Contents

Chapter 7

Sampling Distributions

7.1 Introduction

This chapter begins inferential statistics, the method by which inferences concerning a whole population are made from a sample. Inferential statistics is concerned with estimation and hypothesis testing. Estimation uses the data from samples to provide estimates of various characteristics of samples, especially the population mean and the proportion of successes in the population. In hypothesis testing, an hypothesis concerning the nature of the population is made. This hypothesis may concern the nature of the whole distribution or the value of a specific parameter. For example, it may be possible to test whether a distribution of grades in a class can be considered to be more or less normally distributed. Alternatively, if a value of a particular parameter, such as the mean, is hypothesized, the data provided from a sample can often be used to test whether the value of the parameter is as hypothesized or not.

Inferential Statistics. Inferences concerning a population make use of the principles of probability. Under certain conditions, conclusions concerning the distribution of a population or concerning the values of particular parameters, can be made with a certain probability. For example, on the basis of a sample, the mean for a population may be estimated to be within a specific range with probability 0.90. Alternatively, an hypothesis may be proved on the basis of a sample, but there may be a probability of 0.95 that this conclusion is incorrect. All inferences concerning the nature of a population have a probability attached to them.

Since all statistical inferences are based on probability, this means that such conclusions are never absolutely certain. Rather, statistical proofs and statements are always stated with some degree of uncertainty. This uncertainty is quantifiable in a probability sense, and may be extremely small. An hypothesis might be proven with only 0.0001 chance of being incorrect. The manner in which probabilities are interpreted in statistical inference will be discussed in detail in the following sections and chapters. Suffice it to say at this point that the probability based method of making statistical conclusions has proven extremely useful, and allows researchers to deal with uncertainty in a realistic and meaningful manner.

Since conclusions in inferential statistics are based on probability, this means that inferential statistics can be used when the conditions for probability which were outlined in Chapter 6 are satisfied. While these conditions are not always exactly satisfied in practice, researchers attempt to match these principles as closely as possible. Some of the ways this can be done are outlined in the following paragraphs.

Random Sampling. The circumstance in which the conditions for probability can be most closely matched is usually considered to be in random sampling. A random sample is a method of sampling such that each member of the population has an equal chance of being selected. If this condition is truly satisfied when selecting a sample, then the principles of probability of Section 6.2 are satisfied. These conditions may be modified in order to allow for unequal probabilities of selection of different members of a population, as in stratified or cluster sampling. As long as the probability of selection for each member of a population can be determined, the methods of inferential statistics can be applied. In this textbook, only the case of random sampling is discussed. Most of Chapter 7 is concerned with random sampling, and the nature of samples drawn on the basis of random selection.

Experimental Design. Another circumstance where the principles of probability can be closely matched is in experimental design. In those research areas where either inanimate objects or people can be given different experimental treatments, then this method may be more practical than surveys and sampling. In agricultural experiments, areas of a field crop may be randomly assign different levels of fertilizer, rainfall, cultivation, or other treatments. In psychological experiments, a wide range of subjects is selected, and different tests or treatments are randomly assigned to these

subjects, often without the researcher even being aware of which treatment is being applied to each subject. These random assignments are carried out at least partly to allow application of the principles of probability and statistical inference in experimental design. While randomized types of experimental design are important in many research areas, these methods are not examined in any detail in this textbook. In the social sciences, extensive analysis of statistical inference in experimental design can be found in textbooks of behavioural statistics in psychology.

Probabilistic Models. A third way in which the principles of probability are applied is to use probabilistic models as a means of attempting to explain social phenomena. The binomial and the normal probability distributions are the two main distributions which will be used in this textbook, but there are many other probability based models which are used in the social sciences. Each of these models is based on certain clearly stated assumptions. If the model yields results which come close to matching conditions found in the real world, then the model may provide an explanation of reality. This may be no more than a statistical explanation, although in some cases researchers may regard the model as explaning the real social phenomenon being examined. In this case, the assumptions on which the model is based may be regarded as approximating the real world processes which produce the observed phenomenon.

When using a model in inferential statistics, the researcher begins by assuming that the model explains the social phenomenon. The assumptions of the model, and the processes by which the model works are assumed to correctly explain reality. The phenomenon in question is observed, and the probability of the phenomenon occurring in the manner it does is determined on the basis of the model. If this probability is extremely low, then the model may be considered to be inappropriate, and not capable of providing an adequate explanation of what was observed. Alternatively, the processes involved in the model may work as hypothesized, but the assumptions on which the model was based do not approximate reality. In either case, the discrepancy between the assumptions, the model, and the real world can be further studied, and knowledge concerning the social phenomenon may be developed.

On the other hand, if the probability associated with the observed result is fairly large, then the model, and the assumptions on which the model is based, may be regarded as providing an explanation of reality. The model

is then likely to become acceptable as providing an explanation of the social phenomenon,

Outline of Chapter. This chapter begins with a short discussion of sampling, in particular representative and random sampling. Section 7.4 discusses the behaviour of statistics such as the sample mean, standard deviation and proportion, when there are repeated random samples from a population. The manner in which these statistics behave under repeated random sampling is referred to as the sampling distribution of each of these statistics. The sampling distribution provides a means of estimating the potential sampling error associated with a random sample from a population. This is demonstrated in Sections 7.5 and 7.6. Sampling distributions also lay the basis for the estimation and hypothesis testing of chapters 8-10.

7.2 Representative Samples

As noted in Chapter 2, a sample can be regarded as any subset of a population. An observation concerning one member of a population can be a sample of that population. Such a sample is not considered to be a very good or useful sample in most circumstances. In order to obtain a better sample, it is usually recommended that a researcher obtain a **representative sample**. There are various definitions of exactly what a representative sample might be. Roughly speaking, a representative sample can be considered to be a sample in which the characteristics of the sample reasonably closely match the characteristics of the population.

A sample need not be representative in every possible characteristic. As long as a sample is reasonably representative of a population in the characteristics of the population being investigated, this is usually considered adequate. For example, if a researcher is attempting to determine voting patterns, as long as the sample is representative of voting patterns in the population as a whole, then this may be adequate. Such a sample might not be representative of the population in characteristics such as height, musical preference, or religion of respondents. However, if any of these latter characteristics do influence voting patterns, then the researcher should attempt to obtain a sample which is reasonably representative in these latter variables as well. Religion may be associated with political preference, and if the sample does not provide a cross section of religious preferences, then the sample may provide a misleading view of voting patterns.

Example 7.2.1 Representativeness of a Sample of Toronto Women

The data in Table 7.1 comes from Michael D. Smith, "Sociodemographic risk factors in wife abuse: Results from a survey of Toronto women," **Canadian Journal of Sociology** *15 (1), 1990, page 47. Smith's research involved a telephone survey using*

> a method of random digit dialing that maximizes the probability of selecting a working residential number, while at the same time producing a simple random sample ... .

Variable	Sample (%)	Census (%)
Age		
20-24	21	18
25-34	44	50
35-44	35	32
Total	100	99
(n)	(490)	(753,320)
Ethnicity		
Canadian, British, Irish	70	74
Italian	4	6
Portugese	2	2
Greek	2	1
Other	23	17
Total	101	100
(n)	(588)	(3,253,350)

Table 7.1: Comparison of Sample Characteristics with Census Data

In order to determine the representativeness of the sample, the characteristics of the sample of Toronto women were compared with the characteristics of all Toronto women, based on data obtained from the 1986 Census of Canada. In the article, Smith comments that the data

> reveal a close match between the age distributions of women in the sample and all women in Toronto between the ages of 20 and 44. This is especially true in the youngest and oldest age brackets. ... Comparing the sample and population on the basis of ethnicity is even cruder because different measures of ethnicity were used; ... Nevertheless, the ethnic distributions ... were surprisingly similar.

While Smith provides no statistical tests to back his statements, a comparison of the percentage distributions of the sample and the population, based on the Census, is provided in Table 7.1. The sample distributions appear to provide a close match to the Census distributions for the characteristics shown. Smith argues

> As far as can be determined on the basis of these limited data, the sample was roughly representative of Toronto women.

The implication of Smith's analysis is that this is a good sample, and that many of the results from this research from this sample can be taken to represent the situation with respect to women as a whole in Toronto. One of Smith's major findings is that "low income and marital dissolution are strongly and consistently related to abuse."

This example shows the importance of attempting to determine the representativeness of a sample. If the sample is not representative in some of the relevant characteristics, then the applicability of the research results to a whole population could be questioned. In terms of the method used, this example compares whole distributions of various characteristics for the sample and the population. That is, no summary measures of population characteristics such as the mean, were used by Smith. Rather, the percentage distributions of sample and population were compared. In Chapter 10, the chi square test will provide a method of testing whether the two distributions are really as close as claimed by Smith.

Another method of determining the representativeness of a sample is to compute summary statistics from a sample, and compare these with the same summary measures, or parameters, from the population as a whole. The summary measurs used are usually the mean and various proportions. These are illustrated in the following examples.

