Contents

Chapter 9

Hypothesis Testing

9.1 Introduction

Hypothesis testing is a formal method of using data from a sample to test a specific hypothesis concerning a population. In this chapter, various statements or hypotheses are made concerning the mean or proportion of a population. Data is obtained from a random sample of the population, and a statistic is computed from this data. Based on this statistic, the statement concerning the mean or proportion will be rejected if the sample statistic is quite different from the population parameter. If the sample statistic is close to the value hypothesized in the statement, it is likely that this statement will not be rejected, but will be adopted as a working hypothesis.

As an example of this procedure, suppose that an economic downturn has occurred in the economy, and this appears to be associated with an increase in poverty. A researcher is studying a large city which has been affected by this recession in the economy. Community groups and trade unionists are fairly sure that the recession has caused an increase in poverty in the city. The researcher wishes to determine whether the community groups and trade unionists are correct or not, and determine the extent to which poverty has increased in this city. Based on the results of this research, the researcher intends to make some suggestions concerning how poverty might be reduced or alleviated.

The above statements concerning poverty and an increase in poverty are fairly general statements. One of the tasks of the researcher is to determine the meaning of poverty, measure the extent of poverty, and determine whether or not there has been an increase in poverty. Some of these issues

involve definitional problems of the type introduced in Chapter 2. Once poverty has been fairly clearly defined, the researcher must find a way of testing whether the statements that poverty has increased are correct or not. This is where sampling and hypothesis testing can be useful. If the researcher is able to conduct a random sample of the population in the city, and is able to ask appropriate questions concerning family and individual incomes, resources and needs, then some answers concerning whether or not poverty has increased can be provided by the researcher. Two possible examples of hypothesis tests of the type used in this chapter might be as follows.

1. The researcher might attempt to determine the mean income of residents of the city, and see how much this has changed. In order to make a claim that there has been a decline, there needs to be some benchmark against which a change can be tested. Since the Census gathers information on income every 5 or 10 years, this could be the benchmark figure used by the researcher. The hypothesis that the researcher might test is that the mean income of residents of the city has not declined since the last Census was conducted. Data concerning incomes of city residents is obtained from a random sample, and these incomes would be corrected for price changes over the period since the last Census. If the mean income of the people surveyed in the sample is a lot less than the mean income of all city residents based on the last Census, then the researcher can reject the hypothesis that there has been no decline in the mean income of city residents. In contrast, if the random sample shows a very small decline in mean income since the Census, the researcher may not have sufficient evidence to conclude that incomes have declined for all city residents.

   This hypothesis test is referred to as a test of a population mean. Since this hypothesis tests whether the mean income has declined or not, the researcher might be criticized on the grounds that this is not a direct measure of poverty. For example, if the researcher concludes that the mean income has not declined in the city, that may be because the well off have experienced an increase in income at the same time as the less well off have had declining incomes.

2. A more direct test of poverty could be provided by testing whether or not the proportion of the city population that is below the poverty line has increased. Again, the Census would provide a benchmark

measure of the proportion of the population which was in poverty at an earlier date. Suppose that the Census had shown that 0.15, or 15%, of the city population had poverty level incomes at the time of the Census. The researcher could hypothesize that there was no decline in the proportion of the population with poverty level incomes since the time the Census had been conducted.

The random sample of the city population would provide sample data which could be used to test whether or not poverty had increased. Suppose that in the sample, there were 0.25, or 25%, of the city population now in poverty. This is a much larger proportion of the population than the 0.15, or 15%, previously in poverty. If the sample size for this sample is reasonably large, an hypothesis test using this sample would likely provide sufficient evidence to show that the incidence of poverty in the city had increased. In constrast, suppose that the sample showed that there were 0.17. or 17%, of the population in poverty. Given this data, it is likely that the method of sampling and conducting an hypothesis test would not provide sufficient evidence to conclude that there had been an increase in the proportion of the population in poverty.

The procedures just described seem fairly logical, and consistent with how we confirm or deny statements in ordinary life. The short general description just provided shows though, that the method of hypothesis testing is quite formal. The specific statements, which are either confirmed or denied on the basis of data from the sample, are of a particular type. In the above examples, the hypotheses to be tested were that there was no change in the mean income, or no change in the proportion of the population in poverty.

When conducting an hypothesis test, the sampling distribution of the sample statistic must be used, in order to determine how large a difference between the sample statistic and the hypothesized value of the parameter must be in order to either confirm or deny the hypothesis. In addition, each conclusion from an hypothesis test has a certain probability of being either correct or not. There are various types of error which can emerge in an hypothesis test, and each conclusion has a probability of being incorrect. This probability is called the level of significance, and an hypothesis test is sometimes called a significance test. It is necessary to develop some understanding of each of these concepts and procedures in order to be able to conduct an hypothesis test.

In addition to tests concerning particular population parameters, this chapter introduces tests concerning two different populations. Random samples from each of two different populations can be used to test whether the means or the proportions for the populations differ. While the hypothesis tests for two populations are conceptually similar to those for one population, introducing a second population increases the complexity of the formulas.

In the above example, suppose that the researcher had no earlier Census results to rely on, but did have information from an earlier sample of the population. Suppose the earlier sample had shown that 14% of the population was in poverty, and the researcher finds that there are now 17% of the population in poverty. The researcher would need to conduct a test for the difference between two percentages, or two proportions (0.14 and 0.17), in order to test whether this increase in sample percentages was sufficient to conclude that the incidence of poverty had increased for the population as a whole.

The method of hypothesis testing outlined in this chapter is widely used in statistical work. In inferential statistics, hypothesis testing is more extensively used than is interval estimation. There are many different types of hypothesis tests, of which only a very few are introduced in this chapter, with more in Chapters 10 and 11. There are many other hypothesis tests, beyond what can be introduced in an introductory textbook. All of these tests use the same principles as the tests developed here. If you develop a good understanding of the method of testing a mean, a proportion, and the difference between two means or proportions, you will be able to understand other types of hypothesis testing as well. As a result, this chapter is important for understanding inferential statistics.

Chapter Outline. This chapter begins with the test of a population mean, for a large random sample. The principles of hypothesis testing are introduced throughout the following section. The first time through the following section, you may have difficulty understanding the manner in which an hypothesis test is conducted, and how the test is to be interpreted. As a result, it may be worthwhile to come back to Section 9.2 several times, to review the various aspects of hypothesis testing. Section 9.3 presents hypotheses tests for the mean when the sample size is small, so that the t distribution must be used. In Section 9.4, an hypothesis test for a proportion will be given. You will see that this is essentially the same method as

that used in Section 9.2 for the mean, but with $\hat{p}$ replacing $\bar{X}$. In Section 9.5 a test for the difference between two proportions is discussed. In Section 9.6, tests for the difference between two population means are examined, for both large and small sample sizes. The formulas for conducting these tests are given, and in addition, in Section 9.7 there are results from computer programs to test for the difference between two means. A test for the equality of two variances is also given in Section 9.7. A short discussion of how results from hypothesis tests can be reported when doing research is given in Section 9.9. The last test in the chapter, in Section 9.10, is the paired t test for dependent samples.

Section 9.2.9 shows the similarity between hypothesis testing and interval estimation. Looked at in one way, there is really only one type of inferential statistics, the conclusions may be stated either in the form of an hypothesis test, or in the form of an interval estimate.

9.2 Hypothesis Test for a Mean

This section shows how an hypothesis test for a mean, when the sample size is large, can be conducted. Since this is the first time hypothesis testing has been discussed in this textbook, the various principles of hypothesis testing will be introduced by showing how a test for the mean is conducted. As a result, this section is quite long, containing both actual tests for the mean, and a more general discussion of the principles of hypothesis testing. Testing is introduced in this manner because hypothesis testing is more easily understood by example than on the basis of a theoretical presentation. Once you follow a few examples, then the principles of testing can be applied to other types of hypothesis test. If you find this presentation confusing, go to Example 9.2.1 after you have read part of the way through this section. Example 9.2.1 goes directly through a test, without a full discussion of peripheral points. Then you can return to the more general discussion, and the issues involved in hypothesis testing should become clearer.

9.2.1 Steps in Hypothesis Testing

Hypothesis testing has a set of procedures which are used for each test. These procedures are rather formal, and are much the same for each type of hypothesis test. This section outlines these, and if you are able to follow this set of procedures when encountering a new hypothesis test, then you should have little difficulty carrying out the test. The following description

is built around the test of a mean, but the steps outlined here are general procedures, and each hypothesis test uses these same steps. The following description is given in words, and a diagrammatic presentation of hypothesis testing follows later in Section 9.2.2.

Formulating the Hypotheses. Hypothesis testing begins by stating two hypotheses. If the test is concerned with the value of the population mean, then these hypotheses are statements concerning the value of the population mean. These statements may be based on a hunch, or they may be claims made by another researcher, or by someone else. In any case, someone has made a claim, and this claim must be tested by using some data.

The claim is called an **hypothesis**, and each hypothesis test has two hypotheses, the **null hypothesis** and the the **alternative hypothesis**. The null hypothesis is a specific claim. In the test of a mean, the null hypothesis is a claim that the population mean equals some specific value. The alternative hypothesis, sometimes referred to as a **research hypothesis**, is a more general statement. For the test of a mean, the alternative hypothesis states that the population mean is not equal to the value claimed in the null hypothesis, or that the mean is either greater than, or less than, claimed in the null hypothesis. Each hypothesis test is conducted assuming that the null hypothesis is true. At the end of the test, the conclusion is either to reject the null hypothesis, or not to reject the null hypothesis.

1. Null and Alternative Hypotheses. The null hypothesis is given the symbol H_0, and the research or alternative hypothesis is given the symbol H_1. Suppose that a claim is made that the mean of a population is M. Suppose that μ is the true, but unknown, mean of a population. The null and alternative hypotheses could take the following form.

$$H_0 : \mu = M$$

$$H_1 : \mu \neq M$$

There are other possible forms for the test, but this is the most common and easily understood form. The null hypothesis H_0 states that the true mean of the population is equal to the hypothesized value M. The alternative or research hypothesis, is that the population mean is not equal to M.

As a general rule, the null hypothesis is always an equality, and the alternative hypothesis is an inequality. When conducting an hypothesis test, the null hypothesis is always assumed to be true, and the assumption adopted when conducting the test is that this equality holds. In the test of a population mean, the assumption of the null hypothesis is that the population mean is, in fact, equal to the hypothesized value M. At the conclusion of the test, the null hypothesis H_0 is either rejected or it is not rejected. If the sample mean differs quite considerably from the hypothesized value M, there may be sufficient evidence to reject the claim that $\mu = M$. The other possibility is that the sample mean is relatively close to M and then there is not sufficient evidence from the sample data to reject the null hypothesis.

2. The Test Statistic. Once the null and research hypotheses have been stated, the next step in hypothesis testing is to obtain a statistic which can be used to test the claim. The same statistics or point estimates as were used in estimation, are used for hypothesis testing. In the above hypotheses concerning μ, the test statistic is the sample mean $\bar{X}$. This sample mean is obtained from a random sample of members of the population. In an hypothesis test concerning p, the proportion of a population having a particular characteristic, the test statistic is $\hat{p}$, the proportion of the sample having that same characteristic.

3. Distribution of the Test Statistic. Once the test statistic has been determined, it is necessary to know how the test statistic behaves when there are repeated random samples from the population. That is, the sampling distribution of the test statistic must be obtained.

In the case of the sample mean $\bar{X}$, if a random sample of reasonably large sample size has been obtained, then the Central Limit Theorem describes the distribution of the sample mean. That is, when n is large,

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

where μ is the true population mean, σ is the true standard deviation of the population and n is the size of the random sample. This distribution describes how $\bar{X}$ behaves when repeated random samples of the population are drawn.

Since the null hypothesis is assumed to be true, and since the null hypothesis states that $\mu = M$, the sampling distribution for $\bar{X}$ is assumed to

be centred at $\mu = M$. The standard deviation of the sampling distribution of $\bar{X}$ is $\sigma/\sqrt{n}$. The value of σ, the standard deviation of the population from which the sample is drawn is usually unknown, so that the sample standard deviation s is used as an estimate of σ. Thus, if the null hypothesis is assumed to be correct, the sampling distribution of $\bar{X}$ which is used to test the hypotheses is

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

The idea behind hypothesis testing is to obtain the sample mean $\bar{X}$, and to determine how close $\bar{X}$ is to μ. If $\bar{X}$ is quite a distant from μ, then the null hypothesis H_0 is rejected and the alternative hypothesis H_1 is accepted. On the other hand, if $\bar{X}$ is relatively close to the hypothesized mean $\mu = M$, then this null hypothesis H_0 is not rejected. The question that emerges from this is what constitutes a large distance from μ and what is a small distance from μ. This decision is made on the basis of the **level of significance** of the hypothesis test.

4. Level of Significance of the Test. The level of significance of an hypothesis test is much like a confidence level, except that the significance level is a **small number**. Common levels of significance are 0.10, 0.05 or 0.01. These significance levels are given as probabilities, or proportions, rather than as percentages. In the test of the hypotheses that $\mu = M$ or that $\mu \neq M$, the 0.10 significance level is equivalent to the 90% confidence level, the 0.05 level to the 95% confidence level, and so on.

The significance level can be expressed as an area under the normal curve, with the area equal in size to the given level of significance. This area is in the extreme part of the distribution. Take the 0.10 significance level, and let this be the extreme 0.10 of the area under the sampling distribution of $\bar{X}$. The region of rejection of the null hypothesis will be the extreme 0.10 of the area under the sampling distribution, in either of the two tails of this distribution.

The alternative hypothesis is $\mu \neq M$. This hypothesis is accepted only if the value of $\bar{X}$ is considerably greater than M, or if $\bar{X}$ is considerably less than M. The decision concerning what is large and what is small is based on the significance level. If the level of significance is 0.10, then the null hypothesis is rejected if $\bar{X}$ is in the extreme 0.10 of the area under the sampling distribution of $\bar{X}$. If the value of $\bar{X}$ obtained from the sample is not in the extreme 0.10 area of the sampling distribution, then the null

hypothesis H_0 is not rejected.

Notation

For hypothesis testing, the α notation is helpful, and more straightforward than in the case of interval estimation. The symbol α is the first letter of the Greek alphabet, and is written and pronounced as 'alpha.' The level of significance is given the symbol α. If the significance level is 0.10, then $\alpha = 0.10$; if the significance level is 0.05, then $\alpha = 0.05$.

5. Critical Values and Critical Rejection. The region of rejection of the null hypothesis is called the **critical region** for the hypothesis test. The **critical region** is sometimes referred to as the **region of rejection of** H_0, and the two terms are synonymous. The critical region is the extreme part of the sampling distribution, equal in area to the significance level α. If the sample yields a sample mean $\bar{X}$ which falls in the critical region, then the null hypothesis is rejected. If the sample mean does not lie in the region of rejection of H_0, then the null hypothesis is not rejected.

The region of rejection of the null hypothesis is a set of Z values which are consistent with (a) the sampling distribution, with (b) the null and alternative hypothesis, and with (c) the level of significance. For the test of a sample mean, the distribution of $\bar{X}$ is normal, so the region of rejection will be an area associated with the normal curve. The null hypothesis is H_0 and the alternative hypothesis is $H_1 : \mu \neq M$. A region of rejection consistent with the alternative hypothesis is the set of values of $\bar{X}$ which are either a lot less than μ or a lot greater than μ. If the level of significance adopted in the test is $\alpha = 0.10$, then the critical region for testing H_0 has an area equal to 0.10. The region of rejection is all those values of Z in the extreme 0.10 of the normal distribution. This region is at both ends of the distribution, either at the extreme left end of the distribution, or at the extreme right end of the distribution.

If the total area of $\alpha = 0.10$ is to be split between the two ends of the distribution, this means that there is $0.10/2 = 0.05$ of the area in each tail of the distribution. Looking down column B of the table of the normal

distribution gives an area of 0.05 when Z is midway between 1.64 and 1.65. The Z value of 1.645 is usually used in this case. This value of Z which begins the critical region is called the **critical value** of Z. For this test, the critical values of Z are -1.645 and +1.645, and the critical region is all Z values which exceed these critical values in magnitude.

In summary, for the $\alpha = 0.10$ level of signficance, the region of rejection of H_0, or the critical region, is all values of $\bar{X}$ which are farther than $Z = -1.645$ on the left of centre, or farther than $Z = +1.645$ on the right of centre. More generally, the Z value associated with each α is determined from the normal table. The Z value is called the critical value for the hypothesis test. This is then used to define the region of rejection of H_0 in the tails of the sampling distribution.

