Tests of Hypotheses

8

Chapter Outline

In this chapter, we formally introduce the testing of hypotheses, define key terms to help us succinctly communicate the ideas of hypothesis testing, and show how to conduct tests. The statistical ideas used in the tests of hypotheses share the same roots with those used in confidence intervals presented in the previous chapter. Therefore, we do not repeat the details on the distributions of the test statistics that we presented in Chapter 7.

8.1   Preliminaries in Tests of Hypotheses

Hypothesis testing is a way of organizing and presenting evidence that helps us reach a decision. Although the confidence interval and the test of hypothesis can be used to reach the same conclusion, their emphases are different. The confidence interval provides limits that are likely to contain the parameter. These limits can also be used to test a hypothesis, but that is not necessarily the reason why they were created. The test of hypothesis aids in reaching a decision about whether or not we believe that the hypothesized value of the parameter is correct. The use of the test of hypothesis also provides additional information about the decision that is not provided with the confidence interval. Example 8.1 illustrates the basic idea.

Example 8.1

Let us consider a situation associated with the decision to proceed with the marketing of a new drug for reducing cholesterol. This decision was reached because it is unlikely that the greater mean reduction of serum cholesterol observed in a sample of patients receiving a new drug, when compared to the reduction achieved for a sample of patients who received the standard treatment, was due to chance. Or the decision may be for the local health department to allocate more resources to an immunization campaign for childhood diseases. This decision was reached because,

based on the sample proportion immunized, it is unlikely that the proportion of 5-year-old children in the community that have the required immunizations equals the targeted value of 95 percent.

There are negative outcomes associated with making a wrong decision, and these must be weighed carefully. If the decision to market the new drug is wrong — that is, it is not an improvement over the standard treatment — patients may pay more money for no additional benefit or for a treatment that does not work. However, if the decision were not to market and the drug was better, patients would lose by not having access to a better treatment, and the company would lose because it did not realize the profit from this drug. If the health department's decision to conduct an immunization campaign is wrong — that is, the proportion of 5-year-old children immunized in the community is at least 95 percent — scarce resources would be misdirected. Other needy programs would not receive additional resources. However, if the decision were not to conduct the campaign when it was needed, there would be increased risk of unnecessary disease in preschool children.

We use another example to clarify these notions and to lead into the definitions used in tests of hypotheses.

Example 8.2

Suppose two diets are proposed for losing weight. We have 12 pairs of individuals, matched on age (±5 years), sex, initial weight (±10 pounds), and level of exercise. One member of the pair is assigned at random to diet 1 and the other member is assigned to diet 2. Individuals remain on their diets for six weeks and are then reweighed. We wish to determine whether or not the diets are equivalent from a weight loss perspective. Table 8.1 shows how the data — the weight losses for those on diets 1 and 2 and the within pair difference — may be presented.

There are several ways of analyzing these data. We demonstrate a very simple approach here and other approaches will be shown later. We shall examine the proportion of pairs in which the person on diet 1 had the greater weight loss. If the diets do not differ with regards to weight loss, assuming there are no ties in weight loss, the proportion should be 0.50. Deviations from 0.50 suggest that there is a difference in the diets in terms of weight loss. If there are ties in the weight losses, the hypothesis being tested is that the proportion of pairs in which the person on diet 1 had the greater weight loss is the same as the proportion of pairs in which the person on diet 2 had the greater weight loss. Note that we have converted the hypothesis in words into something that we can deal with analytically.

Table 8.1 Weight losses (pounds) by diet for 12 pairs of individuals.

Diet						Pairs						
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
2	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}
Difference	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}	d_{12}

8.1.1   Terms Used in Hypothesis Testing

The statistical terms in hypothesis testing are defined, and their underlying concepts are explained following, based on Example 8.2.

Null and Alternative Hypotheses: The hypothesis being tested is called the null hypothesis and is denoted by H_0. The *null hypothesis* is that π, the proportion of pairs in the population for which persons on diet 1 would show the greater weight loss, is 0.50. The *alternative hypothesis*, denoted by H_a or H_1, to the null hypothesis is that π is not equal to 0.50. In symbols, these hypotheses are

$$H_0: \pi = 0.50 \quad \text{and} \quad H_a: \pi \neq 0.50.$$

We either reject or fail to reject the null hypothesis. If we reject the null hypothesis, we are expressing a belief that the alternative hypothesis is true. If there are ties in the weight losses, the alternative hypothesis is that the proportion of pairs in which the person on diet 1 had the greater weight loss is not equal to the proportion of pairs in which the person on diet 2 had the greater weight loss.

Type I and Type II Errors: If we reject the null hypothesis in favor of the alternative hypothesis, there are two possible outcomes. Either we have correctly rejected the null hypothesis or we have falsely rejected it. Falsely rejecting the null hypothesis is called a Type I error. In this example, the Type I error is claiming that the proportion of pairs for which diet 1 showed the greater weight loss is not equal to 0.50 when, in fact, it is 0.50.

If we fail to reject the null hypothesis, again there are two possible outcomes. Either we have failed to reject the null hypothesis when it should have been rejected or we have correctly failed to reject the null hypothesis. Failing to reject the null hypothesis when it should have been rejected is called a Type II error. The Type II error in this example is claiming that the proportion of pairs for which diet 1 showed the greater weight loss is 0.50 when, in fact, the proportion is different from 0.50. Figure 8.1 shows these four possibilities. The probability of a Type I error is usually labeled α, and the probability of a Type II error is usually labeled β. Ideally we would like to keep both of these probabilities as small as possible, although we usually focus more on the Type I error and its probability.

Our Decision about the Null Hypothesis	Reality:Null Hypothesis Is	
	True	False
True	**Good**	**Type II Error**
False	**Type I Error**	**Good**

Figure 8.1 Possibilities associated with a test of hypothesis.

The Test Statistic: The *test statistic*, the basis for the test of hypothesis, is the number of pairs out of the 12 sample pairs for which those on diet 1 achieved the greater weight loss. Equivalently, the observed sample proportion of pairs for which those on diet 1 achieved the greater weight loss, p, could be used. The test is based on the sign of the difference and, therefore, this particular test is called the sign test. Now that we know

what hypothesis is to be tested and what test statistic is to be used, we must specify the decision rule to be used.

8.1.2 Determination of the Decision Rule

The decision rule specifies which values of the test statistic (or some function of it) will cause us to reject the null hypothesis in favor of the alternative hypothesis. The decision rule is based on the probabilities of the Type I and II errors. The probabilities of Type I and Type II errors are found from consideration of the distribution of the test statistic. In this example, the test statistic follows the binomial distribution. The binomial is used because there are only two outcomes: diet 1 is better or diet 2 is better (again ignoring the possibility of a tie in weight loss). We begin by assuming that the null hypothesis is true — that is, π is 0.50. Because we know that n is 12, we know both parameters of the binomial distribution. The probability distribution of the possible outcomes is shown in the following table and in Figure 8.2.

