Contents

Chapter 9

Hypothesis Testing

9.4 Test of a Proportion

Suppose that a researcher is investigating a characteristic of a population in order to determine the proportion of the population with the characteristic. An interval estimate for the proportion, or a test of a claim concerning the proportion can be used. The method of constructing interval estimates for proportions was given in Chapter 8. In Section 8.5, the extension of the normal approximation to the binomial was used to show that the sample proportion $\hat{p}$ is normally distributed. This sampling distribution has mean p and standard deviation $\sqrt{pq/n}$, where p and q are the proportion of successes and failures, respectively, in the population. The interval estimates for p were

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

where Z is the normal value associated with the confidence level that is selected.

This same sampling distribution, along with the method of hypothesis testing presented in the previous sections, is used to conduct hypothesis tests for a population proportion. A short presentation of the method is provided here, and then several examples are presented.

The test begins by making a claim or a statement concerning the population as a whole. Let p be the true proportion of the members of the population having the characteristic being investigated. Let the hypothesized value of p be p_h. For example, suppose someone claims that the proportion of NDP supporters in Regina is one half. Then the characteristic is support for the NDP and the claim is that $p_h = 0.5$. This claim could then be tested using data from a random sample of Regina adults. As before, the null hypothesis is an equality, with the proportion hypothesized to be equal to the value claimed. The alternative hypothesis will be an inequality, either a two directional inequality, or a one directional inequality in either the positive or negative direction. If the researcher has no idea whether the null hypothesis is true or not, and no suspicion concerning direction, a two directional alternative hypothesis is used, and the hypotheses are

$$H_0 : p = p_h$$

$$H_1 : p \neq p_h$$

If the researcher has some suspicion that the claim overestimates the true proportion, then the alternative hypothesis would be

$$H_1 : p < p_h.$$

If the suspicion was that the claim understated the true proportion, then

$$H_1 : p > p_h$$

would be the alternative hypothesis.

Suppose that a random sample of n members of the the population is selected. Any member of the sample which has this characteristic is defined as constituting a success. If there are X successes in the sample, the proportion of successes is $\hat{p} = X/n$. If n is reasonably large,

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

where $\sigma_{\hat{p}}$ is the standard deviation of the sampling distribution of $\hat{p}$ and is equal to $\sqrt{pq/n}$. This means that

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

This result is obtained from the normal approximation to the binomial, and holds as long as n is large enough to satisfy the condition

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

A level of significance is selected and the Z value is determined from the normal table in Appendix H. If the significance level is α and a two directional alternative hypothesis is being tested, then the critical region is all Z values greater than $Z_{\alpha/2}$ or less than $-Z_{\alpha/2}$. If the sample proportion has a Z value which falls in the critical region, the null hypothesis is rejected, and the alternative hypothesis accepted. If the sample proportion has a Z between the critical values, then the null hypothesis is not rejected.

The final stage of the test is to obtain the Z value from the sample. For any variable, Z is the variable minus its mean, divided by its standard deviation. For the sampling distribution of the sample proportion, the variable is $\hat{p}$. The mean and standard deviation are p and $\sigma_{\hat{p}} = \sqrt{pq/n}$, respectively. Thus the Z value is

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

If this Z value is in the critical region for the test, then the null hypothesis is rjected and the research hypothesis accepted. If Z is not in the critical region, then there is insufficient evidence to reject the null hypothesis.

Based on this description, it can be seen that an hypothesis test for a proportion uses the same approach as does the hypothesis test for a mean with large sample size. The only difference is that $\hat{p}$ replaces $\bar{X}$, p replaces μ, and $\sqrt{pq/n}$ replaces $\sigma/\sqrt{n}$. In the test for a mean, the true population standard deviation σ was unknown, so that s was used as an estimate of σ. In the test of a proportion, p and q are unknown in $\sqrt{pq/n}$. But since a particular value has been hypothesized for p in H_0, this value is used in estimating this standard deviation. That is, the hypothesis test is conducted assuming that the null hypothesis is true. If the null hypothesis is $H_0 : p = p_h$, then q is hypothesized to equal $q_h = 1 - p_h$. These values can be used in the standard deviation so that the Z value for the sample proportion is

$$Z = \frac{\hat{p} - p}{\sqrt{p_h q_h/n}}.$$

Several examples of hypothesis tests for proportions are provided here.

Example 9.4.1 Testing for a Level of Political Support

In the 1988 federal election, the NDP obtained 20% of the popular vote across Canada. The June 1990 Gallup poll reported that 23% of the decided voters said they would vote NDP if an election were to be held that month. The sample size was 1,011. At the $\alpha = 0.01$ *level of significance, could you* *conclude that support for the NDP increased between the 1988 election and June, 1990?*

In the June, 1990 Gallup poll, 32% of those polled were undecided. If these undecided respondents are eliminated, so that the sample size includes only the decided respondents, does this change the conclusion?

Solution. For this problem, let p be the true proportion of Canadians who support the NDP in June of 1990. If there had been no change in this proportion since November 1988, then p would equal 0.20, the proportion of support the NDP had in 1988. Using the notation of the earlier description, $p_h = 0.20$. The null and alternative hypotheses are

$$H_0 : p = 0.20$$

$$H_1 : p > 0.20$$

The alternative hypothesis is a one tailed test because the question asks whether support for the NDP has **increased** or not. The test statistic is $\hat{p}$ and if the sample size is large,

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

For this sample, $n = 1,011$, $\hat{p} = 0.20$ and $\hat{q} = 1 - 0.20 = 0.80$. Using these values to determine whether n is large gives

$$\frac{5}{min(p, q)} = \frac{5}{0.20} = 25$$

The sample size of $n = 1,011$ very considerably exceeds the minimum required value of 25 so $\hat{p}$ should be very close to normally distributed, as noted above.

Using the significance level of $\alpha = 0.01$ gives a critical region of all Z values of 2.33 or more. For the sample,

$$\begin{aligned} Z &= \frac{\hat{p} - p}{\sqrt{pq/n}} \\ &= \frac{0.23 - 0.20}{\sqrt{(0.20 \times 0.80)/1011}} \\ &= \frac{0.03}{0.01258} \\ &= 2.38 > 2.33 \end{aligned}$$

As a result, at the 0.01 level of significance, the Z value from the sample falls within the region of rejection of the null hypothesis. This sample provides sufficient evidence to show that support for the NDP increased from the 1988 federal election through June, 1990. Since $\alpha = 0.01$, this conclusion is made with a small chance of Type I error, and this provides fairly strong evidence that support for the NDP increased over this time period.

Figure 9.1: Test of Level of Support for the NDP

Figure 9.1 shows diagrammatically how the test is conducted. The sampling distribution of $\hat{p}$ is shown as a normal curve, centred at the hypothesized proportion of $p = 0.20$. The standard deviation of this sampling distribution of $\hat{p}$ is $\sqrt{pq/n}$, and for this sampling distribution equals 0.013 The sample proportion of 0.23 is shown near the right end of the diagram, and it has the Z value of 2.38 associated with it. The critical region for the test is the set of Z values of 2.33 or more. This is the set Z values associated with the shaded area of 0.01 at the right end of the diagram. It can be seen *that* $Z = 2.38 > 2.33$ *is in the critical region, so that the null hypothesis can be rejected at the 0.01 level of significance.*

If the undecided are excluded, all that changes is the sample size. The effective sample size becomes $n = 1,011-(0.32\times 1,011) = 1,011-324 = 687$. *This sample size is much greater than the minimum sample size of* $n = 25$ *required to ensure that* $\hat{p}$ *is normally distributed. The hypothesis test is exactly the same as the first test, except that the sample size has been reduced to 687. For this sample size,*

$$\begin{aligned} Z &= \frac{0.23 - 0.20}{\sqrt{(0.20 \times 0.80)/687}} \\ &= \frac{0.03}{0.01526} \\ &= 1.97 \end{aligned}$$

This Z *value is less than 2.33 and is not in the region of rejection so that the null hypothesis is not rejected. Once the undecided are taken out, the proportion of NDP supporters is not enough to conclude that the support for the NDP has increased. This conclusion is made at the 0.01 level of significance.*

Additional Comments. *Note that if the* $\alpha = 0.05$ *level of significance had been used, then the region of rejection of* H_0 *would be all values of* $Z > 1.645$. *Since the value of* Z *associated with the sample is 1.97, this exceeds 1.645 and, at the 0.05 level of significance, this provides evidence that the level of support for the NDP increased.*

Finally, note that all of these conclusions are rather uncertain, because of the large proportion of those polled who said they were undecided. If an election had actually held in June, 1990, many of the undecided would have voted for one of the parties. In this case the NDP might, or might not, have increased their level of support.

Example 9.4.2 Representativeness of Regina Labour Force Survey

An interval estimate was used to examine the representativeness of the Regina Labour Force Survey in Example 8.5.3. In that example, the interval estimate of the proportion of Regina adults with less than a grade 9 education showed that the Survey tended to underrepresent these people. This conclusion was based on a comparison of the proportion of adults with less than a grade 9 education in each of the Survey and the Census.

The data from the 1986 Census of Canada showed that of 132,825 adults aged 15 or over in Regina, 15,240 have completed less than grade 9 education. In the Social Studies 203 Regina Labour Force Survey, of the 937 adults aged 15 and over. 74 had completed less than grade 9 education.

Test whether the Survey underrepresents Regina adults having less than a grade 9 education. Use $\alpha = 0.02$ *significance.*

Solution. *Suppose the Survey were to be exactly representative of all adults with respect to the characteristic of having less than a grade 9 education. Using the Census, there are 15,240 out of 132,825 adults who had less than a grade 9 education. This is a proportion* $15,240/132,825 = 0.115$ *of the adult population of Regina. The null and alternative hypotheses can be stated in words as:*

$$H_0 : \text{The sample is representative of the population.}$$

$$H_1 : \text{The sample underrepresents those with less than grade 9.}$$

Let p *be the proportion of respondents with less than grade 9 education which would be expected in the Survey were to be exactly representative of the Regina population in terms of education. The expected proportion of respondents with less than a grade 9 education, if the Survey is exactly representative, is* $p_h = 0.115$. *The null and alternative hypotheses then become*

$$H_0 : p = 0.115$$

$$H_1 : p < 0.115.$$

The alternative hypothesis states that the Survey underrepresents those with less than a grade 9 education. If this happens, then $p < 0.115$.

The test statistic is the sample proportion of those with less than a grade 9 education, $\hat{p}$. *For this sample,*

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

Since $n = 937$, *and*

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

the sample size is large enough to ensure that

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

With a significance level of $\alpha = 0.02$ *and a one tailed test in the negative direction, the region of rejection of the null hypothesis is all* Z *values in the left 0.02 of the normal distribution. That is,* $Z_{0.02} = -2.05$ *is the critical value for the test. If the sample mean lies more than 2.05 standard deviations below centre, then the null hypothesis is rejected and the alternative hypothesis accepted. If* $Z > Z_{0.02} = -2.05$*, then the null hypothesis is not rejected.*

From the sample, $\hat{p} = 0.0709$ *so that* $\hat{q} = 1 - 0.0709 = 0.9291$*. These are the values which are used in the estimate of* $\sigma_{\hat{p}}$*. The* Z *value for the sample proportion is*

$$\begin{aligned} Z &= \frac{\hat{p} - p}{\sqrt{pq/n}} \\ &= \frac{0.0709 - 0.115}{\sqrt{0.0709 \times 0.9291/937}} \\ &= \frac{-.0441}{\sqrt{0.000070302}} \\ &= \frac{-0.0441}{0.008385} \\ &= -5.26 < -2.05 \end{aligned}$$

This Z *value of -5.26 is considerably less than -2.05, so that the sample mean is well within the region of rejection. The null hypothesis that the Survey represents Regina adults having less than a grade 9 education can be quite strongly rejected. There is strong evidence that the Survey underrepresents Regina adults having less than a grade 9 education.*

Additional Comments. *There is a possibility of Type I error here. That is, the method the Survey uses to select people of different educational levels may be quite adequate. This just happens to be one of those random samples which produces a very small sample of those with less than grade 9 education. But the 0.02 level of significance means that the chance of this being the case is less than 0.02.*

The exact significance level could be interpreted as follows. Assume that the sampling method generally provides a representative sample, or that the *sample is random. If this is the case, then the probability of selecting a $\hat{p}$ of 0.0790 from a population where the proportion is 0.115 is the probability that $Z < -5.26$. This probability is less than 0.000000287, the last entry in the normal table. This is an extremely small probability, and this low probability makes it very unlikely that the assumption of a random sample could be true. This probability is so low that the null hypothesis is very strongly rejected.*

Some of the possible reasons for underrepresenting this group of respondents were given in Example 8.5.3 and are not repeated here.

