Contents

Chapter 8

Estimation

8.1 Introduction

Estimation uses data from a sample to estimate characteristics of a population. Inferential statistics is comprised of estimation and hypothesis testing, so that estimation is one of the main parts of inferential statistics. Estimation provides researchers with a means of making either a point or an interval estimate of a population characteristic. Sampling error was introduced in Chapter 7, showing how inferences concerning the difference between the sample mean and population mean could be obtained. From sampling error, it is only a short step to providing an interval estimate of a population parameter. To each interval estimate, a certain probability is attached, in much the same way that each level of sampling error has a probability. The method of estimation can be used to provide an estimate of a population mean, a population proportion, or some other summary measure of a population.

Interval estimates are often referred to as **confidence intervals**, based on the probability, or degree of confidence, associated with the interval. Suppose large random samples are drawn from a population, and a sample mean is obtained from each sample. Around each sample mean an interval can be constructed, with the sample mean being at the centre of each interval. For example, a 95% confidence interval for a population mean can be constructed using the methods of this chapter. The meaning of the 95% confidence interval is that 95 out of 100 random samples from a population yield interval estimates which contain the true population mean.

The method of estimation can be further extended to to providing es-

timates of the sample size required in order to achieve a given degree of accuracy when estimating a population parameter. Again, this accuracy will have a certain probabilty attached to it. In this chapter this method of determining sample size will be provided for random sampling only.

Chapter Outline. An introduction to the nature of estimation is contained in the next section. This is followed in Section 3 by a discussion of the method of estimating a sample mean using a large random sample. The method of estimation of a population parameter for a small random sample is discussed next. Estimation using a small sample size requires use of the t-distribution, and this distribution is introduced in Section 4. Section 5 shows how interval estimates of population proportions can be determined. The method of determining the required sample size in order to provide an estimate of the population mean or population proportion is contained in Section 5. These are the sample sizes necessary to provide estimates having a specified accuracy and probability.

8.2 Point and Interval Estimates

Estimation is a method of providing the researcher with an idea of the size of a summary measure concerning a population. The summary measure can be any parameter concerning a population. Since much statistical work concerns the common measures of central tendency and variation discussed in Chapter 5, these are the measures which the researcher is usually concerned with estimating. In Chapter 11, the method of estimation is extended to provide estimates of the nature of the relationship among two or more variables.

Estimates of population parameters can be either point estimates or interval estimates. A **point estimate** is a single number, or a statistic, which provides an estimate of a population parameter. For example, the Gallup opinion poll found that the percentage of Canadian adults who supported the Conservative Party in August, 1992 was 21%. This sample value of 21% is a point estimate of the true percentage of all Canadian adults who supported the Conservative Party in August, 1992. In Chapter 7, a random sample of 601 Regina labour force members was given as an example. This sample found that the mean gross monthly pay of these 601 respondents was $2,205 per month. This sample mean is a point estimate of the true mean gross monthly pay of all Regina labour force members.

Measure	Parameter	Point Estimate
Mean	μ	$\bar{X}$
Variance	σ^2	s^2
Standard Deviation	σ	s
Proportion	p	$\hat{p}$

Table 8.1: Parameters and Point Estimates

Ordinarily a point estimate is the statistic which corresponds to the population parameter. That is, the sample mean is used as a point estimate of the true mean, the sample proportion is a point estimate of the true proportion, and so on. Table 8.1 gives the commonly used population parameters and their respective point estimates.

It is possible that a point estimate of a population parameter will be a number other than the corresponding statistic of Table 8.1. For example, if the distribution of a variable is close to symmetric, the sample median might be used to provide an estimate of the population mean. It may be that information for determining the median is available, but the data necessary to calculate the mean is not available. Such a situation is not the most desirable, because the sample mean is generally regarded as a better estimator of the population mean than is the sample median. But where such situations emerge, it may be necessary to use point estimates different from those given in Table 8.1.

Statisticians have devised various properties of estimators in order to determine which are the best estimators for each population parameter. In random sampling, the point estimates of Table 8.1 are generally regarded as the best estimates of the corresponding population parameters.

When reporting summary measures of populations, point estimates may be the only data provided. In the popular media, reports of characteristics of populations are unlikely to give more than means or proportions. These are often sufficient to provide a good description of the main characteristics of a population. However, where these characteristics are obtained from

samples, each statistic has some sampling error associated with it. That is, each random sample of a population yields a slightly different set of cases in the sample, and this produces a different sample mean or proportion in each sample. This sampling distribution is used to construct an interval estimate. The interval estimate constructed gives the researcher an idea of how variable these statistics from samples are, and where the value of the population seems likely to lie.

An **interval estimate**, or a **confidence interval**, is an interval constructed around the point estimate, or the statistic. This interval is based on the sampling distribution of the statistic. Each interval has a probability, or a **confidence level**, associated with it. Since the sampling distribution assumes that random samples of a population have been obtained, these confidence intervals are also based on the assumption of random sampling from the population. The interval estimate is interpreted as meaning that there is a certain confidence level, or probability, that these intervals contain the true value of the population parameter.

A description of, and rationale for, the construction of an interval estimate for the mean is given in the following section. The interpretation of what this interval estimate means is also provided. The examples of interval estimates should help you understand how this method can be used to make statements concerning the population parameters.

Note: In the following discussion, the terms **interval estimate** and **confidence interval** are used interchangeably. Each interval has a particular confidence level attached to it, so that the interval may be referred to as a confidence interval. This confidence interval could more fully be called a **confidence interval estimate**.

8.3 Interval Estimate of the Mean.

An interval estimate of the population mean begins by taking a random sample of the population for which the mean is to be estimated. Suppose the characteristic of the population which is being investigated is measured by variable X, and the true mean and standard deviation of X in the population are μ and σ respectively. Neither μ nor σ are known, and the shape of the distribution of the variable X is also unknown. This is why the researcher must obtain a sample, and attempt to provide estimates of characteristics of a population.

A random sample of size n gives the researcher a set of randomly selected values of the variable, $X_1, X_2, X_3, \cdots, X_n$. From these n values, the sample mean $\bar{X}$ can be computed as

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}.$$

In addition, the sample standard deviation can be computed as

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

Since each random sample which is drawn from this population produces a different set of X_is, each random sample has a different $\bar{X}$ associated with it. Each of the sample means $\bar{X}$ can be regarded as a point estimate of the true populatin mean μ. The Central Limit Theorem of Chapter 7, showed that if the sample sizes of these random samples are reasonably large, then

$$\bar{X} \text{ is Nor } (\mu, \frac{\sigma}{\sqrt{n}}).$$

That is, the standard deviation of the sample means is $\sigma/\sqrt{n}$, and the sampling distribution of $\bar{X}$ is a normal distribution. This distribution formed the basis for the discussion of sampling error earlier, and this same distribution is now used to present interval estimates.

The idea of a confidence interval estimate of the population mean μ is to construct an interval around the sample mean $\bar{X}$. The interval is constructed so that it is wide enough that there is reasonable confidence that this interval does contain the true mean μ. The level of confidence is ordinarily quite a high level, such as 90%, 95% or 99%. For this introductory discussion, the 95% confidence level, the most commonly used confidence level, will often be used as an example.

In this chapter, the confidence level will be given the symbol C, or $C\%$, so that this explanation will be based on $C = 95\%$. The meaning of a confidence level is that $C\%$ of the random samples will yield intervals such that μ is within the interval. This means that occasionally a random sample will be selected such that the interval associated with the sample mean from this sample does not contain μ. But if $C\%$ of the intervals contain μ, then only $(100 - C)\%$ of of the intervals do not contain μ. In the case of a $C = 95\%$ confidence level, only $100 - 95 = 5\%$ of the random samples will have intervals which do not contain μ.

The confidence level is really a probability. That is, if many random samples are selected from a population with mean μ, the probability is $C/100$ that a interval contains the population mean μ. In the case of the 95% interval estimate, the probability is $95/100 = 0.95$ that a random sample of the population yields an interval which contains μ. The probability is only $1 - 0.95 = 0.05$ that a random sample is selected which yields an interval which does not contain μ. This is why the confidence level is set at quite a high level. By selecting a high confidence level, the probability is quite large that a sample will be selected which does contain the true mean μ.

The normal distribution for $\bar{X}$ is used to determine the size of the interval. For $C = 95\%$ confidence, 95% of the area under the normal distribution lies within 1.96 standard deviations of the mean. That is, the middle 95% of the area under the normal curve is equivalent to an area of 0.95 in the centre, or $0.95/2 = 0.475$ on each side of the centre of the normal curve. From column A of Appendix A, this area is associated with $Z = 1.96$. Since the distribution is symmetric, going out 1.96 standard deviations on each side of centre gives the middle 0.95 or 95% of the area under the normal curve, that is, $Z = 1.96$ is the Z value associated with the 95% confidence level.

Now imagine an interval is constructed around $\bar{X}$, going out 1.96 standard deviations on each side of $\bar{X}$. Since the standard deviation of $\bar{X}$ is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

a distance of 1.96 standard deviations on either side of $\bar{X}$ is

$$\pm 1.96 \frac{\sigma}{\sqrt{n}}.$$

The interval estimate around $\bar{X}$ is thus

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

or expressed in interval form this is

$$\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \ , \ \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

The 95% confidence interval estimate of μ is thus $\bar{X}$ plus or minus 1.96 times $\sigma_{\bar{X}}$, the standard deviation of $\bar{X}$.

This interval, from

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}$$

to

$$\bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

has a probability of 0.95 of containing μ. The proof of this is given in Section 8.3.3. The method of constructing and interpreting this interval is discussed in the following paragraphs.

Figure 8.1 provides a diagrammatic illustration of the 95% interval estimate of the mean. Two diagrams are presented in order to avoid cluttering a single diagram. The diagram at the top of the page gives the sampling distribution of the sample mean, and shows the area associated with the confidence level of 95%. The area under the normal curve between $Z = -1.96$ and $Z = +1.96$ is 0.95, or 95% of the area under the normal curve. Since the standard deviation of $\bar{X}$ is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

a distance of 1.96 standard deviations on either side of the mean is equivalent to a distance of $1.96(\sigma/\sqrt{n})$ on either side of μ, in units of $\bar{X}$.

Now suppose a random sample of size n is taken from the population. Also suppose that the particular value of $\bar{X}$ which occurs is $\bar{X}_1$ in the diagram at the bottom of the page. Remember that μ is unknown, and μ is shown on the diagram only to illustrate the manner in which the interval does provide an estimate of the true population mean. As can be seen in the diagram at the bottom, $\bar{X}_1$ is close to μ but not exactly equal to μ. The method of interval estimation is to begin with $\bar{X}_1$ and construct an interval around $\bar{X}_1$. The 95% interval goes out from $\bar{X}_1$ a distance of $1.96 \times \sigma_{\bar{X}}$ on either side. This is the interval constructed around $\bar{X}_1$ which has the same width as the interval in the top diagram. For the 95% confidence interval, this is the interval associated with an area of 0.95 in the middle of the distribution. From the diagram, and this explanation, the interval from $\bar{X}_1 - 1.96\sigma_{\bar{X}}$ to $\bar{X}_1 + 1.96\sigma_{\bar{X}}$ in the distribution of $\bar{X}$ can be seen to be the same width as the interval between $Z = -1.96$ and $Z = +1.96$ in the standardized normal distribution.

Based on these diagrams, it can be seen that interval estimation for the mean begins by selecting the confidence level. The Z value which corresponds to this confidence level is then determined. The sample mean is

Figure 8.1: 95% Interval Estimate for the Population Mean

obtained from the random sample. The interval estimate is an interval centred at the sample mean, going out from this mean Z times the standard deviation of the sample mean. This interval has probability equal to the confidence level of containing the true mean.

Confidence Limits. When estimating the population mean, the limits on the interval are $\bar{X}_1 - 1.96\sigma_{\bar{X}}$ and $\bar{X}_1 + 1.96\sigma_{\bar{X}}$. The first of these values $\bar{X}_1 - 1.96\sigma_{\bar{X}}$ is sometimes called the **lower confidence limit** of the estimate and $\bar{X}_1 + 1.96 \times \sigma_{\bar{X}}$ is termed the **upper confidence limit** of the estimate. Together these two values are called the **confidence limits** of the interval estimate.

In order to illustrate how an interval estimate is constructed, an example is given next. Following this example, more of the principles of interval estimation are discussed.

Example 8.3.1 Study Patterns of University of Regina Students

A survey of 494 undergraduate students at the University of Regina, conducted by a student in Sociology 404 in the Winter 1988 semester, showed that the mean number of hours students spent studying per week was 18.8 hours with a standard deviation of 13.1 hours. Assume that the survey is a random sample of all University of Regina undergraduate students. Obtain the 90 per cent and the 99 per cent interval estimate for the mean number of hours studied per week for all University of Regina undergraduate students.
Solution. *In order to determine this interval, let μ be the true mean number of hours studied per week for all University of Regina undergraduate students. Let σ be the true standard deviation of the number of hours per week studied by all University of Regina undergraduates. Since μ is unknown, the sample mean $\bar{X}$ is used as a point estimate of μ, and $\bar{X} = 18.8$ hours per week is the best estimate of the mean weekly study time of undergraduates.*

Since $n = 494$ is a large random sample of University of Regina students, $\bar{X}$ is normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. Since σ is unknown, $s = 13.1$ is used as an estimate of σ, again since n is large. For the 90 per cent interval estimate, the appropriate Z value is 1.645 since an area of 90% in the middle of the normal distribution is associated with $Z = 1.645$. That is, if 90% or 0.90 of the area is in the middle, there is $0.90/2 = 0.45$ of the area on each side of centre. From column A of the normal table in Appendix A, an area of 0.45 on one side of centre is midway

between $Z = 1.64$ and $Z = 1.65$, so that a Z of 1.645 is used here. The interval estimate is then:

$$\bar{X} \pm Z\frac{s}{\sqrt{n}} = 18.8 \pm 1.645 \times \frac{13.1}{\sqrt{494}}$$

$$18.8 \pm (1.645)(0.589) = 18.8 \pm 0.970$$

The 90 per cent interval estimate is thus (17.8, 19.8) if all the values are rounded to the nearest one tenth of an hour. Using the survey result, we can be relatively certain that the true mean is somewhere between 17.8 and 19.8 study hours per week.

For the 99 per cent interval estimate, there is 99% or 0.99 of the area in the middle of the normal distribution, so that there is $0.99/2 = 0.495$ on each side of centre. From column A of the normal table, the appropriate Z value is 2.575. The interval estimate is then:

$$\bar{X} \pm Z\frac{s}{\sqrt{n}} = 18.8 \pm 2.575 \times \frac{13.1}{\sqrt{494}}$$

$$18.8 \pm (2.575)(0.589) = 18.8 \pm 1.52$$

The 99 per cent interval estimate is thus (17.3, 20.3) if the values are rounded to the nearest 0.1 hour.

Note that this 99% interval is somewhat wider than the 90% interval because it is necessary to go farther from $\bar{X}$ in order to be 99% confident that the interval contains μ. The meaning of these interval will become clearer in the following sections, but what can be said is that the probability is 0.90 that $\bar{X} \pm 1.0$ contains μ. In terms of the sampling error of Chapter 7, the probability is 0.90 that the sampling error of the mean in this sample does not exceed 1.0 hours per week. That is, a random sample of 494 undergraduates has a probability of 0.90 of yielding a sample mean which differs by less than 1.0 hours per week from the true mean. In addition, the probability is 0.99 that the estimate of the true mean is incorrect by no more than 1.5 hours per week.

Meaning of the Interval. The method of confidence intervals ensures that C% of the random samples taken from a population will yield interval estimates which contain the true mean. In Figure 8.1, 95% of the random samples will yield sample means between $\mu - 1.96\sigma_{\bar{X}}$ and $\mu + 1.96\sigma_{\bar{X}}$. Each sample mean within this range will produce an interval estimate such that

the interval contains μ. Thus 95% of the intervals constructed in this manner contain μ.

It is possible that a random sample from the population yields an interval which does not contain μ. In the bottom part of Figure 8.1, sample mean $\bar{X}_2$ lies a considerable distance from μ. If the interval

$$(\bar{X}_2 - 1.96\sigma_{\bar{X}} \ , \ \bar{X}_2 + 1.96\sigma_{\bar{X}})$$

is constructed, it can be seen that this interval will not contain μ. The value of $\bar{X}_2$ is so distant from μ, that even an interval of 1.96 standard deviations on either side of this sample mean does not contain μ. By examining the top diagram though, it can be seen that this will not happen very often. Only when $\bar{X} < \mu - 1.96\sigma_{\bar{X}}$ or when $\bar{X} > \mu + 1.96\sigma_{\bar{X}}$ will an interval of this width not contain μ. From the diagram, it can be seen that this will happen only in 5% of the random samples. 95% of the random samples are within these limits of 1.96 standard deviations on each side of the mean. Only a total of $100 - 95 = 5\%$ of the sample means fall more than 1.96 standard deviations from the true mean μ.