No. of Times Diet 1 Is Better	Probability	No. of Times Diet 1 Is Better	Probability
0	0.0002	7	0.1934
1	0.0030	8	0.1208
2	0.0161	9	0.0537
3	0.0537	10	0.0161
4	0.1208	11	0.0030
5	0.1934	12	0.0002
6	0.2256		

What values of the test statistic would cause us to reject the null hypothesis that π is 0.50 in favor of the alternative hypothesis? Large deviations from six pairs for which diet 1 was better — that is, large deviations from π of 0.50 — are suggestive that the diets have different effects. Thus, either very large or very small values of the test statistic would cause us to question the null hypothesis. As we can see from Figure 8.2, it is highly unlikely to observe either very large or very small values of the test statistic if π is really 0.50.

One- and Two-Sided Tests: The test we are considering is called a *two-sided test*, since either large or small values of the test statistic cause us to question the truth of
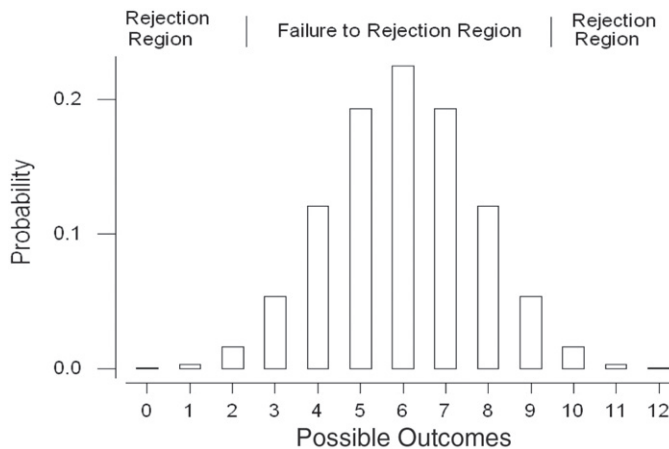

Figure 8.2 Bar chart showing the binomial probability distribution for $n = 12$ and $\pi = 0.5$.

the null hypothesis. A *one-sided test* occurs when only values in one direction cause us to question the null hypothesis. For example, if we were the developers of diet 1, we might only be interested in whether or not diet 1 was better than diet 2, not whether or not it was worse than diet 2. If this were the situation, the null hypothesis remains that π is equal to 0.50, but the alternative hypothesis becomes that π is greater than 0.50. In symbols, this is

$$H_0:\ \pi = 0.50 \quad \text{versus} \quad H_a:\ \pi > 0.50.$$

In this case, only large values of the test statistic would cause us to reject the null hypothesis in favor of the alternative hypothesis.

Use of a one-sided test makes it easier to detect departures from the null hypothesis in the indicated direction — that is, π greater than 0.50. However, the use of a one-sided test means that if the departure is in the other direction — that is, π is less than 0.50 — it won't be detected.

Calculation of the Probabilities of Type I and Type II Errors: Suppose that we decide to reject the null hypothesis whenever we observe a test statistic of 0 or 12 pairs — that is, the values of 0 and 12 form the rejection or critical region. The values from 1 to 11 then form the failure to reject or acceptance region. The probability of a Type I error, is thus the probability of observing 0 or 12 pairs in which diet 1 had the greater weight loss when π is actually 0.50. From the preceding probability mass function, we see that is 0.0004. That's great! There is almost no chance of making this error, and this is almost as small as we can make it. Of course, we could decide never to reject the null hypothesis, and, then, there would be zero probability of a Type I error. That is unrealistic, however.

We are pleased with this decision rule because it has an extremely small probability of a Type I error. However, what is the value of β, the probability of a Type II error, associated with this decision rule? To be able to calculate β, we have to be more specific about the alternative hypothesis. The preceding alternative hypothesis is quite general in that it only says π is not equal to 0.50. However, just as we used a specific value, the value 0.50, for π in calculating the probability of a Type I error, we must specify a value of π other than 0.50 to be used in calculating the probability of a Type II error. We must move from the general alternative to a specific alternative hypothesis to be able to calculate a value for β. This means that there is not merely one β associated with the decision rule; rather, there is a value of β corresponding to each alternative hypothesis.

What specific value of π should be used in the alternative hypothesis? We should have little interest in the alternative that π is 0.51 instead of the null hypothesis value of 0.50. The difference between 0.51 and 0.50 is of little practical interest. For all practical intent, if π is really 0.51, there is little difference in the diets. As the value of π departs more and more from 0.50, the ability to detect these departures becomes more important. We may not all agree at which point π differs enough from 0.50 to be important. Some may say 0.60 is different enough, whereas others may say that π must be at least 0.70 for the difference to be important. Most would certainly agree that we should reject the equality of the diets if diet 1 provides for greater weight loss in 80 percent of the pairs.

Let us assume that π is really 0.80, not 0.50, and find the value for β. The binomial distribution for an n of 12 and a proportion of 0.80 is shown below.

No. of Times Diet 1 Is Better	Probability	No. of Times Diet 1 Is Better	Probability
0	0.0000	7	0.0532
1	0.0000	8	0.1328
2	0.0000	9	0.2363
3	0.0001	10	0.2834
4	0.0005	11	0.2062
5	0.0033	12	0.0687
6	0.0155		

Type II error is failing to reject the null hypothesis when it should be rejected. Since our decision rule is to reject only when we observe a test statistic of 0 or 12, we will fail to reject for the values of 1 through 11. The probability of 1 through 11 when π is actually 0.80 is 0.9313 (= 1 − 0.0000 − 0.0687). Therefore, use of this decision rule yields an α of 0.0004 and a β of 0.9313. The probability of the Type I error is very small, but the probability of the Type II error, corresponding to the value of 0.80 for π, is quite large.

8.1.3   Relationship of the Decision Rule, α and β

If we change our decision rule to reject the null hypothesis more often, we will increase α but decrease β — that is, there is an inverse relation between α and β. For example, if we increase the rejection region by including values 1 and 11 in addition to 0 and 12, the value of α becomes 0.0064 (= 0.0002 + 0.0030 + 0.0030 + 0.0002). These probabilities are found from the probability distribution based on the value for π of 0.50. The new value for β, based on this expansion of the rejection region, and using 0.80 for π, is 0.7251 (= 1 − 0.2062 − 0.0687). The probability of a Type I error remains quite small, but the probability of a Type II error is still large.