Example 7.2.2 Survey of Farms in a Saskatchewan Rural Municipality

A study of farm families, conducted by researchers at the University of Regina in 1988, examined a sample of 47 farms in the rural municipality of Emerald, Saskatchewan. The mean cultivated acreage per farm for these 47 farms was 1068.8 acres. The 4 farms that produced flax, in the survey of Emerald, had flax yields of 381.0, 279.4, 127.0 and 381.0 kilograms per acre, for a mean flax yield of 291.1 kilograms per acre.

In **Agricultural Statistics 1989**, *the Economics Statistics section of Saskatchewan Agriculture and Food gives data concerning the characteristics of all farms in various areas of the province. This publication states that the mean number of cultivated acres per Saskatchewan farm was 781 acres in 1986. The same publication reports that in Crop District 5b, of which the rural municipality of Emerald is a part, the mean yield of flax in 1989, for farms that produced flax, was 400 kilograms per acre.*

Based on these respective means, it is apparent that the sets of means are considerably different. The mean flax yield of the four farms is over 100 kilograms per acre below the mean for all farms Crop District 5b. This may mean that a sample of size 4 is too small to yield a very representative mean for the farms in this area, or that Emerald as a whole has a somewhat different mean flax yield than does Crop District 5b generally.

The sample mean cultivated acreage is about 300 acres greater than mean cultivated acreage for all Saskatchewan farms. Again, this could be because the farms in Emerald are larger than the Saskatchewan average farm size, or because the sample is not exactly representative of farms in Emerald. It is clear though, that the characteristics of this sample should not be taken as being representative of the farms in Saskatchewan as a whole.

Example 7.2.3 Gallup Poll Results

Ordinarily the exact degree of representativeness of Gallup poll results is not known. A sample of approximately 1000 Canadian adults is taken each month, and the percentage of these adults who support each of the political parties in Canada is reported. Just before the 1988 federal election, Gallup polled over 4000 Canadian adults. A few days later the federal election was held. Table 7.2 gives the actual results from the 1988 election, along with the Gallup poll conducted on Noveber 19, 1988.

	Per Cent Supporting:			
	PC	Liberal	NDP	Other
Election	43%	32%	20%	5%
Nov. 19, 1988	40%	35%	22%	3%

Table 7.2: 1988 Election Results and Nov. 19, 1988 Gallup Poll Result

The percentage supporting each political party can be seen to be quite close to the actual result in the 1988 federal election. Gallup slightly underestimated the percentage of the Canadian electorate who voted Conservative and overestimated the percentage who voted Liberal and NDP. In terms of the representativeness of the sample though, Gallup appears to have obtained quite a good sample. Estimates of the popular vote for each political party can be reasonably accurately obtained on the basis of opinion polls. In Canada, it is much more difficult to predict the standings in terms of the number of seats in Parliament obtained by each political party. This is because each seat is determined independently of other seats, and the overall popular vote across Canada may be an inaccurate indicator of the results in any particular constituency.

In Examples 7.2.2 and 7.2.3, the notion of sampling error can be used to consider the representativeness of each sample. Roughly speaking, the sampling error is the difference between the value of a statistic in a sample and the corresponding value of this statistic if a survey of the whole population were to be conducted. In the political prefercence example, the last Gallup poll before the election predicted that 40% of the electorate would vote Conservative, while 43% of those who voted cast their vote for the Conservatives. The sampling error associated with this Gallup poll is thus $43 - 40 = 3$ percentage points. In the survey of farms in Example 7.2.2, the 4 farms in Emerald give a sampling error for mean flax yield in Crop District 5b of $291.1 - 400 = -108.9$, approximately -110 kilograms per acre.

When the sampling error is relatively small, as in the case of the Gallup poll, then the sample can be said to be reasonably representative of the population. Where the sampling error is larger, as in the case of the farm survey, the sample is not representative of the population.

Measure	Parameter	Statistic	Sampling Error
Mean	μ	$\bar{X}$	$\lvert\bar{X} - \mu\rvert$
Standard Deviation	σ	s	$\lvert s - \sigma\rvert$
Proportion	p	$\hat{p}$	$\lvert\hat{p} - p\rvert$

Table 7.3: Statistics, Parameters and Sampling Error

Sampling Error. In order to define sampling error, the distinction between statistics and parameters is useful. Table 7.3 gives the most commonly used parameters and statistics, along with the sampling error. Recall that the parameter, or population value, is the true value of the summary measure for the population as a whole. The statistic is the corresponding summary measure based on data from a sample. The mean cultivated acreage reported for all Saskatchewan farms in the 1986 Census was $\mu = 781$ acres. The sample of $n = 47$ farms in the Emerald rural municipality reported a sample mean cultivated acreage of $\bar{X} = 1068.8$ acres. While the Emerald sample was not intended to be representative of the whole province of Saskatchewan, if this sample were to be used to estimate the true mean for all Saskatchewan, the sampling error would be

$$\bar{X} - \mu = 1068.8 - 781 = 287.7$$

or approximaterly 290 acres.

Note that the sampling errors in Table 7.3 are given as **absolute values**. The absolute value $|X|$ of a number X is the magnitude of the number, without reference to whether it is positive or negative. For example, the absolute value of 5 is written $|5|$ and is equal to 5. The absolute value of -5 is written $|-5|$ and is also equal to 5. That is, the absolute value of any number represents the magnitude of the number, without reference to whether it is positive or negative.

For the Gallup poll result, the parameter is the true proportion of the population which voted Conservative in the 1988 election. Let this true proportion be p, and the sample proportion be $\hat{p}$. Since the election results are known, $p = 0.43$. On the basis of the November 18 sample, the estimate of the proportion who would vote Conservative is $\hat{p} = 0.40$. The sampling

error is

$$|\hat{p} - p| = |0.40 - 0.43| = |-0.03| = 0.03.$$

That is, the sampling error is a proportion 0.03, or 3 percentage points.

Based on these considerations of sampling error, the representativeness of a sample can be more carefully defined. A sample with a small sampling error in the characteristic being investigated is considered representative of the population. A sample with a larger sampling error in this characteristic is considered to be a less representative sample. A sample cannot usually be considered to be exactly representative of a population because there is almost always some sampling error associated with any sample. What a good sampling method does is reduce the sampling error to a low level.

The concept of sampling errror is relatively straightforward, and illustrates the nature of representativeness of a sample. The difficulty in discussing sampling error is that the values of parameters concerning a population are generally not known. That is, μ, σ or the population proportion p, are not known by the researcher. If these parameters were known, then there would be little need for sampling in the first place. The examples given above, where the population values and the sample values are both known, are unusual. Often some of the population values can be determined from a Census, or administrative data sources. But is unlikely that the parameters for all the variables which the researcher wishes to investigate are known.

Since the values of the parameters are not known, the size of the sampling error cannot be determined. This is certainly the case before the sample is selected and before the results from the sample have been analyzed. But even after $\bar{X}$, s and $\hat{p}$ have been determined, the parameters which correspond to these statistics are not known. As a result, the exact size of the sampling error cannot be determined. For example, even though $\bar{X}$ can be determined, $|\bar{X} - \mu|$ cannot be determined because μ is unknown. While the researcher can be sure that there is some sampling errror associated with each sample, the extent of this error is not certain.

The principles of probability become important at this point. While the exact size of sampling error cannot be determined, probabilities can be attached to various levels of this sampling error. For example, in the Gallup poll, Gallup usually states that a national sample of 1,000 respondents has a sampling error of no more than 4 percentage points associated with it. Gallup states that this level of sampling error is not exceeded in 19 out of 20 samples. Stated in terms of probability, this means that the probability is $19/20 = 0.95$ that the sample proportion $\hat{p}$ differs from the true proportion

p by not more than 4 percentage points, or 0.04. Symbolically,

$$P(\ |\hat{p} - p| < 0.04\ ) = 0.95$$

The calculations which show this are given later in this chapter in Example 7.6.2. Similarly, when estimating the sample mean μ, the exact size of the sampling error $|\bar{X} - \mu|$ will never be known. But under certain conditions, the probability that this sampling error does not exceed a particular value can be determined.

The following section discusses sampling, emphasizing random sampling and other types of probability samples. In probability based sampling it is possible to determine probabilities of various levels of sampling error. Section 7.3 leads to Section 7.4.1 which discusses the manner in which $\bar{X}$ behaves when random samples are drawn from a population. Following the results of that section, it is possible to determine different levels of sampling error, along with their associated probabilities.

7.3 Sampling

As noted earlier, any subset of a population can be regarded as a sample of that population. As a result, some samples are very poor samples, being quite unrepresentative of the population. Other methods of sampling may provide much better samples, where a reasonably representative sample may be obtained. This section briefly discusses a few principles of sampling, concentrating on random sampling.

Samples can be divided into probability based samples, and nonprobability samples. Probability based samples are those sampling methods where the principles of probability are used to select the members of the population for a sample. Random sampling is an example of a probability based method of sampling. The advantage of using a probability based technique of sampling is that the principles of probability can be used to quantify and place limits on uncertainty. In particular, probabilities of various levels of sampling error can be determined if a sample is selected using the principles of probability. The statistical inference of the following chapters is based on the assumption that the sample is selected on the basis of some of the principles of probability. Before discussing random sampling, a few comments concerning nonprobability sampling are made.