The meaning of the interval estimate should become clearer when the situation in Figure 8.2 is examined. This figure uses a few of the random samples drawn from the population of Regina Labour Force Survey Respondents presented in Section **??** of Chapter 7. The true mean gross monthly pay of this population of respondents is $\mu = \$2,352$ and the standard deviation is $\sigma = \$1,485$. The random samples selected from this population were each of size $n = 50$. The theoretical sampling distribution of the sample means $\bar{X}$, based on the Central Limit Theorem, is

$$\bar{X} \text{ is Nor } \left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Given the above mean and standard deviation of the population of Regina Labour Force Survey Respondents, with a random sample of $n = 50$, the sampling distribution of the sample mean is

$$\bar{X} \text{ is Nor } (\$2,352 \ , \ \$210)$$

as shown in Section 7.5. This normal distribution is given at the top of Figure 8.2. In addition, the area under the normal curve associated with the 95% confidence intervals is given. This is the area under the curve with 1.96 standard deviations of the mean, or between $Z = 1.96$ and $Z = +1.96$.

Since the population standard deviation is $\sigma = 1,485$ and $n = 50$, the standard deviation of the sampling distribution of $\bar{X}$ is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1,485}{\sqrt{50}} = 210.$$

The bottom part of Figure 8.2 shows the interval estimates associated with the first 8 random samples given in Table 7.5, and sample number 158. The interval estimates of μ are constructed around $\bar{X}$ plus or minus 1.96 standard deviations of $\bar{X}$, that is

$$\pm 1.96 \frac{\sigma}{\sqrt{n}} = 1.96 \times 210 = 411.6$$

or \$412 on each side of $\bar{X}$.

The first random sample of $n = 50$ respondents gave a sample mean of $\bar{X}_1 = \$2,205$. The interval estimate around $\bar{X}_1$ is thus

$$\bar{X}_1 \pm 1.96 \frac{\sigma}{\sqrt{n}} = 2,205 \pm (1.96 \times 210) = 2,205 \pm 412.$$

The 95% interval estimate associated with $\bar{X}_1 = \$2,205$ is thus $2,205 - 412 = 1,793$ to $2,205 + 412 = 2,617$ or \$1,793 to \$2,617.

The interval is usually reported with the lower confidence limit first, and the upper confidence limit second, placed in brackets with a comma between the limits.

Thus the 95% confidence interval for the mean using the sample mean from the first sample is

$$(\$1,793\ ,\ \$2,617).$$

This is the first interval pictured in the bottom section of the diagram. It can be seen that this interval does contain μ, so that this interval is one of the 95% of those which contain the true mean.

The second random sample in Table 7.5 yielded a sample mean of $\bar{X}_2 = \$2,641$. The method of constructing the 95% confidence interval for this sample mean is exactly the same as for the first sample mean. The distance of 1.96 standard deviations on each side of the sample mean is ± 412. The interval for this sample is

$$\bar{X}_2 \pm 412 = 2,641 \pm 412$$

or $2,641 - 412 = 2,229$ to $2,641 + 412 = 3,053$. For sample 2, the 95% confidence interval is

$$(\$2,229, \$3,053).$$

Figure 8.2: 9 Sample Means and their Corresponding Interval Estimates

The sample mean $\bar{X}_2$ is larger than the population mean, but once again the interval contains μ. Random sample 2 is one of the 95% of random samples which leads to an interval estimate for μ which contains μ.

The same method is used for each of the successive sample means. For sample 3, $\bar{X}_3 = \$2,128$ and the 95% interval estimate is

$$(\$1,716\ ,\ \$2,540).$$

The intervals for the next 5 samples of Table 7.5 are given below this. Note how the sample mean, and the associated interval, differs in each case. But each of the first 8 sample means is close enough to the true mean μ that each interval estimate contains μ. The sample which is farthest from the true mean is sample 7 where $\bar{X}_7 = \$2,703$. But even this sample has an interval (\$2,291 , \$3,115) which contains μ near the lower end of this interval.

In order to see that some random samples do result in intervals which do not contain μ, take sample number 158 where $\bar{X}_{158} = \$1,863$. Using the same method as earlier, the 95% interval estimate for this sample is

$$(\$1,451\ ,\ \$2,275)$$

and this interval does not contain μ. This whole interval lies to the left of the true mean μ. If all the random samples in Table **??** are examined, it can be seen that only samples 55, 59, 65, 102, 107, 158 and 171 have 95% interval estimates which do not contain μ. This amounts to only 7 out of 192 random samples, or about 3.6% of the random samples. Just over 96% of the samples yield 95% interval estimates which contain the true mean.

The fact that approximately 95% of the random samples in Table 7.5 contain the true mean demonstrates the meaning of the 95% confidence interval. This method of random sampling and construction of interval estimates results in the intervals containing the true mean in C% of the cases. Only (100-C)% of the samples will be associated with C% confidence intervals which do not contain the true population mean.

Since the true mean of the population is not known, and since only one sample is usually taken, the researcher cannot be absolutely certain that the confidence interval contains the mean. At the same time, the researcher can be reasonably certain that the interval does contain the true population mean, because C% of the random samples will yield C% confidence intervals which contain the true mean.

Note that the statement concerning a particular interval is not a probability statement. In general, 95% of the intervals contain μ. A particular

interval that is constructed either contains μ or it does not contain μ. For example, in Figure 8.2 each of samples 1-8 contains μ, and only sample 158 was associated with an interval which did not contain μ. When dealing with a specific random sample, since μ is unknown, the researcher does not know whether the interval construted from the sample mean contains μ or not. This uncertainty is always associated with an interval estimate, and there is no way to quantify this uncertainty for a specific interval estimate. However, the uncertainty can be quantified in a probability interpretation for interval estimates in general. That is, for the 95% confidence level, it is meaningful to say that 95% of the random samples will yield interval estimates which contain the true mean μ. A probability statement can also be attached to these intervals. That is, with 95% confidence level, the probability is 0.95 that

$$\bar{X} \pm 1.96\sigma_{\bar{X}}$$

contains μ. If the size of $\sigma_{\bar{X}}$ can be determined, then this can be included in the probability statement as follows. In the example of the estimate of mean gross monthly pay, $\sigma_{\bar{X}} = 210$ so that $1.96\sigma_{\bar{X}} = 412$. It is then meaningful to say that the probability is 0.95 that

$$\bar{X} \pm 412$$

contains μ. This is alternatively stated as

$$P(\bar{X} \pm 412 \text{ contains } \mu) = 0.95$$

But when a specific value for $\bar{X}$ is given, and the specific confidence interval is determined, then either μ is in this interval or it is not. A probability statement should not be attached to a specific interval, such a statement should only be used for the general case.

Confidence Levels. Most of the discussion so far has used a 95% confidence level. Any other level of confidence, C%, could have been used instead of this. As noted earlier, the level of confidence is usually a large value, close to 100%. Confidence levels of 90% and 99% are also common, but levels such as 88% or 93% could be used. The level of confidence is often given in the problem or question, or may be assigned by those commissioning the research.

The choice of confidence level is arbitrary, but is chosen as a level close to 100% so that there is a high degree of confidence that the interval does

contain the population parameter being investigated. The level of confidence can never be 100%, because with random sampling, there is always some degree of uncertainty concerning where the value of the parameter is. An interval estimate could be 100% certain only if the very general statement is made that the parameter is between $-\infty$ and $+\infty$.

The Central Limit Theorem can be used when making estimates of the population mean from random samples of reasonably large sample size. The sample mean is normally distributed so that the appropriate Z associated with each confidence level can be obtained from the normal table in Appendix A. The method used is the same as that used earlier in this chapter. That is, if the confidence level is C%, this is the area in the middle of the normal distribution. The Z value associated with $C/2$ of the area on each side of centre is the appropriate Z value for the interval.

Some of the more common confidence levels, and the Z value which corresponds to each of these levels are given in Table 8.2. You can confirm each of these Z values by checking them in the table of the normal distribution in Appendix A.

Confidence Level	Z value
80%	1.28
85%	1.44
90%	1.645
95%	1.96
99%	2.575
99.5%	2.81

Table 8.2: Confidence Levels and Z Values

Unknown Population Standard Deviation. As noted in Chapter 7, the value of the standard deviation of the population is required in order to determine the standard deviation of the distribution of the sample mean. If the Central Limit Theorem holds, then

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

is the standard deviation of the sampling distribution of $\bar{X}$. Since μ is being estimated, and determination of σ requires knowledge of μ, it appears that there is no way of estimating σ. What is usually done when estimating μ from random samples is to collect the data and compute the sample standard deviation s at the same time that $\bar{X}$ is being calculated. If the sample size is reasonably large, s is generally regarded as providing a fairly close estimate of σ. While s will not be exactly equal to σ, it can be shown mathematically that s is a good estimator or σ. As n becomes larger, the sample standard deviation s will become closer and closer to σ. As with many of the approximations used in statistics, a larger n is associated with a closer approximation. As a rough rule of thumb, a random sample of size $n \geq 30$ is usually regarded as adequate in this approximation.

Note though that this approximation is potentially another source of error when making estimates of μ. If a random sample is composed of many atypical cases, with a value of $\bar{X}$ which is considerably different than μ, it may be that s is not a good estimate of σ. While this should be remembered when conducting estimates, where the interval estimate is intended to provide only an approximate idea of the error associated with sampling, a small error in s may not be very misleading.

8.3.1 Interval Estimates of the Mean, Large n

The discussion of the previous section outlined the method by which estimation is carried out. This section quickly reviews the rationale of estimation, outlines in point form the steps required to obtain estimates, and provides some examples of confidence intervals.

Estimation begins, and may end, with a point estimate. Suppose a researcher has conducted a survey and has a random sample from a population with unknown mean μ and standard deviation σ. Then the sample mean $\bar{X}$ is a point estimate of μ, and s can also be regarded as a point estimate of σ. If the sample size n in this sample is small, then the small sample method of Section 8.4 should be consulted. If the sample size n of the random sample is reasonably large, then $\bar{X}$ provides a good point estimate of μ.

Any time that estimates of population parameters are obtained from samples, even very well constructed random samples, there will be some sampling error associated with the estimates. A different random sample would yield a different set of cases in the sample. This would lead to a different estimate of $\bar{X}$, and the latter can be regarded as just as good an estimate as the first sample mean. In order to obtain an idea of the potential

variability in estimates of the population mean, the sampling distribution of $\bar{X}$ should be used. This will allow the researcher to obtain interval estimates of the population mean.

If the population mean is being estimated, and the samples are large random samples, the Central Limit Theorem can be used to describe the sampling distribution of $\bar{X}$. According to this theorem,

$$\bar{X} \text{ is Nor } \left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

This result allows the researcher to use the normal curve to describe the sampling distribution of $\bar{X}$. The next stage is for the researcher to decide what level of confidence he or she plans to use for the interval estimate. Let the confidence level chosen be C%, where C is a relatively large number, at least 80 or more, and usually 90 or more. This confidence level is equated with an area in the middle of the normal distribution. The Z values in the normal table which are associated with this area are then determined. The appropriate interval estimate is then

$$\bar{X} \pm Z\frac{\sigma}{\sqrt{n}}$$

where Z is the standardized normal value associated with an area of C% in the middle of the normal distribution. The C% interval estimate is then

$$\left(\bar{X} - Z\frac{\sigma}{\sqrt{n}}\ ,\ \bar{X} + Z\frac{\sigma}{\sqrt{n}}\right)$$

The probability is $C/100$ that this interval contains the true population mean μ. The values of $\bar{X}$ and n can be determined from the sample. σ is unknown, but if n is reasonably large, then the sample standard deviation s can be obtained from the sample, and this provides a reasonably close estimate of σ, for purposes of determining the interval estimate. Z is given on the basis of the confidence level C and the normal curve. When all these numbers are place in the last formula, a specific interval results. While it is best not to make probability statements concerning this specific interval, what can be said is that C% of the intervals constructed in this manner will contain μ.

These steps involved in interval estimation can be outlined as follows.

1. Obtain $\bar{X}$, s and n from the sample.

2. If n is large, the sampling distribution of $\bar{X}$ is normal, using the Central Limit Theorem of Chapter 7.

3. Determine the confidence level C, either by choosing a level, or using the confidence level given in the problem.

4. Use the normal table to find the Z value which gives an area of C% in the middle of the normal distribution.

5. Put all these values in

$$\left(\bar{X} - Z\frac{\sigma}{\sqrt{n}}\ ,\ \bar{X} + Z\frac{\sigma}{\sqrt{n}}\right)$$

and this is the C% confidence interval.

Two examples of confidence interval estimates for the population mean follow.

Example 8.3.2 Mean Individual Income by Province of Canada, 1985

The data in Table 8.3 comes from Statistics Canada's General Social Survey of 1986. The values of `TINC`*, total income, represent the total income from all sources, of survey respondents in dollars, for the 1985 calendar year.*

Assuming that the survey is a random sample of the residents of each of the provinces, derive a 95 per cent interval estimate for the mean income of residents of Saskatchewan. Do the same for the residents of Alberta. On the basis of these results, can you conclude that mean income for all Alberta residents is greater than the mean income for all Saskatchewan residents?

Solution. *Let μ be the true mean income of all Saskatchewan residents and let σ be the true standard deviation of the income of all Saskatchewan residents. Since μ is unknown, the sample mean $\bar{X}$ is used as a point estimate of μ. Since there are $n = 612$ residents of Saskatchewan surveyed, and assuming that this sample is random, this is a large random sample of Saskatchewanians. This $\bar{X}$ is normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. Since σ is unknown, $s = 14,837$ is used as an estimate of σ, again since n is large. For the 95 per cent interval estimate, the appropriate Z value is 1.96. The interval estimate is then:*

$$\bar{X} \pm Z\frac{s}{\sqrt{n}} = 15,768 \pm 1.96 \times \frac{14,837}{\sqrt{612}}$$

Summaries of TINC TOTAL INCOME
By levels of PROV PROVINCE OF RESIDENCE

Variable		Mean	Std Dev	Cases
For Entire Population		15389	14156	11720
PROV	Newfoundland	12006	11583	425
PROV	P.E.I.	13703	12624	141
PROV	Nova Scotia	14640	13729	424
PROV	New Brunswick	12248	12316	1341
PROV	Quebec	15139	13582	4052
PROV	Ontario	17152	15371	2713
PROV	Manitoba	15505	14095	519
PROV	Saskatchewan	15768	14837	612
PROV	Alberta	16949	14958	613
PROV	British Columbia	16742	14821	880

Total Cases = 11720

Table 8.3: Statistics of Total Income, Individuals, by Province, 1985

$$15,768 \pm (1.96)(599.75) = 15,768 \pm 1,176$$

The 95 per cent interval estimate is thus ($14,592, $16,944). This might best be rounded off to ($14,600, $16,900), given the approximations involved in the determination of this interval estimate.

For residents of Alberta, the method is the same, except that $n = 613$, $\bar{X} = 16,949$ *and* $s = 14,958$. *The 95 per cent interval is:*

$$\bar{X} \pm Z\frac{s}{\sqrt{n}} = 16,949 \pm 1.96 \times \frac{14,958}{\sqrt{613}}$$

$$16,949 \pm (1.96)(604.15) = 16,949 \pm 1,184$$

The 95 per cent interval estimate is thus ($15,765, $18,133). This could be rounded off to ($15,800, $18,100).

There is a fair overlap to the two intervals. In Alberta, the interval estimate implies that mean income of all Alberta residents could be as low as $15,800, and the estimate for Saskatchewan shows that the mean income of all Saskatchewan residents could be higher than this, as high as $16,900. While these results both depend on the particular samples given here, and the choice of the 95% confidence level, there appears to be considerable chance that the mean income of the Saskatchewan residents could be as high as that of Alberta residents. While the evidence generally points in the other direction, with the sample mean income being lower for Saskatchewan than Alberta, there is still some chance that the mean income for all Saskatchewan residents could be as high as the mean income for all Alberta residents.

Example 8.3.3 Head Size

In the 1800s and in the early part of this century, measurements of brain and head size were popular among some social scientists who were attempting to show that Caucasians, especially those higher in occupational status, were more intelligent than people of other 'races' or lower occupational status. Stephen Jay Gould, in his book **The Mismeasure of Man**, *critiques these points of view, and re-analyses some of the earlier data. Table 8.4 is data from a table on page 109 of Gould's book. Gould found this data in Ernest A. Hooton's book* **The American Criminal**. *Gould argues that Hooton misinterpreted this data and that* "most mean differences between occupational groups are statistically insignificant."

		Head Circumference in Millimetres	
Vocational Status	Sample Size	Mean	Standard Deviation
Professional	25	569.9	9.5
Semiprofessional	61	566.5	11.7
Clerical	107	566.2	11.4
Trades	194	565.7	11.1
Public Service	25	564.1	12.5
Skilled Trades	351	562.9	11.2
Personal Services	262	562.7	11.3
Laborers	647	560.7	7.6

Table 8.4: Mean and Standard Deviation of Head Circumference for People of Varied Occupational Status

Using the data in Table 8.4, derive 94% confidence interval estimates for (i) the true mean head circumference for all those of Semiprofessional vocational status and (ii) for the mean head circumference for all those of Personal Service vocational status. Do your results support Gould's statement? Explain.