If the decision rule is to reject for values of the test statistic of 0 to 2 and 10 to 12, then α's value is increased to 0.0386 (= 2 [0.0161 + 0.0030 + 0.0002]) and the value of β is reduced to 0.4417 (= 1 − 0.0687 − 0.2062 − 0.2834). The probability of a Type I error is still reasonable, whereas, although the probability of Type II error is much smaller than previously, it is still quite large. However, a further change in the decision rule to include the values of the test statistic of 3 and 9 increases the value of α to 0.1460 (= 0.0386 + 2[0.0537]), which is now becoming large.

What Are Reasonable Values for α and β? There are no absolute values that indicate that the probability of error is too large. It is a matter of personal choice, although convention suggests that an α greater than 0.10 is unacceptable. Most investigators set α to 0.05, and some set it to 0.01. There is less guidance for the choice of β. It again is a matter of personal choice. However, the implications of the Type II error play a role in how large a β can be tolerated. A value of 0.20 for β is used frequently in the literature. Investigators often ignore the Type II error because (1) the hypothesis has been framed such that the Type I error is of much greater interest than the Type II error, or (2) it is often difficult to find the value of β.

Ways to Decrease β Without Increasing α: We were in a bind when we left the example above. The value of β was too large and, if we tried to reduce it by further

enlargement of the rejection region, we made α too large. One way of decreasing β without increasing α is to change the alternative hypothesis or to increase the sample size.

(1) Changing Alternative Hypothesis: The specific alternative hypothesis that we had used previously in calculating β was that π was equal to 0.80. We selected the value of 0.80 because if diet 1 performed better for 80 percent of the pairs, we believed this indicated a very important difference between the diets. If we are willing to change what we consider to be a very important difference, we can reduce β. For example, by increasing the value of π in the alternative hypothesis from 0.80 to 0.90, β will decrease. However, this means that we no longer consider it to be important to detect that π was really 80 percent instead of the hypothesized 50 percent. We will focus on the test's ability to detect a very large difference — that is, the difference between 0.90 and 0.50 — and not worry that the test has a small chance of detecting smaller differences.

The following shows the probability mass function for the binomial with a sample size of 12 and a proportion of 0.90.

No. of Times Diet 1 Is Better	Probability	No. of Times Diet 1 Is Better	Probability
0	0.0000	7	**0.0038**
1	0.0000	8	**0.0213**
2	0.0000	9	**0.0853**
3	**0.0000**	10	0.2301
4	**0.0000**	11	0.3766
5	**0.0001**	12	0.2824
6	**0.0004**		

If we again use the rejection region of 0 to 2 and 10 to 12, the probability of the Type I error is still 0.0386, since that was calculated based on π being 0.50. However, β is the probability of not rejecting that π is 0.50 when it is actually 0.90. This probability is the sum of the probabilities of the outcomes 3 through 9 in the preceding distribution, and that sum is 0.1109. Now both the values of α and β are reasonable.

By changing the alternative hypothesis, we have not changed the value of β for the alternative of π being 0.80. The β-value corresponding to a π of 0.80 and a rejection region of 0 to 2 and 10 to 12 remains 0.4417. What has changed is what we consider to be an important difference. If a lesser difference is considered to be important, the probability of the Type II error for that value of π can be calculated. Table 8.2 shows

Table 8.2   Probability of Type II error and power for specific alternative hypotheses based on a rejection region of 0 to 2 and 10 to 12.

Alternative Hypothesis	Probability of Type II Error	Power
$\pi = 0.55$	0.9507	0.0493
$\pi = 0.60$	0.9137	0.0863
$\pi = 0.65$	0.8478	0.1522
$\pi = 0.70$	0.7470	0.2530
$\pi = 0.75$	0.5778	0.4222
$\pi = 0.80$	0.4416	0.5584
$\pi = 0.85$	0.2642	0.7358
$\pi = 0.90$	0.1109	0.8891
$\pi = 0.95$	0.0195	0.9805

the values of the Type II errors for several values of the alternative hypothesis based on a rejection region of 0 to 2 and 10 to 12.

The probability of a Type II error decreases as the value of π used in the alternative hypothesis moves farther away from its value in the null hypothesis. This makes sense, since it should be easier to detect greater differences than smaller ones. As this table shows, there is a very high chance of failing to detect departures from 0.50 less than 0.30 to 0.35 in magnitude. The last column in Table 8.2 is *power*, the probability of rejecting the null hypothesis when it should be rejected — that is, when the alternative hypothesis is true. From the table we can see that power is 1 minus β. Power is often used in the literature when discussing the properties of a test statistic instead of using the probability of a Type II error. From the values in Table 8.2, it is possible to create a *power curve* — that is, to graph the values of power versus the values of π used in the alternative hypothesis. Figure 8.3 shows a portion of the power curve for values of π greater than 0.50. Statisticians use power curves to compare different test statistics.
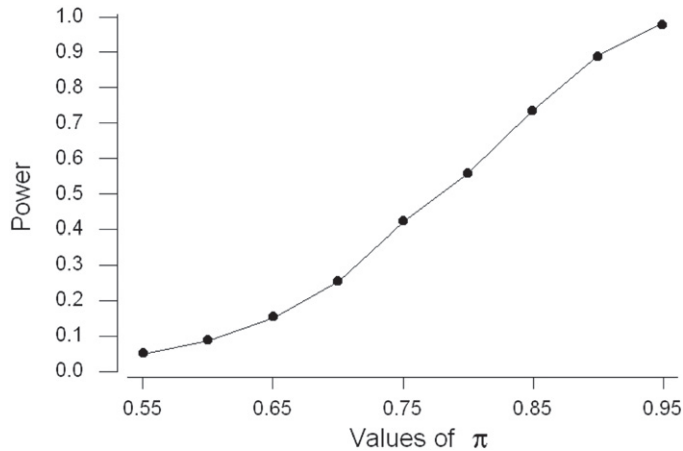

Figure 8.3   Portion of the power curve for π values greater than 0.5.

The preceding trade-off as a way of reducing β may not be very satisfactory. We still may feel that 80 percent is very different from 50 percent. As an alternative, we could increase the sample size instead of changing the alternative hypothesis.

(2) Increasing the Sample Size: None of the calculations shown so far have required the observed sample data. All these calculations are preliminary to the actual collection of the data. Therefore, if the probabilities of errors are too large, we can still change the experiment. As just mentioned, increasing the sample size is one way of decreasing the error probabilities, but doing this increases the resources required to perform the experiment. There is a trade-off between the sample size and the error probabilities.

Suppose we can afford to find and follow 15 pairs instead of the 12 pairs we initially intended to use. We still use the binomial distribution in the calculation of the error probabilities where π remains 0.50, but now n is equal to 15. The binomial probability mass function with these parameters is shown next.

