean is best in none of these three situations but works reasonably well in all three*. That is, $\overline{X}_{tr(10)}$ does not suffer too much in comparison with the best procedure in any of the three situations. ■

More generally, recent research in statistics has established that when estimating a point of symmetry μ of a continuous probability distribution, a trimmed mean with trimming proportion 10% or 20% (from each end of the sample) produces reasonably behaved estimates over a very wide range of possible models. For this reason, a trimmed mean with small trimming percentage is said to be a **robust estimator**.

Until now, we have focused on comparing several estimators based on the same data, such as $\overline{X}$ and $\widetilde{X}$ for estimating μ when a sample of size n is selected from a normal population distribution. Sometimes an investigator is faced with a choice between alternative ways of gathering data; the form of an appropriate estimator then may well depend on how the experiment was carried out.

Example 7.9 Suppose a type of component has a lifetime distribution that is exponential with parameter λ so that expected lifetime is $\mu = 1/\lambda$. A sample of n such components is selected, and each is put into operation. If the experiment is continued until all n lifetimes, $X_1, \ldots, X_n$, have been observed, then $\overline{X}$ is an unbiased estimator of μ.

In some experiments, though, the components are left in operation only until the time of the rth failure, where $r < n$. This procedure is referred to as **censoring**. Let Y_1 denote the time of the first failure (the minimum lifetime among the n components), Y_2 denote the time at which the second failure occurs (the second smallest lifetime), and so on. Since the experiment terminates at time Y_r, the total accumulated lifetime at termination is

$$T_r = \sum_{i=1}^{r} Y_i + (n-r)Y_r$$

We now demonstrate that $\hat{\mu} = T_r/r$ is an unbiased estimator for μ. To do so, we need two properties of exponential variables:

1. The memoryless property (see Section 4.4) says that at any time point, remaining lifetime has the same exponential distribution as original lifetime.
2. If $X_1, \ldots, X_k$ are independent, each exponentially distributed with parameter λ, then min $(X_1, \ldots, X_k)$ is exponential with parameter $k\lambda$ and has expected value $1/(k\lambda)$. See Example 5.28.

Since all n components last until Y_1, $n-1$ last an additional $Y_2 - Y_1$, $n-2$ an additional $Y_3 - Y_2$ amount of time, and so on, another expression for T_r is

$$T_r = nY_1 + (n-1)(Y_2 - Y_1) + (n-2)(Y_3 - Y_2) + \cdots + (n-r+1)(Y_r - Y_{r-1})$$

But Y_1 is the minimum of n exponential variables, so $E(Y_1) = 1/(n\lambda)$. Similarly, $Y_2 - Y_1$ is the smallest of the $n-1$ remaining lifetimes, each exponential with parameter λ (by the memoryless property), so $E(Y_2 - Y_1) = 1/[(n-1)\lambda]$. Continuing, $E(Y_{i+1} - Y_i) = 1/[(n-i)\lambda]$, so

$$\begin{aligned} E(T_r) &= nE(Y_1) + (n-1)E(Y_2 - Y_1) + \cdots + (n-r+1)E(Y_r - Y_{r-1}) \\ &= n \cdot \frac{1}{n\lambda} + (n-1) \cdot \frac{1}{(n-1)\lambda} + \cdots + (n-r+1) \cdot \frac{1}{(n-r+1)\lambda} = \frac{r}{\lambda} \end{aligned}$$

Therefore, $E(T_r/r) = (1/r)E(T_r) = (1/r) \cdot (r/\lambda) = 1/\lambda = \mu$ as claimed.

As an example, suppose 20 components are put on test and $r = 10$. Then if the first ten failure times are 11, 15, 29, 33, 35, 40, 47, 55, 58, and 72, the estimate of μ is

$$\hat{\mu} = \frac{11 + 15 + \cdots + 72 + (10)(72)}{10} = 111.5$$

The advantage of the experiment with censoring is that it terminates more quickly than the uncensored experiment. However, it can be shown that $V(T_r/r) = 1/(\lambda^2 r)$, which is larger than $1/(\lambda^2 n)$, the variance of $\overline{X}$ in the uncensored experiment. ■

Reporting a Point Estimate: The Standard Error

Besides reporting the value of a point estimate, some indication of its precision should be given. The usual measure of precision is the standard error of the estimator used.

DEFINITION

The **standard error** of an estimator $\hat{\theta}$ is its standard deviation $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$. If the standard error itself involves unknown parameters whose values can be estimated, substitution of these estimates into $\sigma_{\hat{\theta}}$ yields the **estimated standard error** (estimated standard deviation) of the estimator. The estimated standard error can be denoted either by $\hat{\sigma}_{\hat{\theta}}$ (the ^ over σ emphasizes that $\sigma_{\hat{\theta}}$ is being estimated) or by $s_{\hat{\theta}}$.

Example 7.10
(Example 7.2 continued)

Assuming that breakdown voltage is normally distributed, $\hat{\mu} = \overline{X}$ is the best estimator of μ. If the value of σ is known to be 1.5, the standard error of $\overline{X}$ is $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 1.5/\sqrt{20} = .335$. If, as is usually the case, the value of σ is unknown, the estimate $\hat{\sigma} = s = 1.462$ is substituted into $\sigma_{\overline{X}}$ to obtain the estimated standard error $\hat{\sigma}_{\overline{X}} = s_{\overline{X}} = s/\sqrt{n} = 1.462/\sqrt{20} = .327$ ■

Example 7.11 (Example 7.1 continued) The standard error of $\hat{p} = X/n$ is

$$\sigma_{\hat{p}} = \sqrt{V(X/n)} = \sqrt{\frac{V(X)}{n^2}} = \sqrt{\frac{npq}{n^2}} = \sqrt{\frac{pq}{n}}$$

Since p and $q = 1 - p$ are unknown (else why estimate?), we substitute $\hat{p} = x/n$ and $\hat{q} = 1 - x/n$ into $\sigma_{\hat{p}}$, yielding the estimated standard error $\hat{\sigma}_{\hat{p}} = \sqrt{\hat{p}\hat{q}/n} = \sqrt{(.6)(.4)/25} = .098$. Alternatively, since the largest value of pq is attained when $p = q = .5$, an upper bound on the standard error is $\sqrt{1/(4n)} = .10$. ■

When the point estimator $\hat{\theta}$ has approximately a normal distribution, which will often be the case when n is large, then we can be reasonably confident that the true value of θ lies within approximately 2 standard errors (standard deviations) of $\hat{\theta}$. Thus if measurement of prothrombin (a blood-clotting protein) in 36 individuals gives $\hat{\mu} = \bar{x} = 20.5$ and $s = 3.6$ mg/100 ml, then $s/\sqrt{n} = .60$, so "within 2 estimated standard errors of $\hat{\mu}$" translates to the interval 20.50 ± (2)(.60) = (19.30, 21.70).

If $\hat{\theta}$ is not necessarily approximately normal but is unbiased, then it can be shown (using Chebyshev's inequality, introduced in Exercises 43, 77, and 135 of Chapter 3) that the estimate will deviate from θ by as much as 4 standard errors at most 6% of the time. We would then expect the true value to lie within 4 standard errors of $\hat{\theta}$ (and this is a very conservative statement, since it applies to *any* unbiased $\hat{\theta}$). Summarizing, the standard error tells us roughly within what distance of $\hat{\theta}$ we can expect the true value of θ to lie.