6. Conclusion of the Test. The final step in conducting the test is to determine whether the sample mean is in the region of rejection or not. The random sample gives the values of $\bar{X}$, s and n. Since the sampling distribution for $\bar{X}$ in the test is

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

the Z value associated with the sample mean can be determined.

Recall that a Z value is a standardized value, and represents a number of standard deviations in the standardized normal distribution. It is obtained by subtracting the mean from the variable, and dividing this difference by the standard deviation. In general

$$Z = \frac{\text{Variable} - \text{Mean of Variable}}{\text{Standard Deviation of Variable}}$$

and this produces a Z value with mean 0 and standard deviation 1. For the test of a mean, the variable is $\bar{X}$, and the mean of $\bar{X}$ is μ. Since μ has been hypothesized to equal M, the mean of $\bar{X}$ is M for purposes of conducting the test. The standard deviation of $\bar{X}$ is $\sigma/\sqrt{n}$, and this gives the Z value for the sample mean as

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

For the sample mean $\bar{X}$, this tells how many standard deviations from the hypothesized mean μ the sample mean is.

Since σ is usually unknown, the sample standard deviation s is used as an estimate of σ. The estimate of the standard deviation of $\bar{X}$ is $s/\sqrt{n}$ and

this is used in determining Z. Since μ has been hypothesized to equal M, and the test is carried out assuming that the null hypothesis is true, the Z associated with the sample mean $\bar{X}$ is approximated by

$$Z = \frac{\bar{X} - M}{s/\sqrt{n}}.$$

If this Z is larger in magnitude than the critical value of Z, then the null hypothesis is rejected. If this Z is not in the critical region, then the null hypothesis is not rejected. In the case of the 0.10 level of significance, where the critical values are -1.645 and +1.645, if

$$Z < -1.645 \text{ or } Z > 1.645$$

the null hypothesis H_0 is rejected, and the alternative hypothesis is accepted. This decision is made at the $\alpha = 0.10$ level of significance. On the other hand, if

$$-1.645 < Z < 1.645$$

then the null hypothesis is not rejected at the $\alpha = 0.10$ level of significance.

This completes the test. If $\bar{X}$ is far enough from μ to be in the critical region, then the hypothesized mean does not represent the mean of the population. If $\bar{X}$ from the sample is closer to μ, so that it is not in the region of rejection, then the null hypothesis concerning the value of the true mean is not rejected.

Each of these conclusions is made with a certain probability, so that the conclusion may be in error. Most conclusions from hypothesis tests are correct conclusions. But each test has associated with it a chance that the conclusion is incorrect. If the null hypothesis is rejected, then there is a chance, equal in size to the significance level, α, that that this conclusion is incorrect. If the null hypothesis is not rejected, the probability of error is not so easily determined, although some comments concerning this type of error are provided later in this section.

An example of an hypothesis test is now presented. Following this example, there is further discussion of the principles of hypothesis testing.

Example 9.2.1 Study Hours of Undergraduates

Suppose that a group of students claims that undergraduates study an average of 20 hours per week. A survey of 494 undergraduate students at the University of Regina, conducted by a student in Sociology 404 in the Winter

1988 semester, was discussed in Example 8.3.1. This example showed that the mean number of hours students spent studying per week was 18.8 hours with a standard deviation of 13.1 hours. On the basis of this data, test the claim made by the students. Use the $\alpha = 0.10$ *level of significance.*

Solution. *The first step in conducting an hypothesis test is to be clear concerning which characteristic of the population is being investigated. The group of students is making a claim concerning the true mean hours studied per week for undergraduates. This true, but unknown, mean is the population parameter being tested. Let* μ *be the true mean number of hours per week studied for all undergraduate students. From the data for the sample of 494 undergraduates,* $\bar{X} = 18.8$ *and this is less than the 20 hours per week that students claim. Based on this sample mean, it would be tempting to conclude that the claim was incorrect. The problem with jumping to this conclusion at this point is that this is a random sample, and the sample mean is subject to sampling error. A different random sample would yield a different sample mean. The question is whether this sample mean, along with the sample standard deviation and sample size provides sufficient evidence to deny the students' claim that the mean hours per week studied is 20.*

Since the question is whether the claim can be supported or not, the claim could be proven incorrect if $\bar{X}$ *is either a lot greater than 20, or a lot less than 20. That is, the alternative hypothesis is that the mean hours per week studied is not equal to 20. The null hypothesis is that the mean hours studied per week equals 20. The hypotheses can be written as follows:*

$$H_0 : \mu = 20$$

$$H_1 : \mu \neq 20$$

Since the hypotheses concern the population mean, the test statistic which is used in this test is the sample mean $\bar{X}$*. Since* $n = 494$ *is much larger than 30, the Central Limit Theorem holds and the sampling distribution of* $\bar{X}$ *is*

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

A significance level of $\alpha = 0.10$ *has been specified, so that the region of rejection of* H_0 *is the extreme 0.10 of the area under the normal curve. Since this area is split between the two tails of the distribution, there is*

$0.10/2 = 0.05$ *of the area in each tail of the distribution. Using column B of the normal table in Appendix H, an area of 0.05 in one tail of the distribution is associated with* $Z = 1.645$. *For the* $\alpha = 0.10$ *level of significance, these are the critical values of* Z *the region of rejection is all* Z *values of less than* -1.645 *or greater than* $+1.645$.

Based on the sample data, $n = 494$, $\bar{X} = 18.8$ *and* $s = 13.1$. *Since* σ, *the true standard deviation of study hours for all undergraduates, is unknown but* n *is large,* s *can be used as an estimate of* σ. *The* Z *value is* $\bar{X}$ *minus* μ, *divided by* $s/\sqrt{n}$. *Since* $\mu = 20$ *is the null hypothesis, the* Z *value for the sample mean is*

$$\begin{aligned} Z &= \frac{\bar{X} - \mu}{s/\sqrt{n}} \\ &= \frac{18.8 - 20}{13.115/\sqrt{494}} \\ &= \frac{-1.2}{13.1/22.226} \\ &= \frac{-1.178}{0.589} \\ &= -2.04 \end{aligned}$$

Since $Z = -2.04 < -1.645$ *this* Z *value and the sample mean are in the region of rejection of* H_0. *As a result, the null hypothesis* H_0, *stating that the mean hours per week studied equals 20, can be rejected. The claim made by the group of students can be rejected at the 0.10 level of significance. The alternative hypothesis* H_1 *that the true mean hours per week studied by undergraduates is not equal to 20 is the hypothesis which is accepted as a result of this test.*

Additional Notes on this Test. *Note that* H_0 *can fairly clearly be rejected since the* Z *value of -2.04 is considerably less than -1.645 and falls well inside the region of rejection of the null hypothesis* H_0. *This is the manner in which hypothesis testing works. A strict criterion concerning rejection or nonrejection of the null hypothesis is adopted. Once the significance level is adopted, and the sample obtained, then the region of rejection of* H_0 *is strictly defined. If* $\bar{X}$ *falls in the region of rejection, then* H_0 *is rejected.*

There are several possible errors that could emerge as a result of making this conclusion. Some of these will be discussed in more detail later, but it is worth observing the possible errors associated with this test at this point.

In terms of errors ordinarily associated with hypothesis testing as a method of making inferences, there is a chance that the claim made by the group of students is correct, even though the test has led to rejection of this hypothesis. When rejecting a null hypothesis, there is always some chance that the null hypothesis really is correct. This is referred to as Type I error, the error of rejecting the null hypothesis when it is actually true. (Type I error is discussed in Section 9.2.4). The chance of this type of error is the level of significance. The hypothesis test begins by assuming that $\mu = 20$. *If this is really the case, there is a 0.10 chance that a random sample from a population with* $\mu = 20$ *could yield a sample mean* $\bar{X}$ *that lies more than 1.645 standard deviations from centre. Yet the rule which has been adopted for the test is that any time that* $\bar{X}$ *is farther than 1.645 standard deviations from centre, the null hypothesis is rejected. This means that 0.10, or 10%, of the random samples from a population where* $\mu = 20$ *will yield sample means which are in the critical region. But each time a sample mean is in the critical region, the hypothesis* $H_0 : \mu = 20$ *will be rejected. There is thus chance of 0.10 that the null hypothesis will be rejected even though this hypothesis is correct. Each hypothesis test has the potential of this type of error built into it. Note that ordinarily this error is not actually made, but where the significance level is* α, *the probability is* α *that this error has occurred if the decision is to reject the null hypothesis.*

The other sources of possible error in the conclusion stem from the nature of the sample. The sample was not random, students may not accurately report the number of hours they spend studying, and so on.

Finally note the similarity of this result to the interval estimate of Example 8.3.1. The mean of $\bar{X} = 18.8$ *hours was associated with a 90% interval estimate of (17.8, 19.8) hours studied per week. Since this interval estimate does not include the hypothesized value of 20 hours per week, this interval estimate is consistent with the conclusion that the true mean is not equal to 20 hours per week. If the interval had included 20 hours, then it would seem that 20 hours per week could be the true mean.*

The conclusion of both the hypothesis test and the interval estimate is that the mean hours studied per week by all undergraduate students is not equal to 20 hours per week. The interval estimate was made with 90% confidence, and the hypothesis test with 0.10 significance. These are equal levels, the 90% of the area for the confidence level being in the middle of the normal curve, with 10% in the tails. This 10% in the tails equals the 0.10 area in the tails of the distribution associated with the 0.10 significance level for the hypothesis test. In Section 9.2.3, you will see that this test is called

a two tailed test, in that both tails of the distribution are part of the region of rejection. Later, in Section 9.2.9 the similarity of hypothesis testing and interval estimation will be discussed.

9.2.2 Diagrammatic Presentation of Hypothesis Testing

It may be easier to understand hypothesis testing if the sampling distribution is given in a diagram. A test of hypothesis for a sample mean is presented diagramatically in this section. The first diagram shows how the hypothesis test of Example 9.2.1 can be pictured. After this, the complete α notation involved in hypothesis testing is provided. In later sections, diagrammatic presentations for tests other than the population mean are provided.

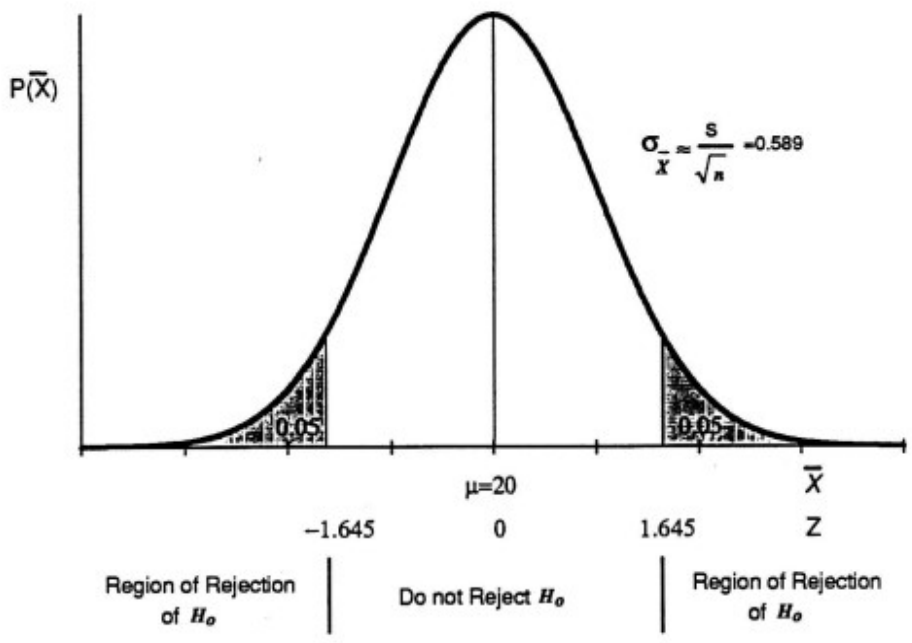

Figure 9.1: Test of Study Hours for Students

Begin with a variable X which has an unknown mean of μ and unknown standard deviation σ. Let a random sample with a large sample size n be selected from this population. Then the Central Limit Theorem can be used to show that the sample mean $\bar{X}$ has a normally distributed sampling

distribution with mean μ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. This result is true, regardless of the nature of the distribution of X. The only assumptions required are that the sample be random, and that n be reasonably large. If these are true,

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

This is pictured in Figure 9.1, with the normal sampling distribution for $\bar{X}$ shown. The values of $\bar{X}$ are shown along the horizontal axis of with μ being the mean of this distribution. The height of the normal curve represents the probability of occurrence for the different values of $\bar{X}$.

In Example 9.2.1 the claim was made that μ, the mean of the population, is $\mu = 20$. This is the null hypothesis H_0. Since the test is carried out under the assumption that the null hypothesis is true, the normal distribution for $\bar{X}$ is shown centred at $\mu = 20$. The standard deviation of the sampling distribution of $\bar{X}$ is not easily pictured in the diagram, but is shown on the upper right of the diagram as being equal to $\sigma_{\bar{X}} = 0.589$.

Since the alternative hypothesis is that $\mu \neq 20$, the null hypothesis is rejected if the Z value is in either the upper or lower tail of the distribution. For the 0.10 level of significance, the region of rejection of H_0 is all Z values of less than -1.645 or greater than 1.645. This is shown at the bottom of the diagram. As can be seen there, the null hypothesis, H_0, is not rejected if $\bar{X}$ gives a Z value between -1.645 and 1.645.

From Example 9.2.1, the sample mean is 18.8, and this is equivalent to $Z = -2.04$. This value of Z is less than -1.645, and thus is in the region of rejection of H_0. The conclusion is that the null hypothesis can be rejected and the alternative hypothesis accepted, at the 0.10 level of significance.

The General Case. The general case, along with all the notation necessary for illustrating hypothesis testing, is given in Figure 9.2. Suppose that the researcher is testing an hypothesis concerning the true mean μ for a population. Someone claims that $\mu = M$, and the researcher wishes to test this claim. Since nothing further is known concerning the population mean, the alternative hypothesis H_1 is that the true mean μ does not equal M. This can be summarized in the hypotheses:

$$H_0 : \mu = M$$

$$H_1 : \mu \neq M$$

An hypothesis test begins by assuming that the null hypothesis is true. If $\mu = M$, then the sampling distribution of $\bar{X}$ is centred at $\mu = M$. Note that this is only an assumption, made for purposes of conducting the test. At the conclusion of the test, this claim that $\mu = M$ is either rejected or is not rejected.

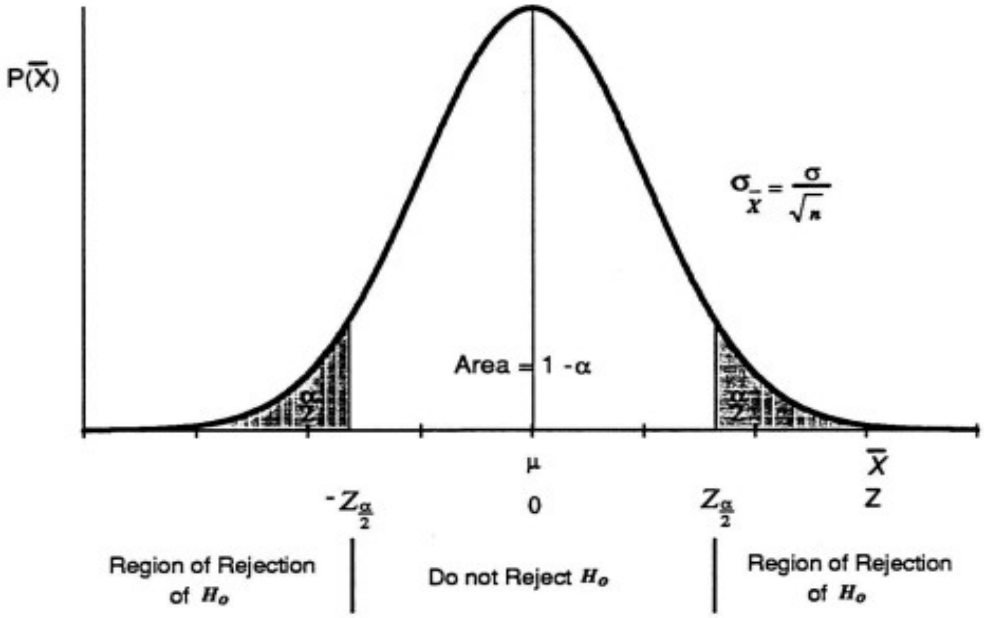

Figure 9.2: Test of a Mean, Large Sample Size

The alternative hypothesis H_1, that μ is not equal to M, can be accepted if the mean from the sample is either a lot greater than M or a lot less than M. On the other hand, if $\bar{X}$ from the sample is close to M, the null hypothesis is not rejected. This divides the set of possible values of $\bar{X}$ into two groups. The values of $\bar{X}$ far from $\mu = M$, in the tails of the distribution, are considered to be the values which lead to rejection of H_0, and acceptance of H_1. But if the value of $\bar{X}$ is one of those values which is not in one of the tails of the distribution, then the claim H_0 that the sample mean is M is not rejected.