No. of Times Diet 1 Is Better	Probability	No. of Times Diet 1 Is Better	Probability
0	**0.0000**	8	0.1964
1	**0.0005**	9	0.1527
2	**0.0032**	10	0.0917
3	**0.0139**	11	0.0416
4	0.0416	12	**0.0139**
5	0.0917	13	**0.0032**
6	0.1527	14	**0.0005**
7	0.1964	15	**0.0000**

Let us use a rejection region of 0 to 3 and 12 to 15. If we do this, the probability of a Type I error is 0.0352 (= 2 [0.0005 + 0.0032 + 0.0139]). The probability of a Type II error, based on the alternative that π is 0.80, uses the binomial distribution with parameters 15 and 0.80 and this probability mass function is now shown.

No. of Times Diet 1 Is Better	Probability	No. of Times Diet 1 Is Better	Probability
0	0.0000	8	**0.0139**
1	0.0000	9	**0.0430**
2	0.0000	10	**0.1031**
3	0.0000	11	**0.1876**
4	0.0000	12	0.2502
5	**0.0001**	13	0.2309
6	**0.0007**	14	0.1319
7	**0.0034**	15	0.0352

The probability of failing to reject a null hypothesis when it should be rejected — that is, of being in the acceptance region (values 4 to 11), when π is 0.80 — is 0.3518 (= 0.0001 + 0.0007 + 0.0034 + 0.0139 + 0.0430 + 0.1031 + 0.1876). The probability of a Type I error, 0.0352, is similar to its preceding value, 0.0386, when we considered this same alternative hypothesis. The probability of a Type II error has decreased from 0.4417 above when n was 12 to 0.3518 now for an n of 15. A further increase in the sample size can reduce β to a more acceptable level. For example, when n is 20, use of values 0 to 5 and 15 to 20 for the rejection region leads to an α of 0.0414 and a β of 0.1958.

8.1.4   Conducting the Test

The procedure used in conducting a test begins with a specification of the null and alternative hypotheses. In this example, they are

$$H_0: \pi = 0.50 \quad \text{versus} \quad H_a: \pi \neq 0.50.$$

We must decide on the significance level to be used in conducting the test. The *significance level* is the probability of a Type I error that we are willing to accept. We use the conventional significance level of 0.05 in this example.

Based on the preceding calculations, we have decided to increase the sample size to 20. We will reject the null hypothesis if the value of the test statistic is from 0 to 5 or from 15 to 20. Use of this sample size and decision rule keeps the probability of a Type I error less than 0.05 and also keeps β reasonably small when considering large departures from the null hypothesis. With discrete data, the probability of a Type I error usually does not equal the significance level exactly. The decision rule used with discrete

Table 8.3   Weight losses (pounds) by diet for 20 pairs of individuals.

Pairs	Diet 1	Diet 2	Difference
1	9	7	2
2	4	6	−2
3	11	9	2
4	7	12	−5
5	−4	3	−7
6	13	8	5
7	6	5	1
8	3	−1	4
9	8	14	−6
10	10	8	2
11	8	6	2
12	9	8	1
13	14	15	−1
14	11	7	4
15	5	7	−2
16	−3	4	−7
17	6	−2	8
18	7	4	3
19	13	10	3
20	9	5	4

data is chosen so that it results in a probability of a Type I error being as close as possible to and less than the desired significance level. The data are collected and shown in Table 8.3.

There are 13 pairs for which persons on diet 1 had the greater weight loss. As 13 does not fall into the rejection region of 0 to 5 or 15 to 20, we fail to reject the null hypothesis in favor of the alternative hypothesis at the 0.05 significance level. The observed result is not statistically significant.

***The* p-*value*:** Another statistic often reported is the *p-value* of the test, the probability of a Type I error associated with the smallest rejection region that includes the observed value of the test statistic. Another way of stating this is that the *p*-value is the level at which the observed result would just be statistically significant. In this example, since we are conducting a two-sided test, the smallest rejection region including the observed result of 13 is the region from 0 to 7 and 13 to 20. Examination of Table B2 for an *n* of 20 and a π of 0.50 yields a probability of being in this region of 0.2632 (= 2[0.0370 + 0.0739] + 0.0414). The value of 0.0414 is the value associated with the region from 0 to 5 and 15 to 20, and to that we have added the probabilities associated with the outcomes 6, 7, 13, and 14. The *p*-value is thus 0.2632.

Some statisticians do not believe in the decision rule approach to testing hypotheses. They believe that the *p*-value provides information regardless of whether or not the hypothesis is rejected. The *p*-value tells how likely the observed result is, assuming that the null hypothesis is true. For example, these statisticians see little difference in *p*-values of 0.05001 and 0.04999, although in the first case we would fail to reject the null hypothesis at the 0.05 significance level, whereas in the second case we would reject the null hypothesis. For these statisticians, the key information to be obtained from the study is that there is roughly 1 chance in 20 that we would have obtained the observed result if the null hypothesis were true. Using the *p*-value in this way is very reasonable.

8.2   Testing Hypotheses about the Mean

Suppose we wish to analyze the Digoxin clinical trial data shown in Table 3.1. However, before performing the analyses, we wish to determine whether or not the population represented by the sample of 200 patients differs from the national adult population as far as systolic blood pressure is concerned. Therefore, we first test the hypothesis that the mean systolic blood pressure for the patients in the Digoxin clinical trial is the same as the national average.

From the calculations in Chapter 3, we know that the sample mean, based on 199 patients (one missing value), is 125.8 mmHg. Based on national data (Lee and Forthofer 2006), we take the national average to be 122.3 mmHg. The test of hypothesis about the population mean, just like the confidence interval, uses the normal distribution if the population variance is known or the t distribution if the variance is unknown. We first assume that the variance is known.

8.2.1   Known Variance

In Chapter 7, when we formed the 95 percent confidence interval for the population mean, we assumed that the population standard deviation was 20 mmHg or that the variance was 400 mmHg. We shall use that value in the test of hypothesis about the population mean. The null and alternative hypotheses are

$$H_0: \mu = \mu_0 \quad \text{and} \quad H_a: \mu \neq \mu_0$$

where μ_0 is 122.3 mmHg in this example. To be able to compare the test results with the confidence interval from Chapter 7, we conduct the test at the 0.05 significance level.

