STATISTICS FOR ECONOMISTS: A BEGINNING

John E. Floyd
University of Toronto

July 2, 2010

PREFACE

The pages that follow contain the material presented in my introductory quantitative methods in economics class at the University of Toronto. They are designed to be used along with any reasonable statistics textbook. The most recent textbook for the course was James T. McClave, P. George Benson and Terry Sincich, *Statistics for Business and Economics*, Eighth Edition, Prentice Hall, 2001. The material draws upon earlier editions of that book as well as upon John Neter, William Wasserman and G. A. Whitmore, *Applied Statistics*, Fourth Edition, Allyn and Bacon, 1993, which was used previously and is now out of print. It is also consistent with Gerald Keller and Brian Warrack, *Statistics for Management and Economics*, Fifth Edition, Duxbury, 2000, which is the textbook used recently on the St. George Campus of the University of Toronto. The problems at the ends of the chapters are questions from mid-term and final exams at both the St. George and Mississauga campuses of the University of Toronto. They were set by Gordon Anderson, Lee Bailey, Greg Jump, Victor Yu and others including myself.

This manuscript should be useful for economics and business students enrolled in basic courses in statistics and, as well, for people who have studied statistics some time ago and need a review of what they are supposed to have learned. Indeed, one could learn statistics from scratch using this material alone, although those trying to do so may find the presentation somewhat compact, requiring slow and careful reading and thought as one goes along.

I would like to thank the above mentioned colleagues and, in addition, Adonis Yatchew, for helpful discussions over the years, and John Maheu for helping me clarify a number of points. I would especially like to thank Gordon Anderson, who I have bothered so frequently with questions that he deserves the status of mentor.

After the original version of this manuscript was completed, I received some detailed comments on Chapter 8 from Peter Westfall of Texas Tech University, enabling me to correct a number of errors. Such comments are much appreciated.

J. E. Floyd
July 2, 2010

Chapter 4

Statistical Sampling: Point and Interval Estimation

In the previous chapter we assumed that the probability distribution of a random variable in question was known to us and from this knowledge we were able to compute the mean and variance and the probabilities that the random variable would take various values (in the case of discrete distributions) or fall within a particular range (in the case of uniform distributions). In most practical applications of statistics we may have some reason to believe that a random variable is distributed according to a binomial, Poisson, normal, etc., distribution but have little knowledge of the relevant parameter values. For example, we might know what n is in the case of a binomial distribution but know nothing about the magnitude of p. Or we may suspect that a variable is normally distributed by have no idea of the values of the parameters μ and σ. The practical procedure for finding information about these parameters is to take a sample and try to infer their values from the characteristics of the sample.

4.1 Populations and Samples

Let us first review what we learned about populations and samples in Chapter 1. A population is the set of elements of interest. It may be finite or infinite. Processes, mechanisms that produce data, are infinite populations. In terms of the analysis of the previous chapter, populations are the complete set of outcomes of a random variable. And a process is a mechanism by which outcomes of a random variable are generated. The population of outcomes of a particular random variable is distributed according to some probability

distribution—possibly but not necessarily binomial, Poisson, normal, uniform, or exponential. The parameters of the population are the parameters of its probability distribution. As such, they are numerical descriptive measures of the population. A census is a listing of the characteristics of interest of every element in a population. A sample is a subset of the population chosen according to some set of rules. *Sample statistics* are numerical descriptive measures of the characteristics of the sample calculated from the observations in the sample. We use these sample statistics to make inferences about the unobserved population parameters. You should keep in mind that a *statistic* refers to a sample quantity while a *parameter* refers to a population quantity. The sample mean is an example of a sample statistic, while the population mean is an example of a population parameter.

A *sample* is thus a part of the population under study selected so that inferences can be drawn from it about the population. It is cheaper and quicker to use samples to obtain information about a population than to take a census. Furthermore, testing items sampled may destroy them so that tests cannot be conducted on the whole population.

A *probability sample* is one where the selection of the elements from the population that appear in the sample is made according to known probabilities. A *judgment sample* is one where judgment is used to select "representative" elements or to infer that a sample is "representative" of the population. In probability samples, no discretion is allowed about which population elements enter the sample.

The most common sampling procedure is to select a *simple random sample*. A simple random sample is one for which *each possible sample combination* in the population has an equal probability of being selected. Every element of the population has the same probability of occupying each position in the sample. The sampling is without replacement, so that no element of the population can appear in the sample twice.

Note that simple random sampling requires more than each element of the population having the same probability of being selected. Suppose that we select a sample of 10 students to interview about their career plans. It is not enough that every student in the population have an equal chance of being among the 10 selected. Each student must have the same chance of being the first selected, the second selected, the third selected, etc. For example, we could divide the population into males and females (suppose the population contains an equal number of each) and select 5 males and 5 females at random for the sample. Each student would have an equal chance of being in the sample, but the sample combinations that contain an unequal number of males and females would be ruled out. One might wish

to rule these combinations out, but then the sample would not be a simple random sample.

One way to ensure that each possible sample combination has an equal chance of being in the sample is to select the sample elements one at a time in such a way that each element of the population not already in the sample has an equal chance of being chosen. In the case of a finite population, select the first element by giving each of the N population elements an equal chance of being picked. Then select the second sample element by giving the remaining $N - 1$ elements of the population an equal chance of being chosen. Repeat this process until all n sample elements have been selected.

Suppose we have a population of 5000 students that we wish to sample. We could assign each student in the population a number between 0 and 4999 and chose 100 numbers at random from the set of integers in this interval, using the numbers so selected to pick the students to appear in the sample. To choose the numbers randomly we could get a computer to spit out 100 numbers between 0 and 4999 in such a way that each of the 5000 numbers had an equal chance of being selected first and each of the 5000 numbers not yet selected had an equal chance of being selected second, third, etc. Alternatively, we could use a table of random numbers. Such a table might list five-digit numbers in the following fashion:

13284 21244 99052 00199 40578 ..... etc.

The table is constructed so each digit between 0 and 9 has an equal chance of appearing in each of the five positions for each number. We could select our sample as follows from these numbers:

1328, 2122, skip, 0019, 4057, skip, ... etc.

Numbers for which the four digits on the left side yield a number larger than 4999 are simply skipped—they can be treated as not being in the table, so that numbers between 0 and 4999 have an equal chance of being selected and numbers over 4999 have a zero chance of being selected. Any number already selected would also be discarded because we want the probability that an element of the population will be selected more than once to be zero. If the size of the population is, say, 500000, requiring that we select the elements in the sample from 6 digit numbers, we merely take each succession of 6 digits in the table of random numbers as a separate number, so that the above line in the table of random numbers would yield

132842 124499 052001 994057 ..... etc.

The first three numbers would be used to select the corresponding population elements, the fourth number would be skipped, and so on. Random numbers can also be obtained from the table by reading down the columns rather than across the rows, and the selection process can begin anywhere in the table.

When the population is generated by a process, the process itself furnishes the sample observations. Take the case of pairs of shoes coming off an assembly line. To test the quality of the production process we could select a sample of 10 pairs by simply taking the next (or any) 10 pairs off the line. This will give us a simple random sample if two conditions are met: First, each item must have the same probability of being defective as any other item. Second, the probability that any one item is defective must be independent of whether any particular other item is defective. More formally, the n random variables $X_1, X_2, X_3, \ldots X_n$ generated by a process constitute a simple random sample from an infinite population if they are *independently and identically distributed*.