Nonprobability Samples. Statisticians are often reluctant to recommend that a sample be drawn without using some systematic principles of random selection. However, in many circumstances, it is not really possible to select a probability based sample, either because of the nature of the problem, or because of the time or cost factors involved. In these circumstances, it may be necessary, or even desirable, to select a sample on some nonprobability basis. If such selection is done with care, then some of these methods may yield very useful and meaningful results. Further, if the researcher makes some effort to determine how representative the sample is in terms of some of the characteristics being investigated, it may be that a fairly representative nonprobability sample can be obtained.

Some of the types of nonprobability samples are the person in the street interview, the mall intercept, asking for volunteers, surveying friends or acquaintances, quota samples and the snowball sample. If care is taken in attempting to obtain diverse types of people, the person in the street interview may obtain a reasonable cross section of the population. The mall intercept is essentially the same method, stopping people in a particular location, and obtaining some information from them. Asking for volunteers or relying on friends can often yield quite an unrepresentative sample. However, if the researcher is not interested in obtaining a cross section of a population, but is attempting to find people who have relatively uncommon characteristics, then these methods may yield the desired sample. A quota sample begins by deciding how many people of each characteristic are to be selected. In this method, a sample of size 100 may aim at selecting 50 males and 50 females. Then some method of finding these people, such as going from house to house, is used. If care is taken in constructing quotas, and using a systematic approach to finding these people, then this technique can produce quite a representative sample in the characteristics given in the quotas. A snowball sample is a method of beginning with a few people having the desired characteristics, and asking them to suggest others. The sample will grow, or snowball, as more names are suggested for inclusion in the sample. Some of these are discussed in more detail in Chapter 12.

The difficulty with all of these nonprobability methods of obtaining a sample is that statistical inference is not possible. That is, the level of sampling error, or the probability associated with this error cannot be determined. As a result, the uncertainty associated with these samples cannot easily be quantified in any meaningful manner. If the nonprobability sample yields useful information, then this may not be a concern. In addition, by comparing sample results with known characteristics of a whole population,

it may be found that a nonprobability sample is representative of the population. If using nonprobability samples though, care must be taken by the researcher concerning the extent to which research results can be generalized to a larger population.

Random Sampling. Random sampling is a method of sampling whereby each member of a population has an equal chance of being selected in the sample. Not many members of a population would ordinarily be selected in such a sample, but no one in the population is systematically excluded from the list from which the sample is drawn. In addition, each member of the population on this list has exactly the same probability of being included in the sample as does any other member of the population. Where these conditions are satisfied, then the sample is said to be random.

There are various ways in which a random sample may be selected. Each member of the population could be given a ticket, these tickets could be placed in a drum, the drum could be shaken well, and one or more cases could be randomly picked from this drum. Such a sample could be drawn either with or without replacement, both methods satisfy the conditions laid down for random sampling.

When dealing with a list of a population, it is most common to use a **random number table**. Such a table is given in Appendix **??** of this textbook. The list of 1,000 numbers in Appendix **??** was randomly generated using a computer program Shazam. In this table, each of the ten digits, 0 through 9, should occur with approximately equal frequency, but with no discernable patterns of repeated sets of digits. Assuming that these 1,000 integers really are random, they can be used to select a random sample from a population list. If there are N members of the population on the list, each member of a population is given a different identification number, from 1, 2, 3, and so on through to N. Sets of random digits from the table are then used to select the number of cases desired in the sample. In the description at the beginning of Appendix **??**, the method of selecting a sample of size $n = 5$ from the population of $N = 50$ people in Appendix A is described.

One method of sampling which has become common is to obtain a sample on the basis of random digit dialling of telephone numbers. (Example 7.2.1 used this method for the sample of Toronto women). A list of all the telephone numbers within particular telephone exchanges may be obtained, and then random numbers are generated. A sufficient set of random digits is generated so that the particular set of 7 digit telephone numbers yields an

adequate number of responses. This method of selecting telephone numbers is superior to selecting numbers randomly from a telephone book. The latter does not have unlisted telephone numbers and may be out of date.

At the same time, there are several reasons why random digit dialling does not yield a perfectly random sample of a population. First, people who do not have telephones have no chance of being selected, while those with more than one telephone stand a greater chance of being selected than do those with only one telephone. As a sample of households, random digit dialling provides close to a representative sample. As a sample of individuals, this method is less representative. Individuals in households with many members have considerably lower probability of being selected than do individuals in smaller households. In addition, telephone interviews have nonsampling errors. This is the same problem as that encountered in most methods of sampling - nonresponse or refusals. In spite of these difficulties, random digit dialling of telephone numbers has become a popular, and relatively good method of selecting a sample, one which is close to random.

Other Probability Based Sampling Methods. In many circumstances, selecting a random sample is neither feasible nor efficient in terms of time and cost. Modifications of the principles of probability can often be used in such circumstances. For example, when sampling households in cities, the city block seems a natural unit to sample. Some blocks, though, have more households than do other blocks. When selecting blocks for sampling, selection may be made on the basis of **probability proportional to size**. That is, if the number of households on each block can be determined from the Census, then probabilities of selection proportional to this number can be assigned. Blocks with more households then have a greater probability of being selected, and blocks with fewer households have a lower probability of being selected.

Another method of sampling is to **stratify** the population into groups, and then draw random samples from each of the strata. For example, students at a university might be stratified into first year, second year, third and fourth years, and graduate students. Then a random sample of each group, or a sample proportional to the size of each group, might be selected. This is termed a **stratified sample**. In some circumstances, a stratified sample provides a more representative sample than does a random sample.

The other common type of sample is the **cluster sample**. In this method of sampling, the population is divided into many small groupings or clusters.

A random selection of these clusters may then be obtained. In the case of sampling city blocks, each block can be regarded as a cluster of households, and a set of these clusters can be chosen, either randomly or with probability proportional to size.

There are many other methods of sampling using the principles of probability. Some combination of stratified, cluster and random sampling may be used in a **multistage sample**. What is common to all of these methods is that the principles of probability can be used to quantify the uncertainty associated with each sample. This means that inferences concerning sampling error and the nature of population parameters can be obtained from these samples.

The following section returns to random sampling. This is the sampling method assumed throughout the rest of this textbook. The behaviour of the sample mean and the sample proportion obtained from random sampling are discussed in some detail in the rest of this chapter.

7.4 Statistics from Random Samples

The methods of statistical inference used in this textbook are based on the principles of random sampling. Recall that a random sample is a sampling method whereby each member of the population has an equal chance or probability of being selected. Statisticians have conducted detailed investigations of the behaviour of statistics which are obtained on the basis of random sampling. The manner in which a statistic such as the mean is distributed, when many random samples are drawn from a population, is referred to as **the sampling distribution of the statistic**. This section outlines and gives examples of sampling distributions for the sample mean and for the sample proportion. These sampling distributions could be obtained for other types of probability based samples, but these are not provided in this textbook. The discussion which follows assumes the sample is drawn on the basis of random selection.

7.4.1 Central Limit Theorem

A random sample of size n drawn from a population with mean μ for variable X will result in values of the variable $X_1, X_2, X_3, \cdots, X_n$. These different X_is may be quite disparate, taking on values anywhere within the range of

values of the variable X. When the mean

$$\bar{X} = \frac{\sum X_i}{n}$$

of these n values is obtained, the researcher hopes that the $\bar{X}$ will be relatively close to the true mean μ. This sample mean is unlikely to be exactly equal to μ, so that there will be some sampling error $|\bar{X} - \mu|$. If this random sample is close to being a representative sample, the sampling error will be quite small.

Now suppose that the sample is one of those samples which just by chance happens to select a set of values $X_1, X_2, X_3, \cdots, X_n$ which are rather unusual. Even though this sample is random, $\bar{X}$ may differ considerably from μ, and the sampling error associated with this sample may be relatively large, resulting in a quite unrepresentative sample. While the researcher hopes that the latter situation does not occur, there is always some chance that this could happen. What the theorems of statistics can show though, is that if n is reasonably large, the probability of the latter situation occurring is relatively small. The following paragraphs give some of the mathematical properties of statistics such as the mean of a sample. These results are not proven in this textbook, but are explained and illustrated.

In order to consider sampling distributions of statistics, it is necessary to imagine many random samples being taken from a population. It should be noted that the researcher is not usually able to obtain many samples. Ordinarily only one sample is drawn, and the researcher must attempt to provide inferences concerning the population on the basis of this one sample. But to discuss the behaviour of statistics, it is necessary to imagine that there is the possibility of selecting repeated random samples from the same population.

Suppose the researcher is attempting to estimate the true mean μ of the population on the basis of these samples. Some of these samples will yield sample means $\bar{X}$ which are close to μ, while other samples will yield sample means more distant from μ. It can be shown mathematically that if many random samples are taken from a population, the average of the sample means $\bar{X}$ is the true mean of the population μ.