The test statistic is $z \ (= [\bar{x} - \mu_0]/[\sigma/\sqrt{n}\,])$, the standard normal statistic. If the null hypothesis is true, z will follow the standard normal distribution. The rejection region is thus defined in terms of percentiles of the standard normal distribution. For a two-sided alternative, if z is either less than or equal to $z_{\alpha/2}$ or greater than or equal to $z_{1-\alpha/2}$, we reject the null hypothesis in favor of the alternative hypothesis. In symbols, this is

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \quad \text{or} \quad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha/2}.$$

If the test statistic is not in the rejection region — that is,

$$z_{\alpha/2} < z < z_{1-\alpha/2}$$

we fail to reject the null hypothesis in favor of the alternative hypothesis. Let us calculate the test statistic for the systolic blood pressure. The z value is

$$\frac{125.8 - 122.3}{20/\sqrt{199}} = 2.47.$$

Since 2.47 falls in the rejection region — that is, it is greater than 1.96 ($= z_{1-0.025}$) or is less than −1.96 — we reject the null hypothesis. This situation is shown pictorially in Figure 8.4. The p-value for this test is the probability of observing a standard normal variable with either a value greater than 2.47 or less than −2.47. This probability is found to be 0.014.
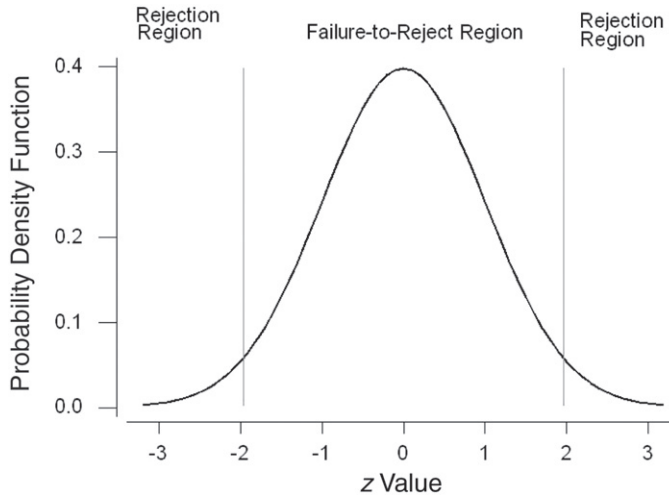
Figure 8.4
Representation of the rejection and failure-to-reject regions in terms of the standard normal statistic for $\alpha = 0.05$.

Equivalence of Confidence Intervals and Tests of Hypotheses: Recall that in Chapter 7 when we found the $(1 - \alpha) * 100$ percent confidence interval for the population mean, we started with the expression

$$\Pr\left\{-z_{1-\alpha/2} < \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right\} = 1 - \alpha.$$

We manipulated this expression, and we obtained the following expression:

$$\overline{x} - z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \overline{x} + z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right).$$

If we replace μ in the middle portion of the preceding first expression by μ_0, the middle portion is the z statistic for testing the hypothesis that μ equals μ_0. Since the confidence interval was derived from this test statistic, this means that if μ_0 is contained in the confidence interval, then the corresponding z statistic must also be in the failure to reject (acceptance) region. If μ_0 is not in the confidence interval, then the z statistic is in the rejection region — that is, it is less than or equal to $-z_{1-\alpha/2}$ or greater than or equal to $z_{1-\alpha/2}$.

In this case, the hypothesized value of 122.3 mmHg is not contained in the 95 percent confidence interval for the population mean. We saw in Chapter 7 (Example 7.1) that the confidence interval ranges from 123.0 to 128.6 mmHg. Therefore, we know that the test statistic will be in the rejection region, and, hence, we will reject the null hypothesis. In addition, using the same logic, from the confidence interval, we know we would reject the null hypothesis for any μ_0 not in the range from 123.0 to 128.6 mmHg.

This same type of argument for the linkage of the test of hypothesis and the corresponding confidence interval can be used with the other tests of hypotheses presented in this chapter. Thus, the confidence interval is also very useful from a test of hypothesis perspective. However, the confidence interval does not provide the p-value of the test, also a useful statistic.

One-Sided Alternative Hypothesis: If we are concerned only when the patients have elevated blood pressure, the null and alternative hypotheses are

$$H_0: \mu = \mu_0 \quad \text{and} \quad H_a: \mu > \mu_0.$$

The test statistic does not change, but the rejection region is a one-sided region now. We reject the null hypothesis in favor of the alternative hypothesis if z is greater than or equal to $z_{1-\alpha}$, or equivalently, if $\overline{x}$ is greater than or equal to $\mu_0 + [z_{1-\alpha}(\sigma/\sqrt{n})]$. A one-sided rejection region is shown in Figure 8.5 in terms of the z test statistic.
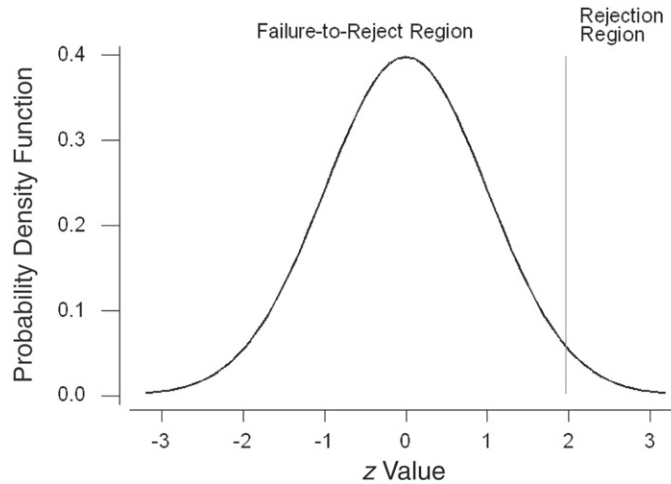

Figure 8.5
Representation of the rejection and failure-to-reject regions for a one-sided alternative of greater than the mean for $\alpha = 0.05$.

If we are concerned only when the patient's blood pressure is too low, the null and alternative hypotheses are

$$H_0: \mu = \mu_0 \quad \text{and} \quad H_a: \mu < \mu_0.$$

We now reject if z is less than or equal to z_α, or equivalently, if $\overline{x}$ is less than or equal to $\mu_0 + [z_\alpha (\sigma/\sqrt{n})]$. In this case, the rejection region in Figure 8.5 moves to the lower end of the distribution.

Power of the Test: Before collecting the data, suppose that we wanted to be confident that, if the systolic blood pressure of patients in the Digoxin clinical trial was substantially more than the national average, we could detect this higher mean blood pressure. By substantially more, we mean 3 percent or more above the national average of 122.3 mmHg. Thus, we wish to conclude that there is a difference between the study subjects and the national average if the study subjects have a population mean of 126.0 mmHg or more. The use of 3 percent is subjective and other values could be used.