The decision concerning how to divide the values of $\bar{X}$ into two groups

is made on the basis of the significance level chosen. The significance level is α, and is equal to the area in the tails of the distribution. In Figure 9.2, a significance level of α has been selected, and the shaded area in the diagram represents this area α. Since α is to be split between the two tails of the distribution, the area in each tail is $\alpha/2$ of the area under the curve. The remaining area of $1 - \alpha$ is the area in the middle of the normal curve.

In order to denote where the area in each tail of the distribution begins, it is necessary to provide some notation for the standardized Z values associated with different points along the horizontal axis. Let $Z_{\alpha/2}$ be the Z value such that $\alpha/2$ of the area under the normal curve lies beyond $Z_{\alpha/2}$. For this test, $Z_{\alpha/2}$ is the critical Z value for the right half of the critical region. On the right, $Z_{\alpha/2}$ represents the Z value at which an area of $\alpha/2$ in the right tail begins. Stated another way, there is $\alpha/2$ of the area which is more than $Z_{\alpha/2}$ standard deviations above the centre of the distribution. Similarly, to the left of centre, the critical Z value is $-Z_{\alpha/2}$, the distance to the left of centre that one must go so that only $\alpha/2$ of the area is less than this. The area under the curve between $-Z_{\alpha/2}$ and $+Z_{\alpha/2}$ is $1 - \alpha$.

The critical region and the decision rule concerning rejection or nonrejection of the null hypothesis can now be illustrated. The **region of rejection of** H_0, or the critical region, is the set of values of $\bar{X}$ which give Z values which lie below $Z = -Z_{\alpha/2}$ or which lie above $Z = +Z_{\alpha/2}$. Note that this region of rejection of H_0 is a **disjointed region**, being composed of the lower $\alpha/2$ and the upper $\alpha/2$ proportions of the possible values of $\bar{X}$. This can also be considered the region of acceptance of H_1, because rejection of H_0 generally mean acceptance of the alternative hypothesis.

The **region of nonrejection of** H_0 is the large set of possible values of $\bar{X}$ in the middle of the distribution. This is associated with the $1 - \alpha$ of the values between $Z = -Z_{\alpha/2}$ and $Z = +Z_{\alpha/2}$. Ordinarily this is referred to only as the region of nonrejection of the null hypothesis, and **not** the region of acceptance of the null hypothesis. As will be shown in Section 9.2.6, a null hypothesis is not often accepted, even when it is not rejected.

The final step in conducting an hypothesis test is to determine where the sample mean lies. As noted earlier, this requires calculating the Z value associated with the sample mean. This is determined by calculating

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

Since μ has been hypothesized to equal M, and since s is usually used to

provide an estimate of σ, the Z value in practice is

$$Z = \frac{\bar{X} - M}{s/\sqrt{n}}.$$

If this value is in the region of rejection of the null hypothesis, then this hypothesis is rejected. If $\bar{X}$ is not in the region of rejection of the null hypothesis, then this hypothesis is not rejected.

The decision rules concerning rejection or nonrejection of H_0 can be summarized as follows:

Decision Rule for Hypothesis Testing

Compute the Z value for the sample mean

$$Z = \frac{\bar{X} - M}{s/\sqrt{n}}.$$

If $Z < -Z_{\alpha/2}$, then reject H_0 and accept H_1.

If $Z > +Z_{\alpha/2}$, then reject H_0 and accept H_1.

If $-Z_{\alpha/2} < Z < +Z_{\alpha/2}$ then do not reject H_0.

The respective areas are illustrated in Figure 9.2 and are labelled as regions of rejection or nonrejection of H_0.

This diagrammatic presentation outlines the method of hypothesis testing for what is refered to as a **two tailed test of the mean**. In the following section, the test refered to as a one tailed test of a mean, is presented. For the latter, the diagram will be altered somewhat, although the basic principles of testing stay the same. The diagram of Figure 9.2 should seem quite similar to the diagrams for interval estimation. It will be seen later, in Section 9.2.9 that the two tailed test of the mean just discussed produces identical conclusions as do the confidence interval estimates.

9.2.3 One and Two Tailed Tests

The hypothesis tests discussed so far have been two tailed tests, tests. In these tests, if $\bar{X}$ is in either the right or in the left tail of the sampling distribution, the null hypothesis is rejected. Similarly, confidence intervals are always constructed so that they are of equal distance on each side of the sample mean. Another type of hypothesis test is a **one directional** or a **one tailed** test. This type of test begins with a one directional inequality for the alternative hypothesis. In a one tailed test, in order to be consistent with the alternative hypothesis, the region of rejection of the null hypothesis is in only one tail of the distribution. Otherwise, the principles of this test are the same as those introduced earlier, but with a specific direction specified for the mean as part of the test.

An hypothesis test for a mean begins with a null hypothesis which is the same as for the tests introduced earlier. Suppose though that the researcher has some evidence, or some suspicion, that the mean falls short of, or exceeds, the value specified in the claim. Then a one directional alternative hypothesis might be used.

To illustrate the method used in a one tailed test, suppose a researcher suspects that μ, the mean of a population, is less than the value M. Then the null hypothesis would still be

$$H_0 : \mu = M$$

but the alternative hypothesis would be

$$H_1 : \mu < M$$

As before, $\bar{X}$ would be the test statistic and, if the sample size is large,

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

If significance level α is selected, the only difference from the earlier tests is that all of the area α is placed in the tail of the distribution specified by the alternative hypothesis. The region of rejection of the null hypothesis is entirely in the left tail of the distribution. The null hypothesis is rejected and the alternative hypothesis accepted if the sample mean lies in the extreme left tail of the sampling distribution.

The critical region is all Z values less than the critical value of Z_α, where Z_α is the Z such that the proportion of the area under the curve less than

Z_α is α. Note that in a one tailed test, the area of α is not split into two, and it lies entirely in one tail of the distribution. For example, if the significance level is $\alpha = 0.05$, the critical Z value is $Z = -1.645$. That is, $Z = -1.645$ is the Z so that there is exactly 0.05 of the area under the normal curve to the left of this. For $\alpha = 0.10$, a one tailed test that $\mu < M$ has a critical region of $Z < -1.28$.

If the alternative hypothesis is in the positive direction, then the region of rejection of H_0 is in the right tail of the distribution. That is, if $\alpha = 0.05$, and the alternative hypothesis is

$$H_1 : \mu > M$$

the null hypothesis is rejected if $\bar{X}$ is sufficiently large so that $Z > 1.645$.

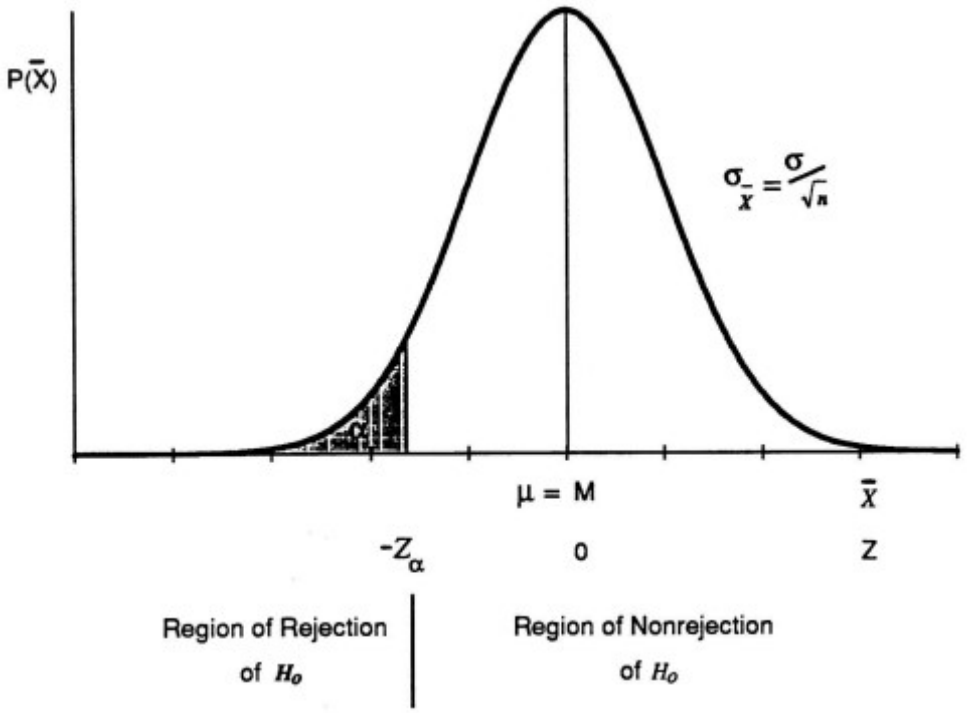

Figure 9.3: Diagrammatic Representation of a One Tailed Test

Figure 9.3 presents a one tailed test diagramatically. The hypotheses in this figure are

$$H_0 : \mu = M$$

$$H_1 : \mu < M$$

and the significance level is α. All of the significance level of α is placed in the left tail of the distribution. The shaded area represents the region of rejection of H_0 and it has area equal to α. This region of rejection is all the Z values less than $-Z_\alpha$. If the Z value associated with the sample is greater than $-Z_{\alpha/2}$, than the null hypothesis that $\mu = M$ is not rejected.

The following example uses the survey of undergraduate study hours presented in Example 9.2.1 to conduct a one tailed test.

Example 9.2.2 Study Hours of Students

The example of the weekly hours that undergraduates spend studying could have been given as a one tailed test. The sample values stay the same as in Example 9.2.1 but the claims might have been presented indicating a direction.

Suppose, as before, that a group of students claims that undergraduates study an average of 20 hours per week. Suppose, in addition, that another group of students views 20 hours as an overestimate of the average study time, and argues that students study less than 20 hours per week. Using the $\alpha = 0.05$ level of significance, test these claims.

Solution. *Again let μ be the true mean of the hours per week that all undergraduates spend studying. The null hypothesis is a always a specific claim, that μ equals a particular value. Of the two claims, the claim that the average is 20 is the specific claim that is tested, so that it becomes the null hypothesis. The alternative hypothesis suggests that students do not study 20 hour per week, on average. The alternative claim is no longer that the mean does not equal 20, but now the alternative claim is that the mean is less than 20. The null and research hypotheses for this test are*

$$H_0 : \mu = 20$$

$$H_1 : \mu < 20$$

The test statistic is the sample mean $\bar{X}$ and, as before, the sample size is large so that $\bar{X}$ is normally distributed with mean μ and a standard deviation

estimated to be $s/\sqrt{n}$. *The significance level is* $\alpha = 0.05$, *and since the alternative hypothesis is that* $\mu < 20$, *this area is entirely in the left tail of the distribution. Looking at column B of the normal table of Appendix H, an area of 0.05 in one tail of the distribution is given by* $Z = 1.645$. *Since this area is in the left tail, the critical region for rejecting* H_0 *is the area given by*

$$Z < Z_\alpha = Z_{0.05} = -1.645.$$

Since the sample data remains the same as in Example 9.2.1, the sample mean $\bar{X}$ *of 18.8 hours per week is associated with* $Z = -1.996 < -1.645$. *This* Z, *and the sample mean* $\bar{X} = 18.8$, *are in the region of rejection of* H_0. *The conclusion of the test is to reject the null hypothesis and accept the research hypothesis. At the 0.05 level of significance, the sample data supports the claim that the mean number of hours studied per week by undergraduate students is less than 20 hours per week.*

The reason a one tailed test was used in this case was that some suspicion concerning the direction of the inequality is suggested. Earlier, in Example 9.2.1, no indication of whether students study more than 20 hours or less than 20 hours was given. In that example, the two tailed test was most appropriate, especially if the researcher had no previous suspicion concerning the direction of the inequality.

Example 9.2.3 Survey of Numerical Abilities of Canadians

In 1989, Statistics Canada conducted a Canada wide survey of literacy levels of Canadians. As part of this survey, **numeracy** *or the numerical abilities of respondents was also measured. This question asks you to examine some of the data concerning numeracy. Numeracy was measured on a three point scale with 1 being the lowest level of numeracy and 3 being the highest. Those respondents who were at level 1 could recognize or locate and recognize numbers in a short text. Those at level 2 could perform simple operations such as addition and subtraction. Those at level 3 could perform sequences of numerical operations which would allow them to carry on everyday needs. (For a fuller description of the methods used and results from this study, see Statistics Canada,* **Adult Literacy in Canada**, *catalogue number 89-525E). Using the 3 point scale of numeracy, the mean level of numeracy for all Canadians was 2.378.*

In Saskatchewan, 374 randomly selected respondents were surveyed and the distribution of their numeracy levels is given in Table 9.1. At the 0.05 level of significance, can you conclude that the mean numeracy level of all

Numeracy Level	Number of Respondents
1	58
2	92
3	224

Table 9.1: Numeracy level of 374 Saskatchewan Adults

Saskatchewan adults exceeds the mean numeracy level for Canadian adults as a whole? If the significance level is reduced to 0.01, does this change your conclusion?

Solution. *The question asks you to use this data to determine whether the mean level of numeracy for all adults in Saskatchewan exceeds the mean level of 2.378 that is given for all Canadians. The true mean numeracy level for all those in Saskatchewan is not known, so let μ represent this true mean numeracy level for Saskatchewan. The null hypothesis is an equality, and it is hypothesized here that the true mean for Saskatchewan equals the Canadian mean of 2.378. The question also asks whether the Saskatchewan mean exceeds the Canadian mean. Given the direction of this question, this test should be conducted as a one tailed test. The null and research hypotheses are:*

$$H_0 : \mu = 2.378$$

$$H_1 : \mu > 2.378$$

The test statistic is the sample mean, and since the Saskatchewan sample of $n = 374$ is large, the sample mean

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

The significance level is $\alpha = 0.05$, and since the test is a one tailed test, in the right tail of the distribution, the region of rejection of H_0 is the upper 0.05 of the normal distribution. The critical value for the test is the familiar $Z = 1.645$. The region of rejection of H_0 is $Z > +1.645$. If $Z < +1.645$, then there is not strong enough evidence to reject the null hypothesis.

Figure 9.4 gives a diagrammatic presentation of this test. The population mean is hypothesized to equal $\mu = 2.378$*, and the standard deviation of the sampling distribution,* $\sigma_{\bar{X}}$ *equals 0.0386. The critical region for the first test is the shaded area representing the upper 0.05 of the possible values for* $\bar{X}$ *in the right tail of the distribution. This begins at* $Z = 1.645$*. For the second test, the shaded area is the dark area in the upper 0.01 of the distribution, beginning at* $Z = 2.33$*.*

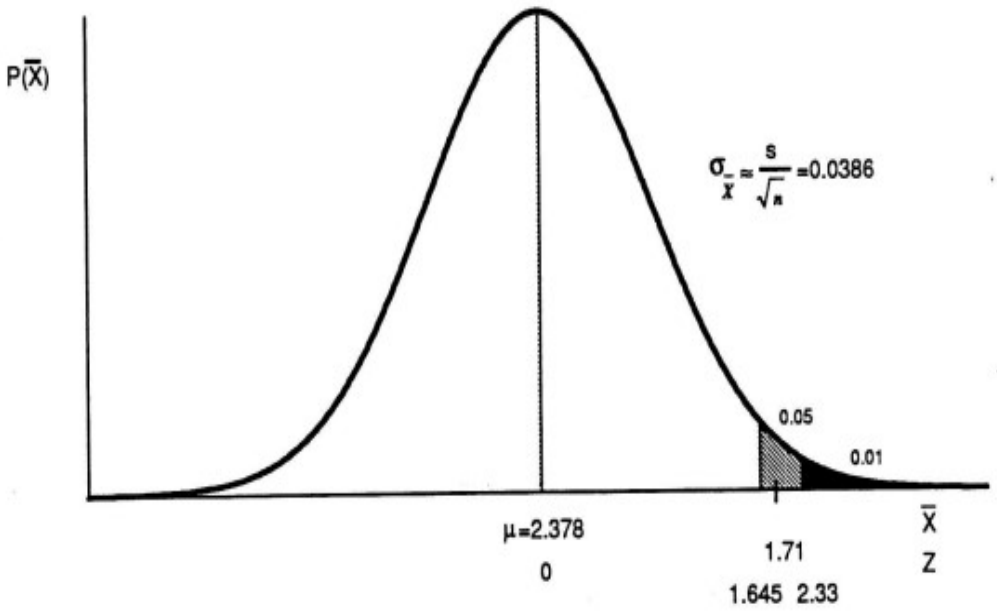

Figure 9.4: Test of Mean Level of Numeracy in Saskatchewan

At this point it is necessary to use the sample data in Table 9.1 to determine the mean and standard deviation of the sample data. This can then be used to determine the Z *value associated with the sample mean.*

Table 9.2 and the following formulas give the calculations for determining the mean and standard deviation. From Table 9.2, $n = 374$*,* $\sum X = 914$*, and* $\sum X^2 = 2,442$*. The mean numeracy level for the Saskatchewan sample is*

$$\bar{X} = \frac{\sum X}{n} = \frac{914}{374} = 2.444$$

X	f	fX	fX^2
1	58	58	58
2	92	184	368
3	224	672	2,016
Total	374	914	2,442

Table 9.2: Numeracy level of 374 Saskatchewan Adults

and the standard deviation is

$$\begin{aligned} s &= \sqrt{\frac{1}{n-1}\left[\sum X^2 - \frac{(\sum X)^2}{n}\right]} \\ &= \sqrt{\frac{1}{373}\left[2,442 - \frac{(914)^2}{374}\right]} \\ &= \sqrt{\frac{1}{373}\left[2,442 - 2,233.7\right]} \\ &= \sqrt{0.559} \\ &= 0.747 \end{aligned}$$

Given this mean and standard deviaton the Z value associated with the sample mean of $\bar{X} = 0.2.44$ can be computed using

$$Z = \frac{\bar{X} - M}{s/\sqrt{n}}$$

$$Z = \frac{2.444 - 2.378}{0.747/\sqrt{374}} = \frac{0.066}{0.0386} = 1.71.$$

Using the data from the Saskatchewan sample, $Z = 1.71 > 1.645$ and the null hypothesis can be rejected, at the 0.05 level of significance. In Figure 9.4, the value of $Z = 1.71$ can be seen to lie just to the right of 1.645, so that $\bar{X}$ is in the critical region for the $\alpha = 0.05$ level of significance. At this level, the alternative hypothesis that the mean numeracy level for Saskatchewan exceeds the Canadian mean can be accepted.