Suppose that several random samples are taken from the same population. From the data obtained in each of these samples a sample mean $\bar{X}$ can be obtained. Each of these individual sample means $\bar{X}$ differs from the true mean μ. But if the mean of the set of all of the different sample means is obtained, this mean of the set of sample means is very close to the true

mean μ of the whole population. If more and more random samples are taken from this population, the mean of the set of sample means gets closer and closer to the true mean of the population. In the limit, if an extremely large number of such random samples are taken from the same population, the mean of the sample means is the true mean of the population.

Not only is the average of the sample means known, but the variability in these sample means can also be determined. The standard deviation of the sample means can be shown to be the standard deviation of the population, divided by the square root of the sample size. Let σ be the standard deviation of the population from which the samples are being drawn. If the samples are randomly selected from this population, each having sample size n, then the standard deviation of the set of sample means is $\sigma/\sqrt{n}$. Again, this property is exactly true only when the number of samples drawn is extremely large. However, as long as the samples are random, the standard deviation of the sampling distribution of $\bar{X}$ can be considered to be $\sigma/\sqrt{n}$.

Standard error. This last property of the mean is often referred to as the **standard error of the mean**, and is given the symbol $\sigma_{\bar{X}}$. That is, if a random sample of size n is selected from a population with standard deviation σ, the standard error of the sample mean $\bar{X}$ is

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

All this means is that the standard deviation of the sample mean, when there is repeated random sampling, is $\sigma/\sqrt{n}$.

The use of 'error' in the term 'standard error' does not mean error in the sense of mistake, or an attempt to mislead. Rather, the standard error is used in the sense of the error associated with random sampling. Historically, the term emerged based on measurements of astronomers. Different astronomers made different estimates of astronomical distances. Since no one measurement was perfect, difference associated with different measurements came to be called measurement errors. Later the term became associated with the error involved in random sampling. The measure of error which mathematicians developed became known as standard error. Remember though that what standard error denotes is the standard deviation of the sampling distribution of the sample mean when there are repeated random samples from a population.

The standard deviation of the sample mean being $\sigma/\sqrt{n}$ means that this standard error has a very desirable property. The standard error of the

sample mean has the square root of the sample size in the denominator. This means that the standard deviation of the sample means is smaller than the standard deviation of the population from which the sample was drawn. This further implies that the sample means are considerably less variable than are the values of the variable themselves. The sample means tend to cluster around the true mean, with most of the sample means being quite close to the true mean.

In addition, the larger the size of the sample, the smaller the standard deviation of the sample means. For example, if $n = 25$, the standard error of the sample mean is $\sigma/\sqrt{25} = \sigma/5$. But if the sample size is enlarged to $n = 100$, this standard error is reduced to $\sigma/\sqrt{100} = \sigma/10$. This implies that the larger the size of the random sample, the less variation in the sample means. Since the mean of the sample means is the true mean of the population, this further implies that a large sample size is associated with sample means which tend to cluster very closely around the true mean.

Not only are the mean and standard deviation of the sample means known, but the type of distribution for $\bar{X}$ can also be determined mathematically. It is possible to prove that if the random samples have a reasonably large sample size, **the sample means are normally distributed**. Recall that the areas under the normal curve can be interpreted as probabilities. Since the mean and standard deviation are both known, and the normal probabilities are known, this implies that it is possible to determine the probability that $\bar{X}$ is any given distance from μ. In Section 7.6, the probabilities associated with various levels of sampling error will be determined. All of these results can be summarized in the following theorem.

The Central Limit Theorem. Let X be a variable with a mean of μ and a standard deviation of σ. If random samples of size n are drawn from the values of X, then the sample mean $\bar{X}$ is a normally distributed variable with mean μ and standard deviation $\sigma/\sqrt{n}$. Symbolically,

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

This result holds for almost all possible distributions of variable X. The only condition for this result to hold is that the sample size n must be reasonably large. Often a sample size as small as $n = 30$ is regarded as being adequate for this theorem to hold.

The Central Limit Theorem is illustrated diagramatically in Figure 7.1. The distribution given at the top of this figure is drawn roughly, in order to

Figure 7.1: Diagrammatic Representation of Central Limit Theorem

illustrate that the population can have practically any type of distribution. In general, the shape of this population distribution will be unknown, and the mean and standard deviation will also be unknown. If random samples of size n are drawn from this population, and n is reasonably large, then the sample means are distributed normally. This is shown in the diagram at the bottom of Figure 7.1. Note that the variable on the horizontal axis of the lower distribution is no longer X but is $\bar{X}$, the sample mean. These sample means vary with each random sample selected, but the middle of this normal distribution for $\bar{X}$ is μ; the distribution of sample means has the same mean as does the population from which the samples were drawn. The normal distribution for the sample means is much more concentrated than the original population distribution. The diagram at the top of Figure 7.1 gives the approximate size of the standard deviation. The normal distribution of sample means in the bottom diagram is much more concentrated, with a considerably smaller standard deviation, and with most cases lying fairly close to the true population mean μ

This theorem is the single most important theorem in statistics. It states that regardless of what a population or a distribution looks like, the sampling distribution of the sample means will be normally distributed, with mean and standard deviation as given above. The only conditions for this to hold are that the samples be random samples with relatively large sample sizes. If the samples are not random, or if the sample size is quite small, then this result may not hold. But in general, the result of the theorem holds, and this means that the nature of the sampling distribution of $\bar{X}$ can be determined. Since the nature of the original distribution of X can be almost any distribution, the normal distribution appears seemingly out of nowhere. Even though nothing is known about the population, the manner in which the sample is distributed can be well understood and described.

In terms of the size of the samples, there is some debate concerning what constitutes a large sample size. If the population from which the sample is originally drawn is close to symmetrically distributed (even though it is not normally distributed), then a random sample of 25 or 30 cases should be sufficient to ensure that the central limit theorem holds. Where the original population is quite asymmetric, skewed either positively or negatively, a considerably larger sample size may be required before this theorem holds. A sample size of over 100 should ensure that the theorem can be used in almost any circumstance. In addition, the larger the sample size, the more closely the theorem holds, so that a random sample of 500 or 1,000 will ensure that the normal probabilities very closely approximate the exact probabilities for

$\bar{X}$. As with any approximation, the theorem can be used even when these minimum conditions do not hold, for example for a sample of only $n = 15$. But in this circumstance, the probabilities calculated on the basis of the normal curve would not provide a very accurate estimate of the correct probabilities associated with values of $\bar{X}$.

Standard Deviation of the Sampling Distribution. One difficulty associated with the use of this theorem in practical situations is that the standard deviation of the population, σ, is not usually known. Since the standard deviation of the sampling distribution of $\bar{X}$ is $\sigma/\sqrt{n}$, this means that this standard deviation of the sampling distribution is not known. It in unlikely that the researcher will ever have an exact idea of σ because a sample is being taken to determine the characteristics of the population. Just as there is sampling error associated with $\bar{X}$, there is also sampling error associated with the sample standard deviation s as an estimate of σ.

In Chapter 8 it will be seen that there are various ways in which the problem of the lack of knowledge of σ can be dealt with. Usually researchers will use the sample standard deviation s as an estimate of σ, even though there is some sampling error associated with this. As long as only a rough estimate of sampling error is needed, then this solution is adequate, at least when the sample size is reasonably large. Thus, in practice,

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

Effect of Sample Size. The standard deviation of the sampling distribution of $\bar{X}$ is

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

so that a larger sample size means a smaller standard deviation for the sampling distribution. This means that the sample means are less variable, from sample to sample, when the sample size is large. This, in turn, implies that there is less sampling error when the sample size is large. This is one of the reasons that researchers prefer random samples with large sample sizes to those with smaller sample sizes.

The effect of sample size on the sampling distribution of $\bar{X}$ is shown in Figure 7.2, where three different sample sizes are given. The possible values of $\bar{X}$ are given along the horizontal, with $\bar{X}$ having a mean of μ. The vertical axis represents the probability of each value of $\bar{X}$.

Figure 7.2: Sampling Distribution of $\bar{X}$ for 3 Sample Sizes

Suppose that three random samples have been selected from a population with mean μ and standard deviation σ. For sample C, with sample size $n = 50$, the sampling distribution is the least concentrated normal distribution, with $\sigma/\sqrt{50}$ as the standard deviation. When the size of the random sample is increased to $n = 200$ in sample B, the standard deviation is considerably reduced, being $\sigma/\sqrt{200}$ in this case. This is the middle of the three distributions. Finally, in sample A, when the sample size is increased again, to $n = 800$, the sampling distribution becomes considerably more concentrated, with

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{800}}.$$

By comparing these three normal distributions, it can be seen that when n is larger, the chance that $\bar{X}$ is far from μ is considerably reduced. A

larger random sample is associated with smaller sampling error, or at least a greater probability that the sampling error is small. In Example 7.6.1 this is illustrated with two different random samples.

7.4.2 Sampling Distribution of a Proportion.

To this point, the discussion in this section has been solely concerned with the behaviour of the mean in random sampling. Another population parameter with which researchers are commonly concerned is the population proportion. Suppose a particular characteristic of a population is being investigated, and the proportion of members of the population with this characteristic is to be determined. Ordinarily the proportion of cases in the sample which take on this characteristic will be used to make statements concerning the population proportion. If the sample is a random sample, and is reasonably large, then the sampling distribution of the sample proportion can be determined. This is based on the binomial distribution, and is a simple extension of the normal approximation to the binomial probability distribution.