The Bootstrap

The form of the estimator $\hat{\theta}$ may be sufficiently complicated so that standard statistical theory cannot be applied to obtain an expression for $\sigma_{\hat{\theta}}$. This is true, for example, in the case $\theta = \sigma$, $\hat{\theta} = S$; the standard deviation of the statistic S, σ_S, cannot in general be determined. In recent years, a new computer-intensive method called the **bootstrap** has been introduced to address this problem. Suppose that the population pdf is $f(x; \theta)$, a member of a particular parametric family, and that data $x_1, x_2, \ldots, x_n$ gives $\hat{\theta} = 21.7$. We now use the computer to obtain "bootstrap samples" from the pdf $f(x; 21.7)$, and for each sample we calculate a "bootstrap estimate" $\hat{\theta}^*$:

First bootstrap sample: $x_1^*, x_2^*, \ldots, x_n^*$; estimate $= \hat{\theta}_1^*$

Second bootstrap sample: $x_1^*, x_2^*, \ldots, x_n^*$; estimate $= \hat{\theta}_2^*$

⋮

Bth bootstrap sample: $x_1^*, x_2^*, \ldots, x_n^*$; estimate $= \hat{\theta}_B^*$

$B = 100$ or 200 is often used. Now let $\bar{\theta}^* = \sum \hat{\theta}_i^*/B$, the sample mean of the bootstrap estimates. The **bootstrap estimate** of $\hat{\theta}$'s standard error is now just the sample standard deviation of the $\hat{\theta}_i^*$'s:

$$S_{\hat{\theta}} = \sqrt{\frac{1}{B-1} \sum \left(\hat{\theta}_i^* - \bar{\theta}^*\right)^2}$$

(In the bootstrap literature, B is often used in place of $B - 1$; for typical values of B, there is usually little difference between the resulting estimates.)

Example 7.12 A theoretical model suggests that X, the time to breakdown of an insulating fluid between electrodes at a particular voltage, has $f(x;\,\lambda) = \lambda e^{-\lambda x}$, an exponential distribution. A random sample of $n = 10$ breakdown times (min) gives the following data:

41.53	18.73	2.99	30.34	12.33	117.52	73.02	223.63	4.00	26.78

Since $E(X) = 1/\lambda$, $E(\overline{X}) = 1/\lambda$, so a reasonable estimate of λ is $\hat{\lambda} = 1/\bar{x} = 1/55.087 = .018153$. We then used a statistical computer package to obtain $B = 100$ bootstrap samples, each of size 10, from $f(x;\,.018153)$. The first such sample was 41.00, 109.70, 16.78, 6.31, 6.76, 5.62, 60.96, 78.81, 192.25, 27.61, from which $\sum x_i^* = 545.8$ and $\hat{\lambda}_1^* = 1/54.58 = .01832$. The average of the 100 bootstrap estimates is $\overline{\lambda}^* = .02153$, and the sample standard deviation of these 100 estimates is $s_{\hat{\lambda}} = .0091$, the bootstrap estimate of $\hat{\lambda}$'s standard error. A histogram of the 100 $\hat{\lambda}_i^*$'s was somewhat positively skewed, suggesting that the sampling distribution of $\hat{\lambda}$ also has this property. ■

Sometimes an investigator wishes to estimate a population characteristic without assuming that the population distribution belongs to a particular parametric family. An instance of this occurred in Example 7.8, where a 10% trimmed mean was proposed for estimating a symmetric population distribution's center θ. The data of Example 7.2 gave $\hat{\theta} = \overline{X}_{\text{tr}(10)} = 27.838$, but now there is no assumed $f(x;\,\theta)$, so how can we obtain a bootstrap sample? The answer is to regard the sample itself as constituting the population (the $n = 20$ observations in Example 7.2) and take B different samples, each of size n, *with* replacement from this population. We expand on this idea in Section 8.5.

Exercises Section 7.1 (1–20)

1. The accompanying data on IQ for first-graders at a university lab school was introduced in Example 1.2.

82	96	99	102	103	103	106	107	108	108	108
108	109	110	110	111	113	113	113	113	115	115
118	118	119	121	122	122	127	132	136	140	146

a. Calculate a point estimate of the mean value of IQ for the conceptual population of all first graders in this school, and state which estimator you used. [*Hint:* $\Sigma x_i = 3753$]

b. Calculate a point estimate of the IQ value that separates the lowest 50% of all such students from the highest 50%, and state which estimator you used.

c. Calculate and interpret a point estimate of the population standard deviation σ. Which estimator did you use? [*Hint*: $\Sigma x_i^2 = 432{,}015$]

d. Calculate a point estimate of the proportion of all such students whose IQ exceeds 100. [*Hint:* Think of an observation as a "success" if it exceeds 100.]

e. Calculate a point estimate of the population coefficient of variation σ/μ, and state which estimator you used.

2. A sample of 20 students who had recently taken elementary statistics yielded the following information on brand of calculator owned (T = Texas Instruments, H = Hewlett-Packard, C = Casio, S = Sharp):

T	T	H	T	C	T	T	S	C	H
S	S	T	H	C	T	T	T	H	T

a. Estimate the true proportion of all such students who own a Texas Instruments calculator.

b. Of the ten students who owned a TI calculator, 4 had graphing calculators. Estimate the proportion of students who do not own a TI graphing calculator.

3. Consider the following sample of observations on coating thickness for low-viscosity paint ("Achieving a Target Value for a Manufacturing Process: A Case Study," *J. Qual. Technol.*, 1992: 22–26):

.83	.88	.88	1.04	1.09	1.12	1.29	1.31
1.48	1.49	1.59	1.62	1.65	1.71	1.76	1.83

Assume that the distribution of coating thickness is normal (a normal probability plot strongly supports this assumption).

a. Calculate a point estimate of the mean value of coating thickness, and state which estimator you used.

b. Calculate a point estimate of the median of the coating thickness distribution, and state which estimator you used.

c. Calculate a point estimate of the value that separates the largest 10% of all values in the thickness distribution from the remaining 90%, and state which estimator you used. [*Hint:* Express what you are trying to estimate in terms of μ and σ]

d. Estimate $P(X < 1.5)$, i.e., the proportion of all thickness values less than 1.5. [*Hint:* If you knew the values of μ and σ, you could calculate this probability. These values are not available, but they can be estimated.]

e. What is the estimated standard error of the estimator that you used in part (b)?

4. The data set mentioned in Exercise 1 also includes these third grade verbal IQ observations for males:

117	103	121	112	120	132	113	117	132
149	125	131	136	107	108	113	136	114

and females

114	102	113	131	124	117	120	90
114	109	102	114	127	127	103	

Prior to obtaining data, denote the male values by $X_1, \ldots, X_m$ and the female values by $Y_1, \ldots, Y_n$. Suppose that the X_i's constitute a random sample from a distribution with mean μ_1 and standard deviation σ_1 and that the Y_i's form a random sample (independent of the X_i's) from another distribution with mean μ_2 and standard deviation σ_2.

a. Use rules of expected value to show that $\overline{X} - \overline{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$. Calculate the estimate for the given data.

b. Use rules of variance from Chapter 6 to obtain an expression for the variance and standard deviation (standard error) of the estimator in part (a), and then compute the estimated standard error.

c. Calculate a point estimate of the ratio σ_1/σ_2 of the two standard deviations.

d. Suppose one male third-grader and one female third-grader are randomly selected. Calculate a point estimate of the variance of the difference $X - Y$ between male and female IQ.

5. As an example of a situation in which several different statistics could reasonably be used to calculate a point estimate, consider a population of N invoices. Associated with each invoice is its "book value," the recorded amount of that invoice. Let T denote the total book value, a known amount. Some of these book values are erroneous. An audit will be carried out by randomly selecting n invoices and determining the audited (correct) value for each one. Suppose that the sample gives the following results (in dollars).

	Invoice				
	1	**2**	**3**	**4**	**5**
Book value	300	720	526	200	127
Audited value	300	520	526	200	157
Error	0	200	0	0	−30

Let $\overline{X}$ = the sample mean audited value, $\overline{Y}$= the sample mean book value, and $\overline{D}$= the sample mean error. Propose three different statistics for estimating the total audited (i.e. correct) value θ — one involving just N and $\overline{X}$, another involving N, T, and $\overline{D}$, and the last involving T and $\overline{X}/\overline{Y}$. Then calculate the resulting estimates when $N = 5{,}000$ and $T = 1{,}761{,}300$ (The article "Statistical Models and Analysis in Auditing,", *Statistical Science*, 1989: 2 – 33 discusses properties of these estimators).