Solution.

Let μ be the true mean head circumference for all those of semiprofessional status and let σ be the true standard deviation of this same occupational group. μ is unknown so that the sample mean $\bar{X}$ provides as a point estimate of μ. Since there are $n = 61$ people of semiprofessional status for which there is data, this can be considered to be a large sample of those of semiprofessional status. Also assume that the sample is randomly selected. Thus $\bar{X}$ is normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. Since σ is unknown, $s = 11.7$ is used as an estimate of σ, again since n is large. For the 94 per cent interval estimate, the appropriate Z value is 1.88. That is, with 94% confidence, there is $0.94/2 = 0.47$ of the area on each side of the mean. An area of 0.47 is associated with $Z = 1.88$, since at that Z, the area in column A of Appendix Table A is 0.4699. The interval estimate is then:

$$\bar{X} \pm Z\frac{s}{\sqrt{n}} = 566.5 \pm 1.88 \times \frac{11.7}{\sqrt{61}}$$

$$566.5 \pm (1.88)(1.498) = 566.5 \pm 2.816$$

The 94 per cent interval estimate is thus (563.7, 569.3).

For those of personal service occupations, the method is the same, except that $n = 262$, $\bar{X} = 562.7$ and $s = 11.3$. The 94 per cent interval is:

$$\bar{X} \pm Z\frac{s}{\sqrt{n}} = 562.7 \pm 1.88 \times \frac{11.3}{\sqrt{262}}$$

$$562.7 \pm (1.88)(0.698) = 562.7 \pm 1.312$$

The 94 per cent interval estimate is thus (561.4, 564.0).

These results are not too conclusive concerning Gould statements. Since we are 94 per cent sure that the first estimate is correct to within plus or minus 2.8 millimetres, and the second to within plus or minus 1.3 millimetres, and since the intervals overlap, there is a small chance that the true means of these two groups are identical. That is, given that there is a small overlap in the intervals, it is possible that the true mean for each of the two groups could be in this overlapping area, around 564. This seems fairly unlikely though. In Chapter 9, an hypothesis test can be used to directly test Gould's statement.

8.3.2 Choosing a Confidence Level.

There are no hard and fast guidelines concerning the choice of a confidence level for a confidence interval. If the confidence level is given in the question, or assigned by someone, then this is the confidence level which should be used. Where no confidence level is given, there are several considerations which guide the choice of confidence level. These are outlined as follows.

1. The first guideline is to make sure you **always report the confidence level**. Each different level gives an interval estimate of different interval width.

2. If you are not sure which confidence level to choose, **use the 95% confidence level**. This has become by far the most common and popular confidence level, and if there is any doubt concerning which level to choose, this level will be widely recognized.

3. If you wish to be **more certain** that the interval contains the true mean μ, select a **high confidence level** If you feel you need not be so certain of the result, select a lower confidence level. Remember that a high confidence level is associated with a wide interval, and a lower level with a narrower interval. A narrow interval makes your results appear more precise. In fact, the result may not be all that precise if the narrow interval is obtained by using a low confidence level. That is, the interval should not be artificially narrowed by selecting too low a confidence level. If you wish to select a low confidence level like 80%, you should make sure that you report the level chosen.

4. If you wish to compare your results with those of other researchers, it is best to **use the same confidence level as that used by other researchers**. If you do this, then your results can easily be compared with results from other research.

5. Some types of problems require a high level of confidence. In issues of **health and safety**, it may be very important to ensure that the mean does not exceed a certain level. For example, if the concentration of certain chemicals or other substances in water supplied to a city exceeds a certain level, then this may be harmful to human health. Samples may be taken from the water supply, and the mean level of the potentially harmful substance determined. Those who use the water suppply would like to be assured that the mean level of the

potentially harmful substance is within the safe level. That is, they would like to be 99% sure, or 99.9% sure that the water is safe. If the 99.9% confidence level for the mean amount of the substance is within the safe level, then it seems fairly certain that the water is safe.

Most social science research is not concerned with matters of life and death, or even of health and safety. For most social science research, a confidence level of 95%, or at most, 99% would seem aequate.

6. There may be many other methodological problems associated with research, or many nonsampling errors. If this is the case, it may not make too much sense to demand a very high confidence level. For example, if there is considered to be up to 15% nonsampling error, then it would seem unnecessary to use more than a 90% confidence level.

The above guidelines are rough guidelines, and it is most common to use a round number such as 90% or 95% confidence.

It would be most desirable to have high confidence level, say 99% confidence, and a narrow interval. But such a result can only be obtained with a large sample size. If the sample size is large enough to use the Central Limit Theorem, but is still not all that large, then the interval may be of considerable width. In the estimates of mean income in Saskatchewan and Alberta in Example 8.3.2, the intervals were approximately $2,000 wide even though the confidence level was only 90% and the sample sizes were over 600 in each case. Given the nature of these populations, there is little that can be done about this. The only way that the interval estimate can be narrowed and, at the same time have larger confidence associated with it, is to have a larger sample size. Obtaining a large sample may be too costly, so that the researcher just has to live with the smaller sample, and the results that are associated with it.

8.3.3 Probability of Interval Estimate.

In this chapter, the meaning of interval estimates and the manner in which interval estimates of the mean are constructed have been discussed. This section shows how an interval estimate and its associated probability is derived from the Central Limit Theorem. If you have difficulty following this section, you can skip it and move on to Section 8.4.1. If you can follow this section, you will obtain a better understanding of interval estimates, and

also how these interval estimates relate to the hypothesis tests of Chapter 9.

In order to provide a systematic notation for this explanation, some new notation is required. The notation in this section may seem a bit strange, and may not seem to be the most straightforward way of presenting the interval estimate. The notation used here is developed in order to be consistent with the notation used in hypothesis testing in Chapter 9. Since estimation and hypothesis testing are similar, it is useful to have the same notation for each.

Beginning from the Central Limit Theorem,

$$\bar{X} \text{ is Nor } \left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

if the sample size n of the random sample is large. μ is the true mean of the population and σ is the true standard deviation of the population. If $\bar{X}$ is being used as the point estimate of μ, then this normal distribution can be used to determine the C% confidence interval. C represents an area in the middle of the normal distribution, usually quite a large area, such as 90%, 95% or 99%.

In terms of notation, the symbol C% is not usually used. Rather the confidence level is defined as

$$(1 - \alpha) \times 100\%.$$

In this expression, α is the first letter of the Greek alphabet, and is called 'alpha'. This expression is just an alternative way of writing the previous confidence level C%. That is

$$(1 - \alpha) \times 100\% = C\%.$$

In order to make C a large number, α must be a small number. With a little experimentation with this expression, it can be seen that when $C = 90\%$, $\alpha = 0.10$. That is, $1 - \alpha = 1 - 0.10 = 0.90$ and

$$(1 - \alpha) \times 100\% = 0.90 \times 100\% = 90\%.$$

For the 95% confidence level, $\alpha = 0.05$ and for the 99% confidence level $\alpha = 0.01$. Whereas C was a percentage, α is a number between 0 and 1, and can be regarded as a proportion or a probability. C was a large value, but α is a small number; the two are complements of each other.

Figure 8.3: α Notation for the Standardized Normal Curve

The value $1 - \alpha$ represents the proportion of the area in the middle of the normal curve. Since the total area under the curve is 1, the area under the curve which is not in the centre is $1 - (1 - \alpha) = \alpha$. That is, if $1 - \alpha$ is the area in the middle of the normal curve, then α is the sum of the areas in the two tails of the distribution. Since the normal curve is symmetric, there is $\alpha/2$ of the area in each tail of the distribution. All of these areas are shown in Figure 8.3. The area in the middle of the curve, between the limits shown, is a proportion of the area $1 - \alpha$, or a percentage $(1 - \alpha) \times 100\%$. This means that there is $(1 - \alpha)/2$ of the area on each side of centre, within the limits shown, and $\alpha/2$ of the area in each tail of the distribution, outside the limits shown.

This notation also provides a means of defining the limits mentioned in the last paragraph. In Figure 8.3, these limits are labelled $-Z_{\alpha/2}$ and $Z_{\alpha/2}$. Since the standardized normal distribution is centred at a mean of $Z = 0$, values of Z to the left of centre are negative, and values to the right of centre

are positive. The Z value of $Z_{\alpha/2}$ is used to denote the Z value associated with an area of $\alpha/2$ beyond it. From the diagram it can be seen that there is an area of $\alpha/2$ to the left of $-Z_{\alpha/2}$ and an equal area of $\alpha/2$ to the right of $Z_{\alpha/2}$.

This awkward notation is quite flexible, and it provides a compete way of labelling and discussing the normal distribution. The values $\alpha/2$ are the values in column C of the normal table in Appendix A, and the associated Z values are those in the table. For example, for an area of 0.95 in the middle of the standardized normal distribution, $1 - \alpha = 0.95$ and $\alpha = 0.05$. Thus $\alpha/2 = 0.025$ represents the area in one tail of the distribution. Looking for a value of 0.025 in column C of Appendix A gives $Z = 1.96$. Thus $Z_{\alpha/2} = Z_{0.025} = 1.96$. This value should by now be recognizable as the familiar Z value of 1.96 associated with the 95% interval estimate.

The values of α can be interpreted as probabilities rather than proportions. If Z is regarded as a random variable with a standardized normal probability distribution, then the areas under the normal distribution represent probabilities of different values of Z occurring. The area in the middle of the normal distribution can be represented in a probability statement:

$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$$

That is, between the limits of $-Z_{\alpha/2}$ and $Z_{\alpha/2}$ there is an area of $1 - \alpha$. If the variable Z varies randomly according to this distribution, then this is also the associated probability of Z being within these limits. Similarly,

$$P(Z < -Z_{\alpha/2}) = \frac{\alpha}{2}$$

and

$$P(Z > Z_{\alpha/2}) = \frac{\alpha}{2}.$$

This can be made more concrete by using the example of the 95% confidence level. Then $(1-\alpha)\times 100\% = 95\%$, $1-\alpha = 0.95$, $\alpha = 0.05$, $\alpha/2 = 0.025$ and $Z_{\alpha/2} = 1.96$. Thus

$$P(Z < -Z_{\alpha/2}) = P(Z < -Z_{0.025}) = P(Z < -1.96) = 0.025$$

and

$$P(Z > Z_{\alpha/2}) = P(Z > Z_{0.025}) = P(Z > 1.96) = 0.025$$

Also, the probability in the middle can be given as:

$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$$

Figure 8.4: α Notation for Interval Estimate of the Mean

$$P(-Z_{0.025} < Z < Z_{0.025}) = 1 - \alpha$$
$$P(-1.96 < Z < +1.96) = 0.95$$

That is, the probability that the normal random variable Z takes on a value between $Z = -1.96$ and $Z = +1.96$ is 0.95.

This notation can now be used to discuss the sampling distribution of $\bar{X}$ and the interval estimates. Figure 8.4 uses the same notation as Figure 8.3. But rather than representing simply the standardized normal distribution, Figure 8.4 represents the sampling distribution of $\bar{X}$. This distribution has a mean of μ and a standard deviation of

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

Suppose that the $(1 - \alpha) \times 100\%$ confidence interval is to be obtained for the estimate of μ. The corresponding Z values are $-Z_{\alpha/2}$ and $Z_{\alpha/2}$. The probability is $1 - \alpha$ that Z takes on a value within these limits. In terms

of the distribution of $\bar{X}$, these limits are $Z_{\alpha/2}$ standard deviations on each side of the true mean μ. These limits are

$$\mu - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

and

$$\mu + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

The possible values of $\bar{X}$ in the sampling distribution of $\bar{X}$ can then be described in a probability statement

$$P\left(\mu - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

While this may appear to be an intimidating expression, all it states is that $1 - \alpha$ is the probability that $\bar{X}$ is within $Z_{\alpha/2}$ standard deviations of the true mean μ.

The above expression may appear to provide the interval estimate, at least in probability form. The difficulty with this expression is that the interval is constructed around μ, rather than around $\bar{X}$. Since μ is unknown, an interval cannot be constructed around it. Fortunately, the above expression can be rearranged to provide the interval estimate. From

$$P\left(\mu - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

the expression inside the probability statement can be rearranged to preserve the inequalities, yet produce an interval constructed around $\bar{X}$. If μ is subtracted from each part of the probability expression inside the probability bracket, and $\bar{X}$ is also subtracted from each part, then this expression still holds and is

$$P\left(-\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

The interval is now stated negatively. If an inequality is multiplied by a negative number, the direction of the inequalities are reversed, in order to preserve the inequalities. That is, $-3 > -5$, but $3 < 5$. Multiplying the expression inside the probability brackets and rearranging gives

$$P\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

The last expression gives the interval estimate, and the probability associated with this interval estimate. It states that the probability is $1 - \alpha$ that the intervals around $\bar{X}$ contain μ. The interval is constructed with a lower confidence limit of

$$\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

and a upper confidence limit of

$$\bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

The $(1 - \alpha)100\%$ confidence interval is thus

$$\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ,\ \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

The only difficulty associated with computing this after the sample has been obtained is that σ is unknown. As long as n is reasonably large, the sample standard deviation s can be used as an estimate of σ, and the interval estimate is:

$$\left(\bar{X} - Z_{\alpha/2}\frac{s}{\sqrt{n}}\ ,\ \bar{X} + Z_{\alpha/2}\frac{s}{\sqrt{n}}\right).$$

The above explanation provides the rationale for the interval estimates. In Chapter 9, it will be seen that interval estimation and hypothesis testing are quite similar. The notation developed in this section will make more sense once the method of hypothesis testing is examined, and this notation is consistent with Chapter 9.

The following section examines the situation when the sample size is small. The method outlined in this section needs to be modified a little when the sample has size less than 30. The method and notation outlined in this section will also be used in Section 8.5, when the interval estimate for a population proportion is obtained.

8.4 Estimate of the Mean, Small n

There are many circumstances where the sample size of the random sample is quite small, so that the Central Limit Theorem cannot be used to describe the sampling distribution of the sample mean. If this is the case, then the t distribution can often be used to describe the sampling distribution of the

sample mean. This section describes how to use the t distribution. Before examining this distribution, a few comments concerning the difficulties associated with small sample sizes are discussed.

A small sample size has many problems associated with it. In Chapter 7 it was seen that the sampling distribution of $\bar{X}$ has a standard error, or standard deviation, of

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

where σ is the standard deviation of the population from which the random sample has been drawn. When the random sample has a large sample size, then the standard error associated with $\bar{X}$ is relatively small. Since the sample size n appears in the denominator of the standard error, the larger the sample size, the smaller the standard error.

When the sample size n is small, this also means that the standard error is relatively large and there is considerable variation in the sampling distribution of $\bar{X}$. This, in turn, implies that the sampling error $|\bar{X} - \mu|$ may be large. It will be seen later in this section that this also produces a wide interval estimate. Thus a statistic obtained from a sample with small sample size, even when the sample is random, may not provide a very accurate estimate of the size of the population parameter.

Small samples are often not random. Many of the small samples selected by researchers use volunteers, students or other captive members of a population. If the sample is selected on this basis, then it becomes difficult to make any inferences from the sample to a larger population. These nonprobability methods of obtaining samples or subjects for an experiment are often necessary. There are often insufficient resources for researchers to select a probability based sample of a population. Since some data concerning a phenomenon is always preferred to no information, small nonprobability samples are often very useful. The caution that should always be attached to them is that the results from these small nonprobability samples should not be generalized too much. Or, if the researcher does wish to generalize the results to larger populations, some effort should be made by the researcher to investigate the representativeness of the sample.

Having discussed these cautions, it is often the case that only a small random sample is available. Alternatively, a researcher may wish to investigate whether a small sample is likely to be a random, or representative, sample of a population. In these circumstances, the researcher may have to use the t distribution.

8.4.1 Student's t distribution.

Student's t distribution describes the sampling distribution of the sample mean of random samples taken from a normally distributed population with unknown standard deviation. The distribution is called **Student's t distribution**, named after Student, the pen name of the statistician who developed this distribution. The distribution is more commonly referred to as merely the **t distribution**.

The t distribution behaves in much the same manner as does the normal distribution, but is more dispersed than the normal distribution. The t distribution is peaked at the centre, and is symmetrical about the centre. The large bulk of the area under the curve of the t distribution is near the centre, with less and less of the area the farther one goes from the centre of the distribution. The curve is asymptotic to the horizontal X axis, always approaching this axis, but never quite touching it. As can be seen in Figure 8.5, the t distribution is somewhat more spread out than the normal distribution.

The horizontal axis of the t distribution is labelled t and this t is interpreted in the same manner as is Z in the standardized normal distribution. That is, the t distribution pictured is a standardized t distribution, so that the mean of the t distribution is 0 and its standard deviation is 1. The values of t represent the number of standard deviations from centre for each point on the horizontal axis. If $t = 1$, this represents a point on the horizontal axis which is 1 standard deviation above the mean. If $t = -1.8$, this is a point 1.8 standard deviations to the left of centre.

