Chapter 7

# PARAMETER ESTIMATION

## 7.1 INTRODUCTION

Let $X_1, \ldots, X_n$ be a random sample from a distribution $F_\theta$ that is specified up to a vector of unknown parameters $\theta$. For instance, the sample could be from a Poisson distribution whose mean value is unknown; or it could be from a normal distribution having an unknown mean and variance. Whereas in probability theory it is usual to suppose that all of the parameters of a distribution are known, the opposite is true in statistics, where a central problem is to use the observed data to make inferences about the unknown parameters.

In Section 7.2, we present the *maximum likelihood* method for determining estimators of unknown parameters. The estimates so obtained are called *point estimates*, because they specify a single quantity as an estimate of $\theta$. In Section 7.3, we consider the problem of obtaining *interval estimates*. In this case, rather than specifying a certain value as our estimate of $\theta$, we specify an interval in which we estimate that $\theta$ lies. Additionally, we consider the question of how much *confidence* we can attach to such an interval estimate. We illustrate by showing how to obtain an interval estimate of the unknown mean of a normal distribution whose variance is specified. We then consider a variety of interval estimation problems. In Section 7.3.1, we present an interval estimate of the mean of a normal distribution whose variance is unknown. In Section 7.3.2, we obtain an interval estimate of the variance of a normal distribution. In Section 7.4, we determine an interval estimate for the difference of two normal means, both when their variances are assumed to be known and when they are assumed to be unknown (although in the latter case we suppose that the unknown variances are equal). In Sections 7.5 and the optional Section 7.6, we present interval estimates of the mean of a Bernoulli random variable and the mean of an exponential random variable.

In the optional Section 7.7, we return to the general problem of obtaining point estimates of unknown parameters and show how to evaluate an estimator by considering its mean square error. The bias of an estimator is discussed, and its relationship to the mean square error is explored.

In the optional Section 7.8, we consider the problem of determining an estimate of an unknown parameter when there is some prior information available. This is the *Bayesian* approach, which supposes that prior to observing the data, information about $\theta$ is always available to the decision maker, and that this information can be expressed in terms of a probability distribution on $\theta$. In such a situation, we show how to compute the *Bayes estimator*, which is the estimator whose expected squared distance from $\theta$ is minimal.

## 7.2 MAXIMUM LIKELIHOOD ESTIMATORS

Any statistic used to estimate the value of an unknown parameter $\theta$ is called an *estimator* of $\theta$. The observed value of the estimator is called the *estimate*. For instance, as we shall see, the usual estimator of the mean of a normal population, based on a sample $X_1, \dots, X_n$ from that population, is the sample mean $\overline{X} = \sum_i X_i/n$. If a sample of size 3 yields the data $X_1 = 2, X_2 = 3, X_3 = 4$, then the estimate of the population mean, resulting from the estimator $\overline{X}$, is the value 3.

Suppose that the random variables $X_1, \dots, X_n$, whose joint distribution is assumed given except for an unknown parameter $\theta$, are to be observed. The problem of interest is to use the observed values to estimate $\theta$. For example, the $X_i$'s might be independent, exponential random variables each having the same unknown mean $\theta$. In this case, the joint density function of the random variables would be given by

$$
\begin{aligned}
&f(x_1, x_2, \dots, x_n) \\
&\quad = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n) \\
&\quad = \frac{1}{\theta} e^{-x_1/\theta} \frac{1}{\theta} e^{-x_2/\theta} \cdots \frac{1}{\theta} e^{-x_n/\theta}, \qquad 0 < x_i < \infty, i = 1, \dots, n \\
&\quad = \frac{1}{\theta^n} \exp\left\{ -\sum_1^n x_i/\theta \right\}, \qquad 0 < x_i < \infty, i = 1, \dots, n
\end{aligned}
$$

and the objective would be to estimate $\theta$ from the observed data $X_1, X_2, \dots, X_n$.