6. Consider the accompanying observations on stream flow (1000's of acre-feet) recorded at a station in Colorado for the period April 1–August 31 over a 31-year span (from an article in the 1974 volume of *Water Resources Res.*).

127.96	210.07	203.24	108.91	178.21
285.37	100.85	89.59	185.36	126.94
200.19	66.24	247.11	299.87	109.64
125.86	114.79	109.11	330.33	85.54
117.64	302.74	280.55	145.11	95.36
204.91	311.13	150.58	262.09	477.08
94.33				

An appropriate probability plot supports the use of the lognormal distribution (see Section 4.5) as a reasonable model for stream flow.

a. Estimate the parameters of the distribution. [*Hint*: Remember that X has a lognormal distribution with parameters μ and σ^2 if $\ln(X)$ is normally distributed with mean μ and variance σ^2.]

b. Use the estimates of part (a) to calculate an estimate of the expected value of stream flow. [*Hint*: What is $E(X)$?]

7. a. A random sample of 10 houses in a particular area, each of which is heated with natural gas, is selected and the amount of gas (therms) used during the month of January is determined for each house. The resulting observations are 103, 156, 118, 89, 125, 147, 122, 109, 138, 99. Let μ denote the average gas usage during January by all houses in this area. Compute a point estimate of μ.

b. Suppose there are 10,000 houses in this area that use natural gas for heating. Let τ denote the total amount of gas used by all of these houses during January. Estimate τ using the data of part (a). What estimator did you use in computing your estimate?

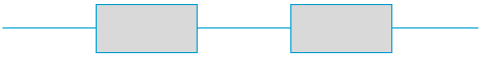
c. Use the data in part (a) to estimate p, the proportion of all houses that used at least 100 therms.

d. Give a point estimate of the population median usage (the middle value in the population of all houses) based on the sample of part (a). What estimator did you use?

8. In a random sample of 80 components of a certain type, 12 are found to be defective.

a. Give a point estimate of the proportion of all such components that are *not* defective.

b. A system is to be constructed by randomly selecting two of these components and connecting them in series, as shown here.



The series connection implies that the system will function if and only if neither component is defective (i.e., both components work properly). Estimate the proportion of all such systems that work properly. [*Hint*: If p denotes the probability that a component works properly, how can P(system works) be expressed in terms of p?]

c. Let $\hat{p}$ be the sample proportion of successes. Is $\hat{p}^2$ an unbiased estimator for p^2? [*Hint*: For any rv Y, $E(Y^2) = V(Y) + [E(Y)]^2$.]

9. Each of 150 newly manufactured items is examined and the number of scratches per item is recorded (the items are supposed to be free of scratches), yielding the following data:

Number of scratches per item	0	1	2	3	4	5	6	7
Observed frequency	18	37	42	30	13	7	2	1

Let X = the number of scratches on a randomly chosen item, and assume that X has a Poisson distribution with parameter λ.

a. Find an unbiased estimator of λ and compute the estimate for the data. [*Hint*: $E(X) = \lambda$ for X Poisson, so $E(\overline{X} = ?)$]

b. What is the standard deviation (standard error) of your estimator? Compute the estimated standard error. [*Hint*: $\sigma_X^2 = \lambda$ for X Poisson.]

10. Using a long rod that has length μ, you are going to lay out a square plot in which the length of each side is μ. Thus the area of the plot will be μ^2. However, you do not know the value of μ, so you decide to make n independent measurements $X_1, X_2, \ldots X_n$ of the length. Assume that each X_i has mean μ (unbiased measurements) and variance σ^2.

a. Show that $\overline{X}^2$ is not an unbiased estimator for μ^2. [*Hint*: For any rv Y, $E(Y^2) = V(Y) + [E(Y)]^2$. Apply this with $Y = \overline{X}$.]

b. For what value of k is the estimator $\overline{X}^2 - kS^2$ unbiased for μ^2? [*Hint*: Compute $E(\overline{X}^2 - kS^2)$.]

11. Of n_1 randomly selected male smokers, X_1 smoked filter cigarettes, whereas of n_2 randomly selected female smokers, X_2 smoked filter cigarettes. Let p_1 and p_2 denote the probabilities that a randomly selected male and female, respectively, smoke filter cigarettes.

a. Show that $(X_1/n_1) - (X_2/n_2)$ is an unbiased estimator for $p_1 - p_2$. [*Hint*: $E(X_i) = n_i p_i$ for $i = 1, 2$.]

b. What is the standard error of the estimator in part (a)?

c. How would you use the observed values x_1 and x_2 to estimate the standard error of your estimator?

d. If $n_1 = n_2 = 200$, $x_1 = 127$, and $x_2 = 176$, use the estimator of part (a) to obtain an estimate of $p_1 - p_2$.

e. Use the result of part (c) and the data of part (d) to estimate the standard error of the estimator.

12. Suppose a certain type of fertilizer has an expected yield per acre of μ_1 with variance σ^2, whereas the expected yield for a second type of fertilizer is μ_2 with the same variance σ^2. Let S_1^2 and S_2^2 denote the sample variances of yields based on sample sizes n_1 and n_2, respectively, of the two fertilizers. Show that the pooled (combined) estimator

$$\hat{\sigma}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

is an unbiased estimator of σ^2.

13. Consider a random sample $X_1, \ldots, X_n$ from the pdf

$$f(x; \theta) = .5(1 + \theta x) \qquad -1 \le x \le 1$$

where $-1 \le \theta \le 1$ (this distribution arises in particle physics). Show that $\hat{\theta} = 3\overline{X}$ is an unbiased estimator of θ. [*Hint*: First determine $\mu = E(X) = E(\overline{X})$.]

14. A sample of n captured Pandemonium jet fighters results in serial numbers $x_1, x_2, x_3, \ldots, x_n$. The CIA knows that the aircraft were numbered consecutively at the factory starting with α and ending with β, so that the total number of planes manufactured is $\beta - \alpha + 1$ (e.g., if $\alpha = 17$ and $\beta = 29$, then $29-17 + 1 = 13$ planes having serial numbers 17, 18, 19, . . ., 28, 29 were manufactured). However, the CIA does not know the values of α or β. A CIA statistician suggests using the estimator $\max(X_i) - \min(X_i) + 1$ to estimate the total number of planes manufactured.

a. If $n = 5$, $x_1 = 237$, $x_2 = 375$, $x_3 = 202$, $x_4 = 525$, and $x_5 = 418$, what is the corresponding estimate?

b. Under what conditions on the sample will the value of the estimate be exactly equal to the true total number of planes? Will the estimate ever be larger than the true total? Do you think the estimator is unbiased for estimating $\beta - \alpha + 1$? Explain in one or two sentences.

(A similar method was used to estimate German tank production in World War II.)

15. Let $X_1, X_2, \ldots, X_n$ represent a random sample from a Rayleigh distribution with pdf

$$f(x;\theta) = \frac{x}{\theta} e^{-x^2/(2\theta)} \qquad x>0$$

a. It can be shown that $E(X^2) = 2\theta$. Use this fact to construct an unbiased estimator of θ based on $\sum X_i^2$ (and use rules of expected value to show that it is unbiased).

b. Estimate θ from the following measurements of blood plasma beta concentration (in pmol/L) for $n = 10$ men.