Once a sample has been selected and observations on the sample elements have been made, the observations constitute a data set and the usual summary measures can be made. If $X_1, X_2, X_3, \ldots X_n$ represent the values of the n sample observations, we have

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{4.1}$$

and

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} \tag{4.2}$$

where $\bar{X}$ and s^2 are the sample mean and variance, and s is the sample standard deviation. These magnitudes are called sample statistics. The population mean, variance and standard deviation—that is, the population *parameters*—are denoted by μ, σ^2 and σ.

4.2 The Sampling Distribution of the Sample Mean

Consider an example of pairs of newly produced shoes coming off an assembly line. We want to verify their quality. The sample space consists of three sample points—neither shoe defective, one shoe defective, both shoes defective. Suppose that the process by which the shoes are manufactured generates the following *population* probability distribution for the three values that the random variable X can take:

x:	0	1	2
$P(x)$:	.81	.18	.01

Note that the population distribution is skewed to the right. Its mean is

$$E\{X\} = \mu = (0)(.81) + (1)(.18) + (2)(.01) = .2$$

and its variance is

$$\sigma^2\{X\} = (-.2)^2(.81) + (.8)^2(.18) + (1.8)^2(.01) = .18.$$

Now suppose that we do not observe the probability distribution for the population and do not know its parameters. We can attempt to make an inference about these parameters, and hence about the probability distribution of the population, by taking a sample. Suppose we take a sample of two and use the sample mean as an estimate of E{X}. There are nine potential samples of two that can be taken from the population. These potential samples and the corresponding sample means together with the probabilities of picking each sample are listed below:

Sample	$\bar{X}$	$P(\bar{X})$
0 0	0.0	$(.81)^2 = .6561$
0 1	0.5	$(.81)(.18) = .1458$
0 2	1.0	$(.81)(.01) = .0081$
1 0	0.5	$(.18)(.81) = .1458$
1 1	1.0	$(.18)^2 = .0324$
1 2	1.5	$(.18)(.01) = .0018$
2 0	1.0	$(.01)(.81) = .0081$
2 1	1.5	$(.01)(.18) = .0018$
2 2	2.0	$(.01)^2 = .0001$
		1.0000

The sum of the probabilities is unity because all possible samples of two that can be drawn from the population are listed. It turns out that the sample mean can take five values— 0, .5, 1, 1.5 and 2. The probabilities that it will take each of these values can be obtained by adding the probabilities associated with the occurrence of each possible sample value in the table above. For example, the probability that the sample mean will be .5 equals .1458 + .1458 = .2916. We thus have

$\bar{X}$:	0	.5	1	1.5	2
$P(\bar{X})$:	.6561	.2916	.0486	.0036	.0001

for which the probabilities sum to unity. This is the exact sampling distribution of $\bar{X}$. It says that there is a probability of .6561 that a sample of two will have mean 0, a probability of .2916 that it will have mean 0.5, and so forth. The mean of the sampling distribution of $\bar{X}$ is

$$E\{\bar{X}\} = (0)(.6561) + (.5)(.2916) + (1)(.0486) + (1.5)(.0036) + (2)(.0001) = .2$$

which is equal to the population mean. The variance of the sample mean is

$$\sigma^2\{\bar{X}\} = (-.2)^2(.6561) + (.3)^2(.2916) + (.8)^2(.0486)$$

$$+(1.3)^2(.0036) + (1.8)^2(.0001) = .09$$

which turns out to be half the population variance.

Now consider all possible samples of three that we could take. These are presented in Table 4.1. The sample mean can now take seven values— 0, 1/3, 2/3, 1, 4/3, 5/3, and 2. The exact sampling distribution of the sample mean (which is obtained by adding up in turn the probabilities associated with all samples that yield each possible mean) is now

$\bar{X}$:	0	1/3	2/3	1	4/3	5/3	2
$P(\bar{X})$:	.531441	.354294	.098415	.014580	.001215	.000054	.000001

The usual calculations yield a mean of the sample mean of $E\{\bar{X}\} = .2$ and a sample variance of $\sigma^2\{\bar{X}\} = .06$. The mean sample mean is again the same as the population mean and the variance of the sample mean is now one-third the population variance.

On the basis of an analysis of the exact sampling distributions of the sample mean for sample sizes of 2 and 3, we might conjecture that the expected value of the sample mean always equals the population mean and the variance of the sample mean always equals the variance of the population divided by the sample size. This conjecture is correct. For a sample of size n consisting of $X_1, X_2, X_3, \ldots, X_n$, the expectation of the sample mean will be

$$\begin{aligned} E\{\bar{X}\} &= E\left\{\frac{1}{n}(X_1 + X_2 + X_3 + \ldots + X_n)\right\} \\ &= \frac{1}{n}(E\{X_1\} + E\{X_2\} + E\{X_3\} + \ldots + E\{X_n\}) \\ &= \frac{1}{n}(n\,\mu) \;=\; \mu \end{aligned} \tag{4.3}$$

Table 4.1: All possible samples of three for the shoe-testing problem.

	$\bar{X}$	$P(\bar{X})$
0 0 0	0	$(.81)^3 = .531441$
0 0 1	1/3	$(.81)(.81)(.18) = .118098$
0 0 2	2/3	$(.81)(.81)(.01) = .006561$
0 1 0	1/3	$(.81)(.18)(.81) = .118098$
0 1 1	2/3	$(.81)(.18)(.18) = .026244$
0 1 2	1	$(.81)(.18)(.01) = .001458$
0 2 0	2/3	$(.81)(.01)(.81) = .006561$
0 2 1	1	$(.81)(.18)(.01) = .001458$
0 2 2	4/3	$(.81)(.01)(.01) = .000081$
1 0 0	1/3	$(.18)(.81)(.81) = .118098$
1 0 1	2/3	$(.18)(.81)(.18) = .026244$
1 0 2	1	$(.18)(.81)(.01) = .001458$
1 1 0	2/3	$(.18)(.18)(.81) = .026244$
1 1 1	1	$(.18)^3 = .005832$
1 1 2	4/3	$(.18)(.18)(.01) = .000324$
1 2 0	1	$(.18)(.01)(.81) = .001458$
1 2 1	4/3	$(.18)(.01)(.18) = .000324$
1 2 2	5/3	$(.18)(.01)(.01) = .000018$
2 0 0	2/3	$(.01)(.81)(.81) = .006561$
2 0 1	1	$(.01)(.81)(.18) = .001458$
2 0 2	4/3	$(.01)(.81)(.01) = .000081$
2 1 0	1	$(.01)(.18)(.81) = .001458$
2 1 1	4/3	$(.01)(.18)(.18) = .000324$
2 1 2	5/3	$(.01)(.18)(.01) = .000018$
2 2 0	4/3	$(.01)(.01)(.81) = .000081$
2 2 1	5/3	$(.01)(.01)(.18) = .000018$
2 2 2	2	$(.01)^3 = .000001$
		1.000000

and the variance of the sample mean will be

$$\begin{aligned}
\sigma^2\{\bar{X}\} &= E\left\{\left[\frac{1}{n}(X_1+X_2+X_3+\ldots+X_n)-E\{\bar{X}\}\right]^2\right\} \\
&= E\left\{\left[\frac{1}{n}(X_1+X_2+X_3+\ldots+X_n)-\mu\right]^2\right\} \\
&= E\left\{\left[\frac{1}{n}(X_1+X_2+X_3+\ldots+X_n)-\frac{n\mu}{n}\right]^2\right\} \\
&= \frac{1}{n^2}E\left\{[(X_1+X_2+X_3+\ldots+X_n)-n\mu]^2\right\} \\
&= \frac{1}{n^2}E\left\{[((X_1-\mu)+(X_2-\mu)+(X_3-\mu)+\ldots+(X_n-\mu)]^2\right\} \\
&= \frac{1}{n^2}\left[\sigma^2\{X_1\}+\sigma^2\{X_2\}+\sigma^2\{X_3\}+\ldots+\sigma^2\{(X_n\}\right] \\
&= \frac{1}{n^2}\left[n\,\sigma^2\right] = \frac{\sigma^2}{n}. \qquad (4.4)
\end{aligned}$$

Note that in the second last line we took advantage of the fact that the sample items were chosen independently to rule out any covariance between X_i and X_j.