The null and alternative hypotheses for this situation are

$$H_0: \mu = \mu_0 \quad \text{and} \quad H_a: \mu > \mu_0.$$

We use a significance level of 0.01. Thus, the rejection region includes all z more than or equal to $z_{0.99}$ — that is, z greater than or equal to 2.326. In terms of $\overline{x}$ the rejection region includes all values of $\overline{x}$ greater than or equal to

$$\mu_0 + z_{1-0.01}\left(\frac{\sigma}{\sqrt{n}}\right) = 122.3 + 2.326\left(\frac{20}{\sqrt{199}}\right) = 125.6.$$

Figure 8.6 shows the rejection and acceptance regions in terms of $\overline{x}$ as well as showing its distribution under the alternative hypothesis.
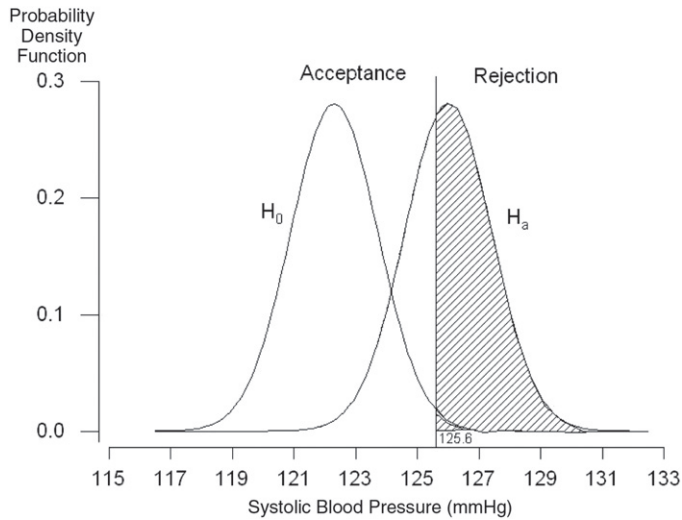
Figure 8.6 Rejection and acceptance regions of testing H_0: $\mu = 122.3$ versus H_a: $\mu = 126.0$ at the 0.01 significance level.

The shaded area provides a feel for the size of the power — the probability of rejecting the null hypothesis when it should be rejected — of the test. Power is the proportion of the area under the alternative hypothesis curve that is in the rejection region — that is, greater than or equal to 125.6 mmHg.

Let us find the power of the test and see if it agrees with our expectations about it based on Figure 8.6. Power is the probability of being in the rejection region — that is, of the sample mean being greater than or equal to 125.6 mmHg, assuming that the alternative hypothesis ($\mu = 126.0$) is true. To find this probability, we convert 125.6 to a standard normal value by subtracting the mean of 126.0 mmHg and dividing by $\sigma/\sqrt{n}$. Thus, the z value is

$$\frac{125.6 - 126.0}{20/\sqrt{199}} = -0.28.$$

The probability of a standard normal variable being greater than or equal to -0.28 is found from Table B4 to be 0.6103 ($= 1 - 0.3897$).

The power of the test is 61 percent. If this value is not large enough, there are several methods of increasing the power. One way is to increase the sample size. For example, let us increase the sample size to 600. Then the z value is

$$\frac{125.6 - 126.0}{20/\sqrt{600}} = -0.49.$$

The probability of a standard normal variable being greater than or equal to -0.49 is 0.6879 ($= 1 - 0.3121$), almost 8 percent larger than the power associated with the sample size of 199.

As was discussed earlier, another way of changing the power is to change the significance level. Let us decrease the significance level to 0.05. Doing this reduces the size of the rejection region. All values of $\bar{x}$ that are greater than or equal to

$$122.3 + 1.645\left(\frac{20}{\sqrt{199}}\right) = 124.6$$

now are in the rejection region. Using this significance level and still using a sample size of 199, the z value becomes

$$\frac{124.6 - 126.0}{20/\sqrt{199}} = -1.06.$$

The probability of a standard normal variable being greater than or equal to -1.06 is 0.855, more than 80 percent, which is often used as the desired level for power in the literature.

Another way of increasing the power is to redefine what we consider to be a substantial difference. If our emphasis were on detecting a blood pressure 5 percent more than the national average, instead of 3 percent more, we would have a higher power. As 5 percent of 122.3 mmHg is 6.1 mmHg, the null and alternative hypotheses become

$$H_0\text{: } \mu = 122.3 \quad \text{and} \quad H_a\text{: } \mu = 128.4.$$

The z statistic becomes

$$\frac{125.6 - 128.4}{20/\sqrt{199}} = -1.97$$

and the probability of a standard normal variable being more than or equal to -1.97 is 0.976. The power associated with the alternative that μ equals 126.0 has not changed, but our emphasis on what difference is important has changed. We have a much higher chance of detecting this greater difference, from 128.4 instead of from 126.0, between the null and alternative hypotheses.

Example 8.3

Let us consider another example of the calculation of power. Suppose that we have reason to suspect that the serum cholesterol level of male college students in a rural town is lower than the national average and we are planning a study to test this. The null and alternative hypotheses for the study are

$$H_0\text{: } \mu = 188 \text{ mg/dL} \quad \text{and} \quad H_a\text{: } \mu < 188 \text{ mg/dL}.$$

We use a value of 30 mg/dL for the standard deviation of serum cholesterol level for male college students.

We must choose a specific value for the mean serum cholesterol level under the alternative hypothesis. We have selected the value of 170 mg/dL, a difference of 18 mg/dL from the national mean, as an important difference that we wish to be able to detect. For this study, our initial plans call for a sample size of 50.

To find the power, we must first find the acceptance and rejection regions. Let us perform the test at the 0.01 significance level. Therefore, the rejection region consists of all values of z less than or equal to $z_{0.01}$ ($= -2.326$). In terms of the sample mean, the rejection region consists of values of $\overline{x}$ less than or equal to

$$\mu_0 + z_{0.01}\left(\frac{\sigma}{\sqrt{n}}\right) = 188 + (-2.326)\left(\frac{30}{\sqrt{50}}\right) = 178.132.$$
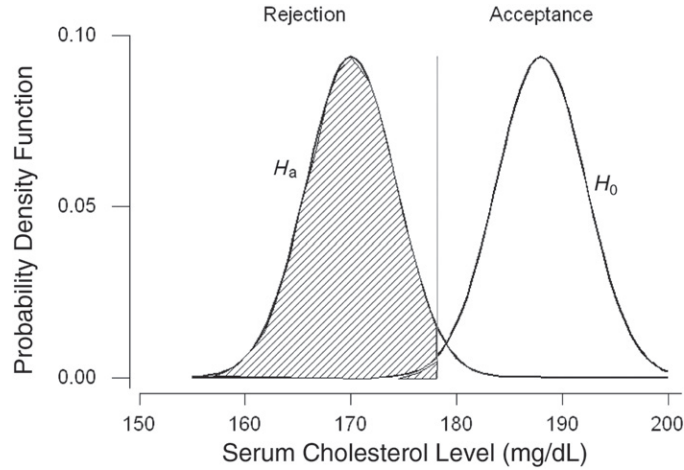
Figure 8.7 shows this situation.