16.88	10.23	4.59	6.66	13.68
14.23	19.87	9.40	6.51	10.95

16. Suppose the true average growth μ of one type of plant during a 1-year period is identical to that of a second type, but the variance of growth for the first type is σ^2, whereas for the second type, the variance is $4\sigma^2$. Let $X_1, \ldots, X_m$ be m independent growth observations on the first type [so $E(X_i) = \mu$, $V(X_i) = \sigma^2$], and let $Y_1, \ldots, Y_n$ be n independent growth observations on the second type [$E(Y_i) = \mu$, $V(Y_i) = 4\sigma^2$]. Let c be a numerical constant and consider the estimator $\hat{\mu} = c\overline{X} + (1-c)\overline{Y}$. For any c between 0 and 1 this is a weighted average of the two sample means, e.g., $.7\overline{X} + .3\overline{Y}$

a. Show that for any c the estimator is unbiased.

b. For fixed m and n, what value c minimizes $V(\hat{\mu})$? [*Hint*: The estimator is a linear combination of the two sample means and these means are independent. Once you have an expression for the variance, differentiate with respect to c.]

17. In Chapter 3, we defined a negative binomial rv as the number of failures that occur before the rth success in a sequence of independent and identical success/failure trials. The probability mass function (pmf) of X is

$$nb(x,r,p) = \begin{cases} \binom{x+r-1}{x} p^r (1-p)^x & x = 0,1,2,\ldots \\ 0 & \text{otherwise} \end{cases}$$

a. Suppose that $r \ge 2$. Show that

$$\hat{p} = (r-1)/(X+r-1)$$

is an unbiased estimator for p. [*Hint*: Write out $E(\hat{p})$ and cancel $x + r - 1$ inside the sum.]

b. A reporter wishing to interview five individuals who support a certain candidate begins asking people whether (S) or not (F) they support the candidate. If the sequence of responses is *SFFSFFFSSS*, estimate p = the true proportion who support the candidate.

18. Let $X_1, X_2, \ldots, X_n$ be a random sample from a pdf $f(x)$ that is symmetric about μ, so that $\widetilde{X}$ is an unbiased estimator of μ. If n is large, it can be shown that $V(\widetilde{X}) \approx 1/\{4n[f(\mu)]^2\}$. When the underlying pdf is Cauchy (see Example 7.8), $V(\overline{X}) = \infty$, so $\overline{X}$ is a terrible estimator. What is $V(\widetilde{X})$ in this case when n is large?

19. An investigator wishes to estimate the proportion of students at a certain university who have violated the honor code. Having obtained a random sample of n students, she realizes that asking each, "Have you violated the honor code?" will probably result in some untruthful responses. Consider the following scheme, called a **randomized response** technique. The investigator makes up a deck of 100 cards, of which 50 are of type I and 50 are of type II.

Type I: Have you violated the honor code (yes or no)?

Type II: Is the last digit of your telephone number a 0, 1, or 2 (yes or no)?

Each student in the random sample is asked to mix the deck, draw a card, and answer the resulting question truthfully. Because of the irrelevant question on type II cards, a yes response no longer stigmatizes the respondent, so we assume that responses are truthful. Let p denote the proportion of honor-code violators (i.e., the probability of a randomly selected student being a violator), and let $\lambda = P(\text{yes response})$. Then λ and p are related by $\lambda = .5p + (.5)(.3)$.

a. Let Y denote the number of yes responses, so $Y \sim \text{Bin}(n, \lambda)$. Thus Y/n is an unbiased estimator of λ. Derive an estimator for p based on Y. If $n = 80$ and $y = 20$, what is your estimate? [*Hint*: Solve $\lambda = .5p + .15$ for p and then substitute Y/n for λ.]

b. Use the fact that $E(Y/n) = \lambda$ to show that your estimator $\hat{p}$ is unbiased.

c. If there were 70 type I and 30 type II cards, what would be your estimator for p?

20. Return to the problem of estimating the population proportion p and consider another adjusted estimator, namely

$$\hat{p} = \frac{X + \sqrt{n/4}}{n + \sqrt{n}}$$

The justification for this estimator comes from the Bayesian approach to point estimation to be introduced in Section 14.4.

a. Determine the mean squared error of this estimator. What do you find interesting about this MSE?

b. Compare the MSE of this estimator to the MSE of the usual estimator (the sample proportion).

7.2 Methods of Point Estimation

So far the point estimators we have introduced were obtained via intuition and/or educated guesswork. We now discuss two "constructive" methods for obtaining point estimators: the method of moments and the method of maximum likelihood. By constructive we mean that the general definition of each type of estimator suggests explicitly how to obtain the estimator in any specific problem. Although maximum likelihood estimators are generally preferable to moment estimators because of certain efficiency properties, they often require significantly more computation than do moment estimators. It is sometimes the case that these methods yield unbiased estimators.

The Method of Moments

The basic idea of this method is to equate certain sample characteristics, such as the mean, to the corresponding population expected values. Then solving these equations for unknown parameter values yields the estimators.

DEFINITION

Let $X_1, \ldots, X_n$ be a random sample from a pmf or pdf $f(x)$. For $k = 1, 2, 3, \ldots$, the ***k*th population moment**, or ***k*th moment of the distribution** $f(x)$, is $E(X^k)$. The ***k*th sample moment** is $(1/n)\sum_{i=1}^{n} X_i^k$.

Thus the first population moment is $E(X) = \mu$ and the first sample moment is $\sum X_i/n = \overline{X}$. The second population and sample moments are $E(X^2)$ and $\sum X_i^2/n$, respectively. The population moments will be functions of any unknown parameters $\theta_1, \theta_2, \ldots$.

DEFINITION

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pmf or pdf $f(x; \theta_1, \ldots, \theta_m)$, where $\theta_1, \ldots, \theta_m$ are parameters whose values are unknown. Then the **moment estimators** $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are obtained by equating the first m sample moments to the corresponding first m population moments and solving for $\theta_1, \ldots, \theta_m$.

If, for example, $m = 2$, $E(X)$ and $E(X^2)$ will be functions of θ_1 and θ_2. Setting $E(X) = (1/n)\sum X_i \ \ (= \overline{X})$ and $E(X^2) = (1/n)\sum X_i^2$ gives two equations in θ_1 and θ_2. The solution then defines the estimators. For estimating a population mean μ, the method gives $\mu = \overline{X}$, so the estimator is the sample mean.

Example 7.13

Let $X_1, \ldots, X_n$ represent a random sample of service times of n customers at a certain facility, where the underlying distribution is assumed exponential with parameter λ. Since there is only one parameter to be estimated, the estimator is obtained by equating $E(X)$ to $\overline{X}$. Since $E(X) = 1/\lambda$ for an exponential distribution, this gives $1/\lambda = \overline{X}$ or $\lambda = 1/\overline{X}$. The moment estimator of λ is then $\hat{\lambda} = 1/\overline{X}$. ■

Example 7.14

Let $X_1, \ldots, X_n$ be a random sample from a gamma distribution with parameters α and β. From Section 4.4 , $E(X) = \alpha\beta$ and $E(X^2) = \beta^2\Gamma(\alpha + 2)/\Gamma(\alpha) = \beta^2(\alpha + 1)\alpha$. The moment estimators of α and β are obtained by solving

$$\overline{X} = \alpha\beta \qquad\qquad \frac{1}{n}\sum X_i^2 = \alpha(\alpha+1)\beta^2$$

Since $\alpha(\alpha+1)\beta^2 = \alpha^2\beta^2 + \alpha\beta^2$ and the first equation implies $\alpha^2\beta^2 = (\overline{X})^2$, the second equation becomes

$$\frac{1}{n}\sum X_i^2 = (\overline{X})^2 + \alpha\beta^2$$

Now dividing each side of this second equation by the corresponding side of the first equation and substituting back gives the estimators

$$\hat{\alpha} = \frac{(\overline{X})^2}{\frac{1}{n}\sum X_i^2 - (\overline{X})^2} \qquad\qquad \hat{\beta} = \frac{\frac{1}{n}\sum X_i^2 - (\overline{X})^2}{\overline{X}}$$

To illustrate, the survival time data mentioned in Example 4.28 is

152	115	109	94	88	137	152	77	160	165
125	40	128	123	136	101	62	153	83	69

with $\bar{x} = 113.5$ and $(1/20)\sum x_i^2 = 14{,}087.8$. The estimates are

$$\hat{\alpha} = \frac{(113.5)^2}{14{,}087.8 - (113.5)^2} = 10.7 \qquad\qquad \hat{\beta} = \frac{14{,}087.8 - (113.5)^2}{113.5} = 10.6$$