Let p be the proportion of members of the population having the characteristic being investigated. This can be stated in terms of the binomial by defining this characteristic as a success. Any member of the population which does not have this characteristic can be considered as a failure.

Suppose random samples of size n taken from this population. Let X be the number of cases selected which have the characteristic of success. Each random sample will have a different number of successes X. The variable X is a random variable, and in Section 6.5 it was shown that the mean of X is np and the standard deviation of X is $\sqrt{npq}$. In addition, if n is reasonably large, then this distribution is normal.

That is, if random samples of size n are selected from a population with p as the proportion of successes, the number of cases X which have the desired characteristic is distributed

$$\text{Nor}(np, \sqrt{npq}).$$

The researcher is not likely interested in the number of successes in n trials, so much as he or she is likely to be concerned with the proportion of successes in n trials. This proportion is a statistic and is given the symbol $\hat{p}$. Since there are X successes in n trials, the proportion of successes in the sample is X/n, so that $\hat{p} = X/n$. Since X is distributed

$$\text{Nor}(np, \sqrt{npq}),$$

it can be seen that $\hat{p}$ will be distributed as is X but divided by n, so that $\hat{p}$ is

$$\text{Nor}(p, \sqrt{pq/n}).$$

This result holds as long as n is reasonably large, and the sample is a random sample. The rule for the size of n is the same as in the normal approximation to the binomial, that is,

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

The standard deviation of the sample proportion, $\sqrt{pq/n}$ can be called the **standard error of the proportion** and is sometimes given the symbol $\sigma_{\hat{p}}$. This symbol is used with σ to denote that it is a standard deviation, and the subscript to denote the statistic for which it is the standard deviation. Thus

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

This result is comparable to the Central Limit Theorem for the mean. The only difference is that the above result is based on the binomial, and the normal approximation to the binomial. What this result shows is that large random samples from a population yield a well known and well understood sampling distribution for the sample proportion. This distribution can be used to provide estimates of the population proportion, test hypotheses concerning the population proportion, or estimate probabilities associated with different levels of sampling error of the population proportion. An example of the latter is given in the following section, with interval estimates and hypothesis tests for proportions in Chapters 8 and 9.

7.5 Example of a Sampling Distribution

This example illustrates the Central Limit Theorem by drawing a large number of random samples from a population which is not normally distributed. Suppose that μ is the mean of a population, σ is the standard deviation of this same population, and random samples of size n are drawn from this population. According to the Central Limit Theorem, the sample mean $\bar{X}$ is normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. This result is true regardless of how the population itself is distributed. The only condition required for this Theorem to be true is that the samples be random samples and that n be reasonably large. In this example, random

samples of size $n = 50$ will be used to illustrate the sampling distribution of the sample mean.

Population of Regina Respondents This example begins with the distribution of gross monthly pay of 601 Regina respondents. The set of 601 pay levels of respondents, along with an identification number for each respondent is given in Appendix **??**. This data was obtained from respondents in the Social Studies 203 Regina Labour Force Survey. These 601 values are grouped into a frequency distribution, along with some summary measures for this distribution, in Table 7.4. Although the data is based on a sample, for this example this distribution of gross monthly pay will be regarded as a distribution for a population. If this sample were to be exactly representative of the population of Regina labour force members, then this distribution could be regarded as describing the distribution of gross monthly pay of all Regina labour force members. Thus it will be assumed that the mean pay level of \$2,352 is a population parameter μ, and the standard deviation of \$1,485 is also a parameter σ.

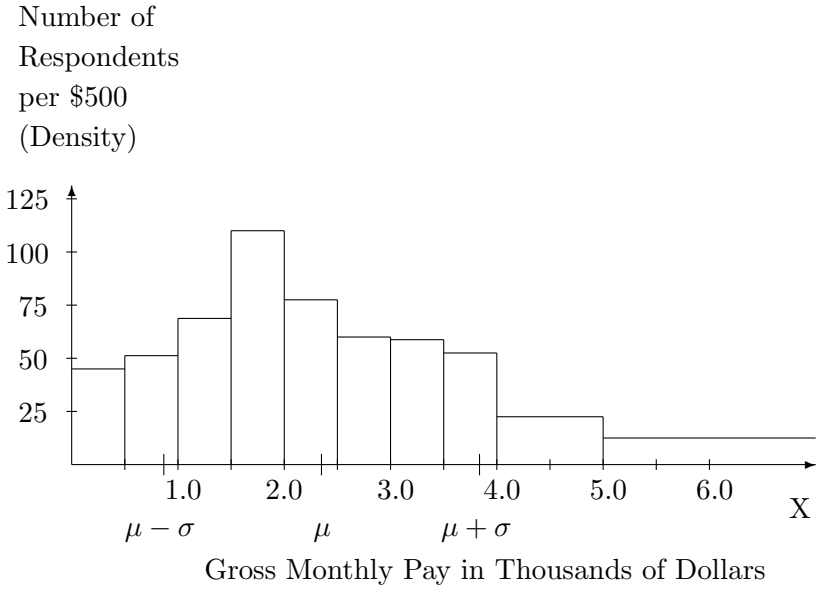

Figure 7.3: Histogram of Distribution of Gross Monthly Pay

The histogram for the frequency distribution of gross monthly pay in Table 7.4 is given in Figure 7.3. A quick examination of Table 7.4 and

Figure 7.3 shows that the distribution of gross monthly pay is not normal. Rather, the distribution peaks at a fairly low pay level, around $1,500-2,000 per month, and then tails off at higher income levels. There are some individuals with quite high pay levels, so that the distribution goes much further to the right of the peak than it does to the left of the peak income level. At these upper income levels though, there are relatively few people. A distribution of this sort is said to be **skewed to the right**; distributions of income and wealth are ordinarily skewed in this manner.

Gross Monthly Pay ($ per month)	f
Less than 500	45
500-999	51
1,000-1,499	69
1,500-1,999	110
2,000-2,499	77
2,500-2,999	60
3,000-3,499	59
3,500-4,000	52
4,000-4,999	46
5,000 and over	32
Total	601

Parameters for Distribution	
Mean	μ =$ 2,352
Median	$ 2,000
Standard Deviation	σ =$1,485
Minimum	$ 50
Maximum	$9,000
Sum	$1,413,316

Table 7.4: Distribution of Gross Monthly Pay, 601 Regina Respondents

For this example, consider these 601 people, with the pay distribution of Table 7.4, as a population from which random samples will be drawn. For this population, the variable X is gross monthly pay, measured in dollars. The mean and standard deviation of gross monthly pay in dollars for this population are

$$\mu = 2,352$$

$$\sigma = 1,485$$

and this population appears to have a distribution which is not normal.

Table 7.5: Mean GMP for 192 Different Random Samples, Each of Size $n = 50$

192 Random Samples From this population of 601 Regina respondents, 192 different random samples have been drawn. Each of these samples is a random sample of the population, and was drawn using the `SAMPLE` command of SPSSX. This command was used 192 different times, each time resulting in a selection of 50 of the Regina respondents. According to the Central Limit Theorem, the distribution of the sample mean, $\bar{X}$, should be normal with mean μ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. Since $\mu = 2,352$ and $\sigma = 1,485$, the sampling distribution of the sample mean should have a standard deviation of

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

rounded to the nearest dollar.

Not all the data for each of the samples selected is given here. Table 7.5 gives the values of the 192 different sample means that were drawn. Each number in Table 7.5 is a sample mean, an average of the 50 different respondents which were chosen in that sample. For example, the first random sample of 50 respondents had a mean pay of \$2,205, the second random sample of 50 respondents had a mean pay of \$2,641, and so on. The last of the 192 random samples drawn had a mean pay of \$2,275 for the 50 respondents in the sample.

Each of the means in Table 7.5 can be regarded as an estimate of the true mean of \$2,352. As can be seen in Table 7.5, most of the means are fairly close to this true mean. The first sample mean of \$2,205 differs from the the true mean by $2,205 - 2,352 = -147$ dollars. The second sample mean of \$2,641 is \$289 greater than the true mean, so that the sampling error for this sample is \$289. In Table 7.6, it can be seen that the minimum mean of Table 7.5 is \$1,777. On the low side, this is the worst estimate of the true mean, being \$575 less than the true mean. On the high side, the sample that was the worst produced a sample mean of \$2,988, \$636 above the true mean. In general though, the sample means provide reasonably close estimates of the true mean, some of the estimates being quite precise, others further away. What this shows is that a random sample of size 50 generally provides a reasonably good estimate of the true mean of a population.