Degrees of Freedom. There was only one standardized normal distribution. In contrast, there are many different standardized t distributions. Each t distribution has a **degree of freedom** associated with it. When working with the sample distribution of sample means, the degree of freedom is the sample size minus one. If d is used as the symbol to represent the degree of freedom associated with the t distribution, and if n is the sample size, then

$$d = n - 1.$$

Appendix **??** gives the table of the t distribution, and this table shows that there are different t values associated with each degree of freedom. For a small degree of freedom, the t distribution is quite dispersed, meaning that it is necessary to go a considerable distance from the mean in order to account for most of the area. When the degrees of freedom are increased,

Figure 8.5: t and Normal Distributions

the t distribution becomes more concentrated around the centre of the distribution, meaning that the t value associated with each area is somewhat smaller than in the case of a smaller degree of freedom. When the degrees of freedom increase beyond 30, the t distribution becomes almost exactly the same as the normal distribution. As the degree of freedom increases even more, the t and the normal distribution become so close to each other, that the t distribution becomes the normal distribution for all practical purposes.

As n becomes large, or as $d = n - 1$ becomes large,

$$t_d(0, 1) \rightarrow \text{Nor}(0, 1)$$

where $t_d(0, 1)$ is the standardized t distribution with mean 0, standard deviation 1, and d degrees of freedom.

Using the t Table. Since there is a different t distribution for each sample size or for each degree of freedom, it is not possible to provide a complete set of t tables. Instead, the table in Appendix **??** gives t values only for selected areas under the curve of the t distribution. These are the most commonly used values, such as the 90%, 95%, 99% confidence levels. The areas associated with significance levels in hypothesis testing are also provided in the table.

The body of the table contains the standardized t value associated with each level of confidence, and each degree of freedom. For example, if a confidence level of 90% is used, and the sample size is $n = 14$, so that there are $d = n - 1 = 13$ degrees of freedom, the t value in the column headed 90%, and row labelled $d = 13$, is 1.77. This means that there is 90% of the area in the middle of the t distribution between $t = -1.77$ and $t = +1.77$. This is the t value which will be used in the interval estimate for the mean, when there is a sample of size $n = 14$.

As another example, suppose that the confidence level is 95%, and the sample size is $n = 28$, so that there are $d = 28 - 1 = 27$ degrees of freedom. Using the 95% confidence level column and row 27, the t value can be seen to be $t = 2.05$. This means that the middle 95% of the t distribution for 27 degrees of freedom is 2.05.

Note how the t values decline from the top to the bottom of the t table. At the top of the t table, where there are relatively few degrees of freedom, the t values are all quite large. This means that for small degrees of freedom, it is necessary to go a considerable distance from centre to obtain any given area under the curve. Where the degrees of freedom are larger, the

corresponding area under the curve is reached somewhat sooner, at smaller t values. However, there is a limit to how small the t values become, and this limit is that provided by the normal distribution. Note how the t values get closer and closer to the corresponding normal values as the degrees of freedom increase. For example, for the 95% confidence level, at 10 degrees of freedom, the t value is 2.228. By the time 27 degrees of freedom have been reached, the t value is 2.05. At 30 degrees of freedom the t value is 2.04, and as the degrees of freedom increase further, the t value gradually declines, until it reaches the familiar value of 1.96, associated with the 95% area under the normal distribution.

When using the t table, you are restricted to using the confidence levels given in the table. If you wish to pick a different confidence level, you could try interpolating between the t values given. For example, if there are 15 degrees of freedom for a sample, and you need to determine the t value associated with the 92% confidence interval, this represents two fifths of the distance between the 90% and 95% levels. The respective t values for those two confidence levels are 1.75 and 2.13, and two fifths of the distance between these is

$$1.75 + \frac{2}{5}(2.13 - 1.75) = 1.75 + 0.15 = 1.90$$

and this provides a rough estimate of the 92% area.

Derivation of the Distribution. The strict conditions for using the t distribution are two:

1. The sample is a random sample of the population, and

2. The population from which the sample is drawn is normal.

Further, the t distribution is most often used to provide confidence intervals for, or hypothesis tests of, the mean. In carrying these out, the assumption is that the variable being studied has an interval or a ratio level scale. As will be seen in the examples, all of these conditions may be relaxed somewhat in actual research studies. You should always be aware of these assumptions though, and consider how the violation of these assumptions might affect the interval estimates or t tests.

The manner in which the t distribution describes the sampling distribution of $\bar{X}$ is discussed in this section. In Chapter 11, when correlation and

regression are discussed, the t distribution is also used, and slightly different conditions will be examined there.

Suppose that a variable X describes a characteristic of a population, and that the mean of X is μ and the standard deviation of X is σ, where neither of these population parameters are known. Further imagine that this population has a distribution which is normal. That is,

$$X \text{ is Nor } (\mu, \sigma).$$

Note that this is **not** the distribution of $\bar{X}$ yet, because samples have not yet been taken. This distribution is the actual distribution of the population. The assumption that the population is normally distributed is a very strong assumption, and one which is not likely to be satisfied in most circumstances. If the variable X is the height of people of one sex, all members of a particular ethnicity, then perhaps X will be normally distributed. If the population is a large set of students who have taken a standardized test, such as the LSAT or GRE, then perhaps the answers will be normally distributed. But many characteristics of populations are not normally distributed, and then this assumption will be violated. For example, distributions of income, wealth, farm size, many test results, and many attitudes, cannot be considered to be normally distributed.

Assume that the distribution of the variable in the population is given by

$$X \text{ is Nor}(\mu, \sigma)$$

and suppose random samples of size n are drawn from this population. Each random sample will yield a different set of X_is in the sample, producing a different $\bar{X}$ for each sample, so that $\bar{X}$ has a sampling distribution. It can be proved mathematically that the sample distribution of the sample means is a t distribution with $d = n - 1$ degrees of freedom. The mean of this sampling distribution is the same as the population mean μ. The standard deviation of the sample means is the sample standard deviation s divided by the square root of the sample size. All this can be summarized as follows.

If

$$X \text{ is Nor } (\mu, \sigma)$$

where μ and σ are unknown, and if random samples of size n are drawn from this population,

$$\bar{X} \text{ is } t_d \left(\mu, \frac{s}{\sqrt{n}}\right).$$

where $d = n - 1$.

When the sample size n is small, say less than $n = 30$, the t distribution should be used. If $n > 30$, then the t values become so close to the standardized normal values Z that the Central Limit Theorem can be used to describe the sampling distribution of $\bar{X}$. That is,

$$t \to Z \text{ as } n \to \infty.$$

This means that the t distribution is likely to be used only when the sample size is small. For larger sample sizes, $\bar{X}$ may still have a t distribution, but if the sample size is large enough, the normal values are so close to the t values that the normal values are ordinarily used.

8.4.2 Interval Estimate for the Mean.

The t distribution for $\bar{X}$ can be used to provide interval estimates of the population mean μ. The method of constructing interval estimates is exactly the same for this small sample method as it is for the large sample method. If the population mean μ is to be estimated, and the sample is a random sample of size n with sample mean $\bar{X}$ and sample standard deviation s, and if the population from which this sample is drawn is normally distributed, then

$$\bar{X} \text{ is } t_d\left(\mu, \frac{s}{\sqrt{n}}\right).$$

where $d = n - 1$.

In order to obtain an interval estimate, the researcher picks a confidence level C% and uses this to determine the appropriate t value from Appendix **??**. For d degrees of freedom, let t_d be the t value such that C% of the area under the t curve lies between $-t_d$ and t_d. The C% confidence interval is then

$$\bar{X} \pm t_d \frac{s}{\sqrt{n}}$$

or in interval form,

$$\left(\bar{X} - t_d \frac{s}{\sqrt{n}}\ ,\ \bar{X} + t_d \frac{s}{\sqrt{n}}\right).$$

Note that this is exactly the same as the confidence interval for the mean when the sample size is large, with the only difference being that t_d replaces the Z value. One other minor difference is that s is used in this formula, rather than σ. The latter was used when presenting the formula for the

interval estimate in the case of the large sample size. But even in the case of a large sample size, σ is almost always unknown, so that in practice, s is used as an estimate of σ in that formula.

The interpretation of the confidence interval estimate is also the same as earlier. That is, C% of the the intervals

$$\bar{X} \pm t_d \frac{s}{\sqrt{n}}$$

contain μ if random samples of size $n = d+1$ are drawn from the population. Any specific interval which is constructed will either contain μ or it will not contain μ, but the researcher can be confident that C% of these intervals will be wide enough so that μ will be in the interval.

8.4.3 Examples of Interval Estimates, Small n

This section contains several examples of interval estimates using the t distribution when the sample size is small. In these examples, the strict conditions laid down earlier may not always hold, but the t distribution can often yield useful interval estimates, even when the assumptions are violated.

Example 8.4.1 Flax Yields

A study of farms in the rural municipality of Emerald, Saskatchewan, was discussed in Example **??**. *This survey found that the 4 farms that produced flax had flax yields of 381.0, 279.4, 127.0 and 381.0 kilograms per acre. Assuming this is a random sample of all farms in the Crop District of which Emerald is a part, obtain the 95% interval estimate for mean flax yield.*

Solution. *When the data is given in this form, as a list of all the values of the variable, the first step in obtaining the interval estimate is to obtain the point estimate. The population value being estimated is the mean flax yield of all flax growing farms in the Crop District. Let this true mean be μ. If X is the variable representing flax yield in kilograms per acre, then $\bar{X}$ will be the point estimate. In order to determine $\bar{X}$, the formulae of Chapter 5 will be used. Since the interval estimate is required, and the sample standard deviation s must be calculated as part of this, both the mean and standard deviation are calculated here. Table 8.5 and the following paragraph give the calculations for determining these statistics.*

X	X^2
381.0	145,161.00
279.4	78,064.36
127.0	16,129.00
381.0	145,161.00
1,168.4	384,515.36

Table 8.5: Calculations for Mean and Standard Deviation of Flax Yield

From Table 8.5,

$$\Sigma X = 1,168.4$$

$$\Sigma X^2 = 384,515.36$$

and $n = 4$. *The mean value of flax yield for these 4 farms is*

$$\bar{X} = \frac{\Sigma X}{n} = \frac{1,168.4}{4} = 292.1$$

Given the data in the table, the variance is

$$s^2 = \frac{1}{n-1}\left[\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right] = \frac{1}{3}\left[384,515.36 - \frac{1168.4^2}{4}\right]$$

$$s^2 = \frac{1}{3}\left[384,515.36 - \frac{1,168.4}{4}\right]$$

$$s^2 = \frac{384,515.36 - 341,289.64}{3} = \frac{43,225.72}{3} = 14,408.573$$

The standard deviation is

$$s = \sqrt{14,408.573} = 120.036$$

or 120.0 kilograms per acre.

The point estimate of μ, *the true mean flax yield in the Crop District, is* $\bar{X} = 292$ *kilograms per acre. The interval estimate can be obtained using the t distribution. That is, the sample is assumed to be random, the mean*

and standard deviation of the population are unkown, and the sample size is small. If it is assumed that the population is normal, then

$$\bar{X} \text{ is } t_d \left(\mu, \frac{s}{\sqrt{n}}\right).$$

where $d = n - 1$.

For this example, $d = n-1 = 4-3 = 3$, $\bar{X} = 292.1$ *and* $s = 120.0$. *Given 3 degrees of freedom, and the 95% confidence level, the appropriate t value from Appendix* **??** *is 3.182. That is, with only 3 degrees of freedom, it is necessary to go out from the centre of the distribution a distance of 3.182 t values, or 3.162 standard deviations, in order to account for the middle 95% of the t distribution. The interval estimate for* μ *is constructed in the same manner as interval estimates earlier. The interval is constructed around* $\bar{X}$, *plus or minus the t value times the standard deviation of* $\bar{X}$. *This is*

$$\bar{X} \pm t_d \frac{s}{\sqrt{n}}$$

and with the values for this sample, this is

$$292.1 \pm 3.182 \frac{210.0}{\sqrt{4}}$$

$$292.1 \pm (3.182 \times 105.0)$$

$$292.1 \pm 334.1$$

Thus the 95% interval estimate for the true mean flax yield is $(-42.0, 626.2)$ *acres. Since negative yields are not possible, this might be made into the interval* $(0, 626)$, *rounding to the nearest integer. Since this is such a wide interval, this sample tells us little concerning the true mean flax yield in the Crop District.*

Additional Comments. *Note that there are three reasons why the interval is so wide in this example. First, the sample size of* $n = 4$ *is very small, so that* $s/\sqrt{n}$, *the standard deviation of* $\bar{X}$ *does not have a very large value for n in the denominator. Second, the t value is quite large, so the interval goes out from* $\bar{X}$ *a considerable distance in each direction. If the sample size had been larger, then the Z value might have been used, and this would put* $Z = 1.96$ *into the interval estimate, rather than* $t = 3.182$. *Using the Z value would produce a narrower interval. The third reason the interval is so wide is that s is fairly large. The 4 farms sampled have quite different*

flax yields. Since this standard deviation of $s = 120$ is used as the estimate of σ, this implies that the flax yield in the crop district has large variation. Any time the population has considerable variability, the interval estimate will be quite wide.

One problem with this sample is that the sample may not be random. In addition, all the farms sampled were in Emerald, only a small part of the Crop District. For purposes of estimating the true mean flax yield in the District as a whole, these farms may not be representative.

One further issue which should be considered here is whether the assumption that the population from which the sample is drawn is normally distributed or not. The assumption is that variable X is a normally distributed variable. Since X represents the flax yield on each farm, the assumption is that the flax yield per farm is normally distributed across farms. While it is unlikely that this assumption is exactly satisfied, this is not all that unreasonable as an assumption. Flax yields will differ from farm to farm based on factors such as the fertility of the soil, cultivation practices, amount of fertilizer used, and weather and other climate considerations. Across all farms growing flax, these factors may balance out, with some farms having above average yield, and others having below average yield, but most farms clustered around the average yield. If this reasoning is correct, then the assumption of normally distributed flax yield could be close to being satisfied.

In summary, the sample of 4 farms is really too small to provide a good estimate of the true mean flax yield in the Crop District. What this example shows is the difficulty of providing a very accurate estimate of a population parameter when the sample size of a sample is very small.

Example 8.4.2 Explanations of Unemployment

The 1985 Edmonton Area Study, conducted by the Population Research Laboratory at the University of Alberta, surveyed 385 Edmonton adults concerning various possible explanations for the existence of unemployment in Canada. This study is examined in H. Krahn et. al., "Explanations of Unemployment in Canada," **International Journal of Comparative Sociology**, *XXVIII, 3-4, 1987, pp. 228-236. One of the explanations given was "Many people are unemployed because they are unwilling to move to places of work" and respondents were asked whether they agreed or disagreed with this explanation. The responses were given on a 7 point scale, with 1 being 'strongly disagree' and 7 being 'strongly agree.' I drew a random*

sample of 6 of these respondents, and their responses were 2, 4, 1, 5, 6, 1. Using these 6 responses, derive a 90% and a 99% interval estimate for the true mean response for all Edmonton adults.

Solution. *Define X as the variable giving the responses to this explanation. X can potentially take on integer values from 1 through 7. Let the true mean of the response to this question for all Edmonton adults be μ. Assume that the distribution of responses to this explanation for unemployment to be normally distributed among all Edmonton adults. If this sample of $n = 6$ respondents represents a random sample from Edmonton, then the sample mean $\bar{X}$ has a t distribution with mean μ and standard deviation $s/\sqrt{n}$, and with $d = n - 1$ degrees of freedom. That is,*

$$\bar{X} \text{ is } t_d\left(\mu, \frac{s}{\sqrt{n}}\right).$$

where $d = n - 1$.

*Using the sample data from this sample, the values of X are 2, 4, 1, 5, 6, and 1. There are $n = 6$ values, and the formulae for the mean and standard deviation can be used to show that $\bar{X} = 3.167$ and $s = 2.137$. For this sample, there are $d = n - 1 = 6 - 1 = 5$ degrees of freedom. For the 90% interval estimate, the t value is obtained from Appendix **??** in the column headed 90% and row $d = 5$. This t value is $t_d = 2.015$. The interval estimates are provided using*

$$\bar{X} \pm t_d \frac{s}{\sqrt{n}}$$

With the values for this sample, this is

$$3.167 \pm 2.015 \frac{2.137}{\sqrt{6}}$$

$$3.167 \pm 2.015 \frac{2.137}{2.449}$$

$$3.167 \pm (2.015 \times 0.872)$$

$$3.167 \pm 1.758$$

The 90% interval estimate for the true mean opinion is $(1.409, 4.925)$. These values should be rounded so that the 90% interval estimate for the true mean opinion level of all Edmonton adults is $(1.4, 4.9)$. Recall that 1 represents strongly disagree with this explanation, 7 represents strongly agree, and by

implication 4 would be a relatively neutral view with regard to this explanation. The point estimate is 3.2, on the disagree side of centre. The interval estimate is mostly on the disagree side of the explanation. As a result, this sample appears to show that Edmonton adults, on average, slightly disagree with this explanation.