These estimates of α and β differ from the values suggested by Gross and Clark because they used a different estimation technique. ■

Example 7.15 Let $X_1, \ldots, X_n$ be a random sample from a generalized negative binomial distribution with parameters r and p (Section 3.6). Since $E(X) = r(1-p)/p$ and $V(X) = r(1-p)/p^2$, $E(X^2) = V(X) + [E(X)]^2 = r(1-p)(r - rp + 1)/p^2$. Equating $E(X)$ to $\overline{X}$ and $E(X^2)$ to $(1/n)\sum X_i^2$ eventually gives

$$\hat{p} = \frac{\overline{X}}{\frac{1}{n}\sum X_i^2 - (\overline{X})^2} \qquad \hat{r} = \frac{(\overline{X})^2}{\frac{1}{n}\sum X_i^2 - (\overline{X})^2 - \overline{X}}$$

As an illustration, Reep, Pollard, and Benjamin ("Skill and Chance in Ball Games," J. Roy. Statist. Soc. Ser. A, 1971: 623–629) consider the negative binomial distribution as a model for the number of goals per game scored by National Hockey League teams. The data for 1966–1967 follows (420 games):

Goals	0	1	2	3	4	5	6	7	8	9	10
Frequency	29	71	82	89	65	45	24	7	4	1	3

Then,

$$\bar{x} = \sum x_i/420 = [(0)(29) + (1)(71) + \cdots + (10)(3)]/420 = 2.98$$

and

$$\sum x_i^2/420 = [(0)^2(29) + (1)^2(71) + \cdots + (10)^2(3)]/420 = 12.40$$

Thus,

$$\hat{p} = \frac{2.98}{12.40 - (2.98)^2} = .85 \qquad \hat{r} = \frac{(2.98)^2}{12.40 - (2.98)^2 - 2.98} = 16.5$$

Although r by definition must be positive, the denominator of $\hat{r}$ could be negative, indicating that the negative binomial distribution is not appropriate (or that the moment estimator is flawed). ■

Maximum Likelihood Estimation

The method of maximum likelihood was first introduced by R. A. Fisher, a geneticist and statistician, in the 1920s. Most statisticians recommend this method, at least when the sample size is large, since the resulting estimators have certain desirable efficiency properties (see the proposition on large sample behavior toward the end of this section).

Example 7.16 A sample of ten new bike helmets manufactured by a company is obtained. Upon testing, it is found that the first, third, and tenth helmets are flawed, whereas the others are not. Let $p = P$(flawed helmet) and define $X_1, \ldots, X_{10}$ by $X_i = 1$ if the ith helmet is flawed and zero otherwise. Then the observed x_i's are 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, so the joint pmf of the sample is

$$f(x_1, x_2, \ldots, x_{10}; p) = p(1-p)p \cdot \cdots \cdot p = p^3(1-p)^7 \tag{7.4}$$

We now ask, "For what value of p is the observed sample most likely to have occurred?" That is, we wish to find the value of p that maximizes the pmf (7.4) or, equivalently, maximizes the natural log of (7.4).[2] Since

$$\ln[f(x_1, x_2, \ldots, x_{10}; p)] = 3\ln(p) + 7\ln(1-p) \tag{7.5}$$

and this is a differentiable function of p, equating the derivative of (7.5) to zero gives the maximizing value[3]:

$$\frac{d}{dp}\ln[f(x_1, x_2, \ldots, x_{10}; p)] = \frac{3}{p} - \frac{7}{1-p} = 0 \Rightarrow p = \frac{3}{10} = \frac{x}{n}$$

where x is the observed number of successes (flawed helmets). The estimate of p is now $\hat{p} = \frac{3}{10}$. It is called the maximum likelihood estimate because for fixed x_1, . . ., x_{10}, it is the parameter value that maximizes the likelihood (joint pmf) of the observed sample. The likelihood and log likelihood are graphed in Figure 7.5. Of course, the maximum on both graphs occurs at the same value, $p = .3$.

Note that if we had been told only that among the ten helmets there were three that were flawed, Equation (7.4) would be replaced by the binomial pmf $\binom{10}{3}p^3(1-p)^7$, which is also maximized for $\hat{p} = \dfrac{3}{10}$.

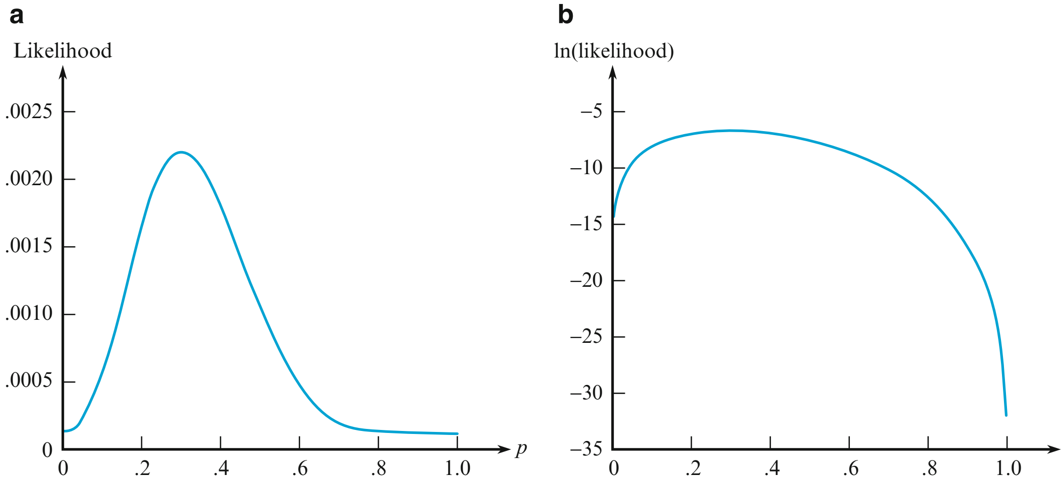


Figure 7.5 Likelihood and log likelihood plotted against p ■

[2] Since $\ln[g(x)]$ is a monotonic function of $g(x)$, finding x to maximize $\ln[g(x)]$ is equivalent to maximizing $g(x)$ itself. In statistics, taking the logarithm frequently changes a product to a sum, which is easier to work with.

[3] This conclusion requires checking the second derivative, but the details are omitted.

DEFINITION

Let $X_1, \ldots, X_n$ have joint pmf or pdf

$$f(x_1, x_2, ..., x_n; \theta_1, ..., \theta_m) \tag{7.6}$$

where the parameters $\theta_1, \ldots, \theta_m$ have unknown values. When $x_1, \ldots, x_n$ are the observed sample values and (7.6) is regarded as a function of $\theta_1, \ldots, \theta_m$, it is called the **likelihood function**. The maximum likelihood estimates $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are those values of the θ_i's that maximize the likelihood function, so that

$$f(x_1, x_2, \ldots, x_n; \hat{\theta}_1, \ldots, \hat{\theta}_m) \geq f(x_1, x_2, \ldots, x_n; \theta_1, \ldots, \theta_m) \text{ for all } \theta_1, \ldots, \theta_m$$

When the X_i's are substituted in place of the x_i's, the **maximum likelihood estimators** (mle's) result.

The likelihood function tells us how likely the observed sample is as a function of the possible parameter values. Maximizing the likelihood gives the parameter values for which the observed sample is most likely to have been generated, that is, the parameter values that "agree most closely" with the observed data.