It should be emphasized that the above calculations of the mean and variance of the sampling distribution are the same regardless of the distribution of the population. For the population above, increasing the sample size from two to three reduced the probability weight at the right tail of the distribution and also at $\bar{X}=0$.

The question immediately arises as to what the distribution of the sample mean will look like if we increase the sample size further. It is not practical to obtain the exact distribution of the sample mean from the above population for sample sizes bigger than three. We have to infer the probability distribution of the sample mean by taking many samples of each size and plotting histograms of the resulting sample means.

4.3 The Central Limit Theorem

Figure 4.1 shows the distribution of the sample means obtained for the shoe-testing problem by taking 1000 samples of $n=2$ (top), $n=3$ (middle) and $n=10$ (bottom). Notice how the range of the sample mean narrows as the sample size increases. Also, with a sample size as large as 10 the modal value ceases to be zero. Figure 4.2 is a continuation of Figure 4.1, showing

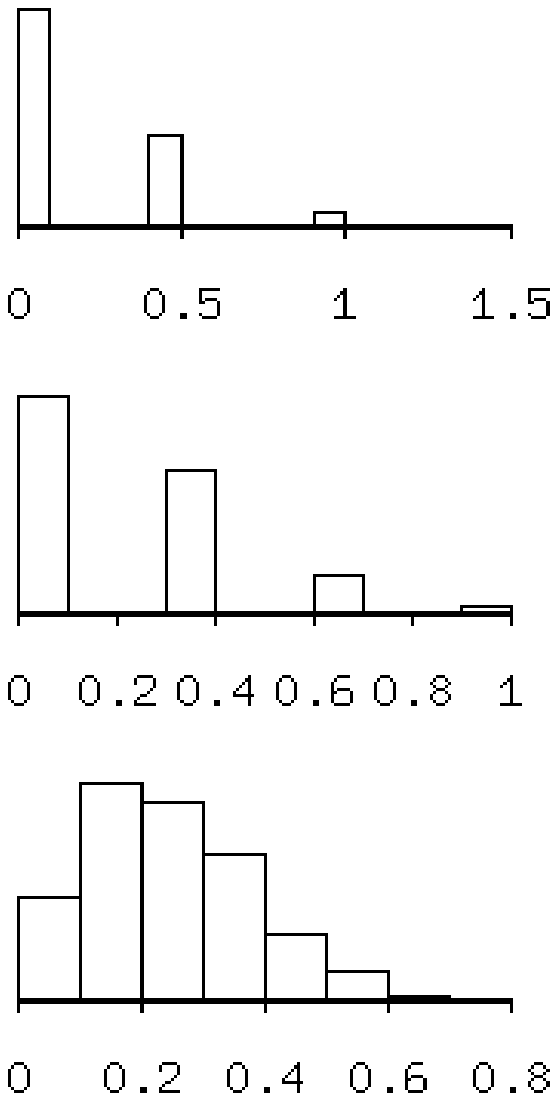

Figure 4.1: Distribution of the Sample Mean for 1000 samples of $n = 2$ (top), $n = 3$ (middle) and $n = 10$ (bottom).

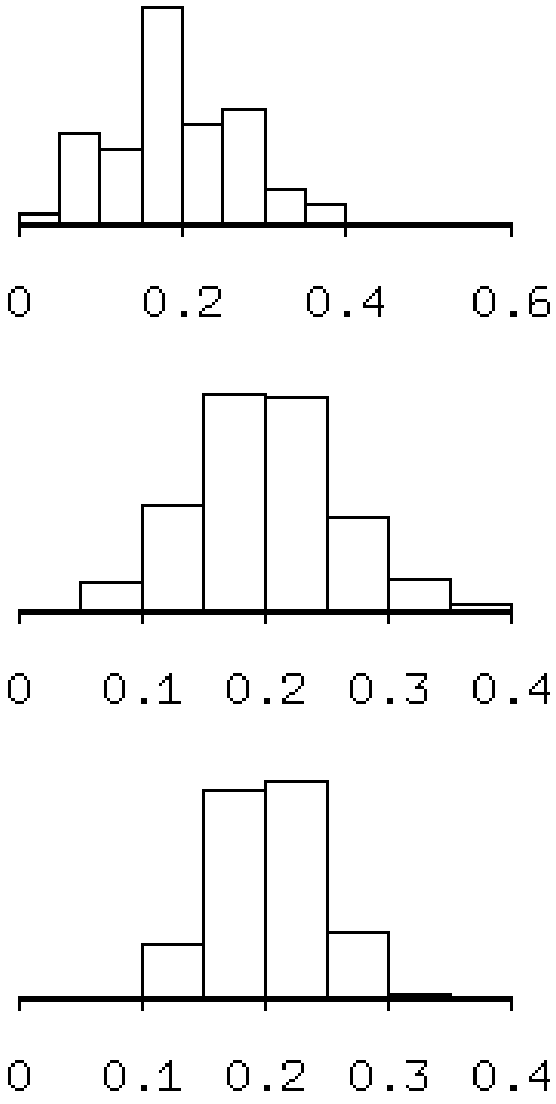

Figure 4.2: Distribution of the Sample Mean for 1000 samples of $n = 30$ (top), $n = 50$ (middle) and $n = 100$ (bottom).

the distribution of the sample means for 1000 samples when $n = 30$ (top), $n = 50$ (middle) and $n = 100$ (bottom). The range of the sample mean again narrows as the sample size increases and the distribution of the sample mean becomes more symmetrical around the population mean, $\mu = .2$.

Figure 4.3 is obtained by superimposing the relative frequencies of the sample means obtained from the 1000 samples of $n = 50$ in the middle panel of Figure 4.2 on a normal probability density function with $\mu = .2$ and $\sigma^2 = 1.8/50 = .0036$. Notice that the sampling distribution of the sample mean does not differ much from the normal distribution when we take account of the fact that the points representing the histogram are the center-points of the tops of its respective bars.

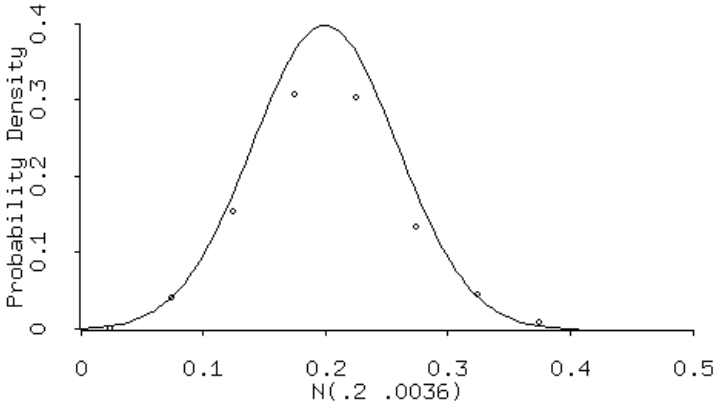

Figure 4.3: Relative frequencies of sample mean from 1000 samples of 50 plotted on normal density function with $\mu = .2$ and $\sigma^2_{\bar{X}} = .0036$.