Example 9.4.3 Test for a Majority

The Regina **Leader-Post** *of September 8, 1992 contained a leading article entitled "Canadians back unity: poll." The article noted*

> *A majority of Canadians outside Quebec back the new constitutional deal but within the province supporters and opponents of the package are in a virtual dead heat, a poll published today suggests.*
>
> *Canadians support the deal by a margin of two to one, the poll indicates.*
>
> *A majority of Canadians surveyed – 51 per cent nationally – say they would vote to approve the reform package while 25 per cent said they would reject it if the Oct. 26 referendum were held today, the poll found.*
>
> *Another 24 per cent had no opinion, said the survey by Environics Research Group of Toronto. .....*
>
> *The poll suggests the deal faces its strongest opposition in Quebec, where 43 per cent of respondents say they would vote in favor and 39 per cent vote against. Another 10 per cent had no opinion.*

The sample surveyed 1,519 adults across Canada over the August 28 to September 1 period. These results can be used to test whether or not a majority of Canadian voters would support the new constitutional deal in the October 26 vote.

For Canada as a whole, and of those who have decided, test whether the new constitutional deal would be approved. (0.01 significance). Assuming that the sample size in Quebec is 380, test whether the support among the *decided voters would be sufficient for the deal to be approved in Quebec. (0.10 significance).*

Solution. *For Canada as a whole, 76% of those polled are decided, so that* $51/76 = 0.671$ *is the proportion of those polled who say they would vote for the deal. The effective sample size, taking out the undecided is* $0.76 \times 1,519 = 1,154$. *For Quebec, 82% have decided, and the degree of support for the deal is* $43/82 = 0.524$*, and the effective sample size is* $0.82 \times 380 = 312$. *This is the data which will be used to conduct the tests.*

Let p *be the true proportion of Canadian voters who support the deal. If the deal is to be approved, just over 50% of the voters would need to vote for the deal. The aim of the test is to determine whether this hypothesis that* $p > 0.5$ *can be proved on the basis of the data. This inequality cannot serve as the null hypothesis, since the null hypothesis must be an equality. The deal would not pass if* p *takes on any value up to and including exactly 50% of voters. In order to determine whether* $p > 0.5$ *is correct or not, assume that* $p = 0.5$. *Then if this hypothesis can be rejected, and* $p > 0.5$ *accepted, there is considerable evidence that the deal would pass. The null and alternative hypotheses are thus*

$$H_0 : p = 0.5$$

$$H_1 : p > 0.5$$

The test statistic is $\hat{p}$*, and the sample size is large enough so that*

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

At the $\alpha = 0.01$ *level of significance, and with a one tailed test in the positive direction, the critical value is* $Z_{0.01} = 2.33$. *The null hypothesis is rejected and the alternative hypothesis accepted if* $Z > 2.33$.

From the sample, $\hat{p} = 0.671$, $n = 1,154$ *and*

$$\begin{aligned} Z &= \frac{\hat{p} - p}{\sqrt{pq/n}} \\ &= \frac{0.671 - 0.500}{\sqrt{(0.5 \times 0.5)/1,154}} \\ &= \frac{0.171}{\sqrt{0.000216638}} \\ &= \frac{0.171}{0.014719} \end{aligned}$$

$$= \quad 11.618 > 2.33$$

and the null hypothesis can be very strongly rejected. The sample proportion of 0.671 is over 11 standard deviations above the hypothesized mean of 0.5, so that there is an extremely small probability of Type I error. Among the decided voters, there is very strong evidence that the deal would pass.

For Quebec, let p *be the true proportion of all Quebec voters who support the deal. The hypotheses are the same as earlier*

$$H_0 : p = 0.5$$

$$H_1 : p > 0.5$$

as are the test statistic and sampling distribution of $\hat{p}$*. The significance level is 0.10 so the critical value of* Z *is 1.28 for a one tailed test in the positive direction.*

From the sample, for Quebec $\hat{p} = 0.524$, $n = 312$ *and*

$$\begin{aligned} Z &= \frac{\hat{p} - p}{\sqrt{pq/n}} \\ &= \frac{0.524 - 0.500}{\sqrt{(0.5 \times 0.5)/312}} \\ &= \frac{0.024}{\sqrt{0.000801282}} \\ &= \frac{0.024}{0.02831} \\ &= 0.848 < 1.28 \end{aligned}$$

and for Quebec, the null hypothesis cannot be rejected at the 0.10 level of significance. Since there is insufficient evidence to reject H_0*, the alternative hypothesis* $p > 0.5$ *cannot be accepted. There is undoubtedly Type II error here, because* p *will not be exactly 0.5 after the vote is taken. What the conclusion of this test implies is that the poll results are inconclusive with respect to Quebec. The evidence very weakly supports passage of the deal in Quebec, but with a sample of only 312 decided voters, 52.4% in favour is just too small a margin to predict passage of the deal.*

Additional Comments. *The results of the tests generally support the statements made in the newspaper. The hypothesis test provides very strong* *evidence that across Canada as a whole, the deal would pass. The tests also show that the result could be very close in Quebec.*

There are likely many other nonsampling errors associated with the results here. First, there are so many undecided across Canada, that the results are unpredictable. If all the undecided moved to the opposed category, then the poll indicates a 51% in favor and 49% against split. Even with a sample of over 1,500 this would be too small a margin to predict victory. Second, it is possible that some voters will change their mind before the October 26 vote. Third, these results come from an hypothetical question, not from an actual vote. People may say something different from what they actually do. In addition, some people will not vote, and exactly which group votes or does not vote could also affect the result. The sample may not be a random sample of Canadian voters and this could invalidate the results.

In summary, the newspaper report gives a good idea of the way Canadians were thinking at the end of August. By October 26, the results could be quite different from what was reported.

9.5 Test for a Difference in Proportions

The discussion of interval estimates and hypothesis tests has involved inferences concerning a single population. This section shifts the discussion somewhat by examining two populations or two groups, and attempts to make some inferences concerning the similarity or difference between the two groups. In Chapter 8, some of the examples of interval estimates provided separate interval estimates for two different groups, and these were used to attempt to make inferences concerning the differences between the groups. In Example 8.5.2 the question of whether there was a difference in the percentage of mothers and fathers using various aggressive disciplinary actions was examined. Interval estimates for each proportion allowed some conclusions about this, but a more complete test of this can be conducted with the methods in this section. This is done in Example 9.5.2.

This section examines a test for the difference between the proportion of successes in two populations. Using random samples from each of two different populations or groups, this test allows the researcher to decide whether the two populations have the same proportions of successes or whether the two populations differ in the proportion of successes. Later in the chapter, in Section 9.6, a test for the difference between the means of the two populations will be discussed.

In carrying out the tests in this section, extra sources of variation in the statistics are introduced, and this increases the complexity of the formulas. That is, if random samples are taken from each of two different populations, then there is variability in the statistics obtained from each of the random samples. In addition, the **difference** between the statistics in the two random samples constitutes an extra source of variation. While each of the tests for the difference between the populations is constructed along essentially the same lines as the tests for a single population, there are more symbols and more complex formulas involved.

Characteristic	Population 1	Population 2
True Proportion	p_1	p_2
Sample Size	n_1	n_2
No. of Successes	X_1	X_2
Proportion of Successes	$\hat{p}_1$	$\hat{p}_2$
Proportion of Failures	$\hat{q}_1$	$\hat{q}_2$

Table 9.6: Notation for Test of a Difference in Two Proportions

In order to conduct the test for a difference between the proportions of two populations having a particular characteristic, some extra notation is necessary. This is given in Table 9.6 and described here. Begin with a particular characteristic that is being investigated. This might be the characteristic the respondent supports a particular political party. Also begin with two populations or two groups. For simplicity, one of the populations is called population 1 and the other population is called population 2. It does not matter which population is referred to as 1 or 2, but once each has been labelled, remain consistent throughout the test. For example, the test might be a test to determine whether support for a particular political party differs in two different provinces of Canada. One of the provinces would be called population 1 and the other would be population 2. Random samples from each province would be used to determine whether the proportion of supporters of the party differ between the provinces.

Let p_1 be the proportion of the members of population 1 which have the characteristic being investigated, and p_2 the proportion of population 2 which has this same characteristic. The true values of these proportions are unknown, and the true difference between them is also unknown. The aim of the samples and tests is to make some inferences concerning these true proportions and the differences between these true proportions.

In order to make inferences concerning the differences in proportions, suppose that two random samples are selected. From population 1 a random sample of size n_1 is selected, and from population 2 a random sample of size n_2 is selected. These samples must be be separate or **independent** random samples. That is, the process of selecting each sample should be quite separate and independent of the process of selecting the other sample. Some comments concerning how this condition may be satisfied, and when it could be violated are contained later in this section.

Suppose that X_1 of the sample respondents in the sample of population 1 have the characteristic being investigated. Since this test is an extension of the binomial distribution, the terminology of the binomial can be used here. Any member of the sample which has the characteristic can be referred to as a success, and any member not having the characteristic being examined is a failure. Thus sample 1 has X_1 successes. Similarly, with sample 2, let there be X_2 successes. The proportions of each sample which have this characteristic can now be determined. For the sample from population 1, the proportion of the sample having the characteristic is

$$\hat{p}_1 = \frac{X_1}{n_1}$$

and for the sample from population 2, the proportion of successes is

$$\hat{p}_2 = \frac{X_2}{n_2}.$$

Any case in the sample which does not have the characteristic being investigated is called a failure. For each of the two groups, the proportion of failures plus the proportion of successes equals 1. If $\hat{q}_1$ and $\hat{q}_2$ represent the proportion of failures in populations 1 and 2, respectively, then

$$\hat{q}_1 = 1 - \hat{p}_1 \quad \text{and} \quad \hat{q}_2 = 1 - \hat{p}_2$$

The two sample proportions $\hat{p}_1$ and $\hat{p}_2$ will differ from each other, with each pair of random samples yielding a different value for the difference between the two sample proportions. By an extension of the normal approximation to the binomial, it can be shown that this difference between the sample proportions has the following normal distribution, when both sample sizes are reasonably large.

$$\hat{p}_1 - \hat{p}_2 \text{ is Nor } \left(p_1 - p_2, \sqrt{(p_1 q_1)/n_1 + (p_2 q_2)/n_2} \right)$$

That is, the variable here is the **difference between the two sample proportions**. This difference has a sampling distribution, because many different random samples are taken from each of the two populations. But this difference of sample proportions $\hat{p}_1 - \hat{p}_2$ has a mean which is the difference between the true values of the population proportions $p_1 - p_2$. The standard deviation of this sampling distribution could be given the symbol $\sigma_{\hat{p}_1 - \hat{p}_2}$ and this standard deviation can be shown to be

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{(p_1 q_1)/n_1 + (p_2 q_2)/n_2}.$$

The sampling distribution of $\hat{p}_1 - \hat{p}_2$ is also a normal distribution as long as both n_1 and n_2 are reasonably large, by the normal approximation to the binomial. The rule for how large these should be is the same as that adopted earlier, that is each n should be large enough so that

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

Conducting the Test. The sampling distribution for $\hat{p}_1 - \hat{p}_2$ can be used to construct an hypothesis test in the conventional manner. The test attempts to determine whether there is any difference between the two proportions or not. As in other test, the null hypothesis is an equality, hypothesizing that there is no difference between the true proportion of successes in the two populations. This can be stated in either the form

$$H_0 : p_1 = p_2$$

or

$$H_0 : p_1 - p_2 = 0$$

Each of these states that there is no difference between the two populations in the proportion of each population which has the characteristic being investigated. The latter form states the hypothesis in the form that will be tested when the difference between the sample proportions is used.

The alternative hypothesis can be any one of three forms, a two directional inequality, or either of the one directional inequalities. In the case of a two direction inequality, the alternative hypothesis is

$$H_1 : p_1 \neq p_2 \quad \text{or} \quad p_1 - p_2 \neq 0.$$

One directional inequalities are used in some of the examples which follow.