The 99% interval estimate uses the same n, $\bar{X}$ and s, but the t value changes. The degree of freedom is still 5 so row 5 of the table in Appendix **??** *is used. But with the column for the 99% interval, $t_d = 4.032$. The interval is given as follows:*

$$3.167 \pm 4.032\frac{2.137}{\sqrt{6}}$$

$$3.167 \pm 4.032\frac{2.137}{2.449}$$

$$3.167 \pm (4.032 \times 0.872)$$

$$3.167 \pm 3.518$$

The 99% interval estimate for the true mean opinion is $(-0.351, 6.685)$. Again, these values should be rounded so that the 99% interval estimate is that the true mean opinion level for all Edmonton adults is $(-0.4, 6.7)$. Since this includes practically the whole range of opinions, this interval estimate tells us little that could not have been said before the sample was selected. Since the t value associated with the 99% interval is so large, practically the whole range of possible opinions is included in the interval estimate. In order to be 99% confident that the estimates contain μ, the whole range of opinions has to be included.

Additional Comments. *These results show some of the problems that can develop if the sample size is small. The 99% interval estimate is so wide as to be useless, and even the 90% interval estimate is quite wide, including practically all the opinions on the diagree side. What this shows is that a random sample of size $n = 6$ is really too small to make any definitive conclusions concerning the nature of opinions among all Edmonton adults.*

The complete Edmonton Area Study contained data from 385 adults surveyed. This data can be used to provide a much more precise interval estimate of the true mean opinion. For the set of all $n = 385$ adults surveyed, $\bar{X} = 4.151$, $s = 1.834$ and the sample size is quite large, so that

$$\bar{X} \text{ is Nor}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

For the 99% interval, the appropriate Z value is 2.575 and the 99% interval would be

$$\bar{X} \pm 2.575\frac{\sigma}{\sqrt{n}}$$

$$4.151 \pm 2.575\frac{1.834}{\sqrt{385}}$$

$$4.151 \pm 2.575\frac{1.834}{19.621}$$

$$4.151 \pm (2.575 \times 0.093)$$

$$4.151 \pm 0.241$$

Based on the sample size of $n = 385$, the 99% interval estimate for the true mean opinion is $(3.9, 4.4)$. The increase in the sample size from 6 to 385 has dramatically reduced the size of the interval, making the estimate of mean opinion level a much more precise estimate. The probability is 0.99 that a random sample of size $n = 385$ will provide an estimate of μ which is correct to within ± 0.241.

In addition, note that with the larger sample size, the Central Limit Theorem can be invoked, so that the Z value can be legitimately used. With the small sample size of only 6, there is some question concerning whether even the t distribution can be used. The t distribution for $\bar{X}$ assumes that the population from which the random sample is drawn is a normally distributed population. In this case, the frequency distribution of the larger sample of 385 respondents allowed me to check whether the opinions for that sample were close to normally distributed or not. Although not shown here, the distribution of opinions is not too far from a normal distribution, although it is not real close to being normal either. Thus the t value may not be quite appropriate.

As a general rule, when the sample size is small, it is always advisable to use the t value from the t distribution rather than the Z value from the normal distribution when determining the interval estimate.

Since the Z value is much smaller than the t value, using the Z value could produce an interval which appears much narrower than it really should be. That is, the Z value underestimates the size of the interval in the case of small sample size. Even the use of the t value may lead to underestimation of the interval width. But since the t value produces a wider interval than does the Z value, there is less chance of underestimating the interval width if the t value is used.

Example 8.4.3 Self-Esteem of Elderly Women

A situation in which the t distribution may be used is when the number of subjects which can be studied is fairly small. This was the case for a study of 25 elderly women who were residents of a Windsor, Ontario nursing home. Two psychologists at the University of Windsor studied these women, along with a sample of 28 elderly women who were living in their own homes. The results of the study are in Rhonda A. Loomis and Cheryl D. Thomas, "Elderly Women in Nursing Home and Independent Residence: Health, Body Attitudes, Self-Esteem and Life Satisfaction," **Canadian Journal on Aging**, *Vol. 10, No. 3, 1991, pp. 224-231. The authors state*

> One aim of the current study was to determine if elderly women who live in a nursing home differ from those who have remained in their own homes with respect to their self-reported health, body attitudes, self-esteem, or life satisfaction.

The elderly women in the study were asked to give responses to various questions, including the 'Index of Self-Esteem ' (ISE).

> The ISE ... is a 25-item self-report inventory designed to measure the level of self-esteem problems experienced by the respondent. Items (e.g., "I feel that I bore peple") are rated on a five-point scale from 1 (rarely or none of the time) to 5 (most or all of the time) and item ratings are summed to yield a total score between 0 and 100. Scores above 30 ... indicate clinically signficant self-esteem problems.

	Community Group	Nursing Home Group
Mean	29.4	25.1
SD	15.0	14.4
n	28	25

Table 8.6: Means and Standard Deviations (SD) of ISE

The summary statistics for ISE from the study are contained in Table 8.6. Since a higher ISE score indicates more problems, the community women

who lived in their own homes appear to score higher on the index of self esteem than did the women in nursing homes. A higher ISE score indicates lower self-esteem, so that the community group appears to show lower self-esteem. However, the authors claim that this difference is not statistically significant. While proof of this must wait until Chapter 9, interval estimates can be used to show much the same result.

Obtain 95% interval estimates for the true mean of each group. Use these results to comment on whether women in nursing homes and in the community show different levels of self-esteem.

Solution. *It is not clear what the population from which these samples are drawn really is. Imagine though that the community women are a random sample of all elderly women who live in their own homes. Further suppose that the residents of the nursing home represent a random sample of all elderly women in nursing homes. In addition, assume that the index of self-esteem is normally distributed for both groups. Given these assumptions, the t distribution can be used to obtain interval estimates of the true mean for each group. In each case, the sample size is under 30, and the standard deviation of the whole population is not known, so that the t, rather than the normal, distribution must be used in each case.*

For the community women, $n = 28$, so that the degree of freedom is $d = n - 1 = 27$. The appropriate t value for 95% confidence from Appendix **??** *is $t_{27} = 2.052$. If μ_c is the true mean of ISE for all elderly women who live in their own homes, the interval estimate for μ_c is*

$$\bar{X} \pm t_d \frac{s}{\sqrt{n}}$$

$$29.4 \pm 2.052 \frac{15.0}{\sqrt{28}}$$

$$29.4 \pm 2.052 \frac{15.0}{5.292}$$

$$29.4 \pm (2.052 \times 2.834)$$

$$29.4 \pm 5.8$$

The 95% interval estimate for the true mean ISE of community women is $(23.6, 35.2)$.

Let the true mean of ISE for all elderly women who live in nursing homes be μ_n. Since the sample size for these women is $n = 25$, there are

$d = 24$ degrees of freedom, and at the 95% confidence level, the t value is $t_{24} = 2.064$. The interval estimate for μ_n is

$$25.1 \pm 2.064\frac{14.4}{\sqrt{25}}$$

$$25.1 \pm 2.064\frac{14.4}{5}$$

$$25.1 \pm (2.064 \times 2.88)$$

$$25.1 \pm 5.9$$

The 95% interval estimate for the true mean ISE of community women is $(19.2, 31.0)$.

Since each of these intervals has considerable width, the estimates of the respective means of ISE are not very precise. The researchers can be 95% confident that the estimate of μ_c is no more than 5.8 units from its true value, and that the estimate of μ_n is not incorrect by more than 5.9 units. But since the estimates of mean ISE differ by only $29.4 - 25.1 = 4.3$ units, there is not sufficient evidence to conclude that the true means for the two groups are different. If the 5.8 and 5.9 are regarded as sampling error, this sampling error for each group is greater than the estimated difference between the groups.

Also note that the two intervals overlap considerably. This means that the true mean of each group could be in the overlapping region. For example, the true mean μ_n could be as large as 31.0, considerably greater than the sample mean for the community group. Thus the authors appear to be justified in their statement that there is not really a significant difference between the reported self-esteem for the two groups. In Chapter 9, an hypothesis test will be used to test this directly, but the interval estimates provide the same result.

Additional Comments. *Finally, a few comments can be made concerning the possible violation of assumptions here. First, the samples do not appear to be random samples, so that it is not clear how the sample results can be generalized to a large population, or to what population they could be generalized. While this is a problem, it is sometimes useful to treat non-random samples of this sort as if they were random samples. In this case, the results above show that even if the samples were random, the reported means are little different. This is useful information, but care must be taken*

concerning how general this result may be. It could be that if larger random samples were taken, then there would be a difference.

Second, whether or not the distributions of ISE are normal is not discussed by the authors. It would be useful to know a little more concerning how this scale is constructed. Often a scale of this sort is constructed to produce a normally distributed pattern of responses. This may be the case with this scale.

The assumptions may be violated in this example. But since the difference between the two groups is found to be insignificant on the ISE scale, no major claims or conclusions are being made for differences between the two groups. Whatever violation of assumptions there may be, has been interpreted in a cautious fashion. Where problems develop is when assumptions are violated, and stronger conclusions are reported.

8.4.4 Notation for Confidence Levels with t Distribution.

Just as in the case of the normal distribution, a complete notation for the areas under the t curve, the t values, and interval estimates for small n, can be obtained by using the α notation. In you did not follow this when the interval estimate for the mean was discussed, you can skip this section. But if you follow the notation in this section, then you should have a better grasp of interval estimation using the t distribution. The notation of this section is also consistent with the t tests of Chapter 9.

Suppose X is a variable which is normally distributed with mean μ and standard deviation σ. That is

$$X \text{ is Nor } (\mu, \sigma).$$

Suppose though that both μ and σ are unknown. If a random sample of size n is taken from this population, the sample mean $\bar{X}$ has a t distribution with mean μ and standard deviation $s/\sqrt{n}$, $d = n - 1$ degrees of freedom. That is,

$$\bar{X} \text{ is } t_d \ (\mu, \frac{s}{\sqrt{n}}).$$

This result holds regardless of the sample size, but the t distribution is most often used when the sample size is small. When n is larger, this t distribution approaches the standardized normal distribution.

The t distribution can be used to determine the C% confidence interval, where

$$C\% = (1 - \alpha) \times 100\%.$$

Figure 8.6: α Notation for the Standardized t Distribution

where $1 - \alpha$ represents the proportion of the area in the middle of the t distribution. Then α is the sum of the areas in the two tails of the distribution, and since the t distribution is symmetric, there is $\alpha/2$ of the area in each tail of the distribution. All of these areas are shown in Figure 8.6.

Since there is already one subscript on the t value representing the degree of freedom, it is necessary to add another subscript to denote the area. The two subscripts are separated by a comma. For d degrees of freedom, and $\alpha/2$ of the area in one tail of the distribution, let $t_{d,\alpha/2}$ represent the appropriate t value. In Figure 8.6, these limits are labelled $-t_{d,\alpha/2}$ and $+t_{d,\alpha/2}$. Since the standardized t distribution is centred at a mean of $t = 0$, values of t to the left of centre are negative, and values to the right of centre are positive. The value of t of $t_{d,\alpha/2}$ is used to denote the t value associated with an area of $\alpha/2$ beyond it. From the diagram it can be seen that there is an area of $\alpha/2$ to the left of $-t_{d,\alpha/2}$ and an equal area of $\alpha/2$ to the right of $+t_{d,\alpha/2}$.

If these areas are interpreted as probabilities,

$$P(-t_{d,\alpha/2} < t < t_{d,\alpha/2}) = 1 - \alpha$$

That is, between the limits of $-t_{d,\alpha/2}$ and $t_{d,\alpha/2}$ there is an area of $1-\alpha$. If the variable t varies randomly according to this distribution, then this is the associated probability of t being within these limits. Similarly,

$$P(t < -t_{d,\alpha/2}) = \frac{\alpha}{2}$$

and

$$P(t > t_{d,\alpha/2}) = \frac{\alpha}{2}.$$

For a sample mean $\bar{X}$ which has a t distribution with unknown mean μ and standard deviation $s/\sqrt{n}$, the limits on the confidence interval are $t_{d,\alpha/2}$ standard deviations on each side of the true mean μ. These limits are

$$\mu - t_{d,\alpha/2}\frac{s}{\sqrt{n}}$$

and

$$\mu + t_{d,\alpha/2}\frac{s}{\sqrt{n}}$$

The possible values of $\bar{X}$ in the sampling distribution of $\bar{X}$ can then be described in a probability statement

$$P(\mu - t_{d,\alpha/2}\frac{s}{\sqrt{n}} < \bar{X} < \mu + t_{d,\alpha/2}\frac{s}{\sqrt{n}}) = 1-\alpha.$$

The expression in brackets can be rearranged, producing

$$P(\bar{X} - t_{d,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{d,\alpha/2}\frac{s}{\sqrt{n}}) = 1-\alpha$$

and this is the interval estimate in the brackets.

8.5 Estimate of a Proportion

Estimating a population proportion uses the same method as that used for estimating a population mean, with only a few modifications for different notation. Since the latter has been described in detail earlier in this chapter, this section provides only a short description of the method of estimating a population proportion. Examples of these interval estimates for a population proportion are also provided.

Suppose that a researcher wishes to determine the proportion of population members which have a particular characteristic. Let this proportion

be defined as p. This population characteristic is unknown, but suppose a random sample with a large sample size is taken from this population. The sampling distribution of the sample proportion can be determined using the extension of the normal approximation to the binomial discussed in Chapter 7. This is as follows.

Suppose a random sample of size n is taken from the population having p as the true proportion of members with the characteristic in question. Any member of the sample which has this characteristic is defined as being a success. If there are X successes in the sample of n members of the population, the proportion of successes is $\hat{p} = X/n$. For a given sample, $\hat{p} = X/n$ is the point estimate of the true population proportion p.

Each random sample will yield a different value of $\hat{p}$, but if n is reasonably large, the sampling error of the sample proportion $\hat{p}$ is normally distributed with mean p and standard deviation $\sigma_{\hat{p}} = \sqrt{pq/n}$. Symbolically,

$$\hat{p} \text{ is Nor } (\ p \ , \sigma_{\hat{p}} \ )$$

or

$$\hat{p} \text{ is Nor } \left(\ p \ , \sqrt{pq/n} \ \right)$$

This result holds as long as

$$n \geq \frac{5}{\min(p, q)}$$

In order to determine the confidence interval estimate, the same method as used for the mean can now be used. Pick a confidence level C%. Since the distribution of $\hat{p}$ is normal, the Z value corresponding to an area of C% in the middle of the normal curve can be obtained from Appendix A. Using this Z value, an interval estimate for p can be constructed around $\hat{p}$. The lower confidence limit for this interval will be

$$\hat{p} - Z\sqrt{\frac{pq}{n}}$$

and the upper confidence limit is

$$\hat{p} + Z\sqrt{\frac{pq}{n}}$$

The C% confidence interval is thus

$$\hat{p} \pm Z\sqrt{\frac{pq}{n}}$$

or in interval form,

$$\left(\hat{p} - Z\sqrt{\frac{pq}{n}}\ ,\ \hat{p} + Z\sqrt{\frac{pq}{n}}\right).$$

The only difference between this interval estimate and the large sample method for the mean is that $\hat{p}$ replaces $\bar{X}$, and the standard deviation of $\bar{X}$ is replaced with the standard deviation of $\hat{p}$. The standard deviation of $\hat{p}$ is $\sqrt{pq/n}$.

After the sample has been obtained, the only problem in providing the interval estimate for a proportion is to determine the value of p and q in $\sqrt{pq/n}$. As in the discussion of the sampling error for a proportion on page **??** in Chapter 7, there are two choices. If $p = q = 0.5$ is used in $\sigma_{\hat{p}} = \sqrt{pq/n}$, then this provides the maximum possible value for $\sigma_{\hat{p}}$. This is a conservative method, providing an interval estimate which may be a little wider than the interval really should be.

The alternative choice for the estimate of $\sigma_{\hat{p}} = \sqrt{pq/n}$ is to let $p = \hat{p}$ and $q = 1 - p = 1 - \hat{p}$ in this part of the expression. This will produce a somewhat narrower interval. The only danger with using this method is that the interval which is obtained may be a little narrower than the interval really should be. This is only likely to occur is either $\hat{p}$ or $\hat{q}$ is quite different from 0.5. If that is the case, this method may produce an underestimate of the true interval width. (The true interval width would be based on the actual values of p and q, but these values are unknown, these are the values being estimated).

Example 8.5.1 Interval Estimate for the Gallup Poll

*In Example **??** of Chapter 7, the sampling error for Gallup poll estimates was shown to be a little under 4 percentage points, in 19 out of 20 samples. Here a recent poll result is taken, and the respective interval estimates are determined.*

In the August, 1992 poll, Gallup asked the question

> If a federal election were held today, which party's candidate do you think you would favor?