It turns out that the similarity of the histograms to normal distributions as the sample size increases is not accidental. We have here a demonstration of the *Central Limit Theorem*. The Central Limit Theorem says that when the sample size is sufficiently large the sample mean $\bar{X}$ will become approximately normally distributed with mean equal to the population mean and variance equal to the population variance divided by the sample size. And the larger the sample size, the closer the approximation of the sampling distribution of $\bar{X}$ to a normal distribution. This holds true regardless of the distribution of the population provided it has a finite standard deviation.

The fact that the sample mean is normally distributed for large sample

sizes tells us that if the sample size is large enough the sample mean should lie within one standard deviation of the population mean 68% of the time and within two standard deviations of the population mean 95% of the time. The standard deviation referred to here is, of course, the standard deviation of the sample mean, not the standard deviation of the population.

The true standard deviation of the sample mean is $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. Since the population standard deviation is usually not known, we use

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

to provide an estimate of σ. The standard deviation of the sample mean is thus estimated as

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}.$$

The Central Limit Theorem tells us the approximate nature of the sampling distribution of the sample mean when the sample is large and the distribution of the population is either unknown or the population is not normally distributed. If the population happens to be normally distributed the sampling distribution of the sample mean will turn out to be *exactly* normally distributed regardless of the sample size. This follows from two facts—first, that the mean of a sample from a normally distributed population is a linear function of the population elements in that sample, and second, that any linear function of normally distributed variables is normally distributed.

4.4 Point Estimation

The central purpose of statistical inference is to acquire information about characteristics of populations. An obvious source of information about a population mean is the mean of a random sample drawn from that population. When we use the sample mean to estimate the population mean the sample mean we obtain is called a *point estimate* of the population mean.

In general, suppose there is an unknown population characteristic or parameter that we will denote by θ. To estimate this parameter we select a simple random sample $X_1, X_2, X_3, \ldots, X_n$, from the population and then use some statistic S which is a function of these sample values as a point estimate of θ. For each possible sample we could take we will get a different set of sample values, $X_1, X_2, X_3, \ldots, X_n$, and hence a different S. The statistic S is thus a random variable that has a probability distribution which we

call the sampling distribution of S. We call S an *estimator* of θ. When we take our sample and calculate the value of S for that sample we obtain an *estimate* of θ.

Notice the difference between an estimate and an estimator. An *estimator* is a random variable used to estimate a population characteristic. An actual numerical value obtained for an estimator is an *estimate*.

Consider, for example, a trade association that needs to know the mean number of hourly paid employees per member firm, denoted by μ. To estimate this the association takes a random sample of $n = 225$ member firms (a tiny fraction of the total number of firms belonging to the association). The sample mean $\bar{X}$ is used as an *estimator* of μ. The *estimate* of μ is the particular value of $\bar{X}$ obtained from the sample, say, 8.31.

Note that the sample mean is only one possible estimator of the population mean. We could instead use the sample median or, perhaps, the average of largest and smallest values of X in the sample.

It should be evident from the discussion above that we are using

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

as an estimator of the population standard deviation σ. As an alternative we might think of using

$$\hat{s} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}}.$$

Why should we use $\bar{X}$ rather than, say, the sample median, as an estimator of μ? And why should we use s rather than $\hat{s}$ as an estimator of σ?

4.5 Properties of Good Point Estimators

There are essentially three criteria which we use to select good estimators. The problem that arises, of course, is that a particular estimators may be better than another under one criterion but worse than that other estimator under another criterion.

4.5.1 Unbiasedness

An estimator is *unbiased* if the mean of its sampling distribution is equal to the population characteristic to be estimated. That is, S is an unbiased

estimator of θ if

$$E\{S\} = \theta.$$

If the estimate is biased, the bias equals

$$B = E\{S\} - \theta.$$

The median, for example, is a biased estimator of the population mean when the probability distribution of the population being sampled is skewed. The estimator

$$\hat{s}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

turns out to be a biased estimator of σ^2 while the estimator

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

is unbiased. This explains why we have been using s^2 rather than $\hat{s}^2$.

Unbiasedness in point estimators refers to the tendency of sampling errors to balance out over all possible samples. For any one sample, the sample estimate will almost surely differ from the population parameter. An estimator may still be desirable even if it is biased when the bias is not large because it may have other desirable properties.

4.5.2 Consistency

An estimator is a *consistent* estimator of a population characteristic θ if the larger the sample size the more likely it is that the estimate will be close to θ. For example in the shoe-pair testing example above, $\bar{X}$ is a consistent estimator of μ because its sampling distribution tightens around $\mu = .2$ as n increases. More formally, S is a consistent estimator of population characteristic θ if for any small positive value ϵ,

$$\lim_{n\to\infty}(P(|S - \theta| < \epsilon) = 1.$$

4.5.3 Efficiency

The *efficiency* of an *unbiased* estimator is measured by the variance of its sampling distribution. If two estimators based on the same sample size are both unbiased, the one with the smaller variance is said to have greater *relative efficiency* than the other. Thus, S_1 is relatively more efficient than S_2 in estimating θ if

$$\sigma^2\{S_1\} < \sigma^2\{S_2\} \qquad \text{and} \qquad E\{S_1\} = E\{S_2\} = \theta$$

For example, the sample mean and sample median are both unbiased estimators of the mean of a normally distributed population but the mean is a relatively more efficient estimator because at any given sample size its variance is smaller.

4.6 Confidence Intervals

Point estimates have the limitation that they do not provide information about the *precision* of the estimate—that is, about the error due to sampling. For example, a point estimate of 5 miles per gallon of fuel consumption obtained from a sample of 10 trucks out of a fleet of 400 would be of little value if the range of sampling error of the estimate is 4 miles per gallon—this would imply that the fuel consumption of the fleet could be anywhere between 1 and 9 miles per gallon. To provide an indication of the precision of a point estimate we combine it with an *interval estimate*. An interval estimate of the population mean μ would consist of two bounds within which μ is estimated to lie:

$$L \leq \mu \leq U$$

where L is the *lower bound* and U is the *upper bound*. This interval gives an indication of the degree of precision of the estimation process.

To obtain an estimate of how far the sample mean is likely to deviate from the population mean—i.e., how tightly it is distributed around the population mean—we use our estimate of the variance of the sample mean,

$$s_{\bar{x}}^2 = \frac{s^2}{n}.$$

This enables us to say that if the sample is large enough, $\bar{X}$ will lie within a distance of $\pm 2s$ of μ with probability .95.

Take, for example, the above-mentioned trade-association problem where a random sample of 225 firms was selected to estimate the mean number of hourly paid employees in member firms. Suppose the estimators $\bar{X}$ of μ and s of σ yield point estimates $\bar{X} = 8.31$ and $s = 4.80$. Since the sample size is quite large we can reasonably expect that in roughly 95 percent of such samples the sample mean will fall within $2s/\sqrt{n} = 9.60/15 = .64$ paid employees of μ in either direction. It would thus seem reasonable that by starting with the sample mean 8.31 and adding and subtracting .64 we should obtain an interval [7.67 — 8.95] which is likely to include μ.