The extent of variability of the sample means can be examined by looking over the list of all sample means in Table 7.5. Each sample mean provides a different estimate of the true mean, and these sample means vary considerably. Table 7.6 summarizes the set of 192 sample means of Table 7.5 as a sampling distribution. The standard deviation of these 192 sample means

Gross Monthly Pay ($ per month)	f
Under 1,780	1
1,780-1,884	2
1,884-1,988	6
1,988-2,092	11
2,092-2,196	23
2,196-2,300	44
2,300-2,404	33
2,404-2,508	35
2,508-2,612	18
2,612-2,716	13
2,716 and over	6
Total	192

Statistics for Sampling Distribution	
Mean	$2,337
Median	$2,329
Standard Deviation	$206
Minimum	$1,777
Maximum	$2,988
No. of Cases	192

Table 7.6: Distribution of 192 Different Sample Means

is calculated and given in Table 7.6 as $206. This is a little lower than if central limit theorem were exactly true. It was noted above that if the theorem were exactly true, the standard deviation of these sample means would be approximately $210. Again, the fact that the sample size is only $n = 50$ may be the reason that the actual standard deviation is a little different from what would be expected from the theorem.

Finally, the distribution of the sample means is given in Figure 7.5. While this distribution is not the exact normal distribution would be expected from the central limit theorem, this distribution comes close to being symmetrical. In addition, note how concentrated this distribution of sample means is, compared with the original population distribution of Figure 7.3, given again as Figure 7.4. Each of the sample means in Figure 7.5 is quite close to the true mean μ. The chance that any sample mean lies very distant from μ is quite small.

Conclusion. This example shows that when a random sample is drawn from a population, the sample mean generally provides quite a good estimate

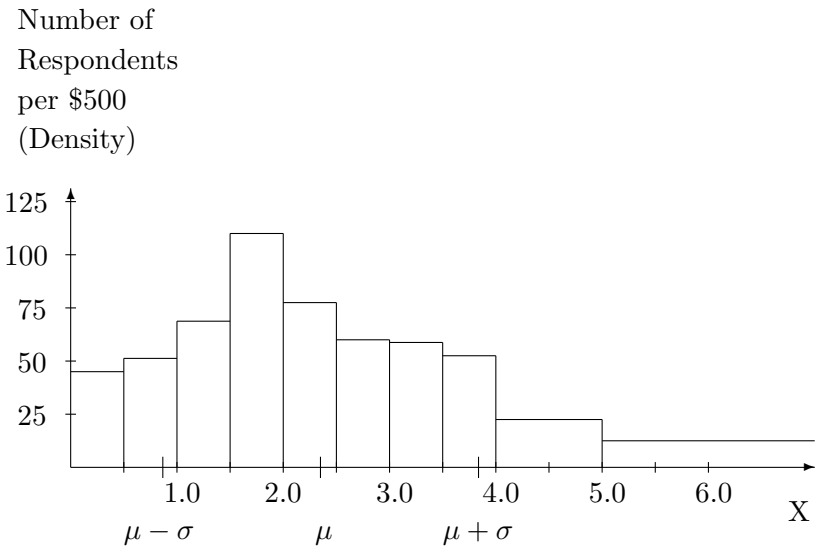

Figure 7.4: Histogram of Distribution of Gross Monthly Pay

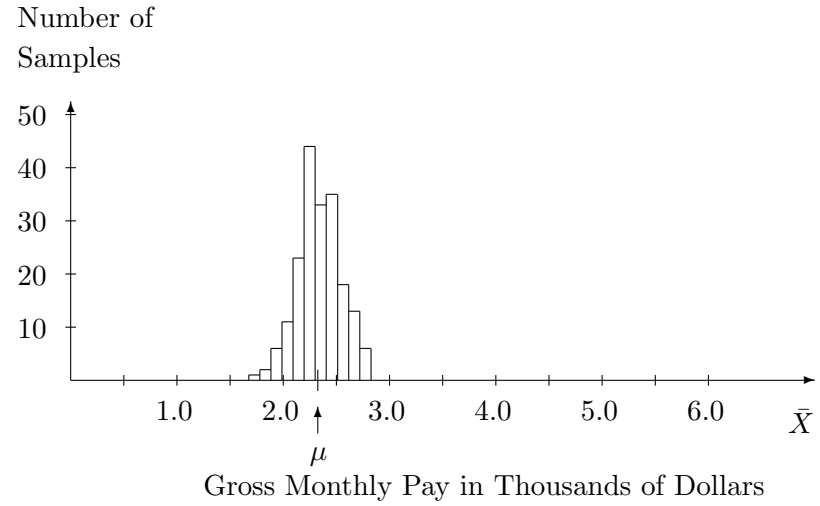

Figure 7.5: Histogram of Sampling Distribution of 192 Sample Means

of the true population mean. Sometimes the sample mean is a little on the low side, sometimes a little on the high side, but the sample mean is generally quite close to the true mean. The standard deviation of the sample means is $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. Since the standard deviation of the original population, σ, is divided by $\sqrt{n}$ in this expression, this implies that the standard deviation of the sample means is much smaller than the standard deviation of the population from which the samples are drawn. In addition, because $\sqrt{n}$ is in the denominator of $\sigma_{\bar{X}}$, the larger the sample size, the smaller the standard deviation of the sample means. Together these mean that if a random sample is taken from a population, a larger sample size is more likely to provide a close estimate of the true population mean than is a smaller sample size.

Finally, the Central Limit Theorem also states that the distribution of sample means is likely to be close to normally distributed. Since the probabilities associated with the normal distribution are well known, the probability of the sample mean differing from the population mean by any given amount can be calculated. This will be seen in the following section, where sampling error is briefly discussed. These properties of the sampling distribution also provide the basis for interval estimates and hypothesis tests concerning the population mean.

7.6 Sampling Error

Each time a sample is taken from a population, there will be some sampling error. The sampling error associated with a sample is the numerical difference between the sample statistic and the corresponding population parameter. Since the value of the latter is unknown, the exact size of the sampling error is not known. If the sample is being used to make inferences concerning the population mean, then the central limit theorem lays the basis for determining the probability of various levels of sampling error. Similarly, when inferences concerning a population proportion are being made, the extension of the normal approximation to the binomial provides the basis for estimating probabilities associated with different sizes of sampling error for a population proportion. In this section, the probability associated with the sampling error for each of the sample mean and the sample proportion will be discussed.

If a reasonably large random sample is taken from a population, the Central Limit Theorem states that the sample mean is normally distributed

with mean μ and standard deviation $\sigma/\sqrt{n}$, where μ is the population mean, σ is the standard deviation of the population, and n is the sample size. The probabilities associated with different sample means $\bar{X}$ can be determined on the basis of this normal distribution.

Suppose that a particular level of sampling error, E, is specified. This will be a value of $|\bar{X} - \mu|$, and can be represented by a distance around the centre of the normal distribution. Whatever value of E is specified, this can be represented by the distance from $\mu - E$ to $\mu + E$. That is, if the sampling error is not to exceed E, then $\bar{X}$ must be between $\mu - E$ to $\mu + E$. If $\bar{X}$ is outside this interval, then the smpling error exceeds E. The area under the normal curve between $\mu - E$ and $\mu + E$ can be determined by using the table of the normal distribution in Appendix **??** If this area is interpreted as a probability, then this is the probability that the sample mean $\bar{X}$ is within these limits. This area further represents the probability that the sampling error does not exceed E. All of this can be written as a probability statement as follows:

$$P(\mu - E < \bar{X} < \mu + E) = P_E$$

where P_E is the area under the normal curve between $\mu - E$ and $\mu + E$. That is, the probability that $\bar{X}$ is between $\mu - E$ and $\mu + E$ is P_E. This also means that

$$P(|\bar{X} - \mu| \leq E) = P_E$$

and this states that the probability is P_E that the sampling error does not exceed E. This is ilustrated diagramatically with an example from the distribution of gross monthly pay of Regina labour force members.

Example 7.6.1 Sampling Error of Mean Gross Monthly Pay

In Section 7.5 the mean gross monthly pay of Regina labour force members was given as $\mu = 2,352$, with a standard deviation of $\sigma = 1,485$, where both measures are in dollars. Even though this data was originally obtained on the basis of a sample, it was assumed that this mean and standard deviation are parameters representing the true gross montly pay of all Regina labour force members. In this example, the probability that the sampling error does not exceed $100 will be determined, first for a sample size of 50, and then for a sample size of 200. This is the probability that in a random sample from this population $\bar{X}$ differs from μ by less than $100.

If a random sample of size $n = 50$ is taken from this population, then the central limit theorem states that the sample means are normally distributed with mean $\mu = 2,352$ and standard deviaton

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1485}{\sqrt{50}} = \frac{1485}{7.07107} = 210.011$$

Rounded to the nearest dollar, the standard error of the sample mean is $\sigma_{\bar{X}} = \$210$. That is,

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

Since the sampling error is not to exceed $E = \$100$ on either side of μ, if the area under the normal curve within plus or minus \$100 of the mean can be determined, then this is the appropriate probability. Since the standard deviation of this normal distribution is \$210, a distance of \$100 on either side of the mean is $100/210 = 0.48$ standard deviations or Z values.

The required area is the area under the normal distribution between $Z = -0.48$ and $Z = +0.48$. This area is $0.1844 + 0.1844 = 0.3688$. Thus the probability that the sampling error is no more than \$100 is approximately 0.37 when the sample size is $n = 50$. Note that neither the value of the mean of the original distribution nor of the sample, was not required in obtaining this estimate. However, the value of the standard deviation of the population was required in order to determine the Z value associated with the sampling error.