For the 0.01 level of significance, all that changes is the region of rejection. This significance level is a much more demanding one, requiring that

stronger evidence be presented in order to reject the null hypothesis. For $\alpha = 0.01$ *and a one tailed test in the positive direction,* $Z_\alpha = Z_{0.01} = 2.33$. *The region of rejection of* H_0 *is all* Z *values of 2.33 or more. If* $Z < 2.33$, *then there is insufficient evidence to reject* H_0 *at the 0.01 level of significance.*

The data from the sample does not change, and earlier it was shown that $\bar{X} = 2.444$ *was associated with* $Z = 1.71$. *As can be seen in Figure 9.4, this value is less than 2.33, and is not in the region of rejection of* H_0. *At the 0.01 level of significance, the conclusion is that the Saskatchewan mean level of numeracy does not exceed the Canadian mean level of numeracy.*

Additional Comments of this Test.

The first problem that might be noted with this test is that the numeracy scale is not an interval level scale, but is only an ordinal scale. In spite of this, the mean and standard deviation can be calculated, and the test carried out. But the results should be interpreted with caution. The conclusion that the mean level of numeracy is greater for Saskatchewan than for Canada as a whole should be interpreted very cautiously. Given the mean, standard deviation, and sample size, the conclusion is correct at the 0.05 level, for a one tailed test. But if a scale which measured numeracy at the interval level was available, then these statistics might have quite different values.

Another problem related to this is the nature of the Survey itself. The Survey is being treated as if it were a random sample of Saskatchewan adults. It is unlikely that it is a random sample, but is likely based on some modification of random sampling. If this is so, then the standard deviation $s/\sqrt{n}$ *may not accurately represent the variability in the sample mean. This could introduce additional error into the conclusion.*

The differing conclusions associated with the different significance levels must also be considered. Ignoring the above problems, and assuming the data is acceptable as analyzed here, the Saskatchewan mean level of numeracy has been shown to exceed the mean Canadian level when the 0.05 level of significance has been used. But at the 0.01 level of significance, the conclusion that the Saskatchewan and the Canadian mean are identical could not be rejected.

Different conclusions at different significance levels are part of the problem of interpreting hypothesis tests, and of choosing a significance level for the test. A significance level which is quite large is associated with a large area in the region of rejection. This means that the Z *is not so large, and the sample mean need not differ all that much from the hypothesized mean in order to reject the null hypothesis. In this example, if* $\alpha = 0.05$, *then*

the Z associated with the sample need only exceed 1.645 in order that H_0 is rejected. But when the significance level is reduced to $\alpha = 0.01$, there must be a larger Z, in order to reject H_0. In this example, the sample mean would have had to differ from the hypothesized mean by over 2.33 standard deviations before the null hypothesis could have been rejected at the 0.01 level of significance.

Which of these levels of signficance should be used in this example is not entirely clear. The 0.05 level of significance is the most common, but some research issues require a more demanding significance level. Since the numeracy level is not a matter of life and death, or of physical health, the 0.05 level of significance for this test is probably adequate. While there are the other problems associated with this data, in terms of the test itself, this conclusion provides evidence that the mean Saskatchewan numeracy level exceeds the mean Canadian level. In Section 9.2.5, the choice of levels of significance is discussed.

9.2.4 Types of Error

When conducting an hypothesis test, there is always the possibility of one of two types of error. These errors are referred to as **Type I Error** and **Type II Error**. These types of error may seem confusing at first, but as you carry out hypothesis tests, you will find that these types of error illustrate the meaning of the tests. The types of error are defined first, and then followed by a discussion and examples.

Suppose that at the end of an hypothesis test, the Z value is in the critical region for the test and the conclusion is that the null hypothesis can been rejected. While this will usually be the correct conclusion, it may be that the null hypothesis is true and the decision to reject the null hypothesis is not the correct conclusion. When the null hypothesis is rejected, even though it should not be, this is Type I error.

Type I Error

Type I Error is the error of rejecting the null hypothesis H_0 when the null hypothesis is true.

The probability of committing Type I error is α, the significance level. Type I error is sometimes called α error.

At the end of an hypothesis test, the conclusion may be that there is insufficient evidence to reject the null hypothesis. Yet it may be that the the null hypothesis H_0 is not true, and it should have been rejected. When this happens, Type II error has been committed.

Type II Error

Type II Error is the error of failing to reject the null hypothesis H_0 when the null hypothesis is false.

The probability of committing Type II error is β. Type II error is sometimes referred to as β error.

The symbol β is the second letter of the Greek alphabet, and is written and pronounced 'beta'. The probability of Type I error is α, and this equals the significance level selected. In contrast, β, the probability of Type II error, cannot usually be determined. While some statements can be made concerning β, its exact level is usually unknown.

Explanation of Types of Error. These two types of error can be explained as follows. A test of a mean with significance level α and

$$H_0 : \mu = M$$

$$H_1 : \mu \neq M$$

will be used to explain and illustrate these types of error. For this test, the critical region is all Z values less than $-Z_{\alpha/2}$ or greater than $Z_{\alpha/2}$. If

$$Z = \frac{Z - M}{s/\sqrt{n}}$$

is in the critical region, then H_0 is rejected, and H_1 accepted. If Z is not in the critical region, then H_0 is not rejected.

The hypothesis test proceeds assuming that the null hypothesis is true. If $\bar{X}$ is more than $Z_{\alpha/2}$ standard deviations from $\mu = M$, then the null hypothesis is rejected. This is usually the correct conclusion, because the sample mean is so far from the hypothesized mean, that the hypothesis $\mu = M$ is not likely to be correct.

Suppose now that the null hypothesis is true, and that $\mu = M$. If this is the case, most of the random samples selected would not have means that fall into the critical region. But just by chance, there would be some of the more unusual random samples which yield means which are in the region of rejection of the null hypothesis. According to the construction of the normal sampling distribution of $\bar{X}$, there would be a proportion α of the random samples which have these unusual sample means. But each time a sample mean which lies in the region of rejection of H_0 occurs, the null hypothesis H_0 is rejected. Thus, in $\alpha\%$ of these samples, the null hypothesis is rejected, even though it is true. This is the origin of the statement that the probability is α that Type I error occurs.

The probability of Type I error can be written as a conditional probability:

$$P \text{ (Type I Error)} = P \text{ (Rejecting } H_0 / H_0 \text{ is true)} = \alpha.$$

That is, if the null hypothesis is true, the conditional probability is α that the null hypothesis is rejected. There is thus a maximum probability of α of making Type I error. The consequence of Type I error will be discussed after explaining Type II error.

Type II error can occur when the null hypothesis is not rejected. This occurs when the sample mean is relatively close to the hypothesized mean, so that Z is not in the region of rejection of H_0. If the null hypothesis is correct, or is very close to being correct, then this is the proper conclusion. But suppose that that the sample mean is quite different from the value $\mu = M$, hypothesized in H_0. By chance, perhaps because the sample size is small, or because an atypical sample has been selected, suppose that the Z

value from the sample is not large enough to be in the region of rejection of H_0. What has happened then is that the sample mean appears to come from a population with mean $\mu = M$, when in fact it comes from a population with a mean different than M. The method of hypothesis testing has been unable to separate the true mean of the population from the hypothesized mean. It seems as if the null hypothesis is true, even though the true mean is not equal to this hypothesized value. This is Type II error.

Type II error can be written as a conditional probability. This is

$$P \text{ (Type II Error)} = P \text{ (Failing to Reject } H_0 / H_0 \text{ is not true)} = \beta.$$

If the null hypothesis is not true, the conditional probability is β that the null hypothesis is not rejected. There is thus a probability of β of making Type II error.

	Actual Situation	
Decision Taken	H_0 True	H_0 Not True
Reject H_0	Type I Error	Correct Decision
Do Not Reject H_0	Correct Decision	Type II Error

Table 9.3: Types of Error in Hypothesis Testing

Table 9.3 summarizes the possible decisions and the types of error associated with hypothesis testing. Note that there are correct decisions in two cases, but that when rejecting the null hypothesis, there can be Type I error. When not rejecting the null hypothesis, there can be Type II error.

Consequences of Types of Error. The situation in Table 9.3 seems symmetrical with respect to the two types of error. But hypothesis testing is ordinarily constructed so that the negative consequences of Type I error are more serious than those for Type II error. In most hypothesis tests, the researcher wishes to keep Type I error to quite a low level.

In order to understand why this should be the case, consider the claim of the group of students in Example 9.2.1. The claim of this group of students was that the mean study hours for all undergraduates was 20 hours per week. The method adopted in hypothesis testing is to construct the test so

that it is relatively difficult for the researcher to reject the null hypothesis. If the null hypothesis is rejected, then this rejection is quite solidly based.

With $H_0 : \mu = 20$, and the $\alpha = 0.10$ level of significance, the region of rejection was all Z values which lie more than 1.645 standard deviations from $\mu = 20$. This means that the researcher must find a sample mean which lies farther than 1.645 standard deviations from the hypothesized mean before the null hypothesis can be rejected. But note that there is a chance of Type I error, that $\mu = 20$ and the sample just happens to be a group of students whose sample mean $\bar{X}$ is more than 1.645 standard deviations from center. While this is unlikely, it could happen in 0.10, or 10%, of the random samples selected from the population.

Now consider how Type II error could occur. Suppose that a random sample of 494 students is taken, and the sample mean is $\bar{X} = 19.5$ hours per week. Then

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{19.5 - 20}{13.1/\sqrt{494}} = \frac{0.5}{0.589} = -0.85$$

and this Z is closer to the mean than 1.645 standard deviations. The null hypothesis is not rejected and the conclusion is that $\mu = 20$ is the mean hours studied per week for undergraduates. But suppose that undergraduates study only an average of 19 hours per week on average so that the true mean is $\mu = 19$. In this case, the hypothesis $H_0 : \mu = 20$ has not been rejected, but this value for μ is incorrect. Type II error has occurred.

Whenever the null hypothesis is not rejected, it is almost inevitable that there is Type II error, although exactly how incorrect such a conclusion really is, may not be all that clear. The researcher may say that this is not a very serious error. All that has happened is that the claim of $\mu = 20$ is off by a bit, and the researcher was unable to distinguish whether the true mean is 19 hours or the hypothesized value of 20 hours. While the null hypothesis may not be exactly correct, it is not all that incorrect either, and perhaps it makes little difference if $\mu = 19$ or $\mu = 20$.

In contrast, the conclusion to reject H_0 results in a more definitive statement, one that concludes that $\mu = 20$ is not correct. Based on this conclusion, the researcher may write something like 'The null hypothesis that $\mu = 20$ is rejected at the 0.10 level of significance.' The researcher has made a fairly definite statement that the true mean is not 20. In contrast, when the conclusion of the test is that the null hypothesis is not rejected, the researcher may not really conclude that $\mu = 20$. He or she may say something like 'The research results do not give a sample mean of exactly 20 hours,

but the sample mean is not different enough from 20 to reject the claim that the mean is 20.' The researcher may not believe that $\mu = 20$ is exactly true, but concludes that the true mean is close to 20. The differences between the hypothesized mean of 20, and the unknown true mean, may be relatively small or inconsequential.

If the conclusion to not reject the null hypothesis has more serious consequences to it, then the researcher can always change the significance level. By increasing the significance level, the researcher can make it somewhat easier to reject the null hypothesis.

The levels of the two types of error vary inversely, so that a reduction in Type II error increases the possibility of Type I error. Recall that Type I error is equal to the significance level. If the researcher wishes to reduce Type I error, then a smaller significance level is chosen. The consequence of this though, is that a higher level of Type II error could occur. When the significance level is reduced, this means that a larger Z is required in order to reject the null hypothesis. But a larger Z increases the set of values of Z which could lead to nonrejection of H_0, making it more difficult to reject H_0.

In order to see this, consider the one tailed test of numeracy levels in Example 9.2.3. When the significance level was 0.05, the region of rejection of H_0 was all Z values exceeding 1.645. When the significance level was reduced to 0.01, the region of rejection was $Z > 2.33$. The latter test is more demanding in that a larger Z is required in order to reject the null hypothesis. The sample mean of $Z = 1.71$ was sufficiently large to reject the null hypothesis at the 0.05 level of significance. But at the 0.01 level of significance, this Z was not large enough to conclude that the mean numeracy level in Saskatchewan was greater than the Canadian level. It is very likely that the mean numeracy level for Saskatchewan is somewhat different that the Canadian average, but this cannot be proved at the 0.01 level of significance. Thus the requirement that Type I error be reduced by lowering α has considerably increased the chance of Type II error.

Types of Error for a Pedestrian. The different consequences of Types I and II error can be illustrated by a practical example, the decision a person takes when deciding when to cross a street where there is a considerable volume of traffic. It is unlikely that the person contemplating crossing the street consciously considers the situation in the terms described here. Implicitly though, each of us makes decisions in many aspects of daily life, and

the considerations we take when making these decisions can be cast in an hypothesis testing model.

Suppose a person contemplates crossing the street at a busy intersection. The null and research hypothesis might be

$$H_0 : \text{It is unsafe to cross the street.}$$

$$H_1 : \text{It is safe to cross the street.}$$

Notice the manner in which the hypotheses have been constructed. The assumption adopted at the beginning is that it is unsafe to cross the street. The pedestrian wishes to make sure that if the street is to be crossed, then it should be safe. The level of Type I error is to be a low number here, so that if a decision to reject H_0 is made, there is minimal error in this decision. The assumptions, decisions, and types of error are summarized in Table 9.4.

The pedestrian begins with the assumption that it is not safe to cross the street. Only if there is sufficient evidence to reject this hypothesis, does the pedestrian take the decision to cross the street. In doing this, the pedestrian gathers evidence by looking each way, and at a certain point decides that the there is so little traffic that it is safe to cross. If the pedestrian is careful, then a correct decision is made, and the conclusion to cross the street is the correct decision, since it really is safe to cross. But it is possible that an error has been made. The pedestrian may not notice a car coming, or may misjudge the speed of a car, or the time it will take to cross. If it is not safe to cross the street, but the pedestrian has decided to cross, an incorrect decision is made. This is Type I error.

	Actual Situation	
Decision Taken	Unsafe to Cross (H_0 True)	Safe to Cross (H_0 Not True)
Reject H_0 and Cross Street	Type I Error	Correct Decision
Do Not Reject H_0, Do not Cross	Correct Decision	Type II Error

Table 9.4: Types of Error Associated with Pedestrian Crossing a Street

If the pedestrian does not reject the null hypothesis, this means that the pedestrian waits at the crossing. If it really is unsafe to cross, then this is

the correct decision. But it may be that it was safe to cross, and yet the pedestrian waits at the crossing longer than necessary. In this case, the null hypothesis has not been rejected, even though it was incorrect. This is Type II error.