The test statistic is the difference between the sample proportions $\hat{p}_1 - \hat{p}_2$, and if both sample sizes are large, this is

$$\hat{p}_1 - \hat{p}_2 \ \text{ is Nor } \left(p_1 - p_2, \sqrt{(p_1q_1)/n_1 + (p_2q_2)/n_2}\right)$$

The level of significance α is selected, and since the test statistic $\hat{p}_1 - \hat{p}_2$ is normally distributed, the normal table can be used to obtain the Z value. Based on this, a region of rejection of the null hypothesis can be obtained. Then the Z value associated with the test statistic is

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{(p_1q_1)/n_1 + (p_2q_2)/n_2}}.$$

That is, Z is the statistic minus its mean, and divided by its standard deviation. If this Z is in the region of rejection of H_0 then the hypothesis of equal proportions is rejected. If Z is not in the critical region, then there is insufficient evidence, at the α level of signficance, to conclude that the true proportions differ.

Diagrammatic Presentation of the Test. Figure 9.2 gives a diagrammatic presentation of the test. The sampling distribution is shown as a normal curve, with the difference of the two sample proportions, $\hat{p}_1 - \hat{p}_2$, on the horizontal axis. The null hypothesis is that the difference in the true values of the proportions $p_1 - p_2$ equals 0. This also corresponds to a Z value of 0 at the centre of the normal curve. The vertical axis gives the probability for each value of the difference of the sample proportions, so it is labelled $P(\hat{p}_1 - \hat{p}_2)$.

If the level of significance is α, then with a two tailed test, the critical region is all Z values in the upper $\alpha/2$ or the lower $\alpha/2$ of the area under the normal curve. The critical Z values and the shaded areas are shown in the tails of the distribution. If $Z_{\alpha/2} < Z < Z_{\alpha/2}$, then the null hypothesis cannot be rejected. But if the Z statistic is less than $-Z_{\alpha/2}$ or greater than $Z_{\alpha/2}$, the null hypothesis is rejected at the α level of significance.

Figure 9.2: Two Tailed Test for a Difference of Two Proportions

Estimating the Z Value. Before examing some examples, a few comments concerning the estimation of the above Z are necessary. The values p_1 and p_2 appear in the formula for Z, but both of these true proportions are unkown. These values first appear in the numerator as a difference $p_1 - p_2$. Now the null hypothesis is that there is no difference between these two proportions. Since an hypothesis test always proceeds as if the null hypothesis is true, this difference can be assumed to be 0 for purposes of determining Z.

In the denominator of Z, the expression $\sqrt{(p_1 q_1)/n_1 + (p_2 q_2)/n_2}$ appears, and this is a little more difficult to estimate. Again since these two proportions are assumed to be equal, an estimate of their common value can be used to provide an estimate of the expression under the square root sign. Let the common value of p_1 and p_2 be p. Again, this common value is not known, so let its estimate be referred to as $\hat{p}$. The usual procedure for estimating the common value p is to let $\hat{p}$ be the weighted average, or mean, of $\hat{p}_1$ and $\hat{p}_2$. The formula for $\hat{p}$ is either

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

or

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

If the numbers of successes X_1 and X_2 are given, then the first of these two formulas is used. The second formula is used if the data comes in the form of the sample proportions, p_1 and p_2. The estimate of the common proportion of failures, q, is

$$\hat{q} = 1 - \hat{p}.$$

This provides an estimate of the average value of the proportions for the two populations, and this can be used in the denominator of Z.

Given these modifications, for purposes of carrying out the test,

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\,(1/n_1 + 1/n_2)}}.$$

Example 9.5.1 Change in Support for a Political Party

In a Gallup poll conducted in October, 1989, 28% of the decided voters surveyed said that if a federal election were held, they would vote Progressive Conservative. In another Gallup poll conducted in October, 1991, 13% of the decided voters surveyed said they would vote Conservative. In a third poll conducted in April of 1992, 16% of decided voters said that they would support the Conservative party. The sample size was 1,034 in October, 1989, 1,022 in October, 1991, and 1,035 in April, 1992. Can you conclude that support for the PCs among all Canadian adults declined between October, 1989 and October, 1991? Further, does the reported shift in support between October, 1991 and April, 1992 provide sufficient evidence to indicate that support for the PCs among all Canadian adults increased over these six months? Use the 0.01 level of significance in each case.

Solution. *The question is whether there is sufficient evidence to conclude that Canadian voters as a whole show less support for the PCs in October, 1991 than they did in October, 1989. Let population 1 be Canadian voters in October, 1989 and population 2 be the population of Canadian voters in October, 1991. Then p_1 is the true proportion of all Canadian voters* *who supported the PCs in October, 1989, and p_2 is the true proportion of all Canadian voters who supported the PCs in October, 1991. The null hypothesis is that over this time, there is no change in the proportion of Canadian voters who supported the PCs. Since the question is whether the proportion of PC supporters has declined over this period, this is a one directional test with the alternative hypothesis indicating a decline in support, that is, $p_2 < p_1$. The null and alternative hypotheses are*

$$H_0 : p_1 = p_2 \text{ or } p_1 - p_2 = 0$$

$$H_1 : p_1 > p_2 \text{ or } p_1 - p_2 > 0$$

The test statistic is $\hat{p}_1 - \hat{p}_2$, and if the sample sizes for both samples, n_1 and n_2, are large, then

$$\hat{p}_1 - \hat{p}_2 \text{ is } Nor\left(p_1 - p_2, \sqrt{(p_1 q_1)/n_1 + (p_2 q_2)/n_2}\right)$$

The minimum p or q is 0.13 and

$$\frac{5}{Minimum\ (p,q)} = \frac{5}{0.13} = 38.5$$

and all the sample sizes are much larger than this, so that the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is closely approximated by the normally distribution.

The data from the first two samples is summarized in Table 9.7. The percentages must be divided by 100, in order that they can be interpreted as proportions for purposes of carrying out the test. Note that X_1 and X_2 are not given, but rather the sample proportions are given in the question. For this sample, $n_1 = 1,034$ and $n_2 = 1,022$, and these are large enough to ensure that $\hat{p}_1 - \hat{p}_2$ is normally distributed as above. The estimate of the values of p and q is provided by $\hat{p}$ and $\hat{q}$ where

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}.$$

and $\hat{q} = 1 - \hat{p}$. For the estimate of the standard deviation of the difference in proportions,

$$\begin{aligned}
\hat{p} &= \frac{(1,034 \times 0.28) + (1,022 \times 0.13)}{(1,034 + 1,022)} \\
&= \frac{289.52 + 132.86}{1,034 + 1,022} \\
&= \frac{422.38}{2,056} \\
&= 0.205
\end{aligned}$$

Characteristic	Population 1 October, 1989	Population 2 October, 1991
True Proportion of Successes	p_1	p_2
Sample Size	1,034	1,022
Sample Proportion	0.28	0.13

Table 9.7: Shift in PC Support, October, 1989 to October, 1991

and $\hat{q} = 1 - 0.205 = 0.795$.

The significance level is $\alpha = 0.01$, *and for a one tailed test, the region of rejection of* H_0 *is all values of Z of greater than 2.33. The value of Z for this test is:*

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\hat{p}\hat{q}\,(1/n_1 + 1/n_2)}}$$

Based on the sample data, $\hat{p}_1 = 0.28$ *and* $\hat{p}_2 = 0.13$.

$$\begin{aligned}
Z &= \frac{0.28 - 0.13 - 0}{\sqrt{0.205 \times 0.795\,(1/1,034 + 1/1,022)}} \\
&= \frac{0.28 - 0.13}{\sqrt{0.162975\,(0.0009671 + 0.0009785)}} \\
&= \frac{0.15}{\sqrt{0.0003171}} \\
&= \frac{0.15}{0.0178068} \\
&= 8.424 > 2.33
\end{aligned}$$

As a result, this Z value is well into in the region of rejection in the right tail of the distribution. At the 0.01 level of significance, the null hypothesis of no change in level of support for the PCs between October, 1989 and October, 1991 can be rejected. and The test shows that there been a very significant decline in support for the PCs over this period. Since the hypothesis of no change in the level of support is rejected at a small *significance level, this conclusion is solidly based. The level of Type I error is less than 0.01.*

For the second part of the question, the method is exactly the same, but population 1 is redefined as the population of Canadian voters in October, 1991, and population 2 becomes the population of Canadian voters in April, 1992. The statistics associated with these two samples are given in Table 9.8. The test is shown diagramatically in Figure 9.3.

Characteristic	Population 1 October, 1989	Population 2 October, 1991
True Proportion of Successes	p_1	p_2
Sample Size	1,022	1,035
Sample Proportion	0.13	0.16

Table 9.8: Shift in PC Support, October, 1991 to April, 1992

The null and alternative hypotheses are

$$H_0 : p_1 = p_2 \quad \text{or} \quad p_1 - p_2 = 0$$

$$H_1 : p_1 < p_2 \quad \text{or} \quad p_1 - p_2 < 0$$

In words, the null hypothesis states that there has been no shift in support for the PCs over these months. The alternative hypothesis is in the negative direction because support for the PCs in October, 1991 (population 1) is lower than in April, 1992 (population 2). The statistic, the sampling distribution, and the level of significance are the same as in the earlier part of the question. In Figure 9.3, the region of rejection at $\alpha = 0.01$ *significance is all* Z *values less than* $Z_{0.01} = -2.33$*. This is the shaded area in the left tail of the distribution. If* $Z > -2.33$*, then the null hypothesis of no difference in the proportion of supporters of the PCs cannot be rejected.*

For the estimate of the standard deviation of the difference in proportions,

$$\hat{p} \quad = \quad \frac{(1,022 \times 0.13) + (1,035 \times 0.16)}{(1,022 + 1,035)}$$

Figure 9.3: Test for an Increase in PC Support, October, 1991 to April, 1992

$$\begin{aligned} &= \frac{132.86 + 165.60}{1,022 + 1,035} \\ &= \frac{298.46}{2,057} \\ &= 0.145 \end{aligned}$$

and $\hat{q} = 1 - 0.145 = 0.855$. *Based on the sample data,* $\hat{p}_1 = 0.13$ *and* $\hat{p}_2 = 0.16$*, and the* Z *value is*

$$\begin{aligned} Z &= \frac{0.13 - 0.16}{\sqrt{0.145 \times 0.855\,(1/1,022 + 1/1,035)}} \\ &= \frac{0.13 - 0.16}{\sqrt{0.123975\,(0.0009785 + 0.0009662)}} \end{aligned}$$

$$\begin{aligned} &= \frac{-0.03}{\sqrt{0.0019447}} \\ &= \frac{-0.03}{0.0440986} \\ &= -0.680 > -2.33 \end{aligned}$$

and this Z value is not in the region of rejection of H_0. At the 0.01 level of significance, there is insufficient evidence to reject the null hypothesis that there was no difference in the level of support for the PCs among all Canadian voters in these two months. The difference of 0.03 in the sample proportions is associated with $Z = -0.68$ in Figure 9.3. As can be seen, this lies considerably to the right of -2.33 and does not lie in the region of rejection of H_0.

The size of the shift in support for the PCs was too small to make any conclusions concerning differences in level of support for the PCs over this period. There is a strong possibility of Type II error here, that the level of support for the PCs really did increase, but the results are not conclusive enough to prove this.

Additional Comments on this Test.

1. Note that several decimals were carried in the denominator of the expression for computing Z. For example, $1/1034 = 0.0009671$ was used. While this may be too many decimals to carry, recall the discussion of significant figures in Chapter 4. After the 0s following the decimal, there are only 4 significant figures. While 4 significant figures may be a little more than necessary, it is generally advisable to carry at least 3 significant figures. This will produce estimates of Z which are fairly precise. As a general rule, carry at least 3 significant figures throughout the calculation and round the answer at the conclusion.

2. In the second part of the question, the shift of 3 percentage points between October, 1991 and April, 1992 was too small a shift to make any major conclusion. This is consistent with the sampling error of Chapter 7 or the interval estimates of Chapter 8. For a sample of about 1,000 respondents, at 99% confidence level, the interval estimate would be approximately $\hat{p} \pm 0.04$. Since an interval of this width occurs for each of the two samples, it is no surprise that a difference of 3 percentage points turns out to be statistically insignificant at the 0.01 level. There may be Type II error here, that there really was an increase in support for the PCs, but it was not large enough to detect over these months.