Example 7.17 Suppose $X_1, \ldots, X_n$ is a random sample from an exponential distribution with parameter λ. Because of independence, the likelihood function is a product of the individual pdf's:

$$f(x_1, \ldots, x_n; \lambda) = (\lambda e^{-\lambda x_1}) \cdot \cdots \cdot (\lambda e^{-\lambda x_n}) = \lambda^n e^{-\lambda \Sigma x_i}$$

The ln(likelihood) is

$$\ln[f(x_1, \ldots, x_n; \lambda)] = n \ln(\lambda) - \lambda \sum x_i$$

Equating $(d/d\lambda)$[ln(likelihood)] to zero results in $n/\lambda - \Sigma x_i = 0$, or $\lambda = n/\Sigma x_i = 1/\bar{x}$. Thus the mle is $\hat{\lambda} = 1/\overline{X}$; it is identical to the method of moments estimator but it is not an unbiased estimator, since $E(1/\overline{X}) \neq 1/E(\overline{X})$. ■

Example 7.18 Let $X_1, \ldots, X_n$ be a random sample from a normal distribution. The likelihood function is

$$\begin{aligned} f(x_1, \ldots, x_n; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/(2\sigma^2)} \cdot \cdots \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/(2\sigma^2)} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\Sigma(x_i-\mu)^2/(2\sigma^2)} \end{aligned}$$

so

$$\ln[f(x_1, \ldots, x_n; \mu, \sigma^2)] = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2$$

To find the maximizing values of μ and σ^2, we must take the partial derivatives of $\ln(f)$ with respect to μ and σ^2, equate them to zero, and solve the resulting two equations. Omitting the details, the resulting mle's are

$$\hat{\mu} = \overline{X} \qquad \hat{\sigma}^2 = \frac{\sum (X_i - \overline{X})^2}{n}$$

The mle of σ^2 is not the unbiased estimator, so two different principles of estimation (unbiasedness and maximum likelihood) yield two different estimators. ■

Example 7.19 In Chapter 3, we discussed the use of the Poisson distribution for modeling the number of "events" that occur in a two-dimensional region. Assume that when the region R being sampled has area $a(R)$, the number X of events occurring in R has a Poisson distribution with parameter $\lambda a(R)$ (where λ is the expected number of events per unit area) and that nonoverlapping regions yield independent X's.

Suppose an ecologist selects n nonoverlapping regions $R_1, \ldots, R_n$ and counts the number of plants of a certain species found in each region. The joint pmf (likelihood) is then

$$\begin{aligned} p(x_1, \ldots, x_n; \lambda) &= \frac{[\lambda \cdot a(R_1)]^{x_1} e^{-\lambda \cdot a(R_1)}}{x_1!} \cdot \cdots \cdot \frac{[\lambda \cdot a(R_n)]^{x_n} e^{-\lambda \cdot a(R_n)}}{x_n!} \\ &= \frac{[a(R_1)]^{x_1} \cdot \cdots \cdot [a(R_n)]^{x_n} \cdot \lambda^{\Sigma x_i} \cdot e^{-\lambda \Sigma a(R_i)}}{x_1! \cdot \cdots \cdot x_n!} \end{aligned}$$

The ln(likelihood) is

$$\ln[p(x_1, \ldots, x_n; \lambda)] = \sum x_i \cdot \ln[a(R_i)] + \ln(\lambda) \cdot \sum x_i - \lambda \sum a(R_i) - \sum \ln(x_i!)$$

Taking $d/d\lambda$ $\ln(p)$ and equating it to zero yields

$$\frac{\sum x_i}{\lambda} - \sum a(R_i) = 0$$

so

$$\lambda = \frac{\sum x_i}{\sum a(R_i)}$$

The mle is then $\hat{\lambda} = \sum X_i / \sum a(R_i)$. This is intuitively reasonable because λ is the true density (plants per unit area), whereas $\hat{\lambda}$ is the sample density since $\sum a(R_i)$ is just the total area sampled. Because $E(X_i) = \lambda \cdot a(R_i)$, the estimator is unbiased.

Sometimes an alternative sampling procedure is used. Instead of fixing regions to be sampled, the ecologist will select n points in the entire region of interest and let y_i = the distance from the ith point to the nearest plant. The cumulative distribution function (cdf) of Y = distance to the nearest plant is

$$\begin{aligned} F_Y(y) &= P(Y \le y) = 1 - P(Y > y) = 1 - P\begin{pmatrix} \text{no plants in a} \\ \text{circle of radius } y \end{pmatrix} \\ &= 1 - \frac{e^{-\lambda \pi y^2} (\lambda \pi y^2)^0}{0!} = 1 - e^{-\lambda \pi y^2} \end{aligned}$$

Taking the derivative of $F_Y(y)$ with respect to y yields

$$f_Y(y; \lambda) = \begin{cases} 2\pi\lambda y e^{-\lambda \pi y^2} & y \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

If we now form the likelihood $f_Y(y_1; \lambda) \cdot \cdots \cdot f_Y(y_n; \lambda)$, differentiate ln(likelihood), and so on, the resulting mle is

$$\hat{\lambda} = \frac{n}{\pi \sum Y_i^2} = \frac{\text{number of plants observed}}{\text{total area sampled}}$$

which is also a sample density. It can be shown that in a sparse environment (small λ), the distance method is in a certain sense better, whereas in a dense environment, the first sampling method is better. ■

Example 7.20 Let $X_1, \ldots, X_n$ be a random sample from a Weibull pdf

$$f(x; \alpha, \beta) = \begin{cases} \dfrac{\alpha}{\beta^\alpha} \cdot x^{\alpha-1} \cdot e^{-(x/\beta)^\alpha} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Writing the likelihood and ln(likelihood), then setting both $(\partial/\partial\alpha)[\ln(f)] = 0$ and $(\partial/\partial\beta)[\ln(f)] = 0$ yields the equations

$$\alpha = \left[\frac{\sum [x_i^\alpha \cdot \ln(x_i)]}{\sum x_i^\alpha} - \frac{\sum \ln(x_i)}{n}\right]^{-1} \qquad \beta = \left(\frac{\sum x_i^\alpha}{n}\right)^{1/\alpha}$$

These two equations cannot be solved explicitly to give general formulas for the mle's $\hat{\alpha}$ and $\hat{\beta}$. Instead, for each sample $x_1, \ldots, x_n$, the equations must be solved using an iterative numerical procedure. Even moment estimators of α and β are somewhat complicated (see Exercise 22).

The iterative mle computations can be done on a computer, and they are available in some statistical packages. MINITAB gives maximum likelihood estimates for both the Weibull and the gamma distributions (under "Quality Tools"). Stata has a general procedure that can be used for these and other distributions. For the data of Example 7.14 the maximum likelihood estimates for the Weibull distribution are $\hat{\alpha} = 3.799$ and $\hat{\beta} = 125.88$. (The mle's for the gamma distribution are $\hat{\alpha} = 8.799$ and $\hat{\beta} = 12.893$, a little different from the moment estimates in Example 7.14). Figure 7.6 shows the Weibull log likelihood as a function of α and β. The surface near the top has a rounded shape, allowing the maximum to be found easily, but for some distributions the surface can be much more irregular, and the maximum may be hard to find.

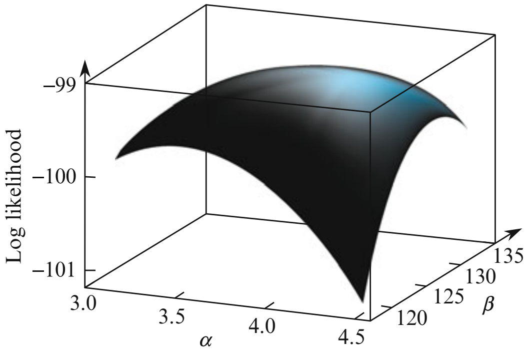


Figure 7.6 Weibull log likelihood for Example 7.20 ■

Some Properties of MLEs

In Example 7.18, we obtained the mle of σ^2 when the underlying distribution is normal. The mle of $\sigma = \sqrt{\sigma^2}$, as well as many other mle's, can be easily derived using the following proposition.