A particular type of estimator, known as the *maximum likelihood* estimator, is widely used in statistics. It is obtained by reasoning as follows. Let $f(x_1, \dots, x_n|\theta)$ denote the joint probability mass function of the random variables $X_1, X_2, \dots, X_n$ when they are discrete, and let it be their joint probability density function when they are jointly continuous random variables. Because $\theta$ is assumed unknown, we also write $f$ as a function of $\theta$. Now since $f(x_1, \dots, x_n|\theta)$ represents the likelihood that the values $x_1, x_2, \dots, x_n$ will be observed when $\theta$ is the true value of the parameter, it would seem that a reasonable estimate of $\theta$ would be that value yielding the largest likelihood of the observed values. In other words, the maximum likelihood estimate $\hat{\theta}$ is defined to be that value of $\theta$ maximizing $f(x_1, \dots, x_n|\theta)$ where $x_1, \dots, x_n$ are the observed values. The function $f(x_1, \dots, x_n|\theta)$ is often referred to as the *likelihood* function of $\theta$.

In determining the maximizing value of $\theta$, it is often useful to use the fact that $f(x_1, \ldots, x_n|\theta)$ and $\log[f(x_1, \ldots, x_n|\theta)]$ have their maximum at the same value of $\theta$. Hence, we may also obtain $\hat{\theta}$ by maximizing $\log[f(x_1, \ldots, x_n|\theta)]$.

**EXAMPLE 7.2a (Maximum Likelihood Estimator of a Bernoulli Parameter)** Suppose that $n$ independent trials, each of which is a success with probability $p$, are performed. What is the maximum likelihood estimator of $p$?

**SOLUTION** The data consist of the values of $X_1, \ldots, X_n$ where

$$X_i = \begin{cases} 1 & \text{if trial } i \text{ is a success} \\ 0 & \text{otherwise} \end{cases}$$

Now

$$P\{X_i = 1\} = p = 1 - P\{X_i = 0\}$$

which can be succinctly expressed as

$$P\{X_i = x\} = p^x(1-p)^{1-x}, \quad x = 0, 1$$

Hence, by the assumed independence of the trials, the likelihood (that is, the joint probability mass function) of the data is given by

$$\begin{aligned} f(x_1, \ldots, x_n|p) &= P\{X_1 = x_1, \ldots, X_n = x_n|p\} \\ &= p^{x_1}(1-p)^{1-x_1} \cdots p^{x_n}(1-p)^{1-x_n} \\ &= p^{\Sigma_1^n x_i}(1-p)^{n-\Sigma_1^n x_i}, \quad x_i = 0, 1, \quad i = 1, \ldots, n \end{aligned}$$

To determine the value of $p$ that maximizes the likelihood, first take logs to obtain

$$\log f(x_1, \ldots, x_n|p) = \sum_1^n x_i \log p + \left(n - \sum_1^n x_i\right) \log(1-p)$$

Differentiation yields

$$\frac{d}{dp} \log f(x_1, \ldots, x_n|p) = \frac{\sum\limits_1^n x_i}{p} - \frac{\left(n - \sum\limits_1^n x_i\right)}{1-p}$$

Upon equating to zero and solving, we obtain that the maximum likelihood estimate $\hat{p}$ satisfies

$$\frac{\sum_{1}^{n} x_i}{\hat{p}} = \frac{n - \sum_{1}^{n} x_i}{1 - \hat{p}}$$

or

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Hence, the maximum likelihood estimator of the unknown mean of a Bernoulli distribution is given by

$$d(X_1, \ldots, X_n) = \frac{\sum_{i=1}^{n} X_i}{n}$$

Since $\sum_{i=1}^{n} X_i$ is the number of successful trials, we see that the maximum likelihood estimator of $p$ is equal to the proportion of the observed trials that result in successes. For an illustration, suppose that each RAM (random access memory) chip produced by a certain manufacturer is, independently, of acceptable quality with probability $p$. Then if out of a sample of 1,000 tested 921 are acceptable, it follows that the maximum likelihood estimate of $p$ is .921. ■

**EXAMPLE 7.2b** Two proofreaders were given the same manuscript to read. If proofreader 1 found $n_1$ errors, and proofreader 2 found $n_2$ errors, with $n_{1,2}$ of these errors being found by both proofreaders, estimate $N$, the total number of errors that are in the manuscript.