One solution to this when working with the Gallup poll is to wait a few more months, and to see what happens to the reported support for each party. After April, 1992, the support for the PCs continued to climb, reaching 22% in July, 1992. A shift from 13% in October, 1991 to 22% would be very close to being significant, at the 0.01 level. In addition, the continued increase in support for the PCs each month provides additional evidence for an increase. Where a trend develops, and it is continued for a considerable period of time, this provides some evidence for a shift in support, even though each month to month shift is not significant statistically.

3. You may wonder why the above test using two proportions is necessary, since Example 9.4.1 was able to test for a shift in political preference by using the simpler test for a single proportion. In the previous example, a single sample was conducted, and the proportion of NDP supporters in this month was compared with the actual proportion of NDP votes in a federal election. That is, the latter proportion was not a sample, but a population value. A test of a single proportion can be used when there is a single sample, and this value is to be compared to a population value. But when the researcher wishes to compare the proportions from two samples, this test for the difference between two proportions must be used.

Example 9.5.2 Child Discipline and Child Abuse

When constructing interval estimates in Chapter 8, a sample of Toronto families was presented showing the various disciplinary actions that parents used on their children. The data are contained in Example 8.5.2 and Table 8.7. Use this data to test whether the percentage of mothers who pushed, grabbed or shoved a child in the year preceding the survey exceeded the percentage of fathers who did the same.

Solution. *From Table 8.7, the percentage of the 89 mothers who said they pushed, grabbed or shoved was 55%, and the percentage of the 48 fathers who said they pushed, shoved or grabbed as 46%. Letting mothers be population 1 and fathers be population 2. Turning the percentages into proportions, Table 9.9 presents the summary statistics for this test.*

For this test, success is defined as a parent having pushed, shoved or grabbed a child in the year before the survey of Toronto families was conducted. The null hypothesis is that there is no difference in the behaviour of mothers and fathers. The alternative hypothesis is that a larger proportion

Characteristic	Population 1 Mothers 1	Population 2 Fathers
True Proportion of Successes	p_1	p_2
Sample Size	89	48
Sample Proportion	0.55	0.46

Table 9.9: Aggressive Disciplinary Action of Parents

of all mothers than fathers have pushed, shoved or grabbed a child. The hypotheses may be stated as:

$$H_0 : p_1 = p_2 \ \ or \ \ p_1 - p_2 = 0$$

$$H_1 : p_1 > p_2 \ \ or \ \ p_1 - p_2 > 0$$

The test statistic is $\hat{p}_1 - \hat{p}_2$, and if the sample sizes for both samples, n_1 and n_2, are large, then

$$\hat{p}_1 - \hat{p}_2 \text{ is } Nor \left(p_1 - p_2, \sqrt{(p_1 q_1)/n_1 + (p_2 q_2)/n_2} \right)$$

The minimum p or q is 0.45 ($\hat{q}_1$ for mothers) and

$$\frac{5}{Minimum\ (p, q)} = \frac{5}{0.45} = 11.1$$

and all the sample sizes are much larger than this, so that the sampling distribution should be normally distributed. The sample of fathers is a bit small, but should be large enough to ensure that the normal distribution describes the sampling distribution reasonably well.

No significance level has been given in the question. The $\alpha = 0.05$ level will be used here. This is a reasonably small level, but not one so small as to set overly demanding conditions concerning the test. For a one tailed test, $Z_\alpha = Z_{0.05} = 1.645$ and the region of rejection at $\alpha = 0.05$ significance is all Z values greater than $Z = 1.645$.

For the estimate of the standard deviation of the difference in proportions,

$$
\begin{aligned}
\hat{p} &= \frac{(89 \times 0.55) + (48 \times 0.46)}{(89 + 48)} \\
&= \frac{48.95 + 22.08}{137} \\
&= \frac{71.03}{137} \\
&= 0.518
\end{aligned}
$$

and $\hat{q} = 1 - 0.518 = 0.482$. *Based on the sample data,* $\hat{p}_1 = 0.55$ *and* $\hat{p}_2 = 0.46$, *and the* Z *value is*

$$
\begin{aligned}
Z &= \frac{0.55 - 0.46}{\sqrt{0.518 \times 0.482\,(1/89 + 1/48)}} \\
&= \frac{0.55 - 0.46}{\sqrt{0.249676\,(0.011236 + 0.0208333)}} \\
&= \frac{0.09}{\sqrt{0.0080069}} \\
&= \frac{0.09}{0.0894815} \\
&= 1.006 < 1.645
\end{aligned}
$$

As a result, this Z *value is not in the region of rejection of the null hypothesis. At the* $\alpha = 0.05$ *level of significance, there is insufficient evidence to reject the null hypothesis that there is no difference in the true proportion of Toronto mothers and fathers who pushed, shoved or grabbed a child in the year before the survey.*

Additional Comments.

1. The conclusions made on the basis of this test confirm the original results suggested by the interval estimates in Example 8.5.2. There the interval estimates were fairly wide so that the interval estimates of the proportions of mothers and fathers who hit a child with an object overlapped very considerably. The method of interval estimation was unable to provide evidence for a difference in behaviour of mothers and fathers.

The test provided in this section makes the same point more strongly. The difference in the percentages for pushing, shoving or grabbing is $55 -$

$46 = 9$ *percentage points. For hitting a child with an object, the difference was only 5 percentage points. Yet even with the larger difference of 9 percentage points in observed behaviour, there is insufficient evidence to prove that all Toronto mothers exhibited more aggressive action than did fathers. While Lenton may be correct in her conclusions, in that the proportions do point in the direction of greater aggressiveness by mothers than by fathers, the evidence is not strong enough to make a conclusion that the behaviour of mothers and fathers is dramatically different. The problem is the small sample size, and the fairly small differences in percentages reported.*

2. The Z value of 1.006 can be used to determine the exact significance level. Round to $Z = 1.01$ *and the area to the right of* $Z = 1.01$ *is 0.1563. Assuming that the null hypothesis of no difference in behaviour of mothers and fathers is correct, the probability of finding a difference of 9 percentage points, with these sample sizes is 0.1563. Since this is not a small probability, the null hypothesis is not rejected.*

3. This test assumes that the samples of mothers and fathers are drawn independently of each other. This may not be the case for this sample. That is, the test assumes that a random sample of all Toronto fathers is taken, and another random sample of all Toronto mothers is taken as well. But if the sampling method that was used involved selecting a random sample of families, and then interviewing both the mothers and the fathers in the same family, this test would not be valid. The latter sampling method is a perfectly legitimate sampling method, it is just that this is not the proper test to use in these circumstances. Some further comments on the independence of samples are contained on page 652.

Example 9.5.3 Political Preferences in Alberta

At the end of Chapter 8, some of the results from the 1991 Alberta Survey, conducted by the Population Research Laboratory at the University of Alberta, were presented in Table 8.9. That table gave the pattern of political preferences by income of the household. The authors commented that NDP supporters tended to have lower incomes than did supporters of other political parties, with the Reform Party supporters being least likely to be low income.

Using the data in that table, test whether (a) there is any difference between the proportion of NDP supporters who are in the lowest income level and the proportion of PC supporters who are in the lowest income category, and (b) whether the proportion of Reform Party supporters in the *lowest income category is less than the proportion of NDP supporters in the lowest income category. Use the 0.05 level of significance in each case.*

Solution. *The data for testing these hypotheses are contained in Table 9.10. The NDP is involved in both tests so it is labelled population 1. There are two population 2s, 2(i) is the PCs for the first part of the question, and 2(ii) refers to the Reform Party for the second part of the question.*

Characteristic	Population 1 NDP	Population 2(a) PC	Population 2(b) Reform Party
True Proportion	p_1	p_2	p_2
Sample Size	198	161	248
No. of Successes	30	21	22
Sample Proportion	0.152	0.130	0.089

Table 9.10: Political Preferences of Alberta Voters

For part (a), the null hypothesis is that for the NDP and the PCs, there is no difference in the proportion of supporters who have the lowest income level. No suspicion concerning the direction of the expected relationship is given, so that a two tailed test can be used here. Let the NDP be population 1 and the PC supporters be population 2. The hypotheses may be stated as:

$$H_0 : p_1 = p_2 \quad \text{or} \quad p_1 - p_2 = 0$$

$$H_1 : p_1 \neq p_2 \quad \text{or} \quad p_1 - p_2 \neq 0$$

The test statistic is $\hat{p}_1 - \hat{p}_2$, and if the sample sizes for both samples, n_1 and n_2, are large, then

$$\hat{p}_1 - \hat{p}_2 \text{ is } Nor \left(p_1 - p_2, \sqrt{(p_1 q_1)/n_1 + (p_1 p_2)/n_2} \right)$$

The minimum p or q is 0.13 ($\hat{p}_1$ for PCs) and

$$\frac{5}{Minimum\ (p, q)} = \frac{5}{0.13} = 38.5$$

and all the sample sizes are much larger than this, so that the sampling distribution should be normally distributed.

For a two tailed test at the 0.05 level of significance, $Z_{\alpha/2} = Z_{0.025} = 1.96$ *and the region of rejection for a two tailed test at* $\alpha = 0.05$ *significance is all* Z *values less than -1.96 or greater than 1.96.*

For the estimate of the standard deviation of the difference in proportions,

$$\begin{aligned} \hat{p} &= \frac{21+30}{161+198} \\ &= \frac{51}{359} \\ &= 0.142 \end{aligned}$$

and $\hat{q} = 1 - 0.142 = 0.858$. *Based on the sample data, the* Z *value is*

$$\begin{aligned} Z &= \frac{0.152 - 0.130}{\sqrt{0.142 \times 0.858\,(1/198 + 1/161)}} \\ &= \frac{0.022}{\sqrt{0.121836\,(0.050505 + 0.0062112)}} \\ &= \frac{0.022}{\sqrt{0.0069101}} \\ &= \frac{0.022}{0.08313} \\ &= 0.2656 < 1.645 \end{aligned}$$

The Z *value from the sample is between -1.96 and +1.96 and thus is not in the region of rejection of the null hypothesis. At the 0.05 level of significance, there is not sufficient evidence to conclude that there is any difference between the NDP and the PCs in the proportion of supporters from the lowest income level.*

For the second test, the NDP continues to be population 1 but population 2 becomes the supporters of the Reform Party. If success is again defined as being in the lowest income category, then $\hat{p}_1 = 0.152$ *and* $\hat{p}_2 = 0.089$. *Since the question is whether there is a lower proportion of low income supporters among the Reform party than among the NDP, the hypotheses are*

$$H_0 : p_1 = p_2 \quad or \quad p_1 - p_2 = 0$$

$$H_1 : p_1 > p_2 \quad or \quad p_1 - p_2 > 0$$

The test is conducted in the same manner as before. The sample sizes are sufficient to assume that $\hat{p}_1 - \hat{p}_2$ *is normally distributed as before. For a one tailed test at the 0.05 level of significance,* $Z_\alpha = Z_{0.05} = 1.645$ *and the region of rejection for this test at* $\alpha = 0.05$ *significance is all* Z *values greater than 1.645.*

For the estimate of the standard deviation of the difference in proportions,

$$\begin{aligned} \hat{p} &= \frac{30+22}{198+248} \\ &= \frac{52}{446} \\ &= 0.117 \end{aligned}$$

and $\hat{q} = 1 - 0.117 = 0.883$. *Based on the sample data, the* Z *value is*

$$\begin{aligned} Z &= \frac{0.152 - 0.089}{\sqrt{0.117 \times 0.883\,(1/198 + 1/248)}} \\ &= \frac{0.063}{\sqrt{0.103311\,(0.050505 + 0.0040323)}} \\ &= \frac{0.063}{\sqrt{0.00563430}} \\ &= \frac{0.063}{0.07506} \\ &= 0.839 < 1.645 \end{aligned}$$

The Z *value from the sample is less than 1.645 and thus is not in the region of rejection of the null hypothesis. At the 0.05 level of significance, there is not sufficient evidence to conclude that the NDP has a larger proportion of lowest income supporters than does the Reform Party.*

Additional Comments.