Of those adults who had decided which party they would favor, 21% supported the PCs, 44% the Liberals, 16% the NDP, 11% the Reform Party, 6% the Bloc Quebecois and 1% other parties. The totals do not add to 100% because of rounding error. Of those polled, 33% were undecided, so these reported percentages are based on the 67% of those polled who were

decided. Obtain the interval estimates for the true proportion of Canadian adults who supported the PCs and Liberals in August, 1992.

Solution. *Since Gallup mentions the sampling error in 19 or 20 samples, the C=95% confidence level will be used here. For each of the political parties, success can be defined as support for that political party.*

For the Conservatives, define success as the characteristic that an adult who is interviewed will support the PC party. Let p be the true proportion of Canadian adults who really do favor the Conservative party. This true population parameter p is unknown, but since the Gallup poll is a random sample of Canadian adults, with a large sample size, the distribution of the sample proportion $\hat{p}$ can be determined on the basis that:

$$\hat{p} \text{ is Nor } \left(p \ , \sqrt{\frac{pq}{n}} \ \right)$$

In this sample, $n = 1,025$, $\hat{p} = 0.21$ and thus $\hat{q} = 1 - \hat{p} = 1 - 0.21 = 0.79$. In order to estimate $\sigma_{\hat{p}} = \sqrt{pq/n}$ either these values, or $p = q = 0.5$ may be used. Using the latter values, the 95% interval estimate is

$$\left(\hat{p} - Z\sqrt{\frac{pq}{n}} \ , \ \hat{p} + Z\sqrt{\frac{pq}{n}} \right)$$

$$\left(0.21 - 1.96\sqrt{\frac{0.5 \times 0.5}{1,025}} \ , \ 0.21 + 1.96\sqrt{\frac{0.5 \times 0.5}{1,025}} \right)$$

$$\left(0.21 - 1.96\sqrt{0.000243902} \ , \ 0.21 + 1.96\sqrt{0.000243902} \right)$$

$$(0.21 - [1.96 \times 0.0156174] \ , \ 0.21 + [1.96 \times 0.0156174])$$

$$(0.21 - 0.0306 \ , \ 0.21 + 0.0306)$$

$$(0.1794 \ , \ 0.2406)$$

Rounding this off, the 95% interval estimate for the true proportion of PC supporters is thus (0.18, 0.24), or between 18% and 24% of Canadian adults.

If $\hat{p}$ and $\hat{q}$ had been used in determining this interval, the interval would have been

$$0.21 \pm 1.96\sqrt{\frac{0.21 \times 0.79}{1,025}}$$

or 0.21 ± 0.0249. Using these values in the standard deviation reduces the size of the interval estimate a little, producing a 95% interval of (0.185, 0.235).

Also recall that the undecided might best be taken out of this estimate, as argued on page **??** *in Chapter 7. If this is done, this reduces the effective sample size to approximately 686. If* $p = q = 0.5$ *is used in* $\sigma_{\hat{p}} = \sqrt{pq/n}$*, then the interval estimate is* $\hat{p} \pm 0.037$*, or approximately plus or minus 4 percentage points. For the PCs, the maximum size of the 95% interval estimate would be (0.17, 0.25). Given this data, it seems quite likely that the true proportion of PC supporters among Canadian adults would be between 17% and 25% in August 1992.*

For the Liberals, the method is the same. Define support for the Liberals as a success. Then $\hat{p} = 0.44$ *and the 95% interval estimates can be constructed in the same manner as for the PCs. If the full sample size of* $n = 1,025$ *is used, then the maximum size of the interval estimate is* 0.44 ± 0.031*. Note that if* $\hat{p}$ *and* $\hat{q}$ *are used in* $\sqrt{pq/n}$*, the interval is reduced in size to only* 0.44 ± 0.030*. Reducing the effective sample size to 686 will have the same effect as it did for the PCs, widening the interval a little.*

For each of the other groups, the interval estimates are approximately the same width. If $p = q = 0.5$*, all the interval estimates for a given sample size and confidence level will be the same width. Only if* $\hat{p}$ *and* $\hat{q}$ *are used in* $\sqrt{pq/n}$ *will be interval estimates be narrowed slightly.*

Based on these considerations, the 95% interval estimates for each proportion are $\hat{p} \pm 0.031$*. This is the same as the sampling error shown in Chapter 7, and the same level as claimed by Gallup.*

Example 8.5.2 Child Discipline and Child Abuse

Table 8.7 is based on Table II of Rhonda L. Lenton, "Techniques of Child Discipline and Abuse by Parents," **Canadian Review of Sociology and Anthropology**, *27 (2), 1990. This table contains the results of a survey of Toronto families, and gives the percentages of mothers and of fathers who used various types of disciplinary action with children in the year before being surveyed. Lenton notes that the sample is a random sample of Toronto families, with 89 mothers and 48 fathers in the sample.*

Referring to this table, Lenton comments

> Mothers appear to be somewhat more likely to have used violent disciplinary techniques with the notable exceptions of two of the most severe acts – withholding food and beating.

Obtain the 90% interval estimate for the proportion of all Toronto mothers who pushed, shoved or grabbed a child in the past year. Do the same

Aggressive Disciplinary Action	Percent of: Mothers	Fathers
Yell	96	94
Ridicule Child	35	52
Verbally Threaten	65	71
Withdraw Emotionally	14	27
Push, Grab, Shove	55	46
Throw something at Child	11	4
Slap or Spank Child	75	58
Hit Child with Object	15	10
Withhold Food	4	8
Beat Child	12	15
Sample Size	89	48

Table 8.7: Aggressive Disciplinary Actions Tried by Parents in Past Year (in percent)

for fathers. Using the percentages in Table 8.7 comment on Lenton's conclusion that mothers appear to have been more likely to have used violent disciplinary techniques.

Solution. *The 90% interval estimates are interval estimates for a proportion. Let p be the true proportion of all Toronto mothers who pushed, shoved or grabbed a child in the past year. The sample of $n = 89$ Toronto mothers shows that 55% of Toronto mothers pushed shoved or grabbed a child in the last year. In terms of a proportion, this is $\hat{p} = 0.55$. Also, $\hat{q}$, the proportion of Toronto mothers who did not push, shove or grab a child in the past year is $\hat{q} = 1 - \hat{p} = 1 - 0.55 = 0.45$. Since*

$$\frac{5}{min(p,q)} = \frac{5}{0.45} = 11.1 < 89$$

the normal approximation to the binomial provides a good estimate of the distribution of $\hat{p}$ so that

$$\hat{p} \text{ is } Nor\left(p, \sqrt{pq/n}\right).$$

For the $(1-\alpha)100\% = 90\%$ *interval estimate,* $Z = 1.645$*, and the interval estimate is as follows. (For the standard deviation of* $\hat{p}$*, in this case* $\hat{p}$ *has been used for* p*, and* $\hat{q}$ *has been used for* q*. Alternatively, a slightly wider interval would have resulted if we let* $p = q = 0.5$*).*

$$\hat{p} \pm Z\sqrt{\frac{pq}{n}} = 0.55 \pm 1.645\sqrt{\frac{0.55 \times 0.45}{89}}$$

$$0.55 \pm 1.645\sqrt{0.0027808} = 0.55 \pm 0.087$$

The 90% confidence interval estimate for the proportion of all Toronto mothers who pushed, shoved or grabbed a child is (0.463, 0.637) or 46% to 64%.

For the fathers, the method is exactly the same, but $\hat{p} = 0.46$*,* $\hat{q} = 0.54$ *and* $n = 48$*. Since*

$$\frac{5}{min(p,q)} = \frac{5}{0.46} = 10.9 < 48$$

the normal approximation to the binomial provides a good estimate of the distribution of $\hat{p}$ *so that*

$$\hat{p} \text{ is } \text{Nor}\left(p, \sqrt{\tfrac{pq}{n}}\right).$$

The 90% interval estimate is

$$\hat{p} \pm Z\sqrt{\frac{pq}{n}} = 0.46 \pm 1.645\sqrt{\frac{0.46 \times 0.54}{48}}$$

$$0.46 \pm \sqrt{0.005175} = 0.46 \pm 0.118$$

The 90% confidence interval estimate for the proportion of fathers who pushed, shoved or grabbed a child in the past year is (0.342,0.578) or 34% to 58%.

The wide size of the interval estimates casts some doubt on Lenton's statements. The sample sizes of mothers and fathers are not all that large, so that even though only the 90% confidence level has been used, the intervals for each of mothers and fathers are plus or minus 9 percentage points or more. Given that the percentage of mothers and fathers who used each technique often differs by less than this, the data here provides only very weak evidence for Lenton's statement. The interval estimates are just too wide to conclude that larger percentages of mothers than of fathers used the more violent techniques.

In addition, it may be questioned whether the sample really was random. While this would require going back to the methodology of the original study, it may be that the sample has some selection problems associated with it.

In conclusion, Lenton may have found evidence for a difference in behaviour of mothers and fathers toward children, but more evidence would be required before this difference could be considered to be well founded. In Chapter 9, an hypothesis test for the difference between mothers and fathers is conducted.

Example 8.5.3 Representativeness of Regina Labour Force Survey

The 1986 Census of Canada gives data showing that of 132,825 adults aged 15 or over in Regina, 15,240 have completed less than grade 9 education. The Social Studies 203 Regina Labour Force Survey also surveys adults aged 15 and over and of the 937 respondents surveyed, 74 have completed less than grade 9 education. Using only the data from the Survey, obtain the 98 per cent interval estimate for the proportion of all Regina adults aged 15 or over, who have completed less than grade 9 education. Based on this result, what conclusion can you draw concerning the representativeness of the Survey?

Solution. *Based on the Survey, the point estimate for the proportion of all Regina adults who have completed less than a grade 9 education is*

$$\hat{p} = \frac{74}{937} = 0.0790.$$

The sample size is $n = 937$. *Using* $\hat{p}$,

$$\frac{5}{min(p,q)} = \frac{5}{0.0790} = 63.3 < 937$$

and the normal approximation to the binomial provides a good estimate of the distribution of $\hat{p}$. *If* p *is the true proportion of all Regina adults who have less than a grade 9 education, then*

$$\hat{p} \text{ is } \mathrm{Nor}\left(p, \sqrt{\tfrac{pq}{n}}\right).$$

For the $(1-\alpha)100\% = 98\%$ *interval estimate,* $Z_{\alpha/2} = Z_{0.01} = 2.33$, *and the interval estimate is as follows. (* $\hat{p}$ *has been used for* p, *and* $\hat{q}$ *has been used for* q *in this case, for calculating the standard deviation of* $\hat{p}$, $\sigma_{\hat{p}}$. *Alternatively,*

a wider interval would have resulted if we let $p = q = 0.5$). The interval estimate is:

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}} = 0.0790 \pm 2.33\sqrt{\frac{0.079 \times 0.921}{937}}$$

$$0.079 \pm 2.33\sqrt{0.00007765} = 0.079 \pm 0.021$$

The 98% confidence interval estimate for the proportion of all Regina adults who have less than grade 9 education is (0.058, 0.100) or 5.8% to 10.0%. Using the result from the Survey would lead a researcher to be fairly confident that the percentage of all Regina adults who have completed less than a grade 9 education is between 6% and 10%.

Data from the Census of Canada is available to provide a separate check on these figures. From the 1986 Census there are 15,240 out of 132,825 adults who had less than a grade 9 education. This is $15,240/132,825 = 0.115$, or 11.5% of the adult population. This is more than the point estimate from the Survey, and is even outside the interval estimate. From this independent check on the figures, it appears that the Regina Labour Force Survey underrepresents those adults who have less than a grade 9 education. This provides useful information for those designing the Survey.

Some reasons for the underrepresentation of those with less than a grade 9 education are hypothesized here. It might be that a considerable proportion of those with less than a grade 9 education are elderly, and elderly people may be missed in a telephone survey. Another consideration might be the reluctance of those with less than a grade 9 education to respond to the Survey interviewers. A further consideration might be that some of these could be people who do not have telephones, or tend to move frequently, and do not have an up to date telephone number in the telephone directory. Whether any of these explanations is likely to explain this underrepresentation cannot be determined on this basis of this data. But these would be some of the factors which the researchers might investigate.

Probability of the Estimate. Before leaving the interval estimate for a proportion, the derivation of the probability associated with the estimate is provided. This was explained earlier in this section, but the complete set of formulae was not given. If you did not follow the derivation of these probabilities earlier in the chapter, you can also skip this section.

Suppose p is the true proportion of a population with a particular characteristic. Take random samples of size n from this population. Let there

be X successes in the sample, that is, X of the respondents have the characteristic being investigated. Then the sample proportion $\hat{p} = X/n$ is a point estimate of p.

In addition, $\hat{p}$ has a mean of p and a standard deviation of $\sigma_{\hat{p}} = \sqrt{pq/n}$. This is based on the binomial distribution, and the characteristics of the binomial probability distribution. In addition, if

$$n \geq \frac{5}{\min(p, q)}$$

then the distribution of $\hat{p}$ is normal, so that

$$\hat{p} \text{ is Nor } (\ p \ , \sigma_{\hat{p}} \ )$$

or

$$\hat{p} \text{ is Nor } \left(\ p \ , \sqrt{pq/n} \ \right).$$

Pick a confidence level $C\% = (1 - \alpha)100\%$. Since $\hat{p}$ has a normal distribution, the Z value associated with the C% confidence interval can be determined from the normal table in Appendix A. Let this Z value be $Z_{\alpha/2}$. Then

$$P\left(p - Z_{\alpha/2}\sigma_{\hat{p}} < \hat{p} < p + Z_{\alpha/2}\sigma_{\hat{p}}\right) = 1 - \alpha$$

and since $\sigma_{\hat{p}} = \sqrt{pq/n}$,

$$P\left(p - Z_{\alpha/2}\sqrt{\frac{pq}{n}} < \hat{p} < p + Z_{\alpha/2}\sqrt{\frac{pq}{n}}\right) = 1 - \alpha$$

This states that the probability is $1 - \alpha$ that the values of $\hat{p}$ fall within $Z_{\alpha/2}$ standard deviations of the true mean p. The expression in the probability brackets can be rearranged so that the interval is constructed around the known value of $\hat{p}$. This expression can be shown to be

$$P\left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{pq}{n}} < p < \hat{p} + Z_{\alpha/2}\sqrt{\frac{pq}{n}}\right) = 1 - \alpha$$

The expression in brackets is the C% interval estimate.

The lower confidence limit for this interval is

$$\hat{p} - Z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

and the upper confidence limit is

$$\hat{p} + Z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

The C% confidence interval is thus

$$\left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{pq}{n}}\ ,\ \hat{p} + Z_{\alpha/2}\sqrt{\frac{pq}{n}}\right)$$

Since p and q are not known, there are two choices concerning the values of p and q in $\sqrt{pq/n}$. A conservative choice is to pick $p = q = 0.5$, and for any given n, this produces the widest possible interval. A narrower interval will result if $p = \hat{p}$ and $q = \hat{q}$ are used in $\sqrt{pq/n}$. This makes the interval estimate look a little more precise, but has the danger that the size of the interval estimate could be an underestimate of the true interval width.

8.6 Sample Size

From the sampling error or from the interval estimate, it is a short step to determining the sample size required to achieve a given accuracy of estimate for a population parameter. This section begins with a short discussion of the general method of determining sample size, and then shows how the sample size can be determined for a mean and for a proportion. Examples of each are provided.

If either a population mean or a population proportion is being estimated, a large random sample from the population ensures that the sample statistic is normally distributed. The Central Limit Theorem ensures this for the mean, and the normal approximation to the binomial provides this result for the proportion. Not only is the distribution of the statistic normal, but in each case, a larger sample size is associated with a smaller standard deviation for the sampling distribution of the sample statistic. For the estimate of the population mean, the standard deviation of the sample mean is $\sigma/\sqrt{n}$, and for the estimate of the population proportion, the standard deviation of the sample proportion is $\sqrt{pq/n}$. By increasing the sample size associated with the random sample, the standard deviation of each of the statistics can be reduced until the standard deviation reaches the desired level.

This is the principle on which the determination of sample size is based. An interval width is specified and a probability that this interval width will

not be exceeded is also specified. These, along with the normal distribution of the statistic, are used to determine the sample size which will satisfy these requirements.

Alternatively, the basis for determining the sample size could be the sampling error, introduced in Chapter 7. There the symbol E was used to denote the sampling error. The normal distribution for $\bar{X}$ or $\hat{p}$ provided the basis for determining the probability that the sampling error would not exceed E. This same method, with a little modification, can be used here.

The method of determining sample size is first to specify the **accuracy of the estimate** desired. The accurace of the estimate is the difference between the sample statistic and the population parameter, and is equivalent to the sampling error of Chapter 7. The accuracy of the estimate is given the symbol E, so that the accuracy for a mean is specified as $\bar{X} - \mu$. For a proportion, the accuracy can be specified as $E = \hat{p} - p$. A confidence level, or a probability, that the accuracy of the estimate will be within E must also be given. Call this probability P_E. Then the formulae of the following sections can be used to determine a sample size n. This will provide a sample statistic so that the value of the statistic does not differ from the value of the population parameter by more than E with probability P_E.