Figure 8.7 Rejection and acceptance regions for testing $H_0: \mu = 94$ versus $H_a: \mu = 100$ at the 0.05 significance level.

Once we know the boundary between the acceptance and rejection regions, we convert the boundary to a z value by subtracting the mean under the alternative hypothesis and dividing by the standard error. For this example, the z value is

$$z = \frac{178.132 - 170}{30/\sqrt{50}} = 1.92.$$

The power of the test is the probability of observing a z-statistic with a value less than or equal to 1.92. From Table B4, we find the power to be 0.9726. This value is consistent with what we would have expected based on Figure 8.7. A study with 50 male college students has an excellent chance of detecting a mean value 18 mg/dL below the national average.

The key point about power is that calculations like in Example 8.3, or like those discussed in the material on confidence intervals, should be performed before any data are collected. These calculations give some indication about whether or not it is worthwhile to conduct an experiment before the resources are actually expended.

8.2.2 Unknown Variance

If the variance is unknown, the t statistic is used in place of the z statistic — that is,

$$t = \frac{\bar{x} - \mu_0}{\left(s/\sqrt{n}\right)}$$

in the test of the null hypothesis that the mean is the particular value μ_0. The rejection region for a two-sided alternative is $(t \leq t_{n-1,\alpha/2})$ or $(t \geq t_{n-1,1-\alpha/2})$.

Suppose that we did not know the value of σ for the systolic blood pressure in the dig200 data or that we were uncomfortable in using the value of 20 mmHg for σ. Then we would substitute s for σ and use the t distribution in place of the z distribution. In this case, the value of the t statistic is

$$t = \frac{125.8 - 122.3}{18.2/\sqrt{199}} = 2.71.$$

To be consistent with the test shown above, we shall also perform this test at the 0.05 significance level. Therefore, t is compared to $t_{198,0.025,}$ which is -1.972, and to $t_{198,0.975}$, which is 1.972. Since 2.71 is in the rejection region, we reject the null hypothesis in favor of the alternative. Not surprisingly, this result is very similar to that obtained when the z statistic was used. The results are similar because there was little difference between the values of s and σ, and, since the sample size is large, the critical values of the t and normal distributions are also close in value. The t test for one mean can be performed by the computer (see **Program Note 8.1** on the website).

8.3   Testing Hypotheses about the Proportion and Rates

In this section we focus on situations for which the use of the normal distribution as an approximation for the binomial distribution is appropriate. In general, these are situations in which the sample size is large.

Example 8.4

In Chapter 7 (Example 7.4) we considered the immunization level of 5-year-olds. The health department took a sample and, based on the sample, would decide whether or not to provide additional funds for an immunization campaign. In Example 7.4 we examined both the 99 percent confidence interval and a one-sided interval. Since the health department will provide additional funds if the proportion of immunization is less than 75 percent, we consider a one-sided test here, considering the following null and alternative hypotheses

$$H_0\text{: } \pi = \pi_0 = 0.75 \quad \text{and} \quad H_a\text{: } \pi < \pi_0 = 0.75.$$

The test statistic for this hypothesis is

$$z = \frac{|p - \pi_0| - 1/(2n)}{\sqrt{p(1-p)/n}}.$$

If $(p - \pi_0)$ is positive, a positive sign is assigned to z; if the difference is negative, a minus sign is assigned to z. The rejection region consists of values of z less than or equal to z_α. This framework is very similar to that used with the population mean, the only difference being the use of the continuity correction with the proportion.

The sample proportion, p, had a value of 0.614 based on a sample size of 140. Thus, the calculation of z is

$$z = \frac{|0.614 - 0.75| - 1/(2\{140\})}{\sqrt{0.614(1-0.614)/140}} = 3.219.$$

Since $(p - \pi_0)$ is negative, the test statistic's value is -3.219. If the test is performed at the 0.01 significance level, values of z less than or equal to -2.326 form the rejection region. Since z is less than -2.326, we reject the null hypothesis in favor of the alternative. The health department should devote more funds to an immunization effort. This conclusion agrees with that reached based on the confidence interval approach in Chapter 7.

The continuity correction can be eliminated from the calculations for relatively large sample sizes because its effect will be minimal. For example, if we had ignored the continuity correction in this example, the value of the test statistic would be -3.306, not much different from -3.219. The computer can be used to analyze these data (see **Program Note 8.2** on the website).

The same procedure applies to the test of crude and adjusted rates. Just as in Chapter 7, we treat rates as if they were proportions. This treatment allows for a simple approximation to the variance of a rate and also gives a justification for the use of the normal distribution as an approximation to the distribution of the rate. Thus, our test statistic has the same form as that used for the proportion.

Example 8.5

Suppose that we wish to test, at the 0.05 significance level, that the 2002 age-adjusted death rate for the American Indian/Alaskan Native male population, obtained by the indirect method of adjustment (using the 2002 U.S. age-specific death rates as the standard), is equal to the 2002 direct adjusted death rate for U.S. white male population of 992.9 per 100,000 (NCHS 2004). The alternative hypothesis is that the rates differ. In symbols, the null and alternative hypotheses are

$$H_0: \theta = \theta_0 = 0.009929 \quad \text{and} \quad H_a: \theta \neq \theta_0.$$

The test statistic, z, for this hypothesis is

$$(\hat{\theta} - \theta_0)/(\text{approximate standard error of } \hat{\theta})$$

where $\hat{\theta}$ is 907.8 per 100,000, the 2002 indirect age-adjusted death rate for the American Indian/Alaskan Native male population. In Chapter 7 we found the approximation to the standard error of $\hat{\theta}$ was 11 per 100,000. If this value of z is less than or equal to -1.96 ($= z_{0.025}$) or greater than or equal to 1.96 ($= z_{0.975}$), we reject the null hypothesis in favor of the alternative hypothesis. The value of z is

$$\frac{0.009078 - 0.009929}{0.00011} = -7.74.$$

Since -7.74 is in the rejection region, we reject the null hypothesis in favor of the alternative hypothesis at the 0.05 significance level. There is sufficient evidence to suggest that the indirect age-adjusted death rate for the American Indian/Alaskan Native male population is significantly different from the U.S. white male rate. The p-value for this test is obtained by taking twice the probability that a z statistic is less than or equal to -7.74; the p-value is less than 0.00001.

As we have previously discussed, this test makes sense only if we view the American Indian/Alaskan Native population data as a sample in time or place.

The tests for the crude rate and for the adjusted rate obtained by the direct method of adjustment have the same form as the preceding.