1. Unlike the two previous examples, the sample proportions were not given in Table 8.9. The actual number of respondents was given in this table. These are the X *values, the actual numbers of successes in the sample. The* *formula for $\hat{p}$ which is based on X_1 and X_2 was used to obtain the estimate of the common proportion of successes for the two populations.*

2. These tests show the difficulty of making decisive conclusions concerning differences of proportions for various subgroups of the population. While the overall sample size for the Alberta survey is 807, there are less than 250 supporters of each of the 4 major political parties in the province. Part (a) showed that a difference of 0.022, or 2.2 percentage points, is not a statistically significant difference at the 0.05 level. For part (b), it appears that there is a major difference between the Reform Party and the NDP, with a difference of 6.3 percentage points. But even this difference turns out to be insignificant statistically at the 0.05 level.

This test does not constitute a complete test of the author's contention that the NDP supporters tend to have lower income than do the supporters of other political parties. There are various other possible tests which could be used. For example, it may be that NDP supporters have a lower mean income than do supporters of other political parties. This can be tested with the method described in the next section.

Independence of the Two Samples. This test for the difference between two proportions assumes that the random samples from each of the two populations are drawn independently of each other. If a random sample is drawn from one of the populations, and then a completely separate random sample is drawn from the second population, this certainly satisfies these conditions. The assumption is that any population member selected in one sample has no influence on the probability of selection for members of the second population.

A random sample of a population which is later split into two or more groups also satisfies this condition. In Example 9.5.3, a random sample of 807 Alberta adults was selected as part of the Alberta Study. This sample was then split into 5 different income groups and 4 political preferences. Each of these income groups, or each of the political parties, can be treated as a separate random sample of that group. Thus the 198 NDP supporters among the respondents can be regarded as a random sample of all Alberta NDP residents. Similarly, the 161 PCs can be regarded as a separate and independent random sample of Alberta PC supporters. Thus the conditions for the test are satisfied in Example 9.5.3.

The assumptions for using this test are violated when the probabilities of selection for members of one population are dependent on, or affected by, the selection of the sample in the other population. For example, if a random sample of parents is selected, and then both the children of these parents, and the parents, are surveyed. Which children are selected depends on which parents have been selected. Another example of a dependent method of sampling occurs when those who are selected in a sample then suggest other names for the researcher to sample. This means that some of those selected depend on having their names suggested by others.

The dependent sampling methods described in the last paragraph are often useful methods, and they can legitimately be used for many research purposes. But the methods of this section should not be used to conduct hypothesis tests for the difference between two proportions, where the samples are dependent samples. In the case of such dependent sampling methods, it might be necessary to investigate other testing procedures in order to obtain meaningful inferences concerning population parameters. The paired t test of Section 9.10 is one example of this. When the samples are independent random samples, or close to being independent random samples, then the test of this section can be used to obtain meaningful inferences concerning two populations.

9.6 Test for a Difference of Two Means

The tests that were used to introduce hypothesis testing were tests of a single mean. When n is large, the test of a single mean uses a normal distribution for the sample mean, and when n is smaller, the t distribution is used to conduct this test. This section shows how tests for a difference between two population means can be conducted for both large sample size and small sample size. The concepts introduced in the last section concerning a test for two different proportions are extended to the test for two different means in this section.

Begin by assuming that random samples are selected from each of two different populations. These samples must be separately and independently selected random samples, so that the cases selected in one population in no way affect selection probabilities in the other population. From each random sample, the sample mean and sample standard deviation can be obtained. These sample statistics are used to test whether there is a difference between the true means of the two different populations.

With two populations and two samples, there is considerabls notation involved. The notation that will be used in this section is summarized in

Table 9.11. The populations are referred to as population 1 and population 2, with subscripts on each of the parameters and statistics to denote which population each of these comes from. All of the population parameters are unknown, and the aim of the test is to provide some inferences concerning how the parameters differ between the two populations. The null hypothesis is of the form

$$H_0 : \mu_1 = \mu_2$$

or as a difference, the equivalent form is

$$H_0 : \mu_1 - \mu_2 = 0.$$

The alternative hypothesis may be an inequality of any one of the three types, not equal to, less than, or greater than. Independent random samples of sizes n_1 and n_2 are taken from each of the two populations, and sample statistics are obtained from the samples.

Characteristic	Population 1	Population 2
True Mean	μ_1	μ_2
True Standard Deviation	σ_1	σ_2
True Variance	σ_1^2	σ_2^2
Sample Size	n_1	n_2
Sample Mean	$\bar{X}_1$	$\bar{X}_2$
Sample Standard Deviation	s_1	s_2
Sample Variance	s_1^2	s_2^2

Table 9.11: Notation for Test of a Difference in Two Population Means

The test statistic which will be used for the test of a difference between the two population means in this section is $\bar{X}_1 - \bar{X}_2$. This is where the difficulty in this test begins. There are various possible sampling distributions for this test statistic, depending on the size of the samples and depending on the assumptions concerning the standard deviations and variances in the two populations. Because of the extra notation associated with two populations and two samples, the formulas in this section are somewhat more complex than in the earlier sections.

This section outlines three possible forms for the test of a difference between two means. Section 9.6.1 gives a test for the difference between two population means when the sample size for each random sample is large. In this case, the difference of sample means is normally distributed. The tests in Section 9.6.2 are similar, except that at least one of the sample sizes is smaller, so that the t distribution must be used. There are two such t tests, one assuming equal population variances and the other assuming different population variances. Some examples of results from the `T-TEST` procedure of the SPSSX computer program and the `TWOSAMPLE` procedure of MINITAB will be given in Section 9.8. These programs can be used to obtain the t and Z values for the difference between the two sample means for data which has been entered on a computer. Since the formulas in this section are sometimes quite time consuming to calculate, these computer procedures can save considerable time when testing for a difference between two mean. Section 9.10 examines the paired t test for the difference between two means. This is an example of how a difference in means can be tested when there are dependent samples.

Independent Random Samples. Before proceeding with these tests, it should be emphasized that the tests of Sections 9.6.1 through 9.8 apply only when independent random samples from each of the two populations are obtained. If you are not clear concerning what this means, reread the note concerning this on page 652 in connection with the test for the difference between two population proportions. The same conditions apply to these tests. Note that one example of a test for dependent samples, the paired t test, is given in Section 9.10.

9.6.1 Test for Different Means, Large Sample Sizes

If the random samples from each of the two populations have large sample sizes, then the difference in sample means is a normally distributed variable. The rule concerning what is a large sample size is the same as used when distinguishing between the t test and the Z test for a single mean, that is, that the sample size exceed 30. For the test of a difference between two means, if both $n_1 > 30$ and $n_2 > 30$, the difference in the sample means has a normally distributed sampling distribution.

The mean of the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is $\mu_1 - \mu_2$. Let the standard deviation of this sampling distribution be given the symbol $\sigma_{\bar{X}_1 - \bar{X}_2}$ with σ being used to denote that it is the standard deviation, and subscript $\bar{X}_1 - \bar{X}_2$ to denote that it is the standard deviation of this statistic. If the two samples are independent random samples, this standard deviation can be shown to equal

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The major difficulty in conducting a test for the difference in two population means is to obtain an estimate of this standard deviation. The problem emerges because both σ_1 and σ_2 are unknown. Let $\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}$ be the estimated standard deviation of $\bar{X}_1 - \bar{X}_2$. Several different estimates of this value will be provided in this section, where these estimates depend on the sample sizes, and on what is assumed about the variances of the two populations.

When the size of each of the two random samples is large, this standard deviation can be reasonably closely estimated by

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Since all parts of this expression are known, this value can be determined from the sample data. This is the estimate of $\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}$ that will be used in this section.

The test begins in the usual way, by hypothesizing equality in the two means. The alternative hypothesis is written here as a two directional inequality, but a one directional inequality in either the positive or negative direction could be used instead.

$$H_0 : \mu_1 = \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 \neq \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 \neq 0$$

are the null and alternative hypotheses.

The test statistic is $\bar{X}_1 - \bar{X}_2$ and if both the sample sizes n_1 and n_2 are large, then

$$\bar{X}_1 - \bar{X}_2 \text{ is Nor } \left(\mu_1 - \mu_2, \sigma_{\bar{X}_1 - \bar{X}_2}\right).$$

The level of significance is selected, and the Z value which defines the critical region is obtained from the normal table of Appendix H. The Z value is the statistic minus its mean and divided by its standard deviation. This gives

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}}.$$

Since the null hypothesis is that the two means are equal, the latter difference, $\mu_1 - \mu_2$ is zero. Using the expression

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

for the standard deviation of the difference in the sample means, this becomes

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

and this is the Z value that will be calculated from the sample data. If this Z is in the critical region, then the null hypothesis is rejected. If it is not in the critical region, then there is insufficient evidence to reject the null hypothesis.

Data from two of the examples in Section 8.3.1 in Chapter 8 will be used as examples of this test. A diagrammatic presentation of the test is given in Example 9.6.2.

Example 9.6.1 Differences in Head Size

Interval estimates for differences in head circumferences for people of different vocational status were constructed in Example 8.3.3 These data were presented by Stephen Jay Gould in **The Mismeasure of Man**, *and were originally collected by Ernest A. Hooton. The means and standard deviations of head circumferences are given in Table 8.4along with the sample sizes for each group. Use this data to test whether the mean head circumference is different for those in semiprofessional occupations and personal service occupations. ($\alpha = 0.10$ significance).*

Solution. *Table 9.12 contains the summary data necessary for conducting this test. These data come from Table 8.4 and have been set up to conform to the notation used for a test of two means. Population 1 is identified as those of semiprofessional vocational status and population 2 is those of personal service status.*

Characteristic	Population 1 Semiprofessional	Population 2 Personal Service
True Mean	μ_1	μ_2
True Standard Deviation	σ_1	σ_2
True Variance	σ_1^2	σ_2^2
Sample Size	$n_1 = 61$	$n_2 = 262$
Sample Mean	$\bar{X}_1 = 566.5$	$\bar{X}_2 = 562.7$
Sample Standard Deviation	$s_1 = 11.7$	$s_2 = 11.3$
Sample Variance	$s_1^2 = 136.89$	$s_2^2 = 127.692$

Table 9.12: Summary Data for Head Circumferences (in millimetres)

Using the notation and data in Table 9.12, the test can be conducted. The null and alternative hypotheses for the test are

$$H_0 : \mu_1 = \mu_2 \text{ or } \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 \neq \mu_2 \text{ or } \mu_1 - \mu_2 \neq 0$$

The null hypothesis states that there is no difference between the mean head circumference of those of semiprofessional and personal service vocational status. The alternative hypothesis states that there is a difference between the mean head circumference of members of these two occupational groupings. The question asked whether there was a difference or not, with no direction indicated for the alternative hypothesis, so that a two tailed test is used here.

The test statistic is $\bar{X}_1 - \bar{X}_2$ *and since both the sample sizes* $n_1 = 61$ *and* $n_2 = 262$ *are above 30,*

$$\bar{X}_1 - \bar{X}_2 \text{ is } Nor \left(\mu_1 - \mu_2, \sigma_{\bar{X}_1 - \bar{X}_2}\right).$$

The level of significance is $\alpha = 0.10$ *and since this is a two tailed test, the critical* Z *value is* $Z_{0.05} = 1.645$. *The critical region is all* Z *values of less than -1.645 or greater than +1.645.*

Using the estimate of the standard deviation of the difference in sample means suggested above, the Z *from the sample data becomes:*

$$\begin{aligned} Z &= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \\ &= \frac{566.6 - 562.7}{\sqrt{3.9/61 + 127.69/262}} \\ &= \frac{3.9}{\sqrt{2.244098 + 0.4873664}} \\ &= \frac{3.9}{1.6527144} \\ &= 2.359 > 1.645 \end{aligned}$$

As a result, the Z *value associated with this difference in sample means is statistically significant at the 0.05 level of significance. The differences in head circumference for these two samples is sufficient to provide evidence that the mean head circumference from all those of semiprofessional and personal service occupations differ.*

Additional Comments. *Gould commented that most of the differences in the means of the different groups are not significant statistically. While this can be shown to be the case for many of the groups, for these two groups there does appear to be a statistically significant difference at the 0.10 level. You could show that each pair of differences for the first 5 groups is not statistically significant, and this would support Gould's statement. But for the two groups examined here, there is a considerable difference in the means.*

Exactly what to make of this difference is not so clear. In his book, Gould shows the many methodological problems associated with measurements of head size, and the problems of interpreting the meaning of these differences. The implication that different head sizes mean different levels of intelligence or ability is certainly not warranted. Whether different head circumferences mean anything more than that some people have different head sizes than do others, is not really clear. This issue is still being debated by many psychologists and other academics.