**SOLUTION** Before we can estimate $N$ we need to make some assumptions about the underlying probability model. So let us assume that the results of the proofreaders are independent, and that each error in the manuscript is independently found by proofreader $i$ with probability $p_i$, $i = 1, 2$.

To estimate $N$, we will start by deriving an estimator of $p_1$. To do so, note that each of the $n_2$ errors found by reader 2 will, independently, be found by proofreader 1 with probability $p_i$. Because proofreader 1 found $n_{1,2}$ of those $n_2$ errors, a reasonable estimate of $p_1$ is given by

$$\hat{p}_1 = \frac{n_{1,2}}{n_2}$$

However, because proofreader 1 found $n_1$ of the $N$ errors in the manuscript, it is reasonable to suppose that $p_1$ is also approximately equal to $\frac{n_1}{N}$. Equating this to $\hat{p}_1$ gives that

$$\frac{n_{1,2}}{n_2} \approx \frac{n_1}{N}$$

or

$$N \approx \frac{n_1 n_2}{n_{1,2}}$$

Because the preceding estimate is symmetric in $n_1$ and $n_2$, it follows that it is the same no matter which proofreader is designated as proofreader 1.

An interesting application of the preceding occurred when two teams of researchers recently announced that they had decoded the human genetic code sequence. As part of their work both teams estimated that the human genome consisted of approximately 33,000 genes. Because both teams independently arrived at the same number, many scientists found this number believable. However, most scientists were quite surprised by this relatively small number of genes; by comparison it is only about twice as many as a fruit fly has. However, a closer inspection of the findings indicated that the two groups only agreed on the existence of about 17,000 genes. (That is, 17,000 genes were found by both teams.) Thus, based on our preceding estimator, we would estimate that the actual number of genes, rather than being 33,000, is

$$\frac{n_1 n_2}{n_{1,2}} = \frac{33{,}000 \times 33{,}000}{17{,}000} \approx 64{,}000$$

(Because there is some controversy about whether some of genes claimed to be found are actually genes, 64,000 should probably be taken as an upper bound on the actual number of genes.)

The estimation approach used when there are two proofreaders does not work when there are $m$ proofreaders, when $m > 2$. Because, if for each $i$, we let $\hat{p}_i$ be the fraction of the errors found by at least one of the other proofreaders $j$, $(j \neq i)$, that are also found by $i$, and then set that equal to $\frac{n_i}{N}$, then the estimate of $N$, namely $\frac{n_i}{\hat{p}_i}$, would differ for different values of $i$. Moreover, with this approach it is possible that we may have that $\hat{p}_i > \hat{p}_j$ even if proofreader $i$ finds fewer errors than does proofreader $j$. For instance, for $m = 3$, suppose proofreaders 1 and 2 find exactly the same set of 10 errors whereas proofreader 3 finds 20 errors with only 1 of them in common with the set of errors found by the others. Then, because proofreader 1 (and 2) found 10 of the 29 errors found by at least one of the other proofreaders, $\hat{p}_i = 10/29$, $i = 1, 2$. On the other hand, because proofreader 3 only found 1 of the 10 errors found by the others, $\hat{p}_3 = 1/10$. Therefore, although proofreader 3 found twice the number of errors as did proofreader 1, the estimate of $p_3$ is less than that of $p_1$. To obtain more reasonable estimates, we could take the preceding values of $\hat{p}_i$, $i = 1, \ldots, m$, as preliminary estimates of the $p_i$. Now, let $n_f$ be the number of errors that are found by at least one proofreader. Because $n_f/N$ is the fraction of errors that are found by at least one proofreader, this should approximately equal $1 - \prod_{i=1}^{m}(1 - p_i)$, the probability that an error is found by at least one proofreader. Therefore, we have

$$\frac{n_f}{N} \approx 1 - \prod_{i=1}^{m}(1 - p_i)$$

suggesting that $N \approx \hat{N}$, where

$$\hat{N} = \frac{n_f}{1 - \prod_{i=1}^{m}(1 - \hat{p}_i)} \tag{7.2.1}$$

With this estimate of $N$, we can then reset our estimates of the $p_i$ by using

$$\hat{p}_i = \frac{n_i}{\hat{N}}, \quad i = 1, \ldots, m \tag{7.2.2}$$

We can then reestimate $N$ by using the new value (Equation 7.2.1). (The estimation need not stop here; each time we obtain a new estimate $\hat{N}$ of $N$ we can use Equation 7.2.2 to obtain new estimates of the $p_i$, which can then be used to obtain a new estimate of $N$, and so on.) ■