PROPOSITION

The Invariance Principle

Let $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ be the mle's of the parameters $\theta_1, \theta_2, \ldots, \theta_m$. Then the mle of any function $h(\theta_1, \theta_2, \ldots, \theta_m)$ of these parameters is the function $h(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$, of the mle's.

Proof For an intuitive idea of the proof, consider the special case $m = 1$, with $\theta_1 = \theta$, and assume that $h(\cdot)$ is a one-to-one function. On the graph of the likelihood as a function of the parameter θ, the highest point occurs where $\theta = \hat{\theta}$. Now consider the graph of the likelihood as a function of $h(\theta)$. In the new graph the same heights occur, but the height that was previously plotted at $\theta = a$ is now plotted at $h(\theta) = h(a)$, and the highest point is now plotted at $h(\theta) = h(\hat{\theta})$. Thus, the maximum remains the same, but it now occurs at $h(\hat{\theta})$. ■

Example 7.21
(Example 7.18 continued)

In the normal case, the mle's of μ and σ^2 are $\hat{\mu} = \overline{X}$ and $\hat{\sigma}^2 = \sum (X_i - \overline{X})^2/n$. To obtain the mle of the function $h(\mu, \sigma^2) = \sqrt{\sigma^2} = \sigma$, substitute the mle's into the function:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \left[\frac{1}{n}\sum (X_i - \overline{X})^2\right]^{1/2}$$

The mle of σ is not the sample standard deviation S, although they are close unless n is quite small. Similarly, the mle of the population coefficient of variation $100\mu/\sigma$ is $100\hat{\mu}/\hat{\sigma}$. ■

Example 7.22
(Example 7.20 continued)

The mean value of an rv X that has a Weibull distribution is

$$\mu = \beta \cdot \Gamma(1 + 1/\alpha)$$

The mle of μ is therefore $\hat{\mu} = \hat{\beta} \cdot \Gamma(1 + 1/\hat{\alpha})$, where $\hat{\alpha}$ and $\hat{\beta}$ are the mle's of α and β. In particular, $\overline{X}$ is not the mle of μ, although it is an unbiased estimator. At least for large n, $\hat{\mu}$ is a better estimator than $\overline{X}$. ■

Large-Sample Behavior of the MLE

Although the principle of maximum likelihood estimation has considerable intuitive appeal, the following proposition provides additional rationale for the use of mle's. (See Section 7.4 for more details.)

PROPOSITION

Under very general conditions on the joint distribution of the sample, when the sample size is large, the maximum likelihood estimator of any parameter θ is close to θ (consistency), is approximately unbiased $[E(\hat{\theta}) \approx \theta]$, and has variance that is nearly as small as can be achieved by any unbiased estimator. Stated another way, the mle $\hat{\theta}$ is approximately the MVUE of θ.

Because of this result and the fact that calculus-based techniques can usually be used to derive the mle's (although often numerical methods, such as Newton's method, are necessary), maximum likelihood estimation is the most widely used estimation technique among statisticians. Many of the estimators used in the remainder of the book are mle's. Obtaining an mle, however, does require that the underlying distribution be specified.

Note that there is no similar result for method of moments estimators. In general, if there is a choice between maximum likelihood and moment estimators, the mle is preferable. For example, the maximum likelihood method applied to estimating gamma distribution parameters tends to give better estimates (closer to the parameter values) than does the method of moments, so the extra computation is worth the price.

Some Complications

Sometimes calculus cannot be used to obtain mle's.

Example 7.23 Suppose the waiting time for a bus is uniformly distributed on $[0, \theta]$ and the results $x_1, \ldots, x_n$ of a random sample from this distribution have been observed. Since $f(x; \theta) = 1/\theta$ for $0 \le x \le \theta$ and 0 otherwise,

$$f(x_1, \ldots, x_n; \theta) = \begin{cases} 1/\theta^n & 0 \le x_1 \le \theta, \ldots, 0 \le x_n \le \theta \\ 0 & \text{otherwise} \end{cases}$$

As long as $\max(x_i) \le \theta$, the likelihood is $1/\theta^n$, which is positive, but as soon as $\theta < \max(x_i)$, the likelihood drops to 0. This is illustrated in Figure 7.7. Calculus will not work because the maximum of the likelihood occurs at a point of discontinuity, but the figure shows that $\hat{\theta} = \max(x_i)$. Thus if my waiting times are 2.3, 3.7, 1.5, .4, and 3.2, then the mle is $\hat{\theta} = 3.7$. Note that the mle is biased (see Example 7.5).

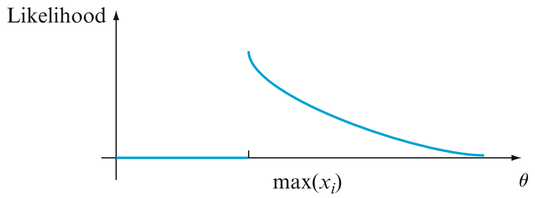


Figure 7.7 The likelihood function for Example 7.23 ■

Example 7.24 A method that is often used to estimate the size of a wildlife population involves performing a capture/recapture experiment. In this experiment, an initial sample of M animals is captured, each of these animals is tagged, and the animals are then returned to the population. After allowing enough time for the tagged individuals to mix into the population, another sample of size n is captured. With X = the number of tagged animals in the second sample, the objective is to use the observed x to estimate the population size N.

The parameter of interest is $\theta = N$, which can assume only integer values, so even after determining the likelihood function (pmf of X here), using calculus to obtain N would present difficulties. If we think of a success as a previously tagged animal being recaptured, then sampling is without replacement from a population containing M successes and $N - M$ failures, so that X is a hypergeometric rv and the likelihood function is

$$p(x;N) = h(x;n,M,N) = \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}$$

The integer-valued nature of N notwithstanding, it would be difficult to take the derivative of $p(x; N)$. However, let's consider the ratio of $p(x; N)$ to $p(x; N - 1)$:

$$\frac{p(x;N)}{p(x;N-1)} = \frac{(N-M) \cdot (N-n)}{N(N-M-n+x)}$$

This ratio is larger than 1 if and only if (iff) $N < Mn/x$. The value of N for which $p(x; N)$ is maximized is therefore the largest integer less than Mn/x. If we use standard mathematical notation $[r]$ for the largest integer less than or equal to r, the mle of N is $\hat{N} = [Mn/x]$. As an illustration, if $M = 200$ fish are taken from a lake and tagged, subsequently $n = 100$ fish are recaptured, and among the 100 there are $x = 11$ tagged fish, then $\hat{N} = [(200)(100)/11] = [1818.18] = 1818$. The estimate is actually rather intuitive; x/n is the proportion of the recaptured sample that is tagged, whereas M/N is the proportion of the entire population that is tagged. The estimate is obtained by equating these two proportions (estimating a population proportion by a sample proportion). ■

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a pdf $f(x; \theta)$ that is symmetric about θ, but the investigator is unsure of the form of the f function. It is then desirable to use an estimator $\hat{\theta}$ that is *robust*, that is, one that performs well for a wide variety of underlying pdf's. One such estimator is a trimmed mean. In recent years, statisticians have proposed another type of estimator, called an *M-estimator*, based on a generalization of maximum likelihood estimation. Instead of maximizing the log likelihood $\Sigma \ln[f(x; \theta)]$ for a specified f, one seeks to maximize $\Sigma \rho(x_i; \theta)$. The "objective function" ρ is selected to yield an estimator with good robustness properties. The book by David Hoaglin et al. (see the bibliography) contains a good exposition on this subject.

Exercises Section 7.2 (21–31)

21. A random sample of n bike helmets manufactured by a company is selected. Let X = the number among the n that are flawed, and let $p = P$ (flawed). Assume that only X is observed, rather than the sequence of S's and F's.

a. Derive the maximum likelihood estimator of p. If $n = 20$ and $x = 3$, what is the estimate?

b. Is the estimator of part (a) unbiased?

c. If $n = 20$ and $x = 3$, what is the mle of the probability $(1 - p)^5$ that none of the next five helmets examined is flawed?