Example 9.6.2 Incomes in Alberta and Saskatchewan

The means and standard deviations of individual incomes, along with sample sizes, obtained for each province of Canada from Statistics Canada's General Social Survey were given in Table 8.3. The data in that table was used in Example 8.3 to see whether incomes in Alberta were different than in Saskatchewan. Use this data to test whether mean Alberta individual income exceeds the mean level of individual income in Saskatchewan. Use the 0.05 level of significance.

Solution. *Table 9.13 contains the summary data necessary for conducting this test. This data comes from Table 8.3 and has been converted into incomes in thousands of dollars. This will make the calculations more managable. In particular, the standard deviations must be squared, and since the standard deviations had 5 figures in the original table, this would have produced squares with 9 or 10 figures, more than many calculators can handle.*

Characteristic	Saskatchewan Population 1	Alberta Population 2
True Mean	μ_1	μ_2
True Standard Deviation	σ_1	σ_2
True Variance	σ_1^2	σ_2^2
Sample Size	$n_2 = 612$	$n_2 = 613$
Sample Mean	$\bar{X}_1 = 15.768$	$\bar{X}_2 = 16.949$
Sample Standard Deviation	$s_1 = 14.837$	$s_2 = 14.958$
Sample Variance	$s_1^2 = 220.137$	$s_2^2 = 223.742$

Table 9.13: Saskatchewan and Alberta Incomes, in thousands of Dollars

Since the suspicion is that Alberta incomes are higher than those in Saskatchewan, the alternative hypothesis is that μ_1 is less than μ_2. The *hypotheses for the test are*

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 < 0$$

The test statistic is $\bar{X}_1 - \bar{X}_2$ *and since both sample sizes* $n_1 = 612$ *and* $n_2 = 613$ *are well above 30,*

$$\bar{X}_1 - \bar{X}_2 \text{ is Nor } \left(\mu_1 - \mu_2, \sigma_{\bar{X}_1 - \bar{X}_2}\right).$$

The level of significance is $\alpha = 0.05$ *and since this is a one tailed test in the negative direction,* $Z_{0.05} = -1.645$ *is the critical value. If* $Z < -1.645$ *then the null hypothesis is rejected, but if* $Z > -1.645$ *the null hypothesis is not rejected.*

Figure 9.4: Test for Lower Mean Income in Saskatchewan than Alberta

Figure 9.4 shows the normal sampling distribution for $\bar{X}_1 - \bar{X}_2$*, with its mean hypothesized to be 0. The region of rejection of* H_0 *is the lower 0.05 of* *the distribution, where the Z value associated with the difference of sample means is less than -1.645.*

Using the estimate of the standard deviation of the difference in sample means suggested above, the Z from the sample data becomes:

$$\begin{aligned}
Z &= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \\
&= \frac{15.768 - 16.949}{\sqrt{220.137/612 + 223.742/613}} \\
&= \frac{-1.181}{\sqrt{0.359701 + 0.3649951}} \\
&= \frac{-1.181}{0.8512908} \\
&= -1.387 > -1.645
\end{aligned}$$

As can be seen in Figure 9.4, $Z = -1.387$ is to the right of $Z = -1.645$ and is not in the critical region for the test. As a result, the Z value associated with this difference in sample means is not statistically significant at the 0.05 level of significance. At 0.05 significance there is insufficient evidence from these two samples to conclude that the mean individual income of individuals in Alberta exceeds the mean individual income of individuals in Saskatchewan.

Additional Comments. *In Example 8.3.2 two interval estimates of mean income were constructed. The 95% interval estimate of mean individual income was ($14,600,$16,900) for Saskatchewan and the corresponding Alberta interval was ($15,800,$18,100). In Example 8.3.2 it was noted that there is a considerable overlap in these two intervals, and this might mean that the data did not provide very strong evidence of a difference in true mean income for the two provinces. This has now been proven more systematically. While the evidence points in the direction of higher mean individual incomes in Alberta than in Saskatchewan, the evidence is not strong. Even though the sample sizes are fairly large, the difference in means income of $15,768 - 16,949 = -1,181$ dollars was not a large enough difference to prove that the mean income in the two provinces differs. This conclusion was made with a one tailed test at the 0.05 level of significance.*

Note that the Z associated with this difference is $Z = -1.39$ and from the normal table, this is associated with an area of 0.0823 in the tail of the *sampling distribution to the left of -1.39. That is, if the null hypothesis is assumed to be true, the probability of obtaining a difference as large as $1,181 with these sample sizes and standard deviations is 0.0823. This is the exact significance level. If you view this as a relatively low probability, then the null hypothesis could be rejected as this level, or at* $\alpha = 0.10$ *it could be rejected. Thus the difference provides weak evidence that Saskatchewan mean income is lower than that for Alberta.*

9.6.2 Test for Different Means, Small Sample Size

If random samples have been taken independently from two populations and one of the sample sizes is small, then it cannot be assumed that the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is normally distributed. If the two populations from which the samples are drawn are themselved normally distributed, then the sampling distribution of $\bar{X}_1 - \bar{X}_2$ has a t distribution with mean $\mu_1 - \mu_2$ and standard deviation $\sigma_{\bar{X}_1 - \bar{X}_2}$. The difficulty with this test is to provide an estimate of the latter standard deviation, and to determine the degrees of freedom associated with this t test. Neither of these has a simple solution, and there is some disagreement among statisticians concerning which estimate of these is preferable.

This section will present two commonly used t tests for the difference between two means. The first is called the **pooled variance** test, and this method assumes that the standard deviations or variances of the two populations are equal. The second method is called the unequal or **separate variance** method, and it assumes that the variances of the two populations are not equal. The formulas presented here may differ slightly from those presented in some other textbooks. Different authors use slightly different formulas, especially for the separate variance method. The formulas presented here seem to be consistent with the formulas used in the SPSS and MINITAB computer programs briefly discussed in Section 9.8.

For each of these methods, the null and research hypotheses are the same as earlier. The test statistic is the difference between the two sample means, and if either sample is small, then the distribution of this statistic is a t distribution. Strictly speaking, each of the samples must be random samples, drawn independently of each other, and the populations from which these samples are drawn must be normally distributed. Then $\bar{X}_1 - \bar{X}_2$ has a t distribution. Recall though that if a t distribution has over 30 degrees of freedom, the t distribution becomes a normal distribution. When the formulas which follow produce more that 30 degrees of freedom, the normal table will usually be consulted, rather than the t table. That is, the t value associated with the difference of two means becomes a Z value.

For this t distribution, $\bar{X}_1 - \bar{X}_2$ is the difference in the two sample means. The sampling distribution of this sample difference of means has a mean which is the difference in the true means of the two populations, $\mu_1 - \mu_2$. Let the standard deviation of this sampling distribution be $\sigma_{\bar{X}_1 - \bar{X}_2}$. There are two different estimates of this standard deviation, and which is used depends on whether the pooled or separate variance method is adopted. Having selected one of these, the significance level is chosen and the critical value and region of rejection of H_0 determined from the t table. The t value associated with the sample data is determined from the sample data as

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}}.$$

Since $\mu_1 - \mu_2 = 0$ by hypothesis, this becomes

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}}.$$

When the two samples are independent random samples, drawn from the two populations with standard deviations σ_1 and σ_2 respectively, it can be shown that

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

In this section, there are different estimates of the denominator depending on whether the pooled or separate variance test is used. The t value will either fall into the critical region or not, and the conclusion concerning the null hypotheses determined as in the earlier tests.

Each of the two approaches to estimating the standard deviation of the difference in the sample means is now examined. An example of each is given.

Pooled Variance t Test. The pooled variance t test begins with the assumption the standard deviations and variances of the two populations are equal to each other. That is $\sigma_1^2 = \sigma_2^2$ or $\sigma_1 = \sigma_2$. Given this assumption, the common variance can given the symbol σ_p^2, and the common standard deviation is σ_p. The problem then becomes one of estimating this common value. Let the true, but unknown, common value of the two variances be denoted by σ_p, and the estimate of the common value be $\sigma_{\hat{p}}$.

The common value of the variances is usually estimated as a mean of the two separate variances, with each variance being weighted by its sample size in order to determine the mean of the two variances. In doing this, the formula used to compute $\hat{\sigma}_p^2$, the estimate of the common variance σ_p^2, is

$$\hat{\sigma}_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and

$$\hat{\sigma}_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

where s_1^2 and s_2^2 are the sample variances from populations 1 and 2, respectively. Now

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

and $\sigma_1^2 = \sigma_2^2 = \sigma_p^2$ for the pooled variance test, so that

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sigma_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \approx \hat{\sigma}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Since the t value is the test statistic minus its mean, divided by its standard deviation, substituting these values gives

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_p \sqrt{1/n_1 + 1/n_2}}.$$

This t value can be computed on the basis of the data from the sample.

The **degrees of freedom** for this test is $n_1 + n_2 - 2$. In the ealier uses of the t distribution, there were $n - 1$ degrees of freedom associated with the estimate of the mean and the t test for a single mean. Since there are two means here, the degrees of freedom associated with this t value is the sum of the two sample sizes minus 2.

Example 9.6.3 Differences in Self-Esteem of Elderly Women

In Example 8.4.3, data concerning the difference between the self-esteem of elderly women who lived in their own homes and those who lived in nursing homes was used to construct interval estimates for each group. These interval estimates appeared to show that there was little difference in the *level of self-esteem for the two groups. Interval estimates do not provide a direct test of differences of means. For the test of a difference of means, this same data can now be used to conduct a direct test.*

Using the data in Table 8.6 can you conclude that there is any difference in the self-esteem of elderly women who live in their own homes and the self-esteem of elderly women who live in nursing homes. Use the 0.10 level of significance.

Solution. *The data from Table 8.6 is presented in the format for a difference between two means in Table 9.14. Population 1 is defined as the community*

Characteristic	Population 1 Community	Population 2 Nursing Home
True Mean	μ_1	μ_2
True Standard Deviation	σ_1	σ_2
True Variance	σ_1^2	σ_2^2
Sample Size	$n_1 = 28$	$n_2 = 25$
Sample Mean	$\bar{X}_1 = 29.4$	$\bar{X}_2 = 25.1$
Sample Standard Deviation	$s_1 = 15.0$	$s_2 = 14.4$
Sample Variance	$s_1^2 = 225.0$	$s_2^2 = 207.36$

Table 9.14: Summary Data for Self-Esteem of Elderly Women

group of elderly women who live in their own homes, and the true mean and standard deviation of the index of self-esteem for all such elderly women are μ_1 *and* σ_1 *respectively. Population 2 is the population of all elderly women who live in nursing homes, with the true mean and standard deviation for all these elderly women being* μ_2 *and* σ_2 *respectively. All of these population parameters are unknown, and the sample sizes are less than 30, so that a t test for the difference in the two population means will be required.*

The sample standard deviations for the two groups can be seen to be *quite similar, with the same being true of the respective sample variances. Thus the pooled variance t test will be used here. A test for the equality of two variances is given later in Section 9.7, and that test can form the basis for deciding whether the pooled or separate variance estimates are to be used.*

In difference form, the null and research hypotheses are

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

The null hypothesis states that there is no difference in the level of self-esteem of elderly women in the two populations. Since the question implies no direction concerning which of the two groups has a higher mean level of self-esteem, the alternative hypothesis is a two directional statement.