**EXAMPLE 7.2c (Maximum Likelihood Estimator of a Poisson Parameter)** Suppose $X_1, \ldots, X_n$ are independent Poisson random variables each having mean $\lambda$. Determine the maximum likelihood estimator of $\lambda$.

**SOLUTION** The likelihood function is given by

$$\begin{aligned} f(x_1, \ldots, x_n|\lambda) &= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!} \\ &= \frac{e^{-n\lambda}\lambda^{\Sigma_1^n x_i}}{x_1! \ldots x_n!} \end{aligned}$$

Thus,

$$\log f(x_1, \ldots, x_n|\lambda) = -n\lambda + \sum_{1}^{n} x_i \log \lambda - \log c$$

where $c = \prod_{i=1}^{n} x_i!$ does not depend on $\lambda$. Differentiation yields

$$\frac{d}{d\lambda} \log f(x_1, \ldots, x_n|\lambda) = -n + \frac{\sum_{1}^{n} x_i}{\lambda}$$

By equating to zero, we obtain that the maximum likelihood estimate $\hat{\lambda}$ equals

$$\hat{\lambda} = \frac{\sum_{1}^{n} x_i}{n}$$

and so the maximum likelihood estimator is given by

$$d(X_1, \ldots, X_n) = \frac{\sum_{i=1}^{n} X_i}{n}$$

For example, suppose that the number of people who enter a certain retail establishment in any day is a Poisson random variable having an unknown mean $\lambda$, which must be estimated. If after 20 days a total of 857 people have entered the establishment, then the maximum likelihood estimate of $\lambda$ is $857/20 = 42.85$. That is, we estimate that on average, 42.85 customers will enter the establishment on a given day. ■

**EXAMPLE 7.2d** The number of traffic accidents in Berkeley, California, in 10 randomly chosen nonrainy days in 1998 is as follows:

$$4, 0, 6, 5, 2, 1, 2, 0, 4, 3$$

Use these data to estimate the proportion of nonrainy days that had 2 or fewer accidents that year.

**SOLUTION** Since there are a large number of drivers, each of whom has a small probability of being involved in an accident in a given day, it seems reasonable to assume that the daily number of traffic accidents is a Poisson random variable. Since

$$\overline{X} = \frac{1}{10} \sum_{i=1}^{10} X_i = 2.7$$

it follows that the maximum likelihood estimate of the Poisson mean is 2.7. Since the long-run proportion of nonrainy days that have 2 or fewer accidents is equal to $P\{X \leq 2\}$, where $X$ is the random number of accidents in a day, it follows that the desired estimate is

$$e^{-2.7}(1 + 2.7 + (2.7)^2/2) = .4936$$

That is, we estimate that a little less than half of the nonrainy days had 2 or fewer accidents. ■

**EXAMPLE 7.2e (Maximum Likelihood Estimator in a Normal Population)** Suppose $X_1, \ldots, X_n$ are independent, normal random variables each with unknown mean $\mu$ and unknown standard deviation $\sigma$. The joint density is given by

$$f(x_1, \ldots, x_n|\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x_i - \mu)^2}{2\sigma^2}\right]$$

$$= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} \exp\left[\frac{-\sum_1^n (x_i - \mu)^2}{2\sigma^2}\right]$$

The logarithm of the likelihood is thus given by

$$\log f(x_1, \ldots, x_n|\mu, \sigma) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{\sum_1^n (x_i - \mu)^2}{2\sigma^2}$$