If we take many large samples and calculate intervals extending two standard deviations of the sample mean on either side of that sample mean for each sample using the estimates of $\bar{X}$ and $s_{\bar{x}}$ obtained, about 95% of these intervals will bracket μ. The probability that any interval so obtained will bracket μ is roughly .95 (actually .9548).

More formally, consider an interval estimate $L \leq \mu \leq U$ with a specific probability $(1 - \alpha)$ of bracketing μ. The probability that a correct interval estimate (i.e., one that actually brackets μ) will be obtained is called a *confidence coefficient* and is denoted by $(1 - \alpha)$. The interval $L \leq \mu \leq U$ is called a *confidence interval* and the limits L and U are called the *lower* and *upper confidence limits*, respectively. The numerical confidence coefficient is often expressed as a percent, yielding the $100\,(1 - \alpha)\%$ confidence interval.

The confidence limits U and L for the population mean μ with approximate confidence coefficient $(1 - \alpha)$ when the random sample is reasonably large are

$$\bar{X} \pm z\,\frac{s}{\sqrt{n}}$$

where $z = z\,(1 - \alpha/2)$ is the $100\,(1 - \alpha/2)$ percentile of the standard normal distribution. The $100\,(1 - \alpha)$ percent confidence interval for μ is

$$\bar{X} - z\,\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + z\,\frac{s}{\sqrt{n}}$$

Note that the confidence interval does *not* imply that there is a probability $(1 - \alpha)$ that μ will take a value between the upper and lower bounds. The parameter μ is not a variable—it is fixed where it is. Rather, there is a probability $(1 - \alpha)$ that the interval will bracket the *fixed* value of μ. The limits $-z\,(1 - \alpha/2)$ and $z\,(1 - \alpha/2)$ are given by the innermost edges of the shaded areas on the left and right sides of Figure 4.4. The shaded areas each contain a probability weight equal to $\alpha/2$. So for a 95% confidence interval these areas each represent the probability weight $(1 - .95)/2 = .05/2 = .025$ and the sum of these areas represents the probability weight .05. The area under the probability density function between the two shaded areas represents the probability weight .95. Note also that the probability $(1 - \alpha)$ is chosen in advance of taking the sample. The actual confidence interval calculated once the sample is taken may or may not bracket μ. If it does, the confidence interval is said to be correct.

What confidence coefficient should be chosen? This question hinges on how much risk of obtaining an incorrect interval one wishes to bear. In the trade-association problem above the 90, 95, and 99 percent confidence intervals are

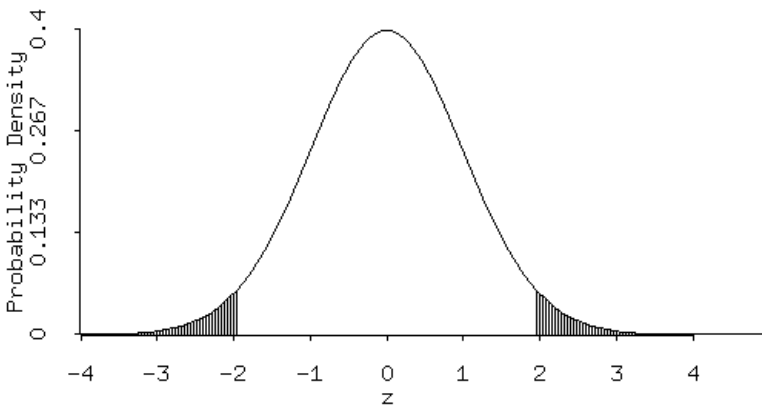

Figure 4.4: The areas $(1-\alpha)$ and $\alpha/2$ (shaded) for a standard normal probability distribution with $\alpha = .05$.

$1-\alpha$	$(1-\alpha/2)$	z	$s_{\bar{x}}$	$zs_{\bar{x}}$	$\bar{X}$	$\bar{X}+zs_{\bar{x}}$	$\bar{X}-zs_{\bar{x}}$
.90	.950	1.645	.32	.5264	8.31	8.84	7.78
.95	.975	1.960	.32	.6272	8.31	8.94	7.68
.99	.995	2.576	.32	.8243	8.31	9.13	7.48

Note that greater confidence in our results requires that the confidence interval be larger—as $(1-\alpha)$ gets bigger, $\alpha/2$ gets smaller and z must increase. We could, of course, narrow the confidence interval at every given level of confidence by increasing the sample size and thereby reducing $s/\sqrt{n}$.

4.7 Confidence Intervals With Small Samples

In making all the above calculations we standardised the sampling distribution of $\bar{X}$, obtaining

$$z = \frac{(\bar{X}-\mu)}{s/\sqrt{n}}$$

and then calculated limits for μ based on values for z in the table of standard normal probabilities. We used s as an estimator of σ. Had we known σ the standardised value would have been

$$z = \frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}} = -\frac{\mu}{\sigma/\sqrt{n}} + \frac{1}{\sigma/\sqrt{n}}\,\bar{X}.$$

Statistical theory tells us that when the population is normally distributed $\bar{X}$ is normally distributed because it is a linear function of the normally distributed X_i. Then the standardised value z is also normally distributed because it is a linear function of the normally distributed variable $\bar{X}$. But when we use s as an estimator of σ the above expression for z becomes

$$z = -\frac{\mu}{s/\sqrt{n}} + \frac{1}{s/\sqrt{n}}\,\bar{X}.$$

Whereas the divisor $\sigma/\sqrt{n}$ is a constant, $s/\sqrt{n}$ is a random variable. This immediately raises the question of the normality of z.

It turns out that the variable

$$\frac{(\bar{X} - \mu)}{s/\sqrt{n}}$$

is distributed according to the t-distribution, which approximates the normal distribution when the sample size is large. The t-distribution is symmetrical about zero like the standardised normal distribution but is flatter, being less peaked in the middle and extending out beyond the standard normal distribution in the tails. An example is presented in Figure 4.5. The t-distribution has one parameter, v, equal to the degrees of freedom, which equals the sample size minus unity in the case at hand. It has mean zero and variance $v/(v-2)$ with $v > 2$.

Because the t-distribution approximates the normal distribution when the sample size is large and because the Central Limit Theorem implies that $\bar{X}$ is approximately normally distributed for large samples, we could use $z = (\bar{X}-\mu)/s_{\bar{x}}$ to calculate our confidence intervals in the previous examples. When the sample size is small, however, we must recognize that $(\bar{X}-\mu)/s_{\bar{x}}$ is actually distributed according to the t-distribution with parameter $v = n-1$ for samples of size n drawn from a normal population. We calculate the confidence interval using the same procedure as in the large sample case except that we now set

$$t = \frac{(\bar{X} - \mu)}{s/\sqrt{n}}$$

and use the appropriate percentile from the t-distribution instead of from the normal distribution.