The test statistic for a difference between the true means is the difference in the sample means $\bar{X}_1 - \bar{X}_2$. *Assuming that the distribution of self-esteem for each group of elderly women is normal, and that these samples are equivalent to random samples from each of the two populations, the sampling distribution of* $\bar{X}_1 - \bar{X}_2$ *is approximated by*

$$\bar{X}_1 - \bar{X}_2 \text{ is } t\ (\mu_1 - \mu_2, \sigma_{\bar{X}_1 - \bar{X}_2}).$$

Since this is the pooled variance test, the t *value will*

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_p \sqrt{1/n_1 + 1/n_2}}$$

where σ_p *is approximated by*

$$\hat{\sigma}_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2^2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

and there are $n_1 + n_2 - 2$ *degrees of freedom.*

The level of significance is $\alpha = 0.10$, *and since this is a two tailed test, an area of 0.10 is split equally between the two tails of the distribution, so that there is 0.05 of the area in each tail of the distribution. The degrees of freedom is* $n_1 + n_2 - 2 = 28 + 25 - 2 = 51$ *degrees of freedom. The* t *table for 50 degrees of freedom and a two tailed test at the 0.10 level of signficance is 1.676. The region of rejection of the null hypothesis is all* t *values of less than -1.676 or greater than +1.676. If the* t *value calculated from the data* *is between -1.676 and +1.676, then the null hypothesis of equal levels of self-esteem cannot be rejected.*

The final stage is the computation of the t value from the data. The first step in this is to compute the standard deviation of the sampling distribution of the difference in the sample means. For the pooled variance estimate, this is

$$\begin{aligned}
\hat{\sigma}_p &= \sqrt{\frac{(28-1)225.0+(25-1)207.36}{28+25-2}} \\
&= \sqrt{\frac{6,075.00+4,976.64}{51}} \\
&= \sqrt{216.69882} \\
&= 14.721
\end{aligned}$$

Note that this standard deviation is a weighted average of the two original standard deviations, and has a value which is between these two standard deviations. Using this common value as the estimate of the standard deviation,

$$\begin{aligned}
t &= \frac{\bar{X}_1 - \bar{X}_2}{\sigma_p\sqrt{1/n_1+1/n_2}} \\
&= \frac{29.4-25.1}{14.721\sqrt{1/28+1/25}} \\
&= \frac{4.3}{14.721\sqrt{0.0357143+0.04}} \\
&= \frac{4.3}{14.721\sqrt{0.0757143}} \\
&= \frac{4.3}{4.0506641} \\
&= 1.062
\end{aligned}$$

and this t value is greater than -1.676 but less than 1.676. As a result, this t value is not in the region of rejection of the null hypothesis. At the 0.10 level of significance, the two sample means do not differ sufficiently to reject the null hypothesis that there is no difference between the two populations. Based on this sample, elderly women living in their own homes and elderly women living in nursing homes have much the same level of self-esteem.

Additional Comments. *1. Note that since the degrees of freedom for the t value is 51, considerably more than 30, it could have been assumed that the distribution of the difference between the sample means is a normally distributed variable. In this case, the region of rejection of the null hypothesis would have been all Z values larger in magnitude than 1.645, that is, less than -1.645 or greater than +1.645. This value can be obtained from the normal table, or by going down the t table to the bottom row. The values in the bottom row of the t table are normal or Z values. For a two tailed test at* $\alpha = 0.10$*, the value in the bottom row of the t table is 1.645.*

2. From the last paragraph, it can be seen that there is often little difference whether a t or Z value is used. The region of rejection for the t distribution is all t values with magnitude above 1.676. Only if t value calculated from the data happened to fall between 1.645 and 1.676 would the conclusions of the t test differ from the conclusions using the Z value.

3. Since a fairly large significance level of 0.10 has been used, it should have been relatively easy to reject the null hypothesis. Yet the hypothesis that the two groups have equal means could not be rejected, so these samples appear to indicate very little difference between the groups. There is some chance that Type II error exists in this test, but it would not seem to be a particularly serious error. If there really had been a large difference in the self-esteem for the two groups of elderly women, then the t value would have been larger. While there is likely some difference between the self-esteem of the two groups, it would seem to be a fairly minimal difference.

Separate Variance t Test. If the standard deviations of the two samples are quite different from each other, then the pooled variance test may produce inaccurate results. The separate variance test described in this section is then used. The separate variance t test begins with the assumption the the standard deviations and variances of the two populations are different from each other. That is $\sigma_1^2 \neq \sigma_2^2$ or $\sigma_1 \neq \sigma_2$. If this is the case, a procedure much like that used in the case of large sample sizes is used. That is, s_1^2 is used to estimate σ_1^2 and s_2^2 can be used to estimate σ_2^2. This gives

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

in the denominator of the t value for the difference of sample means. That is, the distribution of $\bar{X}_1 - \bar{X}_2$ is closely approximated by a t distribution with mean $\mu_1 - \mu_2$ and the standard deviation just given. This produces the t value

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

This is essentially the same expression as the Z value for the difference of two means when both n_1 and n_2 are large.

The **degrees of freedom** associated with this t value is the major problem involved in using this when at least one of the sample sizes is small. Statisticians have suggested two solutions to this. These are as follows.

1. The simplest of the two solutions concerning the choice of the proper degrees of freedom is to determined which is the smaller of $n_1 - 1$ and $n_2 - 1$. The degrees of freedom is then the smaller of these two numbers. That is,

   $$\text{df} = \text{Minimum } (n_1 - 1, n_2 - 1)$$

   While this is an acceptable solution, this sometimes makes the degrees of freedom rather small, and this in turn may make it very difficult to reject a null hypothesis. Remember that the values in the t table are quite large when the degrees of freedom are very small. This implies that the two sample means may have to differ quite considerably before the sample data will produce a t value large enough to reject H_0. When conducting this test by hand, it would be most common to use this procedure for determining the degrees of freedom.

2. The second procedure is to use a more precise determination of the degrees of freedom associated with this t value. The value which is generally given by statisticians for estimating the degrees of freedom is:

   $$\text{df} = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{[1/(n_1 - 1)]\left(s_1^2/n_1\right)^2 + [1/(n_2 - 1)]\left(s_2^2/n_2\right)^2}$$

   This would be quite time consuming to calculate, and this expression is not very commonly calculated by hand. However, this formula is used in some computer programs, including SPSS and MINITAB. In Section 9.8 you will find the separate variance t value, along with a degrees of freedom obtained from this formula. This formula will not generally produce an integer, so the t value for the degrees of freedom which is closest to the value produced from this formula can be used to determine the critical values for the t test.

The following example gives a t test for a difference between two means where one of the sample sizes is quite small. The first of the two procedures for determining the degrees of freedom will be used. In Section 9.8 the example will again be given with the computer output provided for the same data. There the degrees of freedom associated with the second of the two formulas will be given.

Example 9.6.4 Differences in Explanations of Unemployment for Edmonton PCs and Liberals

Some data from the 1985 Edmonton Area Study, conducted by the Population Research Laboratory at the University of Alberta, was given in Example 8.4.2 in Chapter 8. Further data from this survey of Edmonton adults is given here. The explanation of unemployment used in this example is that respondents were asked whether they agreed or disagreed with the statement "Unemployment is high because trade unions have priced their members out of a job" The responses were given on a 7 point scale, with 1 being 'strongly disagree' and 7 being 'strongly agree.' The data in Table 9.15 is a random sample of 40 Conservative Party supporters and 9 Liberal Party supporters who were in the Study.

	Party Supported	
	PC	Liberal
Sample Mean	5.58	3.67
Sample Standard Deviation	1.20	2.29
Sample Size	40	9

Table 9.15: Summary Statistics for a Small Sample of PC and Liberals

Use the data in Table 9.15 to test whether there is a difference in mean opinion level of PC and Liberal supporters concerning whether trade unions have priced their members out of a job. Since PC supporters are often more antagonistic toward trade unions, and might be expected to agree more *strongly with this explanation, than are Liberal supporters, conduct a one tailed test. Use the 0.05 level of significance.*

Solution. *The data and the notation used for the test is summarized in Table 9.16.*

Characteristic	Population 1 PCs	Population 2 Liberals
True Mean	μ_1	μ_2
True Standard Deviation	σ_1	σ_2
True Variance	σ_1^2	σ_2^2
Sample Size	$n_1 = 40$	$n_2 = 9$
Sample Mean	$\bar{X}_1 = 5.58$	$\bar{X}_2 = 3.67$
Sample Standard Deviation	$s_1 = 1.20$	$s_2 = 2.29$
Sample Variance	$s_1^2 = 1.44$	$s_2^2 = 5.2441$

Table 9.16: Data and Notation for Test of Difference in PC and Liberal Opinions

The null hypothesis is that there is no difference between the means of the two populations. The alternative hypothesis is that the PCs are more likely to agree than are Liberals that unemployment is high because unions have priced members out of job. Since agreement is associated with larger values on the 7 point scale, and disagreement with smaller values, the mean opinion for the PCs would be expected to be greater for the PCs than for the Liberals. This produces the hypotheses

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 > 0$$

The test statistic is $\bar{X}_1 - \bar{X}_2$ *and the sample size of Liberals,* $n_2 = 9$ *is* *very small, so that the sampling distribution of* $\bar{X}_1 - \bar{X}_2$ *is approximated by*

$$\bar{X}_1 - \bar{X}_2 \text{ is } t \left(\mu_1 - \mu_2, \sigma_{\bar{X}_1 - \bar{X}_2}\right)$$

In this example, the two standard deviations are 1.20 for the PCs and 2.29 for the Liberals. These are considerably different, with one standard deviation being almost double the other and with an even larger gap between the variances. The separate variance t test should probably be used in this example. This means that

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

The first of the two methods for determining the degrees of freedom is used here, so that

$$df = \text{Minimum } (n_1 - 1, n_2 - 1) = n_2 - 1 = 8.$$

For a one tailed t test at the 0.05 level of significance, the critical t value is 1.860. All t values greater than 1.860 imply rejection of H_0*, and all values less than this mean non rejection of the null hypothesis.*

$$\begin{aligned} t &= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \\ &= \frac{5.58 - 3.67}{\sqrt{1.44/40 + 5.2441/9}} \\ &= \frac{1.91}{\sqrt{0.036 + 0.58268}} \\ &= \frac{1.91}{0.78656} \\ &= 2.428 > 1.860 \end{aligned}$$

This t value falls into the critical region for the test and at the 0.05 level of significance, the null hypothesis can be rejected. The difference in sample means is large enough so that even with the small sample of Liberals, and the small degrees of freedom, the mean level of PC opinion exceeds the mean level of Liberal opinion. Since a larger value means greater agreement, this *result implies that PCs express themselves as more in agreement than are Liberals, that unions have priced their members out of jobs.*

Additional Comments. *This conclusion depends on the assumptions for the test being satisfied. The assumption that the variances are unequal, and the small degrees of freedom, are very cautious assumptions. These assumptions make it more difficult to reject H_0, and even so this hypothesis has been rejected. With respect to these assumptions, if any error have been made, they have been on the cautious side, and the actual differences could be even greater than implied by this test.*

With respect to the randomness of the sample, so long as the Edmonton Area Study as a whole was close to being a random sample, this should not be a problem. The set of 40 PC supporters and 9 Liberal supporters was selected from the Study as a whole by using a random selection procedure in SPSS. The randomness of the sample as a whole implies independence of selection of Liberals and PCers.

One assumption which could cause an incorrect conclusion is the assumption that each of the two populations are normally distributed. Since these are opinion questions, collected on a discrete 7 point scale, they are unlikely to be exactly normally distributed. Further, the assumption that the scale is an interval scale is not satisfied. The responses are measured on an ordinal scale, and yet the mean and standard deviation have been used in this test. Technically this is illegitimate, but it is very commonly done with attitude scales. As long as small differences on these scales are not interpreted as meaning too much, there seems to be little problem with this.

In this example, these samples give large difference in sample means, and the results conform with what we know about PC and Liberal opinions more generally. That is, PC supporters tend to have a more negative view of the effect of trade unions than do Liberals. These results are supported by this test, and can be considered meaningful results.

9.7 Test for a Difference of Variances

All of the statistical inferences in Chapter 8 and to this point in Chapter 9 have concerned means or proportions. This section gives a test for two variances. This test is useful in deciding whether to use the separate or pooled variance t test of the last section, as well as for other purposes. In order to carry out this test, a new distribution, the F distribution is required.

This test requires selection of independent random samples from each of two populations, population 1 and population 2. For purposes of carrying out this test, it is assumed that each of these populations is a normally distributed population. Let the true standard deviations of these populations be σ_1 and σ_2, respectively. It will be seen that the test is more easily organized as a test for variances, rather than standard deviations. The true variances for the two populations are σ_1^2 and σ_2^2, respectively. A random sample of size n_1 is taken from population 1 and a random sample of size n_2 from population 2. The sample standard deviations from these samples are s_1 and s_2, respectively and the sample variances are s_1^2 and s_2^2, respectively.