In order to find the value of $\mu$ and $\sigma$ maximizing the foregoing, we compute

$$\frac{\partial}{\partial\mu} \log f(x_1, \ldots, x_n|\mu, \sigma) = \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2}$$

$$\frac{\partial}{\partial\sigma} \log f(x_1, \ldots, x_n|\mu, \sigma) = -\frac{n}{\sigma} + \frac{\sum_1^n (x_i - \mu)^2}{\sigma^3}$$

Equating these equations to zero yields that

$$\hat{\mu} = \sum_{i=1}^n x_i/n$$

and

$$\hat{\sigma} = \left[\sum_{i=1}^n (x_i - \hat{\mu})^2/n\right]^{1/2}$$

Hence, the maximum likelihood estimators of $\mu$ and $\sigma$ are given, respectively, by

$$\overline{X} \quad \text{and} \quad \left[\sum_{i=1}^{n}(X_i - \overline{X})^2/n\right]^{1/2} \tag{7.2.3}$$

It should be noted that the maximum likelihood estimator of the standard deviation $\sigma$ differs from the sample standard deviation

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

in that the denominator in Equation 7.2.3 is $\sqrt{n}$ rather than $\sqrt{n-1}$. However, for $n$ of reasonable size, these two estimators of $\sigma$ will be approximately equal. ■

**EXAMPLE 7.2f** *Kolmogorov's law of fragmentation* states that the size of an individual particle in a large collection of particles resulting from the fragmentation of a mineral compound will have an approximate lognormal distribution, where a random variable $X$ is said to have a *lognormal* distribution if $\log(X)$ has a normal distribution. The law, which was first noted empirically and then later given a theoretical basis by Kolmogorov, has been applied to a variety of engineering studies. For instance, it has been used in the analysis of the size of randomly chosen gold particles from a collection of gold sand. A less obvious application of the law has been to a study of the stress release in earthquake fault zones (see Lomnitz, C., "Global Tectonics and Earthquake Risk," *Developments in Geotectonics*, Elsevier, Amsterdam, 1979).

Suppose that a sample of 10 grains of metallic sand taken from a large sand pile have respective lengths (in millimeters):

$$2.2,\ 3.4,\ 1.6,\ 0.8,\ 2.7,\ 3.3,\ 1.6,\ 2.8,\ 2.5,\ 1.9$$

Estimate the percentage of sand grains in the entire pile whose length is between 2 and 3 mm.

**SOLUTION** Taking the natural logarithm of these 10 data values, the following transformed data set results

$$.7885,\ 1.2238,\ .4700,\ -.2231,\ .9933,\ 1.1939,\ .4700,\ 1.0296,\ .9163,\ .6419$$

Because the sample mean and sample standard deviation of these data are

$$\bar{x} = .7504, \quad s = .4351$$

it follows that the logarithm of the length of a randomly chosen grain has a normal distribution with mean approximately equal to .7504 and with standard deviation approximately equal to .4351. Hence, if $X$ is the length of the grain, then

$$\begin{aligned} P\{2 < X < 3\} &= P\{\log(2) < \log(X) < \log(3)\} \\ &= P\left\{\frac{\log(2) - .7504}{.4351} < \frac{\log(X) - .7504}{.4351} < \frac{\log(3) - .7504}{.4351}\right\} \\ &= P\left\{-.1316 < \frac{\log(X) - .7504}{.4351} < .8003\right\} \\ &\approx \Phi(.8003) - \Phi(-.1316) \\ &= .3405 \quad \blacksquare \end{aligned}$$

The lognormal distribution is often assumed in situations where the random variable under interest can be regarded as the product of a large number of independent and identically distributed random variables. For instance, it is commonly used in finance as the distribution of the price of a security at some future time. To see why this might be reasonable, suppose that the current price of the security is $s$ and that we are interested in $S(t)$, the price of the security after an additional time $t$. For a large value $n$, let $t_i = it/n$, and consider $S(t_1), \ldots, S(t_n)$, the prices of the security at the times $t_1, \ldots, t_n$. Now, a common assumption in finance is that the ratios $S(t_i)/S(t_{i-1})$ are approximately independent and identically distributed. Consequently, if we let $X_i = S(t_i)/S(t_{i-1})$, then writing