*This is illustrated in the top diagram of Figure 7.6. There the sampling distribution of $\bar{X}$ is the normal curve shown, with μ as the mean of the sampling distribution. A distance of \$100 on each side of μ is shown. The shaded area represents the probability that $\bar{X}$ falls between $\bar{X} - \$100$ and $\bar{X} + \$100$, that is that the sampling error $|\bar{X} - \mu| < 100$. Since the standard deviation $\sigma_{\bar{X}}$ for this distribution is \$210, a distance of \$100 is $100/210 = 0.48$ in terms of Z. From the normal table in Appendix **??**, the area under the normal curve between $Z = -0.48$ and $Z = +0.48$ is 0.3688.*

Standardized Value. Recall that the formula for standardizing a normal variable X with mean μ and standard deviation σ was

$$Z = \frac{X - \mu}{\sigma}.$$

This transformation converts X into a variable Z which has mean 0 and standard deviation 1. Such a transformation is said to **standardize** a variable. This formula can be generalized as follows. Any variable can be

Figure 7.6: Sampling Error for Samples of Size 50 and 200

standardized by subtracting the mean from the variable and dividing this difference by the standard deviation. This produces a new variable with a mean of 0 and a standard deviation of 1. That is, in general

$$Z = \frac{\text{variable} - \text{mean of variable}}{\text{standard deviation of variable}}$$

and this Z has mean 0 and standard deviation 1. In this case, the variable is $\bar{X}$, with mean μ and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. The appropriate formula for standardization in this distribution is

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

and for this Z,

$$Z \text{ is Nor } (0 , 1)$$

Returning to the previous example, but using this new formula for Z, the sampling error is $E = |\bar{X} - \mu|$. The magnitude of this sampling error is the same as the numerator of the expression for determining Z in the new formula. That is, working only with the sampling error to the right of the mean,

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{E}{\sigma_{\bar{X}}} = \frac{E}{\frac{\sigma}{\sqrt{n}}} = \frac{100}{\frac{1485}{\sqrt{50}}} = 0.47$$

and this is essentially the same Z as earlier, with only a slight difference due to rounding error.

In the second part of this example, the sample size is increased to $n = 200$, and all that changes is n. The sampling error is still $E = \$100$, but now the standard deviation of the distribution of the sample mean is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1485}{\sqrt{200}} = \frac{1485}{14.142} = 105.01$$

or \$105. Since one standard deviation is \$105, a sampling error of $E = \$100$ is associated with $Z = 100/105 = 0.95$. The area between $Z = -0.95$ and $Z = +0.95$ is $0.3289 + 0.3289 = 0.6578$. The probability is approximately 0.66 that the sampling error does not exceed \$200 when a random sample of size $n = 200$ is selected from the population. Using the formula for standardization of $\bar{X}$,

$$Z = \frac{E}{\sigma_{\bar{X}}} = \frac{E}{\frac{\sigma}{\sqrt{n}}} = \frac{100}{\frac{1485}{\sqrt{200}}} = 0.95$$

Sample Size (n)	Probability that Sampling Error Does not Exceed \$100
50	0.37
200	0.66
500	0.87
4000	0.00002

Table 7.7: Probability of \$100 Sampling Error with Various Sample Sizes

Note how the probability that the sampling error does not exceed \$100 is increased with the increased sample size. Increasing the size of the random samples from 50 to 200 increases the chance that the sampling error does not exceed \$100 from 0.37 to 0.66. In general, a larger random sample results in an increased probability that the sampling error is no greater than specified. This means that the researcher has more confidence in the results from a larger random sample. It can be shown that for any given probability, the sampling error is smaller, the larger the size of the random sample. The sample mean from a large random sample is likely to be closer to the true population mean compared with the sample mean from a smaller random sample.

Table 7.7 summarizes these results. The determination of the probabilities of the sampling error being no greater than \$100 when the sample sizes are 500 and 4000 are left as an exercise for the student.

Sampling Error for a Proportion. If a random sample of a reasonably large sample size is taken from a population with proportion of successes p, the sampling error of the sample proportion $\hat{p}$ can be determined in much the same manner as in the case of the mean. Based on the extension of the normal approximation to the binomial, 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

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

This result holds as long as

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

Suppose the researcher wishes to set the sampling error at E. Then the probability that $\hat{p}$ is between $p - E$ and $p + E$ can be determined, since the distribution of $\hat{p}$ is known to be normal and the standard deviation can also be determined. As a probability statement, this can be written

$$P(p - E < \hat{p} < p + E) = P_E$$

where P_E is the area under the normal curve between $p - E$ and $p + E$. That is, the probability that $\hat{p}$ is between $p - E$ and $p + E$ is P_E. This also means that

$$P(|\hat{p} - p| \leq E) = P_E$$

The probability that the sampling error for $\hat{p}$ is less than or equal to E is P_E. The sampling error for a sample proportion is illustrated in the following example.

Example 7.6.2 Sampling Error in the Gallup Poll

Each month Gallup Canada, Inc., polls Canadian adults on many issues. The results of these polls are regularly reported in the media. For example, 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. Each of the reported percentages is subject to sampling error. In **The Gallup Report** *of August 13, 1992, Gallup makes the following statement concerning sampling error.*

> Today's results are based on 1,025 telephone interviews with adults, conducted August 6-10, 1992. A national telephone sample of this size is accurate within a 3.1 percentage point margin

> of error, 19 in 20 times. The margins of error are higher for the regions, reflecting the smaller sample sizes. For example, in Quebec 257 interviews were conducted with a margin of error of 6 percentage points, 19 in 20 times.

While Gallup uses slightly different language than has been used in this book, these results can be shown using the concept of sampling error developed here.

First, consider the sampling error associated with the national estimates. Take support for the Conservatives as an example. 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, the distribution of the sample proportion $\hat{p}$ can be determined on the basis that:

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

In this sample, $n = 1,025$, so that if some estimate of p and q are available, then $\sigma_{\hat{p}} = \sqrt{pq/n}$ can be determined. It will be seen a little later that if $p = q = 0.5$ is used as the estimate of p and q in $\sqrt{pq/n}$, then this provides a maximum estimate of the size of the standard deviation of $\hat{p}$ for any given n. Using these values,

$$\sigma_{\hat{p}} = \sqrt{pq/n} = \sqrt{(0.5 \times 0.5)/1,025} = \sqrt{0.0002439} = 0.0156.$$

That is, one standard deviation in the sampling distribution of $\hat{p}$ is 0.0156.

In terms of the probability associated with the sampling error, Gallup says the sampling error is no more than 3.1 percentage points in 19 out of 20 samples. 19 in 20 samples is equivalent to a probability of $19/20 = 0.95$. Gallup is making the claim that the sampling error is 3.1 percentage points, or a proportion $E = 0.031$ with probability 0.95.

Recall that when the normal distribution was being introduced, a Z of 1.96 was associated with the middle 0.95 or 95% of a normal distribution. That is, in order to take account of the middle 0.95 of a normal distribution, it is necessary to go to 1.96 standard deviations below the mean and to 1.96 standard deviations above the mean. This means that a probability of 0.95 is associated with a distance of 1.96 standard deviations on each side of the mean.

Putting together this probability and this Z value, the size of the sampling error can be seen. One standard deviation is $Z = 1.96$ and the size of the standard deviation for the distribution of sample means is 0.0156. In this distribution, 1.96 standard deviations is $1.96 \times 0.0156 = 0.0306$. This is a proportion and converted into percentages is $0.0306 \times 100\% = 3.06\%$ of 3.1 percentage points. This is the margin of error claimed by Gallup. That is, the probability is 0.95 that $\hat{p}$ lies within 3.1 percentage points of the true proportion p. While the error associated with this sample may be larger than 3.1 percentage points in any particular month, 19 in 20 such samples will yield values of $\hat{p}$ which are within 3.1 percentage points of the true proportion of PC supporters p.

Figure 7.7 illustrates this sampling error diagrammatically. Along the horizontal axis is $\hat{p}$, and the normal curve gives the sampling distribution of $\hat{p}$. This is a normal distribution, centred at the true proportion of successes p, and with standard deviation $\sigma_{\hat{p}} = 0.0156$. Since the sampling error of $E = 0.031$ is not exceeded in 0.95, or 19 out of 20, samples, the appropriate area under the normal distribution is the middle 0.95 of the area. From Appendix **??** *it can be seen that this is the area under the normal curve between $Z = -1.96$ and $Z = +1.96$. That is, going out from the centre of the distribution by 1.96 standard deviations in each direction gives this middle 0.95 of the distribution. Since one standard deviation in this distribution is 0.0156, a distance of 1.96 standard deviations on each side of centre is associated with a distance of $1.96 \times 0.0156 = 0.031$ on each side of centre. This is given as the distance from $p - 0.031$ to $p + 0.031$. This interval around p contains 95% of the area under this normal curve, and it can be seen that 19 in 20 samples will yield $\hat{p}$s which are within this range. Thus 19 in 20 samples have a sampling error of 3.1 percentage points or less.*

As an exercise, show that the sampling error for Quebec is 6 percentage points, 19 in 20 times, as claimed by Gallup. All that changes for Quebec is that the sample size n is reduced to only $n = 257$. Otherwise the method used is the same.