22. Let X have a Weibull distribution with parameters α and β, so

$$E(X) = \beta \cdot \Gamma(1 + 1/\alpha)$$
$$V(X) = \beta^2\{\Gamma(1 + 2/\alpha) - [\Gamma(1 + 1/\alpha)]^2\}$$

a. Based on a random sample $X_1, \ldots, X_n$, write equations for the method of moments estimators of β and α. Show that, once the estimate of α has been obtained, the estimate of β can be found from a table of the gamma function and that the estimate of α is the solution to a complicated equation involving the gamma function.

b. If $n = 20$, $\bar{x} = 28.0$, and $\sum x_i^2 = 16{,}500$, compute the estimates. [*Hint*: $[\Gamma(1.2)]^2/\Gamma(1.4) = .95$.]

23. Let X denote the proportion of allotted time that a randomly selected student spends working on a certain aptitude test. Suppose the pdf of X is

$$f(x;\ \theta) = \begin{cases} (\theta+1)x^\theta & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

where $-1 < \theta$. A random sample of ten students yields data $x_1 = .92$, $x_2 = .79$, $x_3 = .90$, $x_4 = .65$, $x_5 = .86$, $x_6 = .47$, $x_7 = .73$, $x_8 = .97$, $x_9 = .94$, $x_{10} = .77$.

a. Use the method of moments to obtain an estimator of θ, and then compute the estimate for this data.

b. Obtain the maximum likelihood estimator of θ, and then compute the estimate for the given data.

24. Two different computer systems are monitored for a total of n weeks. Let X_i denote the number of breakdowns of the first system during the ith week, and suppose the X_i's are independent and drawn from a Poisson distribution with parameter λ_1. Similarly, let Y_i denote the number of breakdowns of the second system during the ith week, and assume independence with each Y_i Poisson with parameter λ_2. Derive the mle's of λ_1, λ_2, and $\lambda_1 - \lambda_2$. [*Hint*: Using independence, write the joint pmf (likelihood) of the X_i's and Y_i's together.]

25. Refer to Exercise 21. Instead of selecting $n = 20$ helmets to examine, suppose we examine helmets in succession until we have found $r = 3$ flawed ones. If the 20th helmet is the third flawed one (so that the number of helmets examined that were not flawed is $x = 17$), what is the mle of p? Is this the same as the estimate in Exercise 21? Why or why not? Is it the same as the estimate computed from the unbiased estimator of Exercise 17?

26. Six Pepperidge Farm bagels were weighed, yielding the following data (grams):

117.6 109.5 111.6 109.2 119.1 110.8

(*Note*: 4 oz = 113.4 g)

a. Assuming that the six bagels are a random sample and the weight is normally distributed, estimate the true average weight and standard deviation of the weight using maximum likelihood.

b. Again assuming a normal distribution, estimate the weight below which 95% of all bagels will have their weights. [*Hint*: What is the 95th percentile in terms of μ and σ? Now use the invariance principle.]

c. Suppose we choose another bagel and weigh it. Let X = weight of the bagel. Use the given data to obtain the mle of $P(X \le 113.4)$. (*Hint*: $P(X \le 113.4) = \Phi[(113.4 - \mu)/\sigma)]$.)

27. Suppose a measurement is made on some physical characteristic whose value is known, and let X denote the resulting measurement error. For an unbiased measuring instrument or technique, the mean value of X is 0. Assume that any particular measurement error is normally distributed with variance σ^2. Let $X_1, \ldots X_n$ be a random sample of measurement errors.

a. Obtain the method of moments estimator of σ^2.

b. Obtain the maximum likelihood estimator of σ^2.

28. Let $X_1, \ldots, X_n$ be a random sample from a gamma distribution with parameters α and β.

a. Derive the equations whose solution yields the maximum likelihood estimators of α and β. Do you think they can be solved explicitly?

b. Show that the mle of $\mu = \alpha\beta$ is $\hat{\mu} = \overline{X}$.

29. Let $X_1, X_2, \ldots, X_n$ represent a random sample from the Rayleigh distribution with density function given in Exercise 15. Determine

a. The maximum likelihood estimator of θ and then calculate the estimate for the vibratory stress data given in that exercise. Is this estimator the same as the unbiased estimator suggested in Exercise 15?

b. The mle of the median of the vibratory stress distribution. [*Hint*: First express the median in terms of θ.]

30. Consider a random sample $X_1, X_2, \ldots, X_n$ from the shifted exponential pdf

$$f(x; \lambda, \theta) = \begin{cases} \lambda e^{-\lambda(x-\theta)} & x \ge \theta \\ 0 & \text{otherwise} \end{cases}$$

Taking $\theta = 0$ gives the pdf of the exponential distribution considered previously (with positive density to the right of zero). An example of the shifted exponential distribution appeared in Example 4.5, in which the variable of interest was time headway in traffic flow and $\theta = .5$ was the minimum possible time headway.

a. Obtain the maximum likelihood estimators of θ and λ.

b. If $n = 10$ time headway observations are made, resulting in the values 3.11, .64, 2.55, 2.20, 5.44, 3.42, 10.39, 8.93, 17.82, and 1.30, calculate the estimates of θ and λ.

31. At time $t = 0$, 20 identical components are put on test. The lifetime distribution of each is exponential with parameter λ. The experimenter then leaves the test facility unmonitored. On his return 24 h later, the experimenter immediately terminates the test after noticing that $y = 15$ of the 20 components are still in operation (so 5 have failed). Derive the mle of λ. [*Hint*: Let $Y =$ the number that survive 24 h. Then $Y \sim \text{Bin}(n, p)$. What is the mle of p? Now notice that $p = P(X_i \geq 24)$, where X_i is exponentially distributed. This relates λ to p, so the former can be estimated once the latter has been.]

7.3 Sufficiency

An investigator who wishes to make an inference about some parameter θ will base conclusions on the value of one or more statistics – the sample mean $\overline{X}$, the sample variance S^2, the sample range $Y_n - Y_1$, and so on. Intuitively, some statistics will contain more information about θ than will others. Sufficiency, the topic of this section, will help us decide which functions of the data are most informative for making inferences.

As a first point, we note that a statistic $T = t(X_1, \ldots, X_n)$ will not be useful for drawing conclusions about θ unless the distribution of T depends on θ. Consider, for example, a random sample of size $n = 2$ from a normal distribution with mean μ and variance σ^2, and let $T = X_1 - X_2$. Then T has a normal distribution with mean 0 and variance $2\sigma^2$, which does not depend on μ. Thus this statistic cannot be used as a basis for drawing any conclusions about μ, although it certainly does carry information about the variance σ^2.

The relevance of this observation to sufficiency is as follows. Suppose an investigator is given the value of some statistic T, and then examines the *conditional distribution of the sample* $X_1, X_2, \ldots, X_n$ given the value of the statistic – for example, the conditional distribution given that $\overline{X} = 28.7$. If this conditional distribution does not depend upon θ, then it can be concluded that there is no additional information about θ in the data over and above what is provided by T. In this sense, for purposes of making inferences about θ, it is *sufficient* to know the value of T, which contains all the information in the data relevant to θ.

Example 7.25 An investigation of major defects on new vehicles of a certain type involved selecting a random sample of $n = 3$ vehicles and determining for each one the value of $X =$ the number of major defects. This resulted in observations $x_1 = 1$, $x_2 = 0$, and $x_3 = 3$. You, as a consulting statistician, have been provided with a description of the experiment, from which it is reasonable to assume that X has a Poisson distribution, and told only that the total number of defects for the three sampled vehicles was four.

Knowing that $T = \sum X_i = 4$, would there be any additional advantage in having the observed values of the individual X_i's when making an inference about the Poisson parameter λ? Or rather is it the case that the statistic T contains all relevant information about λ in the data? To address this issue, consider the conditional distribution of X_1, X_2, X_3 given that $\sum X_i = 4$. First of all, there are only