$$\begin{aligned} S(t) = S(t_n) &= S(t_0) \cdot \frac{S(t_1)}{S(t_0)} \cdot \frac{S(t_2)}{S(t_1)} \cdots \frac{S(t_n)}{S(t_{n-1})} \\ &= s \prod_{i=1}^{n} X_i \end{aligned}$$

we obtain, upon taking logarithms, that

$$\log(S(t)) = \log(s) + \sum_{i=1}^{n} \log(X_i)$$

Thus, by the central limit theorem $\log(S(t))$ will approximately have a normal distribution.

The lognormal distribution has also been shown to be a good fit for such random variables as length of patient stays in hospitals, and vehicle travel times.

In all of the foregoing examples, the maximum likelihood estimator of the population mean turned out to be the sample mean $\overline{X}$. To show that this is not always the situation, consider the following example.

**EXAMPLE 7.2g (Estimating the Mean of a Uniform Distribution)** Suppose $X_1, \ldots, X_n$ constitute a sample from a uniform distribution on $(0, \theta)$, where $\theta$ is unknown. Their joint density is thus

$$f(x_1, x_2, \ldots, x_n|\theta) = \begin{cases} \dfrac{1}{\theta^n} & 0 < x_i < \theta, \quad i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

This density is maximized by choosing $\theta$ as small as possible. Since $\theta$ must be at least as large as all of the observed values $x_i$, it follows that the smallest possible choice of $\theta$ is equal to $\max(x_1, x_2, \ldots, x_n)$. Hence, the maximum likelihood estimator of $\theta$ is

$$\hat{\theta} = \max(X_1, X_2, \ldots, X_n)$$

It easily follows from the foregoing that the maximum likelihood estimator of $\theta/2$, the mean of the distribution, is $\max(X_1, X_2, \ldots, X_n)/2$. ■

## *7.2.1 Estimating Life Distributions

Let $X$ denote the age at death of a randomly chosen child born today. That is, $X = i$ if the newborn dies in its $i$th year, $i \geq 1$. To estimate the probability mass function of $X$, let $\lambda_i$ denote the probability that a newborn who has survived his or her first $i - 1$ years dies in year $i$. That is,

$$\lambda_i = P\{X = i|X > i - 1\} = \frac{P\{X = i\}}{P\{X > i - 1\}}$$

Also, let

$$s_i = 1 - \lambda_i = \frac{P\{X > i\}}{P\{X > i - 1\}}$$

be the probability that a newborn who survives her first $i - 1$ years also survives year $i$. The quantity $\lambda_i$ is called the *failure rate*, and $s_i$ is called the *survival rate*, of an individual who is entering his or her $i$th year. Now,

$$\begin{aligned} s_1 s_2 \cdots s_i &= P\{X > 1\} \frac{P\{X > 2\}P\{X > 3\}}{P\{X > 1\}P\{X > 2\}} \cdots \frac{P\{X > i\}}{P\{X > i - 1\}} \\ &= P\{X > i\} \end{aligned}$$

Therefore,

$$P\{X = n\} = P\{X > n - 1\}\lambda_n = s_1 \cdots s_{n-1}(1 - s_n)$$

Consequently, we can estimate the probability mass function of $X$ by estimating the quantities $s_i$, $i = 1, \ldots, n$. The value $s_i$ can be estimated by looking at all individuals in the

* Optional section.

population who reached age $i$ 1 year ago, and then letting the estimate $\hat{s}_i$ be the fraction of them who are alive today. We would then use $\hat{s}_1\hat{s}_2\cdots\hat{s}_{n-1}(1-\hat{s}_n)$ as the estimate of $P\{X=n\}$. (Note that although we are using the most recent possible data to estimate the quantities $s_i$, our estimate of the probability mass function of the lifetime of a newborn assumes that the survival rate of the newborn when it reaches age $i$ will be the same as last year's survival rate of someone of age $i$.)