More formally, we can state that the confidence limits for μ with confidence coefficient $(1 - \alpha)$, when the sample is small and the population is normally distributed or the departure from normality is not too marked, are

$$\bar{X} \pm t\, s_{\bar{x}}$$

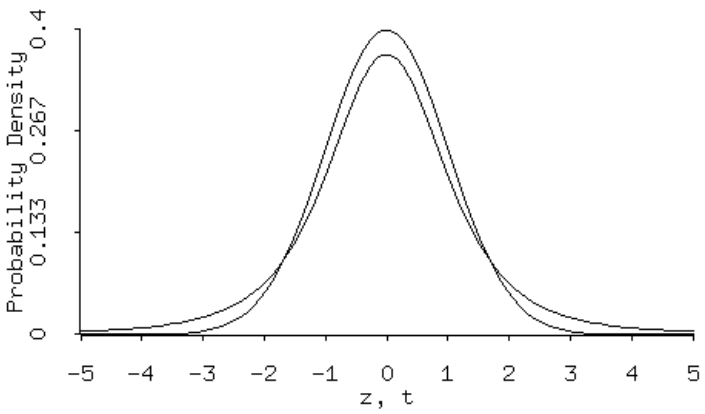

Figure 4.5: A t-distribution compared to the standard normal. The t-distribution is the flatter one with the longer tails.

where $t = t(1 - \alpha/2; n - 1)$. Expressing t in this way means that the value of t chosen will be the one with degrees of freedom $n - 1$ and percentile of the distribution $100(1 - \alpha/2)$.

Now consider an example. Suppose that the mean operating costs in cents per mile from a random sample of 9 vehicles (in a large fleet) turns out to be 26.8 and a value of s equal to 2.5966 is obtained. The standard deviation of the mean is thus $s/3 = .8655$. We want to estimate μ, the mean operating costs of the fleet. For a 90% confidence interval, $t(0.95; 8) = 1.860$. This implies a confidence interval of

$$26.80 \pm (1.8860)(.8655)$$

or

$$25.19 \leq \mu \leq 28.41.$$

Had the normal distribution been used, z would have been 1.645, yielding a confidence interval of

$$26.80 \pm 1.4237$$

or

$$25.38 \leq \mu \leq 28.22.$$

Inappropriate use of the normal distribution would give us a narrower interval and a degree of 'false confidence'.

Notice that the use of the t-distribution requires that the population be normal or nearly so. If the population is non-normal and n is large we can use z and the standard normal distribution. What do we do if the population is non-normal and the sample size is small? In this case we "cross our fingers" and use the t-distribution and allow that the confidence coefficient is now only approximately $1 - \alpha$. This assumes that the t-distribution is *robust*—i.e., applies approximately for many other populations besides normal ones. Essentially we are arguing, and there is disagreement among statisticians about this, that the distribution of $(\bar{X} - \mu)/s_{\bar{x}}$ is better approximated by the t-distribution than the normal distribution when the population is non-normal and the sample size is small.

4.8 One-Sided Confidence Intervals

Sometimes we are interested in an upper or lower bound to some population parameter. For example, we might be interested in the upper limit of fuel consumption of trucks in a fleet. One-sided confidence intervals are constructed the same as two-sided intervals except that all the risk that the interval will not bracket μ, given by α, is placed on one side. We would thus set a single lower confidence interval at $\bar{X} - z(1 - \alpha)s_{\bar{x}}$ instead of $\bar{X} - z(1 - \alpha/2)s_{\bar{x}}$. A single upper-confidence interval is set in similar fashion. Of course, for small samples we would use t instead of z.

4.9 Estimates of a Population Proportion

When the sample size is large the above methods apply directly to point and interval estimation of a population proportion. Suppose that we want to estimate the proportion of voters who will vote yes in the next referendum on whether Quebec should become independent from the rest of Canada. It is natural to take a large sample of voters to determine the sample proportion $\bar{p}$ that are in favour of independence. The Central Limit Theorem tells us that this sample proportion should be normally distributed around the population proportion p if the sample size is large enough. To construct a confidence interval we then need an estimate of the standard deviation of $\bar{p}$. Since the total number of people in the sample voting for independence, X, is distributed according to the binomial distribution with parameters n and p, its variance is $np\,(1 - p)$. The variance of the sample proportion $\bar{p}$ then

equals

$$\begin{aligned} Var\{\bar{p}\} &= Var\{\frac{X}{n}\} = \frac{1}{n^2} Var\{X\} \\ &= \frac{1}{n^2} np(1-p) = \frac{p\,(1-p)}{n}. \end{aligned} \tag{4.5}$$

It is natural to estimate the standard deviation of $\bar{p}$ as the square root of the above expression with $\bar{p}$ substituted for p. When we do so we divide by $n-1$ rather than n. This recognizes the fact that

$$s_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n-1}}$$

turns out to be an unbiased estimator of $\sigma^2_{\bar{p}}$ whereas

$$\tilde{s}_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

is a biased estimator. The $100\,(1-\alpha)$ confidence interval for p therefore becomes

$$\bar{p} \pm z\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}.$$

where z is the value from the standard normal table that will produce the appropriate percentile $100\,(1-\alpha/2)$ for a two-sided confidence interval or $100\,(1-\alpha)$ for a one-sided confidence interval. Suppose that we took a random sample of 1000 voters and found that 350 of them would vote for making Quebec into a separate country. This yields $\bar{p} = .35$ as a point estimate of p. The standard deviation of $\bar{p}$ is estimated to be $\sqrt{(.35)(.65)/999} = .015083$. A two-sided 95% confidence interval for p, for which $z = z(1-\alpha/2) = z(.975) = 1.96$, thus becomes

$$[.35 - (1.96)(.015083)] \leq p \leq [.35 + (1.96)(.025083)]$$

$$.3204 \leq p \leq .3796.$$

4.10 The Planning of Sample Size

If we know the confidence we require in our results we can choose the sample size that will yield that confidence. Resources need not be wasted selecting an excessively large sample while at the same time the risk of choosing an uninformative sample can be avoided. We assume that the sample selected will be reasonably large in absolute value but a small fraction of the population. Let us call the distance between the sample mean and the upper (or lower) confidence limit the half-width (which is half the distance between the upper and lower limits) and denote it by h. The upper limit will then be

$$\bar{X} + h = \bar{X} + z\,\frac{\sigma}{\sqrt{n}}$$

where σ is a value of the population standard deviation picked for planning purposes, so that

$$h = z\,\frac{\sigma}{\sqrt{n}}.$$

Squaring both sides and then multiplying them by n yields

$$n\,h^2 = z^2\sigma^2$$

so that

$$n = \frac{z^2\sigma^2}{h^2}.$$

In formal terms we can thus state that the necessary random sample size to achieve the desired half-width h for the specified confidence coefficient $(1-\alpha)$ for a given planning value of the population standard deviation σ is

$$n = \frac{z^2\sigma^2}{h^2} \tag{4.6}$$

where $z = z(1-\alpha/2)$ and the half-width h represents the deviation of each interval from the sample mean. In the case of a one-sided confidence interval, h would equal the entire interval.

Consider an example. Suppose that a nationwide survey of physicians is to be undertaken to estimate μ, the mean number of prescriptions written per day. The desired margin of error is $\pm.75$ prescriptions, with a 99% confidence coefficient. A pilot study indicated that a reasonable value for the population standard deviation is 5. We therefore have $z = z(1-.01/2) = z(.995) = 2.575$, $h = .75$ and $\sigma = 5$. The proper sample size then equals

$$n = [(2.575)(5)]^2/(.75)^2 = (12.88)^2/.5625 = 165.89/.5625 = 295.$$

The same general principles apply to choosing the sample size required to estimate a population proportion to the desired degree of accuracy. Consider a poll to estimate the results of the next Quebec referendum. How big a sample will we need to estimate the proportion of the voters that will vote for separation to an accuracy of ±2 percentage points, 19 times out of 20? The ratio 19/20 = .95 provides us with $(1 - \alpha)$. We can obtain a planning value of $\sigma_{\bar{p}}$ by noting that $\sqrt{p(1-p)/n}$ will be a maximum when $p = .5$ and using this value of p to obtain the standard deviation of $\bar{p}$ for planning purposes.[1] Thus, a deviation of 2 percentage points or .02 from p must equal $z(1 - \alpha/2) = z(1 - .05/2) = z(.975)$, multiplied by $\sigma_{\bar{p}} = \sqrt{p(1-p)/n} = \sqrt{(.5)(.5)}/\sqrt{n} = .5/\sqrt{n}$. Letting U be the upper confidence limit, we thus have

$$U - \bar{p} = .02 = z\sqrt{\frac{p(1-p)}{n}} = \frac{(1.96)(.5)}{\sqrt{n}} = \frac{.98}{\sqrt{n}},$$

which implies that

$$\sqrt{n} = \frac{.98}{.02} = 49.$$

The appropriate sample size is therefore $(49)^2 = 2401$.