These formulas for sample size assume random selection from a population so that the formulas are formulas for **random samples**. It is possible to develop formulas for the required sample size when other types of sampling, such as stratified or cluster sampling, are used. These are not developed in this textbook, but textbooks on sampling use similar methods to develop such formulas. The latter formulas can become quite complex, but are based on the same principles as used for random sampling.

One other point to consider is that these formulas are not intended as hard and fast rules concerning the sample size which will actually be obtained in a specific survey. The formulas which follow are based on considerations of accuracy and probability. Many other considerations such as time, cost, types of respondents desired in the survey, extent of interference with the population, and so on, are also involved in determining the actual sample size selected. In addition, how many of those people selected actually respond will be an important consideration. If the formulas indicate that a sample of size $n = 500$ be selected, but there is a 20% nonresponse rate, then this sample size should be boosted by 20% so that the number of responses actually obtained is 500. Given these other considerations, the following formulas for sample size give the approximate sample size required so that a random sample achieves the required accuracy with the specified

probability.

8.6.1 Sample Size for Estimation of the Mean

Suppose that a researcher wishes to determine the true mean μ of a population. The researcher knows that if he or she obtains a large, random sample of this population, the Central Limit Theorem can be used to obtain the sampling distribution of $\bar{X}$, the sample mean. Based on this distribution, the probabilities of $\bar{X}$ being various distances from μ can be determined. Suppose the researcher wishes this distance to be less than or equal to E. While there can never be absolute certainty that a random sample can yield an $\bar{X}$ which is less than distance E from the true mean μ, the researcher can be sure of this with a certain probability. If this probability is specified, and the accuracy E is also specified, then the required size of the random sample can be determined as follows.

Let E be the accuracy of the estimate, and let P_E be the specified probability. P_E will ordinarily be a large number, and could be equal to the confidence level. For example, if the researcher ultimately wishes to produce a 95% confidence level for the interval estimate of the mean, then the probability selected is $P_E = 0.95$. A random sample with a large sample size is normally distributed as follows

$$\bar{X} \text{ is Nor } \left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Thus the Z value associated with an area of P_E in the centre of the normal distribution can be determined from Appendix A. For $P_E = 0.95$, this would be the familiar value $Z = 1.96$.

The problem now becomes one of determining which normal curve has an area of P_E in the centre, between $\mu - E$ and $\mu + E$. That is, if the sample mean resulting from the sample is between these limits, the probability will be P_E that the accuracy E will not be exceeded. But since this is a normal distribution, the probability is also P_E that $\bar{X}$ will be within Z standard deviations of the centre, where Z is the value associated with P_E in the centre of the distribution. Since one standard deviation for $\bar{X}$ is $\sigma/\sqrt{n}$, Z standard deviations is $Z(\sigma/\sqrt{n})$. If this latter value is set equal to the accuracy of the estimate, then the sample size can be determined.

Using these considerations, the required sample size will be the sample size n which makes

$$E = Z\frac{\sigma}{\sqrt{n}}.$$

That is, E and Z have been specified by the researcher, and σ is given as the true standard deviation of the population from which the sample is drawn. Different sample sizes yield differnt normal curves, and the problem becomes one of determining the sample size so that the specified accuracy is equal to the number of standard deviations associated with the specified probability. Take the expression

$$E = Z\frac{\sigma}{\sqrt{n}}$$

and solve for n by first squaring each side to get rid of the square root sign

$$E^2 = Z^2\frac{\sigma^2}{n}.$$

Now rearrange this expression so that n is on the left. Multiplying both sides by n, and dividing both sided by E gives

$$n = Z^2\frac{\sigma^2}{E^2}$$

This expression can be rewritten in the following manner to make it a little easier to calculate

$$n = \frac{Z^2\sigma^2}{E^2} = \left(\frac{Z\sigma}{E}\right)^2$$

If a random sample of this size is selected from the population, then the probability is P_E that the sampling error $|\bar{X} - \mu|$ will be less than μ, and the $P_E\%$ confidence interval will be

$$\left(\bar{X} - Z\frac{\sigma}{\sqrt{n}}\ ,\ \bar{X} + Z\frac{\sigma}{\sqrt{n}}\right)$$

where Z is the appropriate value from the normal distribution for an area of P_E in the centre of the distribution.

Example 8.6.1 Sample Size for Estimating Mean Individual Income

In Example 8.3.2 interval estimates for the true mean income of individuals in Alberta and Saskatchewan were constructed. In each case 95% intervals were constructed so that they were just over $\pm\$1,000$ *on each side of the sample mean. Determine the sample size so that mean individual income for each province is determined to accuracy* $E = \$500$*, with (i) probability 0.90, and (ii) with 95% confidence. (For each of Saskatchewan*

and Alberta, the sample standard deviation of individual income was around $15,000.)

Solution. *The formula for determination of the required sample size is*

$$n = \left(\frac{Z\sigma}{E}\right)^2.$$

The values Z, σ and E must be obtained, and then the sample size can be determined. Since the standard deviation of income for each of the two provinces was approximately $15,000, the same sample size would be required for each province. This $s = 15,000$ will be used as the estimate if σ in the formula for sample size. Since an accuracy of 500 is required in each province, then $E = 500$ in each case. For probability 0.90, the Z value can be used because the distribution of sample means will be normal with a large random sample. For probability 0.90, the appropriate Z value is 1.645 since the middle 0.90 of the normal distribution occurs between $Z = -1.645$ and $Z = +1.645$. That is, 0.90 in the middle of the distribution means $0.90/2 = 0.45$ on each side of centre. An area of 0.45 in column B of Appendix A is associated with $Z = 1.645$.

Putting these values in

$$n = \left(\frac{Z\sigma}{E}\right)^2$$

gives

$$n = \left(\frac{1.645 \times 15,000}{500}\right)^2 = (49.35)^2 = 2,435.4$$

Rounding this up, to ensure that the sample size is large enough, a random sample of 2,436 individuals from each of Alberta and Saskatchewan would be required to ensure that mean individual income would be estimated correct to within plus or minus $500. If random samples of this size were obtained in each of the provinces, then the sampling error associated with the estimate of mean individual income would have a probability of 0.90 that it did not exceed $500.

For 95% confidence, all that changes is the Z value. For 95% of the area in the middle of the normal distribution, there is $0.95/2 = 0.475$ of the area on each side of centre. Using column B of the normal table in Appendix A gives $Z = 1.96$ as the Z value associated with 95% confidence. The required sample size is

$$n = \left(\frac{Z\sigma}{E}\right)^2 = \left(\frac{1.96 \times 15,000}{500}\right)^2 = (58.8)^2 = 3,457.44.$$

A sample of size $n = 3,458$ would be required in each province in order that the 95% interval estimate would be $\bar{X} \pm \$500$.

Additional Comments on Sample Size. Before presenting more examples, there are a number of points that should be discussed concerning these estimates of sample size. These are as follows.

1. As noted earlier, the sample sizes refer to **random sampling** only. The derivation of the above formula was based on the method of random selection. Some other sampling methods would give a different formula for the required sample size.

2. Also note that the sample size n obtained using the above formula is the **actual number of cases** for which data is obtained. If there is some nonresponse, the sample sizes indicated by the above formula should be boosted so that there will be n cases for which data is actually obtained.

3. Note how n behaves as the other terms vary. Since

$$n = \left(\frac{Z\sigma}{E}\right)^2$$

   n is increased with an increase in σ or Z, and with an decrease in E. A larger σ represents a population with a larger variation. In order to determine the true mean of a population that is disparate, a relatively large sample size is required. An increase in Z occurs when the probability or confidence level is increased. A 99% confidence level would require $Z = 2.575$, while 90% confidence that the interval contains μ would require only $Z = 1.645$. That is, to be more confident that the interval estimate contains μ, or that the sampling error is small, a larger random sample is required. An decrease in E means greater accuracy, that is, the interval estimate is narrower, or the sampling error is smaller.

   A requirement that the researcher obtain a more accurate estimate means that the required sample size will be larger. In the above example, the inital interval estimate of Example 8.3.2 was accurate to within just over $E = \$1,000$ in each province. This required a sample size of just over 600 cases in each of Alberta and Saskatchewan. Requiring that the accuracy be increased, so that $E = \$500$, boosted the sample size by approximately 4 times, to over $n = 2,400$.

4. Note that the **interval width** associated with the interval estimate is twice the accuracy required. That is, if accuracy E is required, and the sample size for this is n, then the interval estimate resulting from that random sample will be $\bar{X} \pm E$. This is an interval width of $2E$. Sometimes the interval width is specified ahead of time. If W is the require interval width for the interval estimate, then

$$W = 2E$$

or

$$E = \frac{1}{2}E.$$

5. Note that **no reference to the size of the population** has been made in the formula for sample size. All that is contained in the formula is the sample size, the confidence level, the accuracy, and the standard deviation of the population from which the sample is drawn. The size of the population does not influence the required sample size, and the resulting interval estimate depends only on the sample size. This result may seem surprising. For example, Alberta has over twice as many people as does Saskatchewan. But each province requires a random sample of about 2,400 people in order to determine the true mean income correct to within ±$500 with 90% confidence. If the true mean income for all Canada were to be estimated to this accuracy, then only a sample of about $n = 2,400$ would be required for the country as a whole. But this sample of 2,400 people across Canada would provide this level of accuracy only for estimates concerning the country as a whole. The resulting samples in each province would have much smaller sample sizes, meaning that the interval estimates for each province would be much wider than $500. For example, if a random sample of $n = 2,400$ Canadian individuals were to be selected, the number of these who live in Saskatchewan would be only about 100, since less than 4% of the Canadian population live in Saskatchewan. The resulting interval estimate in Saskatchewan would be

$$\bar{X} \pm 1.645\frac{15,000}{\sqrt{100}} = \bar{X} \pm 2,468$$

That is, the interval estimate would be approximately ±$2,500, or about $5,000 wide.

The determination of sample size is based on the accuracy required for the smallest group which needs to be estimated. If a sample is required so that mean income is estimated correct to within \$500 with probability 0.90 for each province of Canada, then a sample of approximately $n = 2,500$ would be required in each province, even in Prince Edward Island.

Now the rule that the population size has no influence on the sample size does not hold when the sample size becomes a considerable proportion of the population. For example, suppose a random sample of graduate students is taken at a university. Suppose there are $N = 500$ graduate students, and the required sample size is $n = 150$. In this case, the sample size is a considerable proportion of the population size. According to textbooks of sampling, the required sample size can be reduced somewhat in these circumstances. As a rough rule of thumb, if the sample size indicated by the formula in this section is less than 10% of the population size, then this formula holds. If the sample size approaches, or passes, 10% of the population size, it is likely that the required sample size can be reduced somewhat. The exact formula in this circumstance can be found in texbooks on sampling.

6. Note that the Z value, rather than the t value is used for determining the sample size. This is equivalent to assuming that the required sample size will be 30 or more. This is almost always the case. If the formula yields a required sample size of less than 30, then a t value could be substituted for the Z value in the above expression. But since the t value has a degree of freedom which depends on the sample size, exactly which degree of freedom to use would have to be a matter of trial and error. It seems unlikely that the sample size would be under 30 in most circumstances. Usually the accuracy and probability specified ahead of time lead to a much larger sample size. Then considerations of time and cost may result in the researcher reducing the sample size to below 30. This would then result in greater sampling error, or lower probabilities. The researcher may have little choice but to accept these uf research resources are quite limited.

7. The formula for sample size is quite simple. The one difficulty with the above formula is that it requires knowledge of σ, the true population standard deviation. Since the population has not yet been investigated, the researcher is very unlikely to know the size of σ. There are

several ways in which this problem can be solved. These are as follows.

(a) In the above example, a sample of the population of each of Alberta and Saskatchewan has already been taken. The sample size of this sample was over 600, a relatively large sample size. This gave an estimated standard deviation of individual income of about \$15,000 for each province. While this is not exactly the true standard deviation of individual income for each province, this is likely to be close. If such a sample is available, perhaps from a different year, or for a slightly different variable then this sample standard deviation can be used as a rough estimate of σ in the formula for sample size.

(b) A similar population, or a similar research issue may have been investigated by someone else. For example, if information concerning the variability of incomes in the other provinces of Canada had not been available, but the researcher knows that the standard deviation of income in each of Alberta and Saskatchewan is around \$15,000, then it seems likely that the standard deviation in the other provinces would have a similar value. In fact, from Table 8.3 the standard deviation of income in the other provinces is less than \$15,000 in all cases except Ontario, and even there it is only slightly greater. The estimate of \$15,000 for σ would not be an unreasonable estimate for each province, or for the country as a whole.

Journals and other publications frequently publish standard deviations as part of the research report. These published standard deviations can often be used in other research studies, so that estimates of required sample sizes can be obtained.

(c) If no information concerning the standard deviation is available, then it may be that the range, or some other measure of variation, can be roughly determined. Recall that the standard deviation is approximately equal to the range divided by 4. If the smallest and largest value of a population can be determined, this rough rule can be used to give a very approximate idea of σ, and a sample size can then be estimated.

If not even the range is available, then a researcher might just try to guess what the range might be. This could then be divided by 4, and some very approximate estimate of required sample size obtained.

(d) When designing the research study, it is possible to construct a sampling method so that the random sample can later be enlarged in size, and yet still be random. Suppose that the researcher underestimates the standard deviation of a population, and the formula indicates that a smaple size of only 150 cases is required. After collecting data on these 150 cases, it is found that the sample standard deviation is considerably more than anticipated, and that a sample of size 250 would be required to achieve the desired accuracy with the specified probability. The researcher can then select another 100 cases, again randomly chosen. The first 150 cases, and the subsequent 100 cases, can be pooled together, and the total of 250 case can be regarded as a random sample of $n = 250$ cases from the population.

(e) Remember that the above formula is only one of the considerations in determining sample size. If the sample size is estimated incorrectly, useful data can still be obtained, the only consequence is that the results will not be as accurate as hoped. Thus the failure of this formula in some circumstances may have minimal negative consequences. Finally, if too many cases have been selected, the results will be more accurate than anticipated. This is good, and the only difficulty this will have caused is that the researcher may have devoted more resources than necessary to the study.

Example 8.6.2 Sample Size for Explanations of Unemployment

In Example 8.4.2 data was presented from an attitude survey of Edmonton residents concerning explanations of the reasons for unemployment in Canada. In that study, 6 people were surveyed, and the standard deviation of opinion was 2.137 for these 6 respondents. What is the sample size required so that the mean attitude can be estimated correct to within plus or minus 0.5 points on the attitude scale, with 98% confidence.

Solution. *This is a problem in determining the required sample size for estimating the mean μ of a population. The accuracy requested is $E = 0.5$. The confidence level is 98%, and from column B of the normal table in Appendix A, an area of $0.98/2 = 0.49$ on each side of centre of the normal curve is associated with $Z = 2.33$. The estimate of σ is 2.137, although since this s was based on a sample of only size 6, this value may not be a very accurate estimate of σ. The required sample size is given by*

$$n = \left(\frac{Z\sigma}{E}\right)^2 = \left(\frac{2.33 \times 2.137}{0.5}\right)^2 = 9.958^2 = 99.17$$

A sample size of about $n = 100$ Edmonton adults would be required to achieve accuracy of plus or minus 0.5 points, with probability 0.98.

Example 8.6.3 Sample Size for Self-Esteem of Elderly Women

In Example 8.4.3 the standard deviations for an index of self-esteem was given for two samples of elderly women. Using the data from this example, what is the sample size so that the 95% interval estimate of self-esteem of elderly women who live in their own homes will be no wider than 3 points on the ISE scale.

Solution. *Again, this is a problem in determining the required sample size for estimating the mean μ of a population. The interval requested is an interval width of 3 points. Since the interval width is twice the accuracy E, or $E = W/2$ $E = 3/2 = 1.5$. The confidence level is 95%, and from column B of the normal table in Appendix A, an area of $0.95/2 = 0.475$ on each side of centre of the normal curve is associated with $Z = 1.96$. From Table 8.6 The estimate of σ for community women is $s = 15.0$. Since the estimated standard deviation for the nursing home group is fairly similar at $s = 14.4$, a value of 14 or 15 for σ would appear to be close to the actual value.*

The required sample size is given by

$$n = \left(\frac{Z\sigma}{E}\right)^2 = \left(\frac{1.96 \times 15.0}{1.5}\right)^2 = 19.6^2 = 384.16$$

A sample size of $n = 385$ elderly women who live in their own homes would be required in order to achieve the interval requested.