The consequences of the two types of error can be seen to be quite different. Committing Type I error by crossing the street when it is unsafe is a dangerous decision, meaning that the pedestrian stands the risk of injury or death. But if Type II error occurs, then this has the relatively minimal consequence of resulting in waiting too long at the corner. The only negative consequence of Type I error is that the pedestrian wastes time.

The situation with respect to hypothesis testing in the social sciences is not as clear cut, nor does it have such different results, as this example. But when conducting tests of hypotheses, most researchers decide what level of Type I error they are willing to have, and then accept the consequences of the Type II error associated with this. There are no hard and fast rules concerning the proper level of significance to adopt in any given situation, although some guidelines are provided in the next section.

Level of Risk. One consideration in deciding on the level of significance, is the degree of risk the individual is willing to accept, and how the consequences of the types of error are judged. In the example of a pedestrian, it is possible to consider various types of approach that different people take toward risk. A very cautious person is likely to adopt a very low level of significance, making almost absolutely certain that the street is safe to cross before deciding to cross. That is, the cautious person reduces the significance level to a very low value, and thus reduces the chance of Type I error to a minimal value. But in doing this, the cautious person may spend too much time waiting at the corner. That is, reducing the level of Type I error increases the Type II error. The consequence of the latter is that there may be an overly long waiting time.

An individual who is willing to take more risks may be willing to take a chance that the street is safe to cross, perhaps when it really is not all that safe. That is, the risk taker is willing to live with a larger significance level, and a higher level of Type I error. As a consequence, the risk taker spends less time waiting. For the person in a hurry, the consequences of a long wait in order to cross a street may have more negative aspects attached to them than for the cautious person. As a result, the risk taker judges Type II error as having more negative consequences than does the cautious person.

Although the risk taker might not wish to contemplate the resulting higher level of Type I error implicit in his or her decision, this consequence does exist. That is, the risk taker has a greater chance of being hit by a vehicle when taking chances in crossing a street.

Finally, the situation could also be described for a suicidal person. If an individual wishes to get hit by a vehicle, then Type I error should be maximized, and Type II error minimized. This would ensure that this individual takes decisions in such a way as to plan to get hit by a vehicle. This is not ordinarily considered to be a rational way to make a decision. But it does illustrate how a different set of goals can lead to quite a different way of making a decision.

9.2.5 Choice of Significance Level

There are no hard and fast rule concerning the level of significance which should be adopted for a particular hypothesis test. In Chapter 8, the same situation occurred with confidence levels. Some guidelines were given there concerning the choice of confidence levels. Many of the same guidelines could be applied to hypothesis testing. These are summarized here.

1. The first rule to remember is that the level of significance should always be reported. For hypothesis testing, you should also report whether the test is a one tailed or a two tailed test.

2. If you are not sure which level to choose, for most social science problems the $\alpha = 0.05$ level of significance can be used. It is the most commonly used level and a level which is usually adequate for tests of hypotheses in the social sciences.

3. If you wish to compare your results with those of other researchers, then adopt the same level as they have used. If you later wish to change the level, this can always be done. But as a first approach, using the same significance level as others have used allows you to compare research results.

4. If you wish to reduce the level of Type I error, then reduce the significance level to a very low level, perhaps to $\alpha = 0.01$, or even to $\alpha = 0.001$. Remember though that this implies a higher level of Type II error. If the negative consequences of Type I error are not so negative, then it is preferable to provide a better balance of Types I and II error by adopting a significance level such as 0.05 or 0.10.

5. The last point suggests that some form of valuation could be placed on the negative consequences of each type of error, and an optimal solution to balancing these types of error could be obtained. This is difficult in the social sciences, but is sometimes possible in decisions in business or economics, and perhaps even in the political arena. If monetary costs can be attached to the different types of error, then it may be possible to evaluate the different costs expected with different levels of significance. The significance level giving the lowest cost can be chosen. In the example of the pedestrian, if time has an important value, then the decision which minimizes the loss of time might be made. In the sociology or psychology though, is difficult to imagine exactly what could be minimized or maximized.

6. Where decisions involve matters of health and safety, or matters of life and death, then the types of error should be adjusted so that the negative consequences are minimized.

   Imagine the sport of parachute jumping. Suppose the null hypothesis is made that a parachute is unsafe, and only if it can be proven safe would you be willing to use it to jump from an airplane. But since the parchute is a mechanical device, there is always some chance that the parachute cannot be proven 100% safe. Type I error here is the error of concluding that the parachute is safe when, in fact, it is unsafe. What level of Type I error would you be willing to accept before jumping? Most of us would like a level of Type I error of under 0.01, that is, less than 1 chance in 100 that the parachute would fail. But one failure per 100 jumps would be too large a probability for most of us. Would you be willing to jump if Type I error is reduced to 0.001? If not, what about 0.0001? Of course, you could decide not to jump at all. But those who do get involved in this sport are implicitly adopting some level of Type I error, because the risk of the parachute failing can never be reduced to exactly 0.

   In the workplace, or dealing with issues of environmental health, we also work with or accept various levels of Type I error for different problems. We may not be aware of these, but they do exist. Proponents of nuclear energy may point to the very low probabilities of Type I error associated with all the safety features built into nuclear power generating stations. But even though the probabilities are very low, the consequences of Type I error are extremely serious. A nuclear

accident could destroy most life in the region of the accident. Thus both the probabilities associated with the error, and the consequences of that error need to be considered when making decisions.

It is sometimes difficult to decide what level of significance is best to adopt. If you have any doubts, stick with the familiar levels such as 0.10, 0.05 and 0.01 that other researchers most commonly use. Another possibility is to report your conclusions for several different levels of significance. It is even possible to avoid making a decision concerning the level of significance. In Section 9.2.7, the method of determining the exact level of significance associated with the sample statistic will be described. This allows the reader to make his or her own decision concerning the implications of this significance level.

9.2.6 Acceptance and Rejection

You may have noticed that the decision to **accept the null hypothesis** has never been made in the discussion so far. The null hypothesis is adopted as a working hypothesis for carrying out the test. If the statistic from the sample is in the critical region, then the null hypothesis is rejected and the research hypothesis is accepted. But when the statistic from the sample is not in the critical region, then the conclusion is that the null hypothesis is not rejected. When concluding the test, what might seem to be the next logical step, that of accepting the null hypothesis, is not usually taken. The conclusion is merely that the null hypothesis is not rejected.

The reason for this is built into the method of hypothesis testing. The researcher wishes to reject claims which have been stated as null hypotheses. By looking on Type I error as a more serious type of error than in Type II error, and by selecting a small level of significance, strong conclusions from research are provided when null hypotheses are rejected.

The decision to not reject a null hypothesis may leave the researcher in a state of uncertainty, and this conclusion is not the most desirable to have. If a researcher conducted a survey, and none of the sample data allow the researcher to reject any of the null hypotheses made, then this research may be regarded as producing little new in terms of findings. All this research would be doing is confirming claims which had previously been made.

It is possible to reverse the direction of these types of error, selecting a much larger significance level, or reversing the null and alternative hypotheses. This is not commonly done, although this may be required for

some types of problems. For example, if the researcher wishes to prove that the sample selected has almost exactly the same characteristics as the population from which the sample is drawn, then the hope is that the sample statistic is very close to the value claimed in the null hypothesis.

For most hypothesis testing, the conventional method is to attempt to reject H_0, and to regard Type I error as being more serious than is Type II error. If the researcher cannot reject the null hypothesis, the researcher should **not** say that the null hypothesis is correct, or that the null hypothesis has been proven. All that should be stated is that the evidence is consistent with the null hypothesis. Alternatively, the researcher may say that the evidence from the sample is not different enough from what has been hypothesized to reject this null hypothesis. The researcher may still not believe that the null hypothesis is correct, and may suggest that further samples, studies, and research be conducted in order to obtain a more decisive conclusion.

A few examples of hypothesis tests for a mean are given here. Following this, a few more principles concerning hypothesis tests are discussed.

Example 9.2.4 Mean Income of Regina Families

According to the 1986 Census of Canada, the mean family income for all Regina families was $41,894 in 1985. The Social Studies 203 Regina Labour Force Survey also asked respondents to state their family income and the summary statistics for the 834 respondents who provided data on their family income was $\bar{X} = \$40,088$ *and* $s = \$23,438$. *The data in this Survey was obtained over the years 1986-1990. Based on the Survey data, could you conclude that there has been a shift in the mean family income in Regina from 1985 to the years from which the Survey data comes? (Use* $\alpha = 0.10$*). Comment on any possible errors there may be in the conclusion.*

Solution. *The question is whether the true mean income of Regina families in the years 1986-90,* μ*, is equal to the 1985 level, or whether it has changed. While the sample mean is below the 1985 level, the question concerning whether the mean income has shifted or not does not give any hint of whether the shift has been up or down, so that a two tailed test should be used. A large shift in either the upward or downward directions would lead to rejection of* H_0*. This can be set up as an hypothesis test as follows:*

$$H_0 : \mu = 41,894$$

$$H_1 : \mu \neq 41,894$$

H_1 states as a research hypothesis that the mean income has shifted, and the null hypothesis H_0 states that there has been no shift in mean income. The test statistic is $\bar{X}$ and since the sample size of $n = 834$ for the Regina sample is reasonably large (greater than 30),

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

For a significance level of $\alpha = 0.10$ and with a two tailed test, the critical region is all Z values of less than -1.645 or greater than +1.645. If Z falls within the critical region, then the null hypothesis H_0 is rejected. If Z is outside the critical region, between -1.645 and +1.645, then the null hypothesis is not rejected.

Based on the sample data, $\bar{X} = 40,088$ and $s = 23,438$. Since σ, the true standard deviation of family income for Regina families, is unknown but n is large, s can be used as an estimate of σ.

$$\begin{aligned} Z &= \frac{\bar{X} - \mu}{s/\sqrt{n}} \\ &= \frac{40,088 - 41,894}{23,438/\sqrt{834}} \\ &= \frac{-1,806}{23,438/28.8791} \\ &= \frac{-1,806}{811.590} \\ &= -2.23 \end{aligned}$$

Since $Z = -2.23 < -1.645$, this Z value and the sample mean are in the critical region. The conclusion is that at the $\alpha = 0.10$ significance level, the results are strong enough to reject the null hypothesis of no shift in the mean income of Regina families. With 0.10 significance, the conclusion is that the mean income has shifted.

Comments on the Conclusion. *There are several possible errors that might be involved in this conclusion. First, any time a null hypothesis is*

rejected, there is always the possibility of Type I error. That is, there is at most an 0.10 chance that the null hypothesis of no shift in mean income is correct, even though there was sufficient evidence to reject this hypothesis.

The other possible errors are errors connected with the sample. There are many possible errors here. The sample may not be a random sample, so that $\bar{X}$ *is not really* $Nor(\mu, \sigma/\sqrt{n})$*. Further, there may be many nonsampling errors such as incomplete coverage, wrong or misleading answers, and so on that can occur in any survey. Finally, the data from the Regina Labour Force Survey has been collected over 3-4 years, and during that time the mean family income may have shifted. This makes the whole question a bit ambiguous concerning what mean is really being measured. While the statistical evidence supports a shift in family income, these nonsampling errors make this conclusion somewhat uncertain. It would be best to have a sample which is conducted at a single point in time in order to be confident that the conclusion is well founded.*

Example 9.2.5 Farm Size in Rural Municipality of Emerald

Example 7.2.2 provided interval estimates for farm size and flax yield in the rural municipality of Emerald. Here some of this same data is used to test hypotheses concerning farm size in this rural municipality. Test whether the mean farm size in Emerald is equal to the mean Saskatchewan farm size of 419 hectares. Use the 0.05 level of significance. For the 47 farms sampled in Emerald, $\bar{X} = 492.1$ *and* $s = 440.5$*. Also test whether the mean cultivated acreage for Emerald exceeds the Saskatchewan mean cultivated acreage of 781. For the 47 farms sampled in Emerald,* $\bar{X} = 1068.8$ *and* $s = 1029.7$*. Use the 0.04 level of significance for the latter test.*

Solution. *The question is whether the true mean farm size for all farms in the rural municipality of Emerald,* μ*, is equal to the mean farm size for all Saskatchewan, or whether it is different. This implies that a tailed test should be used here. This can be set up as an hypothesis test as follows:*

$$H_0 : \mu = 419$$

$$H_1 : \mu \neq 419$$

H_1 *states as a research hypothesis that the mean farm size in Emerald is different than the mean farm size for all Saskatchewan, and the null*

hypothesis H_0 states that the mean farm size in Emerald is equal to the Saskatchewan average. The test statistic is $\bar{X}$ and since the sample size of $n = 47$ for the sample is reasonably large (greater than 30),

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

For a significance level of $\alpha = 0.05$ and with a two tailed test, the region of rejection is all Z values of less than -1.96 or greater than +1.96.

Based on the sample data, $\bar{X} = 492.1$ and $s = 440.5$. Since σ, the true standard deviation for all farms is unknown, but n is large, s can be used as an estimate of σ.

$$\begin{aligned} Z &= \frac{\bar{X} - \mu}{s/\sqrt{n}} \\ &= \frac{492.1 - 419}{440.5/\sqrt{47}} \\ &= \frac{73.1}{440.5/6.856} \\ &= \frac{73.1}{64.2535} \\ \\ &= 1.14 \end{aligned}$$

Since $Z = 1.14 > -1.96$ and $Z = 1.14 < 1.96$, this Z value and the sample mean are not in the critical region. The sample mean for Emerald does exceed the average for all Saskatchewan. But at the $\alpha = 0.05$ significance level, the results are not strong enough to reject the null hypothesis that the mean farm size for all Emerald farms is different from the Saskatchewan average.

Note that there is the possibility of Type II error here, that the true mean for Emerald may differ from the Saskatchewan mean, but that we cannot conclude that these two differ. The two means probably do differ, but the sample mean does not differ sufficiently from the hypothesized mean to make this conclusion. A larger sample might also have helped in distinguishing the farms of Emerald from Saskatachewan farms as a whole.

This test is much the same as the first test, except that this is a one tailed test. Let μ be the mean cultivated acreage for all farms in the rural

municipality of Emerald. The question is whether μ is equal to the mean cultivated acreage for all Saskatchewan farms, or whether it is larger than that for all Saskatchewan farms. This implies that a one tailed test should be used here. The hypotheses are:

$$H_0 : \mu = 781$$

$$H_1 : \mu > 781$$

H_1 states as a research hypothesis that the mean cultivated acreage in Emerald is larger than the mean cultivated acreage for all Saskatchewan, and the null hypothesis H_0 states that the mean cultivated acreage in Emerald is equal to the Saskatchewan average. The test statistic is $\bar{X}$ and since the sample size of $n = 47$ for the sample is reasonably large (greater than 30),

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

For a significance level of $\alpha = 0.04$ and with a one tailed test, the critical region is all Z values greater than $+1.75$.

Based on the sample data, $\bar{X} = 1068.8$ and $s = 1029.7$. Since σ, the true standard deviation for all possible samples, is unknown, but n is large, s can be used as an estimate of σ.

$$\begin{aligned} Z &= \frac{\bar{X} - \mu}{s/\sqrt{n}} \\ &= \frac{1068.8 - 781}{1029.7/\sqrt{47}} \\ &= \frac{287.8}{1029.7/6.856} \\ &= \frac{287.8}{150.1896} \\ &= 1.916 \end{aligned}$$

Since $Z = 1.916 > 1.75$, this Z value and the sample mean are in the critical region for the test. At the $\alpha = 0.04$ significance level, the results are strong enough to reject the null hypothesis that the mean cultivated acreage for all Emerald farms is no different from the Saskatchewan average. We can accept the research hypothesis that the mean cultivated acreage for farms in Emerald exceeds the Saskatchewan average.

9.2.7 Exact Levels of Significance

There are two approaches to choosing the level of significance when testing hypotheses, the traditional method and the method of determining the exact significance level. The method outlined so far in this chapter is the traditional method of picking the level of significance first, and determining the critical region from this. The level of significance may be 0.10, 0.05, 0.01 or some other level. In this traditional approach, statisticians generally suggest that the significance level be selected before examining the data. Once the level of significance has been selected, if the statistic falls in the critical region, the null hypothesis is rejected. If the statistic does not fall into this region, then the null hypothesis is not rejected. This decision is a clear cut one, either the null hypothesis is rejected, or it is not rejected, at the given level of significance.