The use of the survival rate to estimate a life distribution is also of importance in health studies with partial information. For instance, consider a study in which a new drug is given to a random sample of 12 lung cancer patients. Suppose that after some time we have the following data on the number of months of survival after starting the new drug:

$$4, 7^*, 9, 11^*, 12, 3, 14^*, 1, 8, 7, 5, 3^*$$

where $x$ means that the patient died in month $x$ after starting the drug treatment, and $x^*$ means that the patient has taken the drug for $x$ months and is still alive.

Let $X$ equal the number of months of survival after beginning the drug treatment, and let

$$s_i = P\{X > i | X > i-1\} = \frac{P\{X > i\}}{P\{X > i-1\}}$$

To estimate $s_i$, the probability that a patient who has survived the first $i-1$ months will also survive month $i$, we should take the fraction of those patients who began their $i$th month of drug taking and survived the month. For instance, because 11 of the 12 patients survived month 1, $\hat{s}_1 = 11/12$. Because all 11 patients who began month 2 survived, $\hat{s}_2 = 11/11$. Because 10 of the 11 patients who began month 3 survived, $\hat{s}_3 = 10/11$. Because 8 of the 9 patients who began their fourth month of taking the drug (the 9 being all but the ones labelled 1, 3, and 3*) survived month 4, $\hat{s}_4 = 8/9$. Similar reasoning holds for the others, giving the following survival rate estimates:

$$\begin{aligned}
\hat{s}_1 &= 11/12\\
\hat{s}_2 &= 11/11\\
\hat{s}_3 &= 10/11\\
\hat{s}_4 &= 8/9\\
\hat{s}_5 &= 7/8\\
\hat{s}_6 &= 7/7\\
\hat{s}_7 &= 6/7\\
\hat{s}_8 &= 4/5\\
\hat{s}_9 &= 3/4\\
\hat{s}_{10} &= 3/3\\
\hat{s}_{11} &= 3/3
\end{aligned}$$

$$\hat{s}_{12} = 1/2$$
$$\hat{s}_{13} = 1/1$$
$$\hat{s}_{14} = 1/1$$

We can now use $\prod_{i=1}^{j} \hat{s}_i$ to estimate the probability that a drug taker survives at least $j$ time periods, $j = 1, \ldots, 14$. For instance, our estimate of $P\{X > 6\}$ is 35/54.

## 7.3 INTERVAL ESTIMATES

Suppose that $X_1, \ldots, X_n$ is a sample from a normal population having unknown mean $\mu$ and known variance $\sigma^2$. It has been shown that $\overline{X} = \sum_{i=1}^{n} X_i/n$ is the maximum likelihood estimator for $\mu$. However, we don't expect that the sample mean $\overline{X}$ will exactly equal $\mu$, but rather that it will "be close." Hence, rather than a point estimate, it is sometimes more valuable to be able to specify an interval for which we have a certain degree of confidence that $\mu$ lies within. To obtain such an interval estimator, we make use of the probability distribution of the point estimator. Let us see how it works for the preceding situation.

In the foregoing, since the point estimator $\overline{X}$ is normal with mean $\mu$ and variance $\sigma^2/n$, it follows that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{(\overline{X} - \mu)}{\sigma}$$

has a standard normal distribution. Therefore,

$$P\left\{-1.96 < \sqrt{n}\frac{(\overline{X} - \mu)}{\sigma} < 1.96\right\} = .95$$

or, equivalently,

$$P\left\{-1.96\frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < 1.96\frac{\sigma}{\sqrt{n}}\right\} = .95$$

Multiplying through by $-1$ yields the equivalent statement

$$P\left\{-1.96\frac{\sigma}{\sqrt{n}} < \mu - \overline{X} < 1.96\frac{\sigma}{\sqrt{n}}\right\} = .95$$

or, equivalently,

$$P\left\{\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = .95$$

That is, 95 percent of the time the value of the sample average $\bar{X}$ will be such that the distance between it and the mean $\mu$ will be less than $1.96\,\sigma/\sqrt{n}$. If we now observe the sample and it turns out that $\overline{X} = \bar{x}$, then we say that "with 95 percent confidence"

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \tag{7.3.1}$$