The null hypothesis for this test is that the variances of the two populations are equal. The alternative hypothesis is that the two variances are not equal to each other. These hypotheses are

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

The alternative hypothesis is most often a two directional test, since all that may interest the researcher is whether the variances are equal or unequal.

Rather that conduct the test in this form, the ratio of the two variances is usually the form that the test statistic takes. If $\sigma_1^2 = \sigma_2^2$, then $\sigma_1^2/\sigma_2^2 = 1$ and if $\sigma_1^2 \neq \sigma_2^2$ then $\sigma_1^2/\sigma_2^2 \neq 1$. Thus the test is conducted by computing the ratio of the two sample variances, rather than examining the numerical difference between the two variances. In addition, the test is usually constructed so that the larger of the two sample variances is placed in the numerator, and the smaller of the two variances in the denominator.

In order to make the notation straightforward, suppose that the sample from population 1 has a larger sample standard deviation and variance than population 2. (If the latter is not the case, the populations can be renumbered so that population 1 always is the population with the larger variability.) Then the sample statistic is

$$F = \frac{s_1^2}{s_2^2}.$$

This ratio of sample variances has an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. The F distribution is briefly described on the next page.

Before discussing the F distribution, an outline of how this test can be completed is given here. If the null hypothesis of equal variances is correct, then the two sample variances should be relatively close to each other. That is, the ratio of these two sample variances should be relatively close to 1 if H_0 is correct. Since the larger of the two variances is always placed in the numerator, and the smaller variance in the denominator, the F value can never be less than 1. But when the two populations have variances which are quite close to each other, this F value will be only a little above 1.

When the alternative hypothesis of unequal variances is correct, then $F = s_1^2/s_2^2$ will be considerably larger than 1. Since the larger of the two sample variances is in the numerator, and the two sample variances are different, the more unequal the variances of the two populations, the larger the F value will be.

The logic of this test then is to reject the null hypothesis when $F = s_1^2/s_2^2$ is a lot larger than 1. When this F value is just a little above 1, the null hypothesis of equal variances is not rejected. Exactly how large the F value should be, before rejecting the null hypothesis, can be determined from the table of the F distribution. When the computer output is provided in the following section, the exact significance level for this F value will be given.

The F distribution. The F distribution is a theoretical probability distribution which has the general shape given in Figure 9.5. The F values are along the horizontal axis, beginning at 0 and going to $+\infty$ on the right. The vertical axis represents the probability for each F value.

One way of imagining the F distribution, without examining the mathematical formulas used to construct it, is to think of pairs of random samples from a population with variance σ^2. Suppose random sample 1 has sample size n_1 and variance s_1^2, and random sample 2 has sample size n_2 and variance s_2^2. Each of these sample variances can be considered as estimates of σ^2, but because of the randomness of sampling, the two sample variances will differ from sample to sample. If the population from which the samples are drawn is a normally distributed population, the ratio of these two sample variances can be described by the F distribution.

If random samples of size n_1 and n_2 are selected from a population which is normally distributed, the ratio of the sample variances s_1^2/s_2^2 has an F distribution with n_1 and n_2 degrees of freedom. That is,

$$F_{n_1,n_2} = \frac{s_1^2}{s_2^2}$$

This can be seen in an intuitive manner in the following way. If the two samples yield sample variances which are close to being equal, then

Figure 9.5: The F Distribution

$F = s_1^2/s_2^2 \approx 1$. If the sample variance from the first sample is less than that of the second sample, then $F = s_1^2/s_2^2 < 1$. Note that these are variances, and can never be less than zero, because a variance is a sum of squared values. If sample 2 happened to yield quite a large estimate of σ^2, and sample 1 a very small estimate, then this ratio could be close to 0. Now if sample 2 yields a fairly small estimate of σ^2, but sample 1 yields a large estimate, $F = s_1^2/s_2^2 > 1$, with no upper limit on this value. If s_2^2 is quite close to 0, or in the limit, actually equal to 0, then the F ratio could be an extremely large value. Thus the F distribution is asymmetrical, with 0 as the lower limit, 1 as the mean of s_1^2/s_2^2, but then skewed to the right, with no upper limit. The chance of obtaining very large F values is small, so that the F curve approaches the horizontal axis, before F becomes too large. The F distribution is thus considered to be asymptotic to the horizontal axis.

In addition, there are two degrees of freedom, one associated with the numerator and one associated with the denominator. Since there are many possible values for the degrees of freedom, there are many different F distributions. The F table in Appendix K contains only 3 F distributions, one for each of the significance levels 0.10, 0.05 and 0.01. These give the F values that are associated with each pair of degrees of freedom, such that only 0.10, 0.05 or 0.01 of the area under the F curve is in the right tail of the distribution. The degrees of freedom are the sample size for the numerator minus 1 and the sample size for the denominator minus 1. For example, if sample sizes of $n_1 = 6$ and $n_2 = 18$ were selected, the degrees of freedom would be 5 and 17. The degrees of freedom for the numerator are given in the top row. Go across this row to the column headed 5, and then go down to the row labelled 17 in the left column. At the 0.10 level of significance, $F = 2.22$. This can be written

$$F_{5,17;0.10} = 2.22.$$

This means that for an F distribution with 5 and 17 degrees of freedom, there is exactly 0.10 area under the curve to the right of $F = 2.22$. For the same degrees of freedom, at 0.05 significance,

$$F_{5,17;0.05} = 2.81$$

and at 0.01 significance,

$$F_{5,17;0.01} = 4.34.$$

The two examples used in the pooled and separate variance t tests are used here to illustrate how tests for the equality of two variances can be conducted.

Example 9.7.1 Variation in Self-Esteem of Elderly Women

The t test for a difference in the self-esteem levels of elderly women in Example 9.6.3 used the pooled variance approach. This is justified if the variances of the two populations are close to being equal. For the sample of 28 elderly women who lived in their own homes, the standard deviation on the self-esteem scale was 15.0, and the variance was 225.0. For the 25 nursing home women, the standard deviation was 14.4, with a variance of 207.36 on the self-esteem scale. While these standard deviations and variances seem to be close to being equal, test for the equality of the two variances. Use the 0.10 level of significance.

Solution. *The data is summarized in Table 9.17 in the format for a test for equality or inequality of two variances. Note that the information concerning the means has not been included in this table because it is not required for this test.*

Characteristic	Population 1 Community	Population 2 Nursing Home
True Standard Deviation	σ_1	σ_2
True Variance	σ_1^2	σ_2^2
Sample Size	$n_1 = 28$	$n_2 = 25$
Sample Standard Deviation	$s_1 = 15.0$	$s_2 = 14.4$
Sample Variance	$s_1^2 = 225.0$	$s_2^2 = 207.36$

Table 9.17: Variation in Self-Esteem of Elderly Women

Using the notation in Table 9.17, the null and research hypotheses are as follows.

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Note that population 1, the community women, has a larger standard deviation and variance than does population 2. This notation conforms to the setup for the test, where population 1 is defined as the population with the larger variability, and population 2 is defined as the population with lower variability.

The test statistic is the ratio of the two sample variances s_1^2/s_2^2. *As described above, this statistic has an F distribution with* $n_1 - 1$ *degrees of freedom for the denominator, and* $n_2 - 1$ *degrees of freedom for the numerator. That is*

$$F_{n_1-1,n_2-1} = \frac{s_1^2}{s_2^2}.$$

Since $n_1 = 28$ *and* $n_2 = 25$*, the degrees of freedom for this test are 27 and 24. At the 0.10 level of significance, the critical F value is approximately 1.70. That is,*

$$F_{27,24;0.10} \approx 1.70.$$

This means that if the ratio of the variances exceeds 1.70, then the null hypothesis of equal variances is rejected. If this ratio is less than 1.70, then the hypothesis of equal variances is not rejected.

From Table 9.17, the respective variances are $s_1^2 = 225.0$ *and* $s_2^2 = 207.36$ *and the F statistic for these variances is*

$$F = \frac{s_1^2}{s_2^2} = \frac{225.0}{207.36} = 1.09 < 1.70.$$

This ratio of variances produces an F value that is less than the critical F value of 1.70. The conclusion of the test is that the null hypothesis of equal variances cannot be rejected at the 0.10 level of significance.

Additional Comments.

There is the possibility of Type II error here, that the null hypothesis is not true, even though it could not be rejected. It is quite likely that the two populations do have somewhat different variances. But the F value is so close to 1, that these samples tend to support the view that the two variances are very close to each other. Remember that $F = 1$ *is the minimum possible value for the F statistic in this test.*

The assumption that the variances of the two populations are very close to each other seems justified. In Example 9.6.3 the pooled variance t test was used. This assumes that the two variances are close to being equal, so that the common value of the two variances could be used to estimate the standard deviation of the difference in the two sample means. Based on the results of this test of variances, the pooled variance t test seems to be the appropriate test.

Example 9.7.2 Variability of Opinions in Edmonton Area Study

Example 9.6.4 found that there was a statistically significant difference between the mean opinion level of Edmonton PCs and Liberals concerning whether trade unions were responsible for unemployment. In carrying out that test, the separate variance t test was used because the two variances appeared to be different than each other. The standard deviation of the opinion scale for the 40 PCs sampled was 1.20 and the standard deviation for the 9 Liberals was 2.29. Using this information, test whether the variance of opinions differs for Liberal and PC respondents. Test at the 0.01 level of significance.

Solution. *Table 9.18 contains the summary data for this example using the two population notation required for the test. Note that the populations have been redefined so the population with the larger variability (the Liberals) becomes population 1. For purposes of this test, population 2 is the PCs, since the variability in opinions is less for the PCs.*

Characteristic	Population 1 Liberals	Population 2 PCs
True Standard Deviation	σ_1	σ_2
True Variance	σ_1^2	σ_2^2
Sample Size	$n_1 = 9$	$n_2 = 40$
Sample Standard Deviation	$s_1 = 2.29$	$s_2 = 2.20$
Sample Variance	$s_1^2 = 5.2441$	$s_2^2 = 1.44$

Table 9.18: Variation in Opinions of Edmonton PCs and Liberals

If σ_1^2 is the true variance in opinion for all Edmonton PCs, and σ_2^2 is the variance for all Edmonton Liberals, the null and research hypotheses are

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

The test statistic is the ratio of the two sample variances s_1^2/s_2^2 and

$$F_{n_1-1,n_2-1} = \frac{s_1^2}{s_2^2}.$$

Since $n_1 = 9$ and $n_2 = 40$, the degrees of freedom for this test are 8 and 39. At the 0.01 level of significance, the critical F value is approximately 2.99. That is,

$$F_{8,39;0.05} \approx 2.99.$$

This means that if the ratio of the variances exceeds 2.99, then the null hypothesis of equal variances is rejected. If this ratio is less than 2.99, then the hypothesis of equal variances is not rejected.

Figure 9.6: F Distribution for Test of Equality of Variances

The F distribution is shown in Figure 9.6. The critical region for the test is the right $\alpha = 0.05$ *of the distribution. The critical value for the test is* $F = 2.99$*, and there is exactly 0.05 of the area under the F curve which lies to the right of* $F = 2.99$*.*

From Table 9.18, the respective variances are $s_1^2 = 5.2441$ *and* $s_2^2 = 1.44$*. The F statistic for these variances is*

$$F = \frac{s_1^2}{s_2^2} = \frac{5.2441}{1.44} = 3.64 > 2.99.$$

This ratio of variances produces an F value that is greater than the critical F value of 2.99. In Figure 9.6, the F value of 3.64 is right of the critical value of $F = 2.99$*. This means that the data gives an F value which lies in the critical region for the test, and the null hypothesis of equal variances can be rejected at the 0.01 level of significance. At 0.01 significance, it can be concluded that the two population variances differ.*

Using the separate variance t test for Example 9.6.4 appears to have been the proper choice. The sample standard deviations and variances of *the two groups are sufficiently different to reject the equality hypothesis at the 0.01 level of significance. Since this is such a low level of significance, the evidence is very strong that the Liberals and PCs have different variability in opinions.*