An alternative method is to be guided by the exact significance level associated with the Z value from the data. That is, the test is conducted as before, but without selecting a level of significance or a critical region. The sampling distribution of the statistic is determined, and the Z value obtained from the sample data. Then the exact significance associated with this Z can be determined. The conclusion is then not necessarily a hard and fast decision to reject or to not reject H_0. Instead, it may be left to the reader to decide whether the evidence is strong enough to reject H_0.

The method of determining the exact level of signficance is to first obtain the Z value of the sample statistic from the sample data. Let this be Z_c. Then the area in the tail of the distribution, α_c that lies beyond Z_c is obtained from column B of the normal table. This area is referred to as the exact probability in the case of a one tailed test. In the case of a two tailed test, this area is doubled in order to determine the exact level of significance.

The exact level of significance is an area which represents the conditional probability of Z values, assuming the null hypothesis is true. The hypothesis test begins by assuming that the null hypothesis is true. If this is a one tailed test in the positive direction, then the area under the normal curve that lies beyond Z_c represents the probability of obtaining a Z of that size or larger. That is,

$$P\left(\frac{Z > Z_c}{H_0 \text{ is true}}\right) = \alpha_c.$$

If this probability is a very small number, then the null hypothesis is rejected, and the alternative hypothesis accepted. The reader can decide how low this probability should be. If $\alpha_c = 0.055$, then at the $\alpha = 0.05$ level of

significance, the null hypothesis would not be rejected, because the Z value is not quite in the region of rejection. But it is close to the region of rejection, and may provide fairly strong evidence that the null hypothesis is not true.

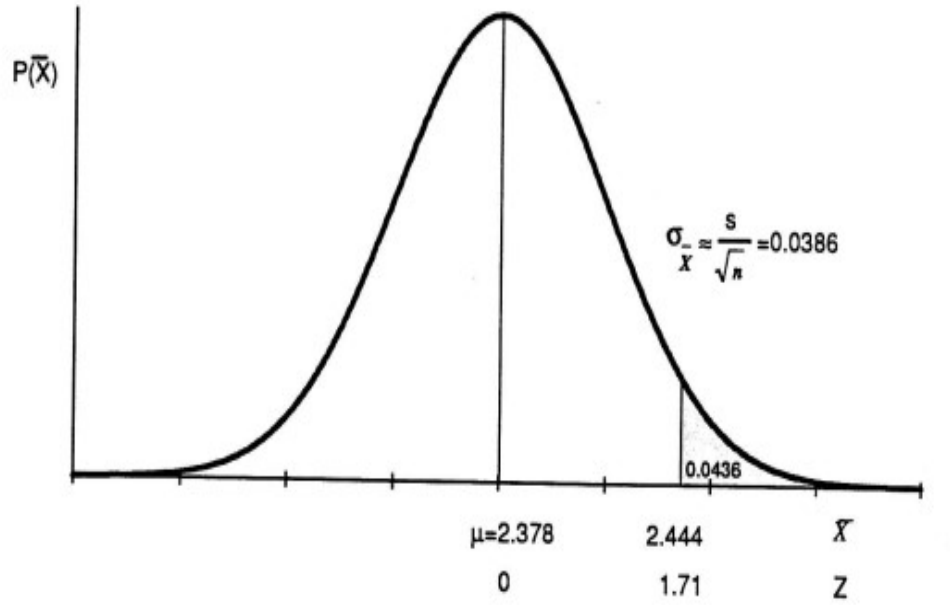

Figure 9.5: Exact Significance for Test of Numeracy Level

Example 9.2.6 Numeracy Levels

In Example 9.2.3, the null hypothesis was that the mean numeracy level for Saskatchewan is 2.378, equal to the Canadian mean level. The sample mean for $n = 374$ *Saskatchewan adults was* $\bar{X} = 2.444$*. The* Z *value which resulted from this was* $Z = 1.71$*. Examining column B of the normal table of Appendix H, there is 0.0436 of the area under the normal curve that lies to the right of* $Z = 1.71$*. Thus the exact significance level is 0.0436. The result might be reported as follows, and is illustrated in Figure 9.5.*

n	$\bar{X}$	Z	One tailed Significance
374	2.444	1.71	0.0436

The sample mean of $\bar{X} = 2.444$ *can be seen to correspond to* $Z = 1.71$ *in Figure 9.5. To the right of* $Z = 0.0436$ *the shaded area is 0.0436 in size. If the 0.05 level of significance is selected, this must include all the shaded*

area, and more. The null hypothesis could be rejected at any significance level just above 0.0436.

If the researcher gives all of this information, then the reader can see how much the sample mean differs from the hypothesized mean, what the Z value is, and what the exact significance associated with that Z is. If the traditional method had been used, all that would have been reported may have been the sample mean and the conclusion that the null hypothesis could be rejected at the 0.05 level of significance.

Example 9.2.7 Study Hours of Students

In Example 9.2.2 the null hypothesis was that the students study a mean of 20 hours per week. This was a one tailed test with the alternative hypothesis being that the study hours were less than 20 hours per week. The sample mean was $\bar{X} = 18.8$ using a random sample of $n = 494$ undergraduates. The Z value was -1.996, or rounded to 2 decimals $Z = 2.00$. The area under the normal curve that lies farther from center than -2.00 is 0.0227. This might be reported:

n	$\bar{X}$	Z	One tailed Significance
494	18.8	-2.00	0.0227

Again, more information is provided than if the researcher reported only that the hypothesis had been rejected at the 0.05 level of significance. Given the 0.0227 area to the left of $Z = -2.00$, it is clear that the null hypothesis could have been rejected at the 0.03, or the 0.025 level, or any other level of signficance down to just above 0.0227.

For a two tailed test, this method must be altered slightly, with the probabilities given above being doubled. That is, the areas given so far represent areas in one tail of the distribution, But if the test is a two direction test, then the significance should be reported as the combined area in the two tails that would lie outside the Z value. That is, if Z_c is the Z value determined from the sample, let $\alpha_c/2$ be the area farther from centre than this. Then the area to the left of $-Z_c$ is $\alpha_c/2$ and the area to the right of Z_c is $\alpha_c/2$. The sum of these areas in the tails of the distribution is α. This is illustrated in the following example.

Example 9.2.8 Mean Farm Size in the Rural Municipality of Emerald

The first hypothesis test in Example 9.2.5 examined whether the mean farm size in the rural municipality of Emerald could be considered different than the mean farm size in all of Saskatchewan. For that test, the null hypothesis was that the mean farm size in Emerald was 419 and the alternative hypothesis was that the mean farm size was not 419. This was a two tailed test, with no prior suspicion concerning the direction of the test. The sample statistics from the sample of $n = 47$ *farms are* $\bar{X} = 492.1$ *and* $s = 440.5$*. The* Z *value computed from these statistics and the null hypothesis is* $Z = 1.14$*. The exact significance associated with this test is the area under the normal curve which lies more than 1.14 standard deviations from the centre of the distribution. This is the area under the normal curve to the right of* $Z = 1.14$*, plus the area under the normal curve that lies to the left of* $Z = -1.14$*. From the normal table, there is 0.1271 of the area in the tail of the normal distribution that lies to the right of* $Z = 1.14$*. Since the normal curve is symmetric, the same area lies to the left of* $Z = -1.14$*. The total area in the two tails of the distribution is* $0.1271 + 0.1271 = 0.2542$*, and this is the exact significance associated with a two tailed test for this data.*

This value can be interpreted as a probability as follows. If the null hypothesis is true, the probability that the sample mean lies 1.14 standard deviations or more from the hypothesized mean is 0.2542. A null hypothesis is ordinarily rejected only if the probability of obtaining the sample mean is quite low, assuming that the null hypothesis is true. In this case, 0.2542 is not all that low, meaning that there is a considerable chance that a sample mean of 492.1 could have been obtained even when the true mean is 419. As a result, the null hypothesis would not ordinarily be rejected on the basis of this sample.

9.2.8 Critical Region in Units of X.

The region of rejection of H_0, or the critical region, for the test of significance, has been identified by using Z values. It is sometimes useful to define this critical region in units of the value of the variable being tested. This is briefly explained in the next paragraph and two examples are given in this section. Figure 9.6 presents the critical region in units of X diagramatically.

Suppose the level of significance is α for the one tailed hypothesis test

$$H_0 : \mu = M$$

$$H_1 : \mu > M$$

The region of rejection of H_0 is all Z values greater than Z_α, since the null hypothesis will be rejected only if $Z > Z_\alpha$ in the right tail of the sampling distribution of $\bar{X}$. This critical region can easily be stated in terms of the units of the variable X. This is based on the sampling distribution of $\bar{X}$:

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

The region of rejection begins at Z_α standard deviations to the right of centre. Since one standard deviation in the sampling distribution of $\bar{X}$ is $\sigma/\sqrt{n}$, then Z_α standard deviations to the right of centre is

$$Z_\alpha \frac{\sigma}{\sqrt{n}}$$

in units of the variable X. Thus the critical region begins at

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

Since μ is hypthesized to equal M, and σ is estimated by s when n is large, the critical region is all values of the sample mean $\bar{X}$ exceeding

$$M + Z_\alpha \frac{s}{\sqrt{n}}.$$

For this test, the conclusions of the hypothesis test are to reject H_0 if

$$\bar{X} > M + Z_\alpha \frac{s}{\sqrt{n}}$$

and do not reject H_0 if

$$\bar{X} < M + Z_\alpha \frac{s}{\sqrt{n}}.$$

If the test is a one tailed test that $\mu < M$, then the critical region would be all values of $\bar{X}$ to the left of

$$M - Z_\alpha \frac{s}{\sqrt{n}}.$$

If the test is a two tailed test, then the significance level α is split equally between the two tails of the distribution, with $\alpha/2$ of the area in each tail. If the critical Z value is $Z_{\alpha/2}$, then the region of rejection of H_0 would be all values of $\bar{X}$ less than

$$M - Z_{\alpha/2}\frac{s}{\sqrt{n}}$$

or values of $\bar{X}$ greater than

$$M + Z_{\alpha/2}\frac{s}{\sqrt{n}}.$$

In the following examples a one tailed test for numeracy level is given first, and a two tailed test for the weekly study hours of students follows.

Example 9.2.9 Numeracy Levels

Examples 9.2.3 and 9.2.6 presented one tailed tests to determine whether the mean Saskatchewan numeracy level exceeded the Canadian mean level. The hypotheses were

$$H_0 : \mu = 2.378$$

$$H_1 : \mu > 2.378$$

and the level of significance was $\alpha = 0.05$. *From the sample,* $n = 374$, $\bar{X} = 2.444$ *and* $s = 0.747$. *The region of rejection for a one tailed test at 0.05 significance is all* Z *values greater than 1.645. With this data, the standard deviation of the sampling distribution of* $\bar{X}$ *is*

$$\frac{s}{\sqrt{n}} = \frac{0.747}{\sqrt{374}} = \frac{0.747}{19.339} = 0.0386.$$

A distance of 1.645 standard deviations above centre is

$$1.645\frac{s}{\sqrt{n}} = 1.645 \times 0.0386 = 0.0635$$

units of numeracy to the right of centre. Since the null hypothesis is that $\mu = 2.378$, *the region of rejection begins at*

$$\mu + 1.645\frac{s}{\sqrt{n}} = 2.378 + 0.0635 = 2.442.$$

That is, sample mean of 2.442 or greater, on the numeracy scale, would lead to rejection of the null hypothesis and acceptance of the research hypothesis.

From this sample, $\bar{X} = 2.444 > 2.442$ *so that the null hypothesis can be rejected at the 0.05 level of significance. Note that this is the same conclusion as that obtained in the previous examples.*

Example 9.2.10 Study Hours of Students

Example 9.2.1 tested whether the study hours of students differed from hypothesized value of $M = 20$ *hours. The level of significance adopted was 0.10, and the alternative hypothesis was a two directional hypothesis. This meant that the region of rejection of* H_0 *was all* Z *values which lie in either the upper* $0.10/2 = 0.5$ *of the distribution, or the lowest 0.05 of the area under the normal curve. An area of 0.05 in the tail of the normal curve is associated with* $Z = 1.645$.

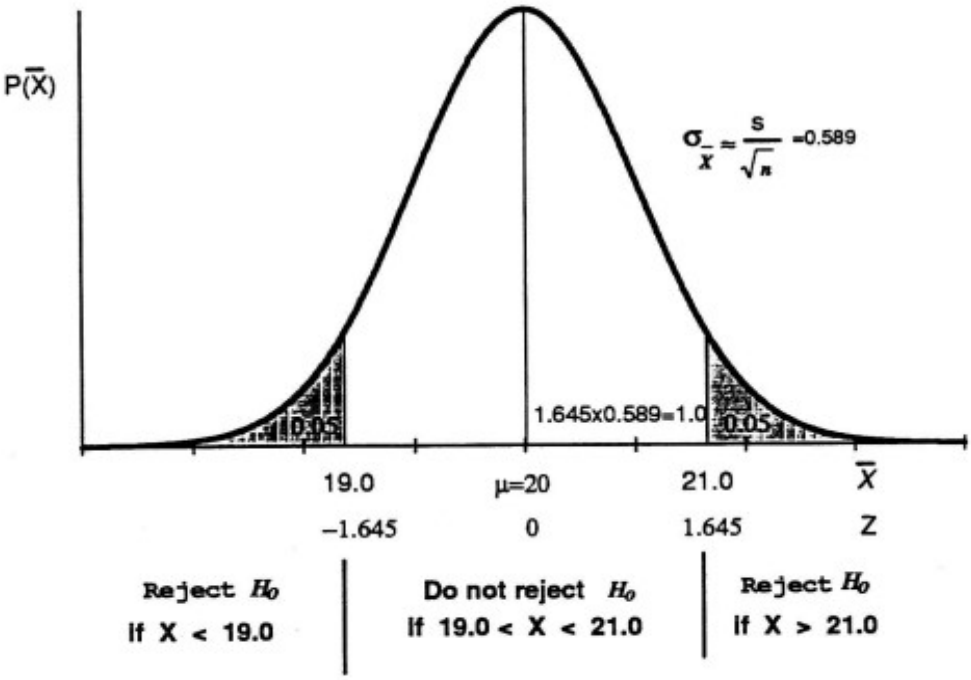

Figure 9.6: Rejection Region for Study Hours in Hours per Week

Figure 9.6 gives the regions of rejection and of nonrejection of H_0 *in terms of both* Z *and hours per week. The region of rejection is all* Z *values which*

are further than 1.645 from the centre of the distribution. This translates into hours of study per week of less than 19.0 or greater than 21.0.

For the sample, $n = 494$, $\bar{X} = 18.8$ *hours and* $s = 13.1$ *hours. Using this data, the critical values are:*

$$M - Z_{0.05}\frac{s}{\sqrt{n}} \text{ and } M + Z_{0.05}\frac{s}{\sqrt{n}}$$

$$20 - 1.645\frac{13.1}{\sqrt{494}} \text{ and } 20 + 1.645\frac{13.1}{\sqrt{494}}$$

$$20 - (1.645 \times 0.589) \text{ and } 20 + (1.645 \times 0.589)$$

$$20 - 0.97 \text{ and } 20 + 0.97$$

$$19.03 \text{ and } 20.97$$

The region of rejection of H_0 *is all sample means of less than 19.0 hours per week or greater than 21.0 hours per week. The sample mean was* $\bar{X} = 18.8$ *and this is in the region of rejection of the null hypothesis. At the 0.10 level of significance, the null hypothesis is rejected, and this test shows that students do not study 20 hours per week.*

9.2.9 Hypothesis Tests and Interval Estimates

In Chapter 8, it was claimed that interval estimation is very similar to hypothesis testing. This section will is devoted to showing that a two tailed hypothesis test at the α level of significance is identical with the $(1-\alpha)100\%$ interval estimate. For example, a two tailed test of hypothesis at the 0.05 level gives essentially the same conclusions as can be obtained with the 95% confidence level.