4.11 Prediction Intervals

Sometimes we want to use sample data to construct an interval estimate for a new observation. Consider the earlier problem of determining the operating costs for a vehicle fleet. Having established a confidence interval regarding the operating costs of vehicles in the fleet, we can use the same evidence to help determine whether a particular vehicle not in the sample meets standards.

Suppose that the vehicle in question is selected independently of our earlier random sample of 9 vehicles. Let the operating costs of this vehicle be X_{new}. And suppose that the population (i.e., the operating costs in cents per mile of all vehicles in the fleet) follows a normal distribution.

Now if we knew the values of μ and σ for the population the calculation of a prediction interval would be very simple. We simply obtain a value of z equal to the number of standard deviations from the mean of a normal distribution that would meet our desired level of confidence—that is,

[1] It can be easily seen that $(.4)(.6) = (.6)(.4) = .24 < (.5)(.5) = .25$ and that values of p less than .4 or greater than .6 yield even smaller values for $p(1 - p)$.

$z = z(1-\alpha/2)$, where $100\,(1-\alpha)$ is our desired level of confidence—and calculate $\mu \pm z\,\sigma$. We would predict that $100\,(1-\alpha)\%$ of the time X_{new} will fall in this interval. If X_{new} does not fall in this interval we can send the vehicle in for service on the grounds that the chance is no more than $100\,\alpha/2$ percent (looking at the upper tail) that its cost per mile is equal to or less than the mean for the fleet.

The problem is that we do not know μ and σ and have to use the sample statistics $\bar{X}$ and s as estimators. To calculate the prediction interval we have to know the standard deviation of X_{new}. The estimated variance of X_{new} is

$$\begin{aligned} s^2\{X_{new}\} &= E\{(X_{new}-\mu)^2\} = E\{[(X_{new}-\bar{X})+(\bar{X}-\mu)]^2\} \\ &= E\{(X_{new}-\bar{X})^2\} + E\{(\bar{X}-\mu)^2\} \\ &= s^2 + \frac{s^2}{n} = [1+\frac{1}{n}]\,s^2. \end{aligned}$$

The prediction interval for X_{new} then becomes

$$\bar{X} \pm t\,s\{X_{new}\}$$

where $t = t(1-\alpha/2; n-1)$ is the 'number of standard deviations' obtained from the t-distribution table for the probability weight $(1-\alpha/2)$ and degrees of freedom $(n-1)$. In the case of a vehicle selected from the fleet,

$$\bar{X} \pm t(.975;8)\,s\{X_{new}\} = 26.80 \pm (2.306)\sqrt{(1+1/9)}\,(2.5966)$$

$$= 26.80 \pm (2.306)(1.05409)(2.5966) = 26.80 \pm 6.31$$

which yields

$$20.49 \leq \mu \leq 33.11.$$

Notice that the prediction interval is much wider than the 95% confidence interval for $\bar{X}$ which would be

$$26.80 \pm (2.306)\frac{s}{\sqrt{n}} = 26.80 \pm (2.306)(.8655) = 26.80 \pm 3.1715$$

or

$$23.63 \leq 26.80 \leq 29.97.$$

This is the case because there are two sources of deviation of X_{new} from μ—the deviation from the sample mean, taken as a point estimate of μ, and the deviation of that sample mean from μ. The confidence interval for the sample mean only includes the second source of deviation.

4.12 Exercises

1. Find the following probabilities for the standard normal random variable z:

a) $P(-1 \leq z \leq 1)$

b) $P(-2 \leq z \leq 2)$

c) $P(-2.16 \leq z \leq .55)$

d) $P(-.42 < z < 1.96)$

e) $P(z \geq -2.33)$

f) $P(z > 2.33)$

2. Suppose that a random sample of n measurements is selected from a population with mean $\mu = 100$ and variance $\sigma^2 = 100$. For each of the following values of n, give the mean and standard deviation of the sampling distribution of the sample mean $\bar{X}$.

a) $n = 4$.

b) $n = 25$.

c) $n = 100$.

d) $n = 50$.

e) $n = 50$.

f) $n = 500$.

g) $n = 1000$.

3. A particular experiment generates a random variable X that has only two outcomes: $X = 1$ (success) with probability $p = 0.6$ and $X = 0$ (failure) with probability $(1 - p) = .4$. Consider a random sample consisting of $n = 3$ independent replications of this experiment. Find the exact sampling distribution of the sample mean.

4. Write down the Central Limit Theorem and explain what it means.

5. The mean and standard deviation of a random sample of n measurements are equal to 33.9 and 3.3 respectively.

a) Find a 95% confidence interval for μ if $n = 100$. (33.2532, 34.5468)

b) Find a 95% confidence interval for μ if $n = 400$.

c) What is the effect on the width of the confidence interval of quadrupling the sample size while holding the confidence coefficient fixed?

6. Health insurers and the federal government are both putting pressure on hospitals to shorten the average length of stay of their patients. In 1993 the average length of stay for men in the United States was 6.5 days and the average for women was 5.6 days (*Statistical Abstract of the United States: 1995*). A random sample of 20 hospitals in one state had a mean length of stay for women in 1996 of 3.6 days and a standard deviation of 1.2 days.

a) Use a 90% confidence interval to estimate the population mean length of stay for women in the state's hospitals in 1996.

b) Interpret the interval in terms of this application.

c) What is meant by the phrase '90% confidence interval'?

7. The population mean for a random variable X is $\mu = 40$. The population variance is $\sigma^2 = 81$. For a (large) random sample of size n drawn from this population, find the following:

a) The expected value and the variance of the sample mean $\bar{X}$ when $n = 36$.

b) The probability that $P(\bar{X} \geq 41)$ in the above case.

c) The probability $P(38.5 \leq \bar{X} \leq 40.5)$ when $n = 64$.

8. A number of years ago, Lucien Bouchard and John Charest were in a tough fight for the premiership of Quebec. How big a simple random sample would have been needed to estimate the proportion of voters that would vote for Bouchard to an accuracy of ± 1 percentage points, 19 times out of 20?

9. One of the continuing concerns of U.S. industry is the increasing cost of health insurance for its workers. In 1993 the average cost of health premiums

per employee was $2,851, up 10.5% from 1992 (*Nation's Business*, Feb. 1995). In 1997, a random sample of 23 U.S. companies had a mean health insurance premium per employee of $3,321 and a standard deviation of $255.

a) Use a 95% confidence interval to estimate the mean health insurance premium per employee for all U.S. companies.

b) What assumption is necessary to ensure the validity of the confidence interval?

c) Make an inference about whether the true mean health insurance premium per employee in 1997 exceeds $2,851, the 1993 mean.