Additional Notes Concerning Sampling Error of $\hat{p}$.

1. Z **Value for** $\hat{p}$. Just as a Z value was determined for $\bar{X}$ in the sampling distribution of the sample mean, so a comparable value can be obtained for the sampling distribution of the sample proportion. Recall that the general

Figure 7.7: Sampling Error for Gallup Poll

format for determining the standardized value Z is

$$Z = \frac{\text{variable} - \text{mean of variable}}{\text{standard deviation of variable}}$$

In the case of the distribution of the sample proportion $\hat{p}$,

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

so that the mean of $\hat{p}$ is p and the standard deviation of $\hat{p}$ is

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

Putting these values into the general formula for Z gives

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

so that for this Z,

$$Z \text{ is Nor}(0, 1)$$

In the case of the sampling error associated with the Gallup poll, the claim is that $E = 0.031$. When making inferences concerning the true population proportion p, the statistic is the sample proportion $\hat{p}$, and the sampling error associated with any given sample is $E = |\hat{p} - p|$. Thus

$$Z = \frac{\hat{p} - p}{\sqrt{pq/n}} = \frac{0.031}{\sqrt{(0.5 \times 0.5)/1,025}} = \frac{0.031}{0.0156} = 1.96$$

so that the sampling error of $E = 0.031$ is associated with $Z = 1.96$. The area under the normal curve within 1.96 standard deviations of the mean is the area between $Z = -1.96$ and $Z = +1.96$. This distance of 1.96 standard deviations in each direction from the mean is associated with an area of 0.95 in the middle of the distribution. The probability of 0.95 is the probability of a sampling error of no more than proportion 0.031, or 3.1 percentage points, in a random sample of size $n = 1,025$.

2. **Maximum Value of $\sigma_{\hat{p}}$.** Earlier it was stated that a good estimate of $\sigma_{\hat{p}}$ could be obtained if $p = q = 0.5$ when substituting p and q into

$$\sigma_{\hat{p}} = \sqrt{pq/n}.$$

It was also claimed that this produces the **maximum** possible value of $\sigma_{\hat{p}}$ for any given n. That is

$$\text{Maximum}(\sigma_{\hat{p}}) = \sqrt{(0.5 \times 0.5)/n} = \sqrt{0.25/n}$$

If you have studied calculus, you will recognize this as a simple problem of maximization. That is, $q = 1 - p$ so that $pq = p(1 - p)$ and the maximum value of this occurs when $p = 1/2$. Taking the derivation of this product with respect to p produces the maximum value of the product. That is,

$$\frac{d(p - p^2)}{dp} = 1 - 2p = 0$$

when $p = 1/2 = 0.5$.

For those unfamiliar with calculus, you can satisfy yourself that $p \times q$ cannot exceed 0.25 if $p + q = 1$. Take a few values such as

$$0.5 \times 0.5 = 0.25$$

$$0.3 \times 0.7 = 0.21$$

$$0.12 \times 0.88 = 0.1056$$

$$0.05 \times 0.95 = 0.0475$$

and it can be seen that p times q cannot be greater than 0.25 when each of p and q is less than 1, and the sum of p and q is 1.

What this means is that the estimate of the standard deviation of $\hat{p}$ is at a maximum when p and q are close to 0.5. With the above set of products, it can also be seen that it is only when p and q are quite far from 0.5 that the product $p \times q$ is much less than 0.25. Only when this latter condition holds is the standard deviation of $\hat{p}$ reduced all that much. Thus the standard deviation of $\hat{p}$ will never be underestimated if $p = q = 0.5$ in

$$\sigma_{\hat{p}} = \sqrt{pq/n}.$$

If anything, using $p = q = 0.5$ will overestimate the standard deviation, and thus overestimate the sampling error. As a result, it is always quite safe to use $p = q = 0.5$ when estimating the sampling error of the sample proportion. You cannot be accused of making the sampling error appear smaller than it really is by doing this. In fact, if $\hat{p}$ is considerably less or considerably more than 0.5, then the sampling error may be somewhat less than determined on the basis of $p = q = 0.5$.

3. **Estimating the Standard Deviation.** In the formulae for the standard deviation of the statistics from samples, the standard deviation of the population from which the sample is drawn always appears in some manner. In the case of the mean, the standard deviation of the sampling distribution of $\bar{X}$ was

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

where n is the sample size, and σ is the standard deviation of the population from which the random sample is drawn. It is very unlikely that the researcher knows σ, since the true population mean μ is unkown, and it is usually necessary to know μ in order to determine σ. One way in which some idea of σ can be obtained is to use the sample standard deviation s. If the random sample has a reasonably large sample size n then,

$$s \approx \sigma$$

for purposes of estimating the standard deviation of the sample mean. If the sample has not be conducted yet, then s may not even be known. Recall that

$$s \approx \frac{\text{Range}}{4}$$

can provide a very rough estimate of the approximate order of magnitude of σ. Another method that can be used is to use the studies of other researchers on other populations, and if the population and variable is reasonably similar to that being investigated, then the standard deviation of another population can be used to provide a rough estimate of σ so that the approximate size of the sampling error of $\bar{X}$ can be determined.

When working with the estimate of the proportion, p, the same difficulty emerges. The standard deviation of the sample proportion $\hat{p}$ is

$$\sigma_{\hat{p}} = \sqrt{\frac{p \times q}{n}}$$

In this case, neither p nor q are known, so this standard deviation cannot be exactly determined. However, point 3 above provides a solution to this. SInce the maximum value of $p \times q$ occurs when $p = q = 0.5$, these values of p and q can be used in the determination of the standard deviation of $\hat{p}$. As noted earlier, these values will never underestimate $\sigma_{\hat{p}}$, if anything, they may overestimate it. This means that these values can always be used, and the researcher can usually obtain a fairly accurate estimate of the standard

deviation of $\hat{p}$. Thus the sampling error associated with a random sample aimed at determining a population proportion can be reasonably closely approximated even before the sample has been conducted.

4. **Effective Sample Size.** While the sample size of the Gallup poll was $n = 1,025$, the effective sample size may be somewhat less. Often a large random sample is selected, but there are many people who refuse to respond when surveyed. Others may respond to some, but not all of the questions. Finally, some may not know the answer to a question, some may be undecided, and others are listed as not responding to a particular question. If any of these situations occur, the **effective sample size** should be considered to be the number of cases for which there is actually some data. In the case of the Gallup survey of Canadian adults concerning the political preferences, the effective sample size is actually considerably less than 1,025. According to **The Gallup Report** of August 13, 1992, 33% of those Canadian adults polled were undecided concerning which political party they would support. This means that 33% of 1,025 or

$$0.33 \times 1,025 = 338.25$$

did not give their political preference. The effective sample size for the August Gallup poll can be seen to be 338 or 339 less than 1,025, that is 686 or 687. This is really the n that should be used in order to determine the sampling error. If you substitute one of these values for n in the formula for sampling error, it can be seen that for probability 0.95, the sampling error is approximately 0.037 or 3.7 percentage points. Thus the sample proportion may have more sampling error associated with it than what Gallup indicates.

Further, the undecided in any poll can produce considerable **nonsampling error**. Unless these undecided respondents are further studied in order to see which party they previously supported, or to which party they are leaning, not much can be said concerning which party would win an election. The difficulty is that most of the undecided may swing to a particular party, rather than being distributed across all political parties. In 1988, 43% of the voting Canadian electorate voted for the PC party. The Gallup poll of August 13, 1992 reports that only 21% would vote PC today. The disparity between these two percentages indicates that a considerable number of today's undecided voters voted PC in the 1988 federal election. Whether these previous PC supporters, many of whom are likely to support either the Reform Party or the Bloc Quebecois now, would swing back to the PCs if an election were to be held in 1992 or 1993, is not clear. Based on these figures

though, there is some evidence that the undecided are disproportionately former PC supporters.

7.7 Conclusion

This chapter has discussed the use of probability in sampling, and has shown how random samples from a population can be used to obtain some idea of the nature of a population. In particular, the sampling distribution of the sample mean and the sample proportion were discussed. There are many more sampling distributions in statistics, and some of these will be introduced in later chapters. In Chapter 8, the t distribution, associated with small random samples, of sample size under $n = 30$, will be used. Inferences concerning whole distributions of a population can be analyzed using the chi square distribution. This will be discussed in Chapter 10.

The sampling distributions of both $\bar{X}$ and $\hat{p}$ are normal as long as the samples are random and the sample size is reasonably large. Since the nature of the normal distribution is well understood, these sampling distributions can be understood as well. These sampling distributions form the basis for the inferential statistics of the following chapters. In Chapter 8, the methods of providing estimates of the population mean and the population proportion will be discussed, based on the results of this chapter. Hypothesis testing using these same results is examined in Chapters 9 and 10.