This can be seen using the method of the last section, and the diagram in Figure 9.7. A null hypothesis begins by assuming that the sample mean takes on a specific value, μ. If this is a two tailed test at the α level of significance, then the critical values for the test are $Z_{\alpha/2}$ standard deviations on each side of centre. Thus the region of rejection of H_0 is all sample means $\bar{X}$ greater

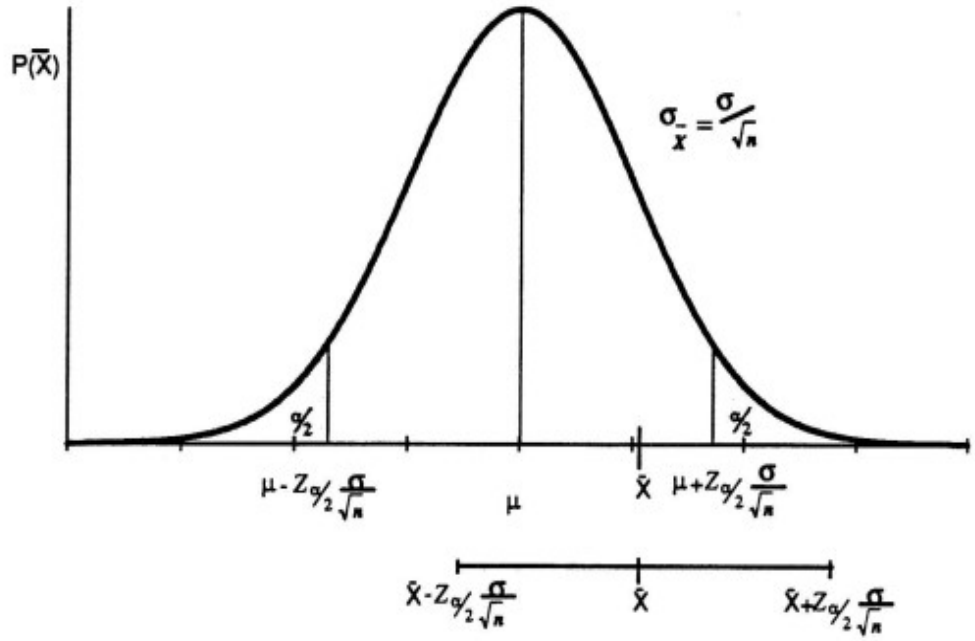

Figure 9.7: Hypothesis Testing and Interval Estimation

than $\mu + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ or less than $\mu - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. This is represented as the shaded portion of Figure 9.7. If $\bar{X}$ falls in the interval

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

then the null hypothesis is not rejected.

The confidence intervals for μ are intervals constructed around $\bar{X}$ plus or minus the appropriate number of standard deviations. The α level of significance is associated with an area of $\alpha/2$ in each tail of the distribution, leaving an area of $1-\alpha$ in the centre. In percentages this is $(1-\alpha)100\%$ of the area under the curve. As can be seen in Figure 9.7, the Z value for the confidence interval is the same as the Z value for the two tailed hypothesis test. That is, the confidence interval estimates are

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

It can be seen that the width of this interval is the same as the width of the region of nonrejection of H_0 in the middle of the distribution.

This means that for any sample mean $\bar{X}$ which lies in the centre of the distribution, in the region where H_0 is not rejected, has an interval which is wide enough to contain the hypothesized mean μ. On the other hand, if $\bar{X}$ lies in the region of rejection of the null hypothesis, then the interval constructed around $\bar{X}$ will not contain μ. In this latter case, the sample mean is so far from μ that even an interval of $Z_{\alpha/2}$ standard deviations on each side of $\bar{X}$ will not contain μ.

The above discussion shows that if $\bar{X}$ is distant from the hypothesized μ, in the region of rejection of H_0, the null hypothesis is rejected and the interval estimate does not contain μ. But if $\bar{X}$ is close enough to μ so that the null hypothesis is not rejected, then the interval estimate also contains μ. Thus interval estimation produces essentially the same result as the hypothesis test. This may become clearer in the following examples.

Example 9.2.11 Study Hours of Students

From Examples 9.2.1 and 9.2.10, the hypotheses for the mean study hours of students are

$$H_0 : \mu = 20$$

$$H_1 : \mu \neq 20$$

Example 9.2.10 showed that for the 0.10 level of significance the critical values for the region of rejection of H_0 are 19.0 to 21.0. That is, 1.645 standard deviations on each side of $\mu = 20$ is

$$\mu \pm Z_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$20 \pm 1.645\frac{13.1}{\sqrt{494}}$$

$$20 \pm 0.97$$

or rounded to the nearest tenth of an hour, this is

$$(19.0\ ,\ 21.0)$$

The 90% confidence interval for any sample mean which falls between 19.0 and 21.0 would contain μ. *The* $(1-\alpha)100\% = 90\%$ *interval estimate is*

$$\bar{X} \pm Z_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$\bar{X} \pm 1.645\frac{13.1}{\sqrt{494}}$$

$$\bar{X} \pm 0.97$$

or $\bar{X}$ *plus or minus 1.0 hour. Any* $\bar{X}$ *just under 21.0 hours, or just over 19.0 hours, would lead to an interval which contains* $\mu = 20$. *For example, if* $\bar{X} = 19.1$, *the interval would be* $(18.1, 20.1)$ *and this contains* $\mu = 20$. *But any sample mean that lies below 19.0 hours, or above 21.0 hours, would not contain the mean* $\mu = 20$. *The sample of 494 undergraduates gave a sample mean of* $\bar{X} = 18.8$ *hours. The 90% interval constructed around this sample mean in Example 8.3.1 was* $(17.8, 19.8)$ *and this does not contain* $\mu = 20$. *The confidence interval of Example 8.3.1 could have been used to reject the null hypothesis that* $\mu = 20$, *at the 0.10 level of signficance, since 20 hours was outside the limits of the 90% interval estimate.*

The similarity of interval estimates and hypothesis test does not extend to one tailed tests of significance. In a one tailed test, all of the α level of significance is placed in one tail of the distribution. In a two tailed test, the α level of significance is split equally between the two tails of the distribution. This is the same procedure that is used in the confidence interval estimates. But confidence intervals are always symmetrical about the sample statistic, so that there is nothing comparable to a one tailed test in the method of interval estimates.

9.2.10 Summary

This section has provided an extended discussion of the principles of hypothesis testing. For each part of the explanation, a test of a mean with a large sample size has been used as the basis for these explanations. Each of these principles can be more applied to other types of hypothesis test. While there are other modifications of hypothesis testing which can be carried out, this section has outlined most of the main principles used in test of hypotheses.

If you have difficulty following some of the tests in the following sections, it may be worthwhile to review some of the principles of testing discussed in this section. In subsequent sections it will be assumed that you have some

familiarity with concepts such as types of error, exact significance levels, and the difference between one and two tailed tests.

9.3 t Test for a Mean, Small n

In Section 8.4.1 of Chapter 8, interval estimates for the mean for a small random sample were discussed. These interval estimates used the t distribution to describe the sampling distribution of $\bar{X}$. The method of constructing the interval estimates using the t distribution was identical with the method used for a large sample size, except that the t value replaced the Z value. This could be done because when the sample size is small, $\bar{X}$ has a t distribution with mean μ, standard deviation $s/\sqrt{n}$ and $d = n - 1$ degrees of freedom. The assumption on which the t distribution is based is that the population from which the sample is drawn is a normally distributed population, and that the sample is random.

This same t distribution can be used to conduct tests of hypotheses concerning population means for small sample sizes. Again the method is the same as the methods just discussed for tests of the mean when the sample size is large. A short description of the assumptions and method when the sample size is small is contained in the following paragraphs. Several examples of the t test for a mean in the case of small random samples follow this discussion.

The t Test. There are many t tests which can be conducted, but the test described in this section is a t test for a single mean when the sample size is small. Suppose that a variable X measures a characteristic of a population for which the true population mean and standard deviation, μ and σ, respectively, are both unknown. In addition, suppose that the population is a normally distributed population. Then,

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

Further suppose that random samples of size n are drawn from this population. Then the sampling distribution of the sample mean is

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

where $d = n - 1$ and s is the standard deviation of the sample.

If the sample size n is small, then $\bar{X}$ is usually described on the basis of the t distribution. 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 $\bar{X}$.

The steps involved in conducting an hypothesis test using the t distribution are the same as adopted earlier for the test of a mean when the sample size is large. The null hypothesis, H_0, states that μ is equal to some specific value. The alternative hypothesis could take the form of a two directional inequality, or it could be a one directional inequality. Using the former, and hypothesizing a value M for the mean,

$$H_0 : \mu = M$$

$$H_1 : \mu \neq M$$

The test statistic is $\bar{X}$ and if the sample is a random sample from a normally distributed population with unknown standard deviation,

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

where $d = n - 1$. A level of significance, α is selected and the t table can be used to obtain the appropriate t value. Since this is a two tailed test, α is split equally between the two tails. Let $t_{\alpha/2}$ be the t value from the t table which defines the critical region for the test. The region of rejection of H_0 is all t values from the sample which are farther from the centre than $t_{\alpha/2}$.

The data from the sample is then used to determine the t value. This t value is a standardized value so that the t value can be constructed as

$$t = \frac{\text{Variable} - \text{Mean of Variable}}{\text{Standard Deviation of Variable}}$$

and this produces a t value with mean 0 and standard deviation 1. For the test of a mean, the variable is $\bar{X}$, and the mean of $\bar{X}$ is μ. Since μ has been hypothesized to equal M, the mean of $\bar{X}$ is M for purposes of conducting the test. For the t distribution, the standard deviation of $\bar{X}$ is $s/\sqrt{n}$. The standardized t value for this t test is

$$t = \frac{\bar{X} - M}{s/\sqrt{n}}$$

If this t value either exceeds $t_{\alpha/2}$ or is less than $-t_{\alpha/2}$, then the null hypothesis is rejected and the alternative hypothesis accepted. If the t value from the sample is between $t_{\alpha/2}$ and $-t_{\alpha/2}$, then the null hypothesis is not rejected.

Using the t table. The t table provides the critical t values for various levels of significance for both a two tailed and a one tailed test. At the top of the t table are various levels of confidence, and these are used when determining the confidence intervals of Chapter 8. For a two tailed hypothesis test, refer to the row labelled '2 Tailed.' This is the row which gives the two tailed signficance levels. For a two tailed test at the $\alpha = 0.05$ level of significance and 12 degrees of freedom, go to the column labelled 0.05 in the '2 Tailed' row, and go down that column until you reach 12 degrees of freedom. You will see a value of 2.179. This is the t value so that there is a total of 0.05 of the area in the two tails of the distribution which lie farther than 2.179 standard deviations from centre. There is $0.05/2 = 0.025$ of the area to the right of $t = 2.179$ and another 0.025 of the area to the left of $t = -2.179$.

In terms of the notation, if $\alpha = 0.05$, then $\alpha/2 = 0.05/2 = 0.025$ and $t_{0.025} = 2.179$ for 12 degrees of freedom. Note that the probability that for a t value with 12 degrees of freedom

$$P(t > 2.179) = 0.025$$

and

$$P(t < -2.179) = 0.025.$$

For a one tailed test, use the row in the t table labelled '1 Tailed', where the values are exactly one half the values in each '2 Tailed' value. In the 1 tailed test, all of the α level of significance is placed in one tail of the distribution, whereas this α is split equally between the two tails in the two tailed test. Thus the t values for the one tailed test correspond to the t values for the 2 tailed test at double the level of significance.

9.3.1 Examples of Hypotheses Tests, Small n

This section contains several examples of hypothesis test using the t distribution when the sample size is small.

Example 9.3.1 Numeracy, a Small Sample

The literacy survey examined in several previous examples gave a mean numeracy level of 2.378 for all Canadians. Recall that numeracy was measured on a 3 point scale with 1 being the lowest level of numeracy and 3 being the highest level. Suppose that a small random sample of 10 Saskatchewan

adults is selected, and their numeracy levels are 3, 3, 3, 3, 3, 3, 3, 2, 1 and 3. If you had only the data from this sample of size 10, could you conclude that the mean level of numeracy of all Saskatchewan adults exceeds the mean Canadian numeracy level? (0.05 level of significance). Discuss any possible errors in your conclusions.

Solution. *Let μ be the true mean numeracy level for Saskatchewan. The null and alternative hypotheses are*

$$H_0 : \mu = 2.378$$

$$H_1 : \mu > 2.378$$

In words, the null hypothesis is that the true mean numeracy level for Saskatchewan equals the Canadian mean level. Since the question asks whether the mean level for Saskatchewan **exceeds** *the mean level for Canada, a one tailed alternative hypothesis in the positive direction is used.*

The test statistic is $\bar{X}$ and if the sample is a random sample from a normally distributed population with unknown standard deviation,

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

where $d = n - 1$. The suitability of these assumptions is discussed later. The level of significance specified is $\alpha = 0.05$. The sample has size 10 so that there are $d = n - 1 = 10 - 1 = 9$ degrees of freedom in the sampling distribution of $\bar{X}$. From the t table, for a one tailed test at the 0.05 level of significance, and 9 degrees of freedom, the t value is 1.833. The region of rejection of H_0 is all t values of 1.833 or more. If the t value from the sample is less than 1.833, then there is insufficient evidence to reject the null hypothesis.

The next step is to obtain the mean, standard deviation, and t value from the sample. Using the formulae for the mean and standard deviation in Chapter 5, $\bar{X} = 2.700$ and the standard deviation $s = 0.675$. Thus the t value for this sample is

$$\begin{aligned} t &= \frac{\bar{X} - \mu}{s/\sqrt{n}} \\ &= \frac{2.700 - 2.378}{0.675/\sqrt{10}} \\ &= \frac{0.322}{0.213} \end{aligned}$$

$$= \quad 1.509 < 1.833$$

Since the t value from this sample of size 10 is not in the region of rejection of H_0, the null hypothesis cannot be rejected. At the 0.05 level of significance, there is insufficient evidence to conclude that the mean level of numeracy in Saskatchewan exceeds the mean level for Canada as a whole.

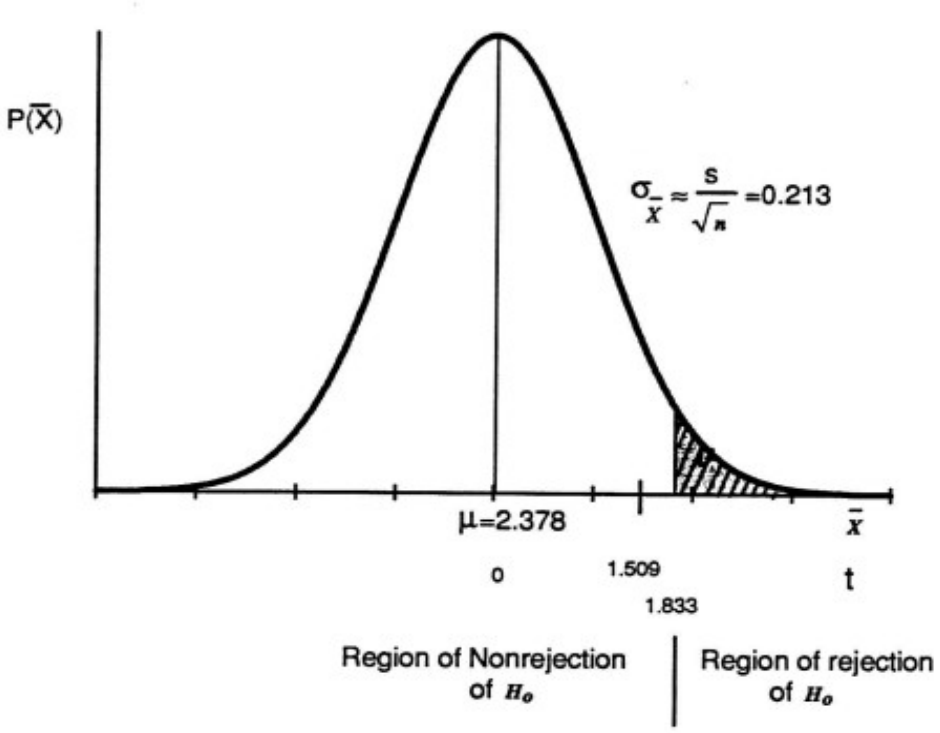

Figure 9.8: Test for Equality of Saskatchewan and Canadian Numeracy

Figure 9.8 shows diagramatically how the test could be conducted. The $\bar{X}$ value is given along the horizontal axis, and is hypothesized to equal $\mu = 2.378$. The standard deviation for this distribution is 0.213 units of numeracy (on the 3 point scale). The vertical axis represents the probability of occurrence of the different values of $\bar{X}$. The t values which correspond to each of the values of $\bar{X}$ are also given below the horizontal axis, with the regions of rejection and of nonrejection of H_0 shown. The shaded area

in the right tail of the t distribution has an area of $\alpha = 0.05$. *For the t distribution with 9 degrees of freedom, this area begins at* $t = 1.833$. *It can be seen in the diagram that the value of the t statistic is 1.509, and this is not in the critical region. As a result, the null hypothesis that the mean level of numeracy for Saskatchewan is no different that the Canadian mean is not rejected.*

Possible Errors in this Test.

One type of error which very likely does occur in this test is Type II error, the error of failing to reject the null hypothesis when it is not true. The conclusion of the test was that the null hypothesis cannot be rejected. This means that the hypothesis that the mean numeracy level for Saskatchewan is the same as for all Canada could not be rejected. It seems very unlikely that the mean numeracy level for all Saskatchewan would be exactly equal to the mean level for Canada. What the test shows is that this sample does not present evidence strong enough to reject this hypothesis. As a result, Type II error likely occurs here.