10. The mean and the standard deviation of the annual snowfalls in a northern city for the past 20 years are 2.03 meters and 0.45 meters, respectively. Assume that annual snowfalls for this city are random observations from a normal population. Construct a 95 percent prediction interval for next year's snowfall. Interpret the prediction interval.

11. Accidental spillage and misguided disposal of petroleum wastes have resulted in extensive contamination of soils across the country. A common hazardous compound found in the contaminated soil is benzo(a)pyrene [B(a)p]. An experiment was conducted to determine the effectiveness of a treatment designed to remove B(a)p from the soil (*Journal of Hazardous Materials*, June 1995). Three soil specimens contaminated with a known amount of B(a)p were treated with a toxin that inhibits microbial growth. After 95 days of incubation, the percentage of B(a)p removed from each soil specimen was measured. The experiment produced the following summary statistics: $\bar{X} = 49.3$ and $s = 1.5$.

a) Use a 99% confidence interval to estimate the mean percentage of B(a)p removed from a soil specimen in which toxin was used.

b) Interpret the interval in terms of this application.

c) What assumption is necessary to ensure the validity of this confidence interval?

4.13 Appendix: Maximum Likelihood Estimators

The *Maximum Likelihood Method* is a general method of finding point estimators with desirable qualities.

Let us proceed by using an example. Suppose we know that the number of annual visits to a dentist by a child is a Poisson random variable X with unknown parameter λ. In a random sample of two children the numbers of visits to the dentist last year were $X_1 = 0$ and $X_2 = 3$.

The idea of maximum likelihood is to choose the value for λ for which it is most likely that we would observe the sample $\{X_1, X_2\}$. We do this by calculating the probability of observing the sample for various values of λ—say, 0, 1, 1.5, 2, 3, etc.—and picking the value of λ that maximizes this probability. The Poisson probability function, defined in equation (3.32), is

$$P(x) \quad = \quad \frac{\lambda^x \, e^{-\lambda}}{x!}.$$

Since the observations are independent of each other, the probability of observing the sample $\{X_1, X_2\}$ is $P(x = X_1)$ times $P(x = X_2)$. From the table of Poisson probabilities we obtain the following probabilities for various values of λ:

λ	$P(x = 0)$	$P(x = 3)$	$P(x = 0)P(x = 3)$
0.0	.0000	.0000	.0000
1.0	.3679	.0613	.0225
1.5	.2231	.1255	.0280
2.0	.1353	.1804	.0244
3.0	.0498	.2240	.0112

The value of λ that maximizes the likelihood of observing the sample in the above table is $\lambda = 1.5$.

We could calculate $P(x = 0)P(x = 3)$ for values of λ between the ones in the table above and plot them to obtain the smooth curve in Figure 4.6. This curve maps the probability density as a function of λ which is called the *likelihood function*. It confirms that 1.5 is the maximum likelihood estimate of λ.

Let us now approach the problem more formally and suppose that we have a set of sample observations X_i from which we want to estimate a parameter θ. There is some probability

$$P(X_1, X_2, X_3, \ldots, X_n; \theta)$$

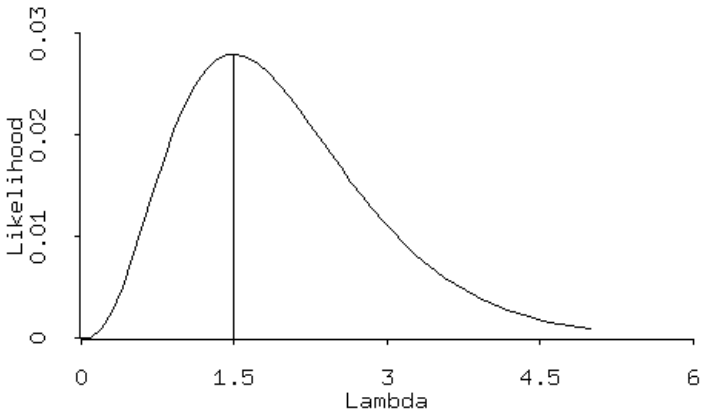

Figure 4.6: The likelihood function for the children-to-the dentist example.

of drawing a particular sample of observations, given the magnitude of the unknown parameter θ. Because the sample observations $X_1, X_2, X_3, \ldots, X_n$ are independent, this probability function equals

$$P(X_1, X_2, X_3, \ldots, X_n; \theta) = P(X_1; \theta)P(X_2; \theta)P(X_3; \theta) \ldots P(X_n; \theta).$$

This product of probabilities, when viewed as a function of θ for given $X_1, X_2, X_3, \ldots, X_n$ is called the *likelihood function*

$$L(\theta) \quad = \quad P(X_1; \theta)P(X_2; \theta)P(X_3; \theta) \ldots P(X_n; \theta). \tag{4.7}$$

We find the value of θ that maximizes $L(\theta)$ either by analytic methods or, when that approach is not feasible, by efficient numerical search procedures.

Consider a Poisson process with unknown parameter λ and select a random sample $X_1, X_2, X_3, \ldots, X_n$. Using the formula for the Poisson probability distribution, the likelihood function can be expressed

$$\begin{aligned} L(\theta) &= \left[\frac{\lambda^{X_1} e^{-\lambda}}{X_1!}\right]\left[\frac{\lambda^{X_2} e^{-\lambda}}{X_2!}\right] \ldots\ldots \left[\frac{\lambda^{X_n} e^{-\lambda}}{X_n!}\right] \\ &= \left[\frac{\lambda^{\sum X_i} e^{-n\lambda}}{X_1! X_2! \ldots X_n!}\right] = \left[\frac{\lambda^{n\bar{X}} e^{-n\lambda}}{X_1! X_2! \ldots X_n!}\right]. \end{aligned} \tag{4.8}$$

To maximize $L(\lambda)$ we differentiate it with respect to λ and find the value for λ for which this differential is zero. Differentiating (using the chain rule

whereby $dxy = xdy + ydx$) we have

$$\begin{aligned}
\frac{dL(\theta)}{d\theta} &= \frac{1}{X_1!X_2!\ldots X_n!}\left[\frac{d}{d\lambda}\left(\lambda^{n\bar{X}}\, e^{-n\lambda}\right)\right] \\
&= \frac{1}{X_1!X_2!\ldots X_n!}\left[\lambda^{n\bar{X}}\frac{d}{d\lambda}\left(e^{-n\lambda}\right) + e^{-n\lambda}\frac{d}{d\lambda}\left(\lambda^{n\bar{X}}\right)\right] \\
&= \frac{1}{X_1!X_2!\ldots X_n!}\left[-\lambda^{n\bar{X}}\, e^{-n\lambda}\, n + e^{-n\lambda}\, n\bar{X}\, \lambda^{n\bar{X}-1}\right] \\
&= \frac{1}{X_1!X_2!\ldots X_n!}\left[n\left(\frac{\bar{X}}{\lambda} - 1\right)\left(\lambda^{n\bar{X}}\, e^{-n\lambda}\right)\right] \qquad (4.9)
\end{aligned}$$

This expression equals zero—i.e., $L(\lambda)$ is a maximum—when

$$\left[\frac{\bar{X}}{\lambda} - 1\right] = 0,$$

which occurs when $\lambda = \bar{X}$. Thus, the sample mean is a maximum likelihood estimator of λ for a random sample from a Poisson distribution. In the children-to-dentist example above, the sample mean is $(0 + 3)/2 = 1.5$, the value of λ that produced the largest value for $L(\lambda)$ in Figure 4.6.