8.6.2 Sample Size for a Proportion

The method of determining the required sample size for estimation of a proportion is the same as for estimation of the mean, with only a few modifications. If a population proportion p is to be estimated, a random sample of the population is taken and the sample proportion $\hat{p}$ is the point estimate of p. Each random sample from the population yields a different estimate of the true proportion p and the sampling distibution of $\hat{p}$ is given by

$$\hat{p} \text{ is Nor } (p, \sqrt{\frac{pq}{n}})$$

using the normal approximation to the binomial.

Again, the level of accuracy required of the estimate must be specified. Let this level be E. Note that proportions are numbers between 0 and 1, so that E should also be a proportion. For example, an accuracy of

plus or minus 4 percentage points would be given by $E = 0.04$, that is a proportion of $4/100 = 0.04$. In addition, a probability or confidence level must be specified. Let this be P_E. Since $\hat{p}$ is normally distributed, a Z value associated with the normal distribution can be obtained.

The required sample size will be the sample size n which makes

$$E = Z\sqrt{\frac{pq}{n}}.$$

That is, E and Z have been specified by the researcher, and the true standard deviation of the sample proportion is pq/n. Different sample sizes yield differnt normal curves, and the problem becomes one of determining the sample size so that the specified accuracy is equal to the number of standard deviations associated with the specified probability. Take the expression

$$E = Z\sqrt{\frac{pq}{n}}$$

and solve for n by first squaring each side to get rid of the square root sign

$$E^2 = Z^2\frac{pq}{n}.$$

Now rearrange this expression so that n is on the left. Multiplying both sides by n, and dividing both sided by E gives

$$n = \frac{Z^2}{E^2}pq$$

This expression can be rewritten in the following manner to make it a little easier to calculate

$$n = \left(\frac{Z}{E}\right)^2 pq$$

If a random sample of this size is selected from the population, then the probability is P_E that the sampling error $|\hat{p} - p|$ will be less than E, and the $P_E\%$ confidence interval will be

$$\left(\hat{p} - Z\sqrt{\frac{pq}{n}}\ ,\ \hat{p} + Z\sqrt{\frac{pq}{n}}\right)$$

where Z is the appropriate value from the normal distribution for an area of P_E in the centre of the distribution.

Example 8.6.4 Sample Size for Child Discipline and Abuse

Example 8.5.2 gave estimates of the prevalence of various techniques of child discipline among a sample of Toronto mothers and fathers. Determine the sample size required so the the proportion of fathers who pushed shoved, or grabbed a child in the year before the survey is estimated correct to within accuracy plus or minus a proportion 0.05. Use the 90% confidence level.

Solution. *For accuracy E, the required sample size is*

$$n = \left(\frac{Z}{E}\right)^2 pq$$

where p is the proportion of successes and $q = 1 - p$. In order to make sure that we have a sample size which is large enough to achieve the required accuracy, let $p = q = 0.5$. This may overestimate the required sample size. Since the interval is to be correct to within a proportion 0.05, or 5 percentage points, $E = 0.05$. For a probability of 0.90 or 90% confidence, $Z = 1.645$ and

$$n = \left(\frac{1.645}{0.05}\right)^2 0.5 \times 0.5 = (32.9)^2 \times 0.25 = 270.6$$

The required sample size is 271. If this many responses are obtained, then the interval estimate will be incorrect by no more than plus or minus 5 percentage points, with probability 0.90. That is, there is 90% confidence that these interval estimates will contain the true proportion p. Note that if the estimate is to be this accurate for mothers as well, another sample of size $n = 271$ of mothers would also be required.

Notes on Sample Size for Proportions. Many of the comments concerning the determination of sample size for a mean also apply to the estimate of sample size for the proportion. The formula is for a random sample only, the n indicated by the formula is the number of cases for which data must be obtained, and so on. In one way though, it is easier to obtain an estimate of the required sample size for a proportion. Earlier, the determination of the sample size for a mean required some idea of the size of the standard deviation of the population from which the sample was drawn. For a proportion, this is not required because $p = q = 0.5$ can be used in the formula. As shown in Chapter 7, the maximum size of pq is obtained when $p = q = 0.5$. In the determination of the sample size using

$$n = \left(\frac{Z}{E}\right)^2 pq$$

p and q can be replaced with 0.5 for each. If this is done, then

$$n = \left(\frac{Z}{E}\right)^2 (0.5)(0.5) = \left(\frac{Z}{E}\right)^2 0.25$$

For any given confidence level, or proability, and given accuracy E, this gives a sample size at least large enough to meet these requirements. If anything, the sample size indicated may be larger than really required.

If there is evidence that p and q differ rather considerably from 0.5, then $\hat{p}$ can be substituted for p, and $\hat{q}$ for q in the formula. The expression for the required sample size then is

$$n = \left(\frac{Z}{E}\right)^2 \hat{p}\hat{q}.$$

This will yield a smaller sample size than when $p = q = 0.5$ is used. The only problem with this formula, is that if $\hat{p}$ and $\hat{q}$ are incorrect, then the sample size indicated by this latter formula may be a little small to achieve the required accuracy and probability.

Example 8.6.5 Common Sample Sizes for a Proportions

If the values $p = q = 0.5$ are used for p and q in the determination of the sample sizes for a proportion, then some of the more common xample sizes can be determined. Using

$$n = \left(\frac{Z}{E}\right)^2 pq$$

and letting $p = q = 0.5$ gives

$$n = \left(\frac{Z}{E}\right)^2 (0.5)(0.5) = \left(\frac{Z}{E}\right)^2 0.25.$$

Table 8.8 gives some of the commonly used sample sizes for a random sample. In each case, the true proportion of members of the population with the characteristic in question is being estimated. The values of E on the left give some of the common levels of accuracy. Each is given in terms of a proportion. For example, the first row gives the sample sizes for estimation of a proportion correct to within plus or minus a proportion 0.05, that is, within ± 5 percentage points. These sample sizes are for the common confidence levels of 90%, 95% and 99%. As an exercise, you should be able to obtain each of these sample sizes using the above formula.

Level of Accuracy (E)	Confidence Level 90%	95%	99%
0.05	271	385	664
0.04	423	601	1,037
0.03	752	1,068	1,842
0.02	1,692	2,401	4,145
0.01	6,766	9,604	16,577

Table 8.8: Sample Sizes for a Proportion, Common Levels of Accuracy and Confidence

Note that the sample size for the Gallup poll of Example 8.5.1 can be closely approximated from this table. Gallup says that their sample is correct to within plus or minus 3.1 percentage points in 19 out of 20 samples. If the level of accuracy $E = 0.03$ is used as being very close to the level claimed by Gallup, and if 95% confidence is used, then the required sample size is $n = 1,068$. This is slightly more than the just over 1,000 cases that Gallup surveys each month. Also note that the sample size of $n = 271$ of Example 8.6.4 can be determined from this table, using 90% confidence and $E = 0.05$.

8.6.3 Notation for Sample Size

The α notation introduced earlier can be used to provide a more complete formula for the determination of sample size. If you had difficulty with the earlier sections using this notation, skip this section.

The α notation was used to determine the probability or confidence level. If a confidence level of $C\% = (1 - \alpha)100\%$ is desired, and if the statistic is normally distributed, then the appropriate Z value is $Z_{\alpha/2}$. This is the Z value in the formula for the determination of the sample size, when estimating either a population mean or a population proportion. Also let the accurace of the estimate be E, and the standard deviation of the population be σ.

Using the arguments developed earlier, for the estimate of the mean,

$$n = \left(\frac{Z_{\alpha/2}\sigma}{E} \right)^2$$

and for the estimate of the proportion,

$$n = \left(\frac{Z_{\alpha/2}}{E}\right)^2 pq.$$

For the latter expression, you have a choice of using $p = q = 0.5$ or $p = \hat{p}$ and $q = \hat{q}$, depending on the circumstances.

8.7 Conclusion

This chapter has presented the method of estimation for the mean of a population and for the proportion in a population. The method is essentially the same in each case. The sample mean provides a point estimate of the population mean, and the sample proportion provides a point estimate of the population proportion. In each case, the interval estimate is the Z or t value multiplied by the standard deviation of the sample statistic. If C% is the confidence level used, then the probability is C that these intervals contain the population parameters.

Interval estimates and the sampling error can be used to provide estimates of the sample size required, before a sample has been selected. The formulas for determining sample size were given in the last section of the chapter.

In the next chapter, the statistics obtained from samples are interpreted in a different manner. The method of hypothesis testing is discussed in Chapter 9. While the method of hypothesis testing may appear to be quite different from the method of interval estimation, the two methods are very similar. In each case, data concerning a population is obtained from a sample, and these sample results are used to provide inferences concerning the characteristics of a population. The similarity of the two methods will be discussed near the end of Chapter 9.

8.8 Additional Problems

Problems on Sample Size. Suggested solutions for the following problems are given at the end of this section.

1. A random sample of the population of Saskatchewan is to be taken to determine the percentage of Saskatchewan residents who favour free trade. The sample is to determine this percentage correct to within ± 2 percentage points with probability 0.94. In order to achieve this degree of accuracy the sample size should be approximately:

   (a) 1414
   (b) 2209
   (c) 553
   (d) 1512
   (e) 45

2. You are administrator of a large agricultural region where the largest farm is 6 sections (1 section=640 acres). How large a sample size would be required so that you can be 90 per cent sure that the mean farm size is correct to within 100 acres?

   (a) 354
   (b) 250
   (c) 16
   (d) 675
   (e) 998

3. According to geneticist John Cohen, the range of IQ scores is comparatively narrow, from 70 to 130. Based on this, how large a sample size would be required so that the mean IQ can be estimated correct to within ± 5 points, at the 95 per cent confidence level?

   (a) 9
   (b) 35
   (c) 554
   (d) 139

(e) 390

4. A sample of 3 acquaintances who were dabbling on the stock market showed that one lost $5,000 last year, one gained $5,000 last year and the third person's holdings did not change in value. If these three people are regarded as representative of all people who played the market, how large a sample size would be required to determine the mean change in value of stocks correct to within $500 with 95 per cent confidence?

   (a) 16
   (b) 385
   (c) 96
   (d) 769
   (e) 400

5. A random sample of Saskatchewan residents is to be taken to determine the percentage who support the revised Constitution of Canada. If the percentage is to be determined so that the confidence interval is no more than 4 percentage points wide, with probability 0.96, the sample size should be approximately:

   (a) 660
   (b) 1900
   (c) 575
   (d) 145
   (e) 2630

Political Preferences in Alberta

The 1991 Alberta Survey conducted by the Population Research Laboratory at the University of Alberta gives the data contained in Table 8.9. The table gives the pattern of political preferences by income of the household. The entries in each cell of the table are the number of respondents with the specified combination of characteristics.

1. The authors of the study comment that "Compared to other parties, NDP supporters reported somewhat lower incomes." They also note

Household Income	Political Voting Preference PC	Liberal	NDP	Reform	Total
$< \$20,000$	21	32	30	22	105
$\$20 - 39,999$	40	52	51	72	215
$\$40 - 59,999$	50	54	61	72	237
$\$60 - 79,999$	27	22	32	40	121
$\$80,000+$	23	40	24	42	129
Total	161	200	198	248	807

Table 8.9: Political Voting Preferences and Incomes of Alberta Voters, 1991, Number of Respondents

that "Only 9% of the Reform groups were in the lowest category." Obtain interval estimates for each group and comment on the conclusions of the authors?

2. Based on the data in Table 8.9, derive the 93% interval estimate for the proportion of all Alberta residents who support the Reform Party. Explain in words the meaning of this interval estimate.

3. What sample size would be required so that the proportion of Reform Party supporters in Alberta could be estimated to within plus or minus 2.5 percentage points, with 85% confidence?

Answers to Problems on Sample Size.

1.
$$n = \frac{Z^2 pq}{E^2} = \frac{1.88^2(0.5)(0.5)}{0.02^2} = 2209$$

2. The smallest possible farm is 0 acres and the largest farm is $6 \times 640 = 3,840$ acres. A rough estimate of the size of the standard deviation of

farm size is the range divided by four or $3,840/4 = 960$.

$$n = \left(\frac{Z\sigma}{E}\right)^2$$

$$n = \left(\frac{1.645 \times 960}{100}\right)^2$$

$$n = (15.792)^2 = 250$$

3. Based on the same formula as in the last question, the rough estimate of the standard deviation is the range divided by 4, or $(130 - 70)/4 = 15$.
$$n = \left(\frac{1.96 \times 15}{5}\right)^2 = 35.$$

4. For this question, the sample of 3 can be used to estimated the standard deviation. The formula for the standard deviation gives $s = \$5,000$. Using the same formula as in the last two questions,
$$n = \left(\frac{1.96 \times 5,000}{500}\right)^2 = 385.$$

5.
$$n = \frac{Z^2pq}{E^2} = \frac{2.052(0.5)(0.5)}{0.02^2} = 2627$$

The sample size which comes closest is 2630.

Solution for Problems Concerning Alberta Political Preferences

1. For each of the NDP and the Reform Party, consider the lowest income group, household income of less than $20,000. In order to determine the proportion, this is the characteristic being investigated. There are 30 out of 198 NDP supporters who are in the lowest income group, so that the point estimate of the proportion of NDP supporters who are low income is $\hat{p_N} = 30/198 = 0.152$. For the Reform Party, the point

estimate of the proportion with low income is $\hat{p_R} = 22/248 = 0.089$. The sampling distribution of the test statistic is normal if

$$n \geq \frac{5}{\min(p, q)}.$$

Or the two $\hat{p}$s given here, $\hat{p_R}$ is the smaller, so that it is sufficient to check to see if the sample sizes are large enough based on this.

$$\frac{5}{0.089} = 56.4$$

and both samples sizes of 198 and 248 exceed this. Thus

$$\hat{p} \text{ is Nor } \left(p \ , \sqrt{pq/n} \ \right).$$

and the interval estimates can be obtained using this. The confidence level is not stated in the question. This estimate will be based on the 95% confidence level, and the Z value is $Z = 1.96$. For the NDP, the interval estimate is

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

$$0.152 \pm 1.96\sqrt{\frac{0.5 \times 0.5}{198}}$$

$$0.152 \pm 1.96\sqrt{0.00126}$$

$$0.152 \pm (1.96 \times 0.0355)$$

$$0.152 \pm 0.070$$

The 95% interval estimate for the true mean proportion of NDP supporters who are low income is (0.072,0.022)

For the Reform Party, the interval estimate is

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

$$0.089 \pm 1.96\sqrt{\frac{0.5 \times 0.5}{248}}$$

$$0.089 \pm 1.96\sqrt{0.0010}$$

$$0.089 \pm (1.96 \times 0.0318)$$

$$0.089 \pm 0.062$$

The 95% interval estimate for the true mean proportion of Reform Party supporters who are low income is (0.027,0.151)

If the sample proportions are used as estimates in the standard deviation, the interval narrows somewhat. For the Reform Party, the interval estimate then becomes

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

$$0.089 \pm 1.96\sqrt{\frac{0.089 \times 0.911}{248}}$$

$$0.089 \pm 1.96\sqrt{0.000327}$$

$$0.089 \pm (1.96 \times 0.018)$$

$$0.089 \pm 0.035$$

The 95% interval estimate for the true mean proportion of Reform Party supporters who are low income is (0.054,0.124). As an exercise, show that this method produces a 95% interval (0.102,0.202) for the NDP.

Before commenting on the conclusions, note that the point estimate of the proportion of PC supporters who are low income is $21/161 = 0.130$ and for the Liberal supporters, there are $32/200 = 0.16$ who are low income. Given the similarity of these point estimates to the point estimate for the NDP, it may be that the NDP supporters are not much less likely to report lower incomes than are supporters of other parties. The Reform Party does appear to have the smallest proportion of low income supporters. But even here the 95% interval estimate is fairly wide, plus or minus 3.5 to 6.2 percentage points, depending on which method is used. While the point estimates appear to support the author's contention, at least with respect to the NDP and Reform Party, the sample sizes are not large enough to make definitive conclusions. The interval estimates are too wide to clearly separate the parties concerning the proportion of low income supporters.

2. In the sample as a whole there are $n = 807$ respondents, and of these, 248 support the Reform Party. The characteristic being investigated in this part is support for the Reform Party, so that the point estimate

of Reform Party supporters is $\hat{p} = 248/807 = 0.307$. Since the sample size is large,

$$\hat{p} \text{ is Nor } \left(p \ , \sqrt{pq/n} \right)$$

and the interval estimate is

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

$$0.307 \pm 1.81\sqrt{\frac{0.5 \times 0.5}{807}}$$

$$0.307 \pm 1.81\sqrt{0.0003098}$$

$$0.307 \pm (1.81 \times 0.0176)$$

$$0.307 \pm 0.032$$

The 93% interval estimate for the true mean proportion of Reform Party supporters who are low income is (0.275,0.339). The researcher can be reasonably confident that the true proportion of Reform Party supporters in Alberta lies within these limits.

3. For 85$ confidence, $Z = 1.44$.

$$n = \frac{Z^2 pq}{E^2} = \frac{1.44^2(0.5)(0.5)}{0.025^2} = \frac{0.5184}{0.000625} = 829.44$$

The sample size required is approximately 830.