Another type of error is possible violation of the assumptions concerning the use of the t distribution. If the sample is truly a random sample, then the assumption of randomness is satisfied. But it seems very unlikely that numeracy as measured on this scale is normally distributed. Table 9.1 gave the distribution of numeracy ability for the sample of 374 Saskatchewan adults, and it can be seen that this distribution is far from normal. As a result, the assumption that the population from which the sample is drawn is normal, is unlikely to be satisfied. This could contribute additional error to the conclusion. However, since the conclusion was to not reject the null hypothesis, this is a cautious conclusion. Violation of the normality assumption is not likely to cause any particular problems here. If the decision had been the stronger one, to reject the null hypothesis, then the violation of normality might have made this an invalid conclusion.

The numeracy scale is no more than an ordinal level scale, but it is treated here as if it was an interval level scale. The determination of the mean and standard deviation, and the use of the t test assume that an interval level scale is being used. Since the results have been interpreted very cautiously, violation of this assumption is not likely to cause serious difficulties.

Additional Comments. *One solution to the problem of Type II error would be to select a larger sample size. Recall Examples 9.2.3 and 9.2.6*

where the sample size was $n = 374$ *and* $\bar{X} = 2.444$. *There* $\bar{X}$ *did not differ as much from the hypothesized mean of* $\mu = 2.378$ *as it does in this example. Yet the null hypothesis was rejected at the 0.05 level of significance.*

Another possible solution would be to increase the level of significance. Suppose that a significance level of $\alpha = 0.10$ *is adopted. At 9 degrees of freedom, the critical region for this test is all t values of 1.383 or more. Since* $t = 1.509 > 1.383$, *the null hypothesis could be rejected at the 0.10 level of significance. In doing this, there may be Type I error, rejecting the null hypothesis when it is not really incorrect. But the main error here is likely to occur because the population is not a normal population. That is, if the null hypothesis is rejected at the fairly large level of 0.10, this is not all that certain a result at the best of times, with a 0.10 level of Type I error.*

In conclusion, it seems best to stick with the weak conclusion that this sample does not tell us much concerning whether the mean level of numeracy differs between Saskatchewan and Canada as a whole. The sample certainly suggests a higher numeracy level for Saskatchewan than for Canada, and there is no evidence that the mean would be lower for Saskatchewan. But the sample mean for this small sample size does not really differ enough from the Canadian mean to conclude that Saskatchewan has a higher mean numeracy level.

Example 9.3.2 Bayley Scores for a Comparison Group of Teen Mothers

Appendix D, contains a data set from a pilot program conducted by Saskatchewan Social Services. This program involved adolescent mothers, with a program of support services being made available to a number of these mothers. Another group of adolescent mothers who did not receive the same support services was also studies, and constituted a comparison, or control group. Bayley scores were obtained for the children of all these mothers, both those who participated in the program, and the control group. The Bayley score is a measure of the level of infant development and has been standardized so that a mean score of 100 would be expected if a cross section of all infants were tested.

Use the t test to test whether the infants of mothers in the control group (Group 2 in Appendix D) have a mean Bayley score which differs from the expected mean of 100. Use the 0.10 level of significance. Also test whether the mean Bayley score for the infants of the mothers in the program exceeds 100.

Solution. *Let μ be the expected mean for the infants of mothers in the Group 2, the control group, assuming that the children in this group represent a broad cross section of all children. The null hypothesis is that $\mu = 100$. Since no direction has been specified for the alternative hypothesis, and since before the data is collected, the researcher does not even know the sample mean score for these children, this is a two tailed test. The hypotheses are*

$$H_0 : \mu = 100$$

$$H_1 : \mu \neq 100$$

In the comparison group, the sample of mothers totalled $n = 13$. Assuming that the distribution of Bayley scores is normal, the sample mean of a random sample of children would have a t distribution with $d = n-1 = 13-1 = 12$ degrees of freedom. At the 0.10 level of significance for a two tailed test, the critical t value is 1.782. A t value of less than -1.782 or greater than $+1.782$ would provide evidence that the children in the comparison group have a different mean level of Bayley score than do children generally. If the t value is not in the critical region, then the children can be regarded as representing a reasonable cross section of children.

From Appendix D, the statistics for the Bayley scores for the children in Group 2 can be obtained. For this sample of 13 children, the mean Bayley score is $\bar{X} = 98.61$ and the standard deviation is $s = 5.39$. The t value for this data is

$$\begin{aligned} t &= \frac{\bar{X} - \mu}{s/\sqrt{n}} \\ &= \frac{98.61 - 100}{5.39/\sqrt{13}} \\ &= \frac{1.39}{1.495} \\ &= 0.930 \end{aligned}$$

and this value is not in the critical region. The null hypothesis that the mean Bayley score for these children differs from the overall average score for all children cannot be rejected at the 0.10 level of significance. While these children have a slightly lower than average Bayley score, the difference is too small to argue that these children have a different mean than children generally do.

For the 28 children whose mothers participated in the project, the mean Bayley score was 109.2 and the standard deviation of the score was 13.16. As an exercise, you can show that the t value for this sample is sufficiently large that the null hypothesis that the mean is 100 can be rejected at the 0.0005 level of significance with a one tailed test.

These results provide evidence that the program had some effect on helping child development. The comparison group, which did not participate in the program, can be regarded as a cross section of ordinary children, with no higher or lower Bayley score than children in the population as a whole. The children whose mothers participated in the project appear to have achieved a somewhat greater level of development, based on the higher than normal Bayley score.

Example 9.3.3 Attitude Survey - Small n

A random sample of 7 Regina adults is asked the question "Is the government more on the side of business than labour in labour relations?" The responses are coded on a 4 point scale with 1 being strongly agree, 2 being agree somewhat, 3 being disagree somewhat and 4 being strongly disagree. The responses of the 7 adults are 1, 1, 3, 1, 2, 2, and 1. Someone claims that people in Regina are equally split between agree and disagree on this issue and the mean response for all Regina adults would be 2.5. Use the sample data to test whether the mean response differs from 2.5. Comment on the conclusion.

Solution.

Let the true mean opinion level of all Regina adults be μ. In doing this, the ordinal attitude scale is being treated as an interval level scale. The null and research hypotheses are

$$H_0 : \mu = 2.5$$

$$H_1 : \mu \neq 2.5$$

The null hypothesis states that the true mean level of opinion is in the middle of the scale. If people were equally split between agree and disagree, then the mean would be 2.5. Since no direction is indicated concerning the test, treat this as a two tailed test, and the alternative hypothesis is that μ is not equal to 2.5.

The test statistic is the sample mean, and if the sample is a random sample from a normally distributed population, then

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

where $d = n - 1$. *The level of significance is not stated in the question, but is selected as* $\alpha = 0.10$ *for this test. The sample has size 7 so that there are* $d = n - 1 = 7 - 1 = 6$ *degrees of freedom in the sampling distribution of* $\bar{X}$. *From the t table, for a two tailed test at the 0.10 level of significance, and 6 degrees of freedom, the t value is 1.943. The region of rejection of* H_0 *is all t values greater than 1.943 or less than -1.943. If the t value from the sample is betweem -1.943 and +1.943, then there is insufficient evidence to reject the null hypothesis.*

The next step is to obtain the mean, standard deviation, and t value from the sample. Using the formulae for the mean and standard deviation in Chapter 5, $\bar{X} = 1.571$ *and the standard deviation* $s = 0.787$. *The t value for this sample is*

$$\begin{aligned} t &= \frac{\bar{X} - \mu}{s/\sqrt{n}} \\ &= \frac{1.571 - 2.500}{0.787/\sqrt{7}} \\ &= \frac{-0.929}{0.297} \\ &= -3.123 < -1.943 \end{aligned}$$

and the conclusion is that the null hypothesis is rejected. The sample mean of 1.571 has a considerable distance from the hypothesized mean, being well into the region of rejection of the null hypothesis. At the 0.05 level of significance, the conclusion on the basis of this sample is that Regina residents are not split between agree and disagree on this issue.

Comments. *There is a possibility of Type I error in this conclusion. The null hypothesis has been rejected, but there is a chance of 0.05 that with this result, Regina adults are equally split between agree and disagree on this issue. This seems unlikely since the t value is fairly large, but this is always a possibility. The exact significance level cannot be determined from the t table, but with 6 degrees of freedom, a t value of 3.123 is very close to the t*

value of 3.143 in the 0.02 column, for a two tailed test. Thus the probability of obtaining a t value of which is more than 3.123 standard deviations from centre is approximately 0.02. Since this is a fairly small probability, the chance of Type I error seems fairly minimal.

The other types of error that could be associated with this conclusion are that the sample may not be random and the distribution of the attitude scale for the population may not be normal. In addition, the attitude scale is being treated as an interval level scale even though it is only ordinal. All of these should be remembered when interpreting the results of this test. While the test of significance in itself is fairly conclusive, any of these problems could make the results from the test misleading.

Example 9.3.4 Global Warming

Global warming and the greenhouse effect have become of serious concern in recent years. Over the years from 1950 to 1980, global temperatures appear to have fallen slightly. Over these years, the global temperature varied considerably from year to year, with the average yearly temperatures being distributed in a pattern that may not have been all that different from a normal distribution of temperatures. The mean for these years was 15.0 degrees Celsius. Table 9.5 contains average temperature data for the 1980s. This data is based on a diagram in **The Economist** *of July 9, 1988, page 80.*

Year	Mean Temperature in Degrees Celsius
1981	15.30
1982	15.36
1983	15.05
1984	15.34
1985	15.10
1986	15.07
1987	15.23
1988	15.40

Table 9.5: Mean Global Temperature in the 1980s

Based on the data in Table 9.5, conduct the following tests:

1. *If the data for only the years 1986, 1987 and 1988 was available, could you conclude that mean global temperature has risen above its level of 1950-80?*

2. *If you now use the temperature data for the 8 years in the 1980s for which data is given, would your conclusion change?*

3. *Write a short comment on the results and possible errors in the results.*

Solution. *This problems differs a little from the previous examples because the data is not a sample, but a complete set of values for the years shown. This problem uses the t distribution more as a model of expected temperature variation, than as a sampling distribution. Some of the modifications in the manner this is approached are discussed as the test is conducted, with other part discussed later.*

1. *Let μ be the mean temperature assuming that weather conditions did not change after 1980. If this is assumed, then for the years 1986-88, μ would be 15.0 degrees. This is the null hypothesis. The alternative hypothesis is that the mean temperature has risen and the mean temperature was greater in 1986-88 than it was from 1950-80. Let μ be the mean temperature consistent with this alternative hypothesis, that is, that $\mu > 15$. The hypotheses to be tested are*

$$H_0 : \mu = 15$$

$$H_1 : \mu > 15$$

The test statistic is the sample mean $\bar{X}$ and

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

where $d = n - 1$. Whether assumptions for use of this distribution are satisfied or not will be discussed in part (3). The level of significance is not stated in the question but the $\alpha = 0.05$ level will be used. The sample is a sample of 3 years so that $n = 3$ and there are $d = n - 1 = 3 - 1 = 2$ degrees of freedom in the sampling distribution of $\bar{X}$. From the t table, for a one tailed test at the 0.10 level of significance, and 2 degrees of freedom, the t value is 2.912. The region of rejection of H_0 is all t values greater than 2.912.

The next step is to obtain the mean, standard deviation, and t value from the sample. Using the formulas for the mean and standard deviation in Chapter 5, for the 3 years 1986-88, $\bar{X} = 15.233$ *and the standard deviation* $s = 0.165$*. The t value for this sample is*

$$\begin{aligned} t &= \frac{\bar{X} - \mu}{s/\sqrt{n}} \\ &= \frac{15.223 - 15.000}{0.165/\sqrt{3}} \\ &= \frac{0.223}{0.0953} \\ &= 2.341 < 2.912 \end{aligned}$$

Based on this set of three years, and using a 0.05 level of significance and a one tailed test, there is insufficient evidence to conclude that there has been an increase in the mean temperature. The null hypothesis that the conditions of 1950-80 have remained the same over the 1986-88 years is not rejected.

2. *If all the years 1981-88 are used as data points, then the sample mean* $\bar{X} = 15.231$*, the sample standard deviation* $s = 0.140$ *and the sample size is* $n = 8$*. The hypotheses stay the same although the null hypothesis now is that climate conditions did not change in 1981-88 from the conditions of 1950-80. The alternative hypothesis is that the climate conditions did change to produce an increase in the mean temperature.*

   The rest of the test is identical in nature, except that the t value changes. For a one tailed test with $\alpha = 0.05$ *and* $d = n - 1 = 8 - 1 = 7$ *degrees of freedom, the t value becomes 1.895. The region of rejection of* H_0 *is all t values greater than* $t_{0.05} = 1.895$*. From the sample,*

$$\begin{aligned} t &= \frac{\bar{X} - \mu}{s/\sqrt{n}} \\ &= \frac{15.231 - 15.000}{0.140/\sqrt{8}} \\ &= \frac{0.231}{0.0495} \\ &= 4.667 > 1.895. \end{aligned}$$

and the null hypothesis can be rejected. The t value from the sample is 4.667 standard deviations to the right of centre, considerably more than the 1.895 required in order to reject the null hypothesis. At the 0.05 level of significance, the data from the years 1981-88 provides fairly strong evidence for the alternative hypothesis that the mean temperature has risen relative to the 1950-80 years.

With 7 degrees of freedom, a t value of 4.667 is close to the t value of 4.783 in the 0.001 column of the t table. I used the SHAZAM computer program to determine the exact significance level associated with $t = 4.667$ *for 7 degrees of freedom, and this was 0.00115, between 0.005 and 0.001, as indicated in the t table. This small probability provides a strong indication that mean temperature has risen, assuming the correctness of the model on which this test is based.*

3. *There are a considerable number of assumptions built into this test. First, the types of error in hypothesis testing are both illustrated here. In (1), the null hypothesis that climate conditions have not changed could not be rejected, even though it appears that mean temperatures were considerably greater over the year 1986-88. The sample of only 3 years was too small a sample size to be able to reject the null hypothesis. It appears very likely that Type II error has been committed in (1). Exactly what level of error is not clear, but after looking at (2), it appears that* H_0 *is not correct, so that the artificially low sample of only 3 years in (1) produced Type II error.*

   In (2), there is a chance of Type I error occurring. That is, it may be that the climate did not really change, there just happened to be a string of 8 years of above average temperature, and these happened to be enough above average that the null hypothesis could be rejected. The chance of this is below 0.05, and may be as low as 0.00115, so it seems unlikely that the null hypothesis is correct. But there is always the small chance that it could be correct.

   With respect to other errors, these are likely fairly minimal. Temperature is a well constructed interval level scale, and meteorological observations are quite accurate, so there would appear to be little measurement error here. Where a problem might develop is in the suitability of the model. This is discussed in the following.

Suitability of the t distribution. *The t distribution requires random sampling from a normal population in order that the sample means have*

a t distribution. Both the assumptions of random sampling and a normal distribution are violated here. The sample is certainly not a random sample, but is a set of consecutive observations. This is not a sample, and it is not random. In addition, there may not be independence of successive observations. If climate conditions of one year carry over and influence the climate conditions of the next year, then if one of these years has higher than normal temperature, the next year may also. Thus even if weather conditions are unchanged, it may be no accident that all the temperatures are above average. This could be followed by a string of below normal temperature years. Further, this latter observation also makes the assumption of normality subject to question. The observations are not independent events, so a normal distribution of temperatures from year to year may not be a correct way of describing the population of temperatures.

Having recognized all these weaknesses of the test, it may still be worthwhile to use the model as a first, and rough, test for the null and alternative hypotheses. Imagine first that there is little carryover of temperatures from one year to the next, so that the temperatures of successive years are roughly independent of each other. Although the temperatures of 1950-80 are not given here, the distribution of yearly temperatures over this period may not have been all that different than a normal distribution. If the model is one of a normal distribution of yearly temperatures, the test provides an estimate of the chance that there could be a string of 8 years of temperatures this much above the mean. What the exact probability at the end of (2) provides is is an estimate of this probability. That is, assuming unchanging climate conditions after 1980, the probability of 8 observations of the sort observed from 1981-88 is 0.00115. Since this is a very low probability, the assumption of unchanged climate conditions does not seem appropriate. The evidence appears to be that climate conditions had changed in a manner that produced warmer temperatures after 1980.

Given the possible violation of assumptions, these probabilities may be incorrect, and meteorologists could provide a better model than the relatively simple t distribution used here. But this is a quick, although rough, way of indicating that there may have been a shift in climate conditions.