8.4   Testing Hypotheses about the Variance

In Chapter 7 we saw that $(n-1)s^2/\sigma^2$ followed the chi-square distribution with $n-1$ degrees of freedom. Therefore, we shall base the test of hypothesis about σ^2 on this statistic. The null and alternative hypotheses are

$$H_0: \sigma^2 = \sigma_0^2 \quad \text{and} \quad H_a: \sigma^2 \neq \sigma_0^2.$$

We shall define X^2 to be equal to $(n-1)\, s^2/\sigma^2$. When X^2 is greater than or equal to $\chi^2_{n-1,1-\alpha/2}$ or when X^2 is less than or equal to $\chi^2_{n-1,\alpha/2}$, we reject H_0 in favor of H_a.

For a one-sided alternative hypothesis, for example, $H_a: \sigma^2 < \sigma_0^2$, the rejection region is $X^2 \leq \chi^2_{n-1,\alpha}$. If the alternative is $H_a: \sigma^2 > \sigma_0^2$, the rejection region is $X^2 \geq \chi^2_{n-1,1-\alpha}$.

Example 8.6

Returning to the vitamin D in milk example discussed in Chapter 7, suppose we wish to test the hypothesis that the producer is in compliance with the requirement that the variance be less than 1600. We doubt that the producer is in compliance, and, therefore, we shall use the following null and alternative hypotheses:

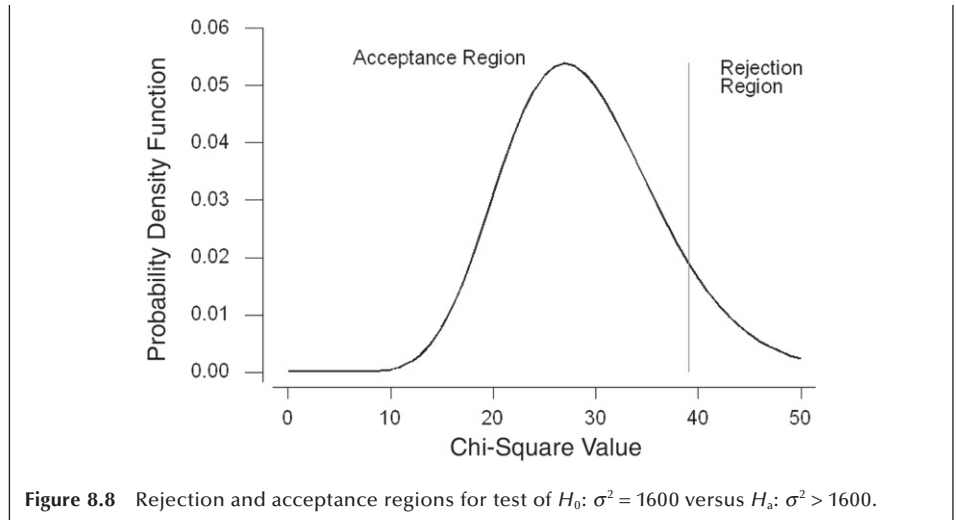$$H_0: \sigma^2 = 1600 \quad \text{and} \quad H_a: \sigma^2 > 1600.$$

Since this is a one-sided test, we are implicitly saying that the null hypothesis is that the population variance is less than or equal to 1600 versus the alternative that the variance is greater than 1600. We shall perform the test at the 0.10 significance level. Thus, the test statistic, X^2, which equals

$$\frac{(n-1)s^2}{\sigma_0^2}$$

is compared to $\chi^2_{29,0.90}$. If X^2 is greater than or equal to 39.09, obtained from Table B7, we reject the null hypothesis in favor of the alternative hypothesis. Using the value of 1700 for s^2 and 30 for n from Chapter 7, the value of X^2 is

$$\frac{29(1700)}{1600} = 30.81.$$

Since X^2 is not in the rejection region, we fail to reject the null hypothesis. There is not sufficient evidence to suggest that the producer is not in compliance with the variance requirement. This is the same conclusion reached when the confidence interval approach was used in Chapter 7. Figure 8.8 shows the rejection and acceptance regions for this test.

Figure 8.8   Rejection and acceptance regions for test of H_0: $\sigma^2 = 1600$ versus H_a: $\sigma^2 > 1600$.

As was mentioned in Chapter 7, the chi-square distribution begins to resemble the normal curve as the degrees of freedom becomes large. Figure 8.8 is a verification of that fact. From this figure, we also see that the p-value for the test statistic is large — approximately 0.40.

8.5   Testing Hypotheses about the Pearson Correlation Coefficient

In Chapter 7, we saw that the z' transformation, $z' = (1/2) \ln[(1 + r)/(1 - r)]$, approximately followed a normal distribution with a mean of $(1/2) \ln[(1 + \rho)/(1 - \rho)]$ and a standard error of $1/\sqrt{n-3}$. Therefore, to test the null hypothesis of H_0: $\rho = \rho_0$ versus an alternative hypothesis of H_a: $\rho \neq \rho_0$, we shall use the test statistic, λ, defined as $\lambda = (z' - z_0')\sqrt{n-3}$ where z_0' is $(1/2) \ln[(1 + \rho_0)/(1 - \rho_0)]$. If λ is less than or equal to $z_{\alpha/2}$ or greater than or equal to $z_{1-\alpha/2}$, we reject the null hypothesis in favor of the alternative hypothesis.

There is often interest as to whether or not the Pearson correlation coefficient is zero. If it is zero, then there is no linear association between the two variables. In this case, the test statistic simplifies to

$$\lambda = z'\sqrt{n-3}.$$

Example 8.7

Table 8.4 shows infant mortality rates for 1988 and total health expenditures as a percentage of gross domestic product in 1987 for selected 21 countries. It is thought that there should be some relation between these two variables. We translate these thoughts into the following null and alternative hypotheses:

$$H_0: \rho = 0.00 \quad \text{and} \quad H_a: \rho \neq 0.00.$$

Table 8.4   1988 infant mortality rates and 1987 health expenditures as a percentage of gross domestic product for selected countries.[a]

Country	1988 Infant Mortality Rate[b]	1987 Health Expenditures as Percentage of GDP
Japan	4.8	6.8
Sweden	5.8	9.0
Finland	6.1	7.4
Netherlands	6.8	8.5
Switzerland	6.8	7.7
Canada	7.2	8.6
West Germany	7.5	8.2
Denmark	7.5	6.0
France	7.8	8.6
Spain	8.1	6.0
Austria	8.1	7.1
Norway	8.3	7.5
Australia	8.7	7.1
Ireland	8.9	7.4
United Kingdom	9.0	6.1
Belgium	9.2	7.2
Italy	9.3	6.9
United States	10.0	11.2
New Zealand	10.8	6.9
Greece	11.0	5.3
Portugal	13.1	6.4

[a]Infant mortality rates are from National Center for Health Statistics, 1992, Table 25, and health expenditures are from National Center for Health Statistics, 1991, Table 104.
[b]Infant mortality rates are deaths to infants under 1 year of age per 1000 live births.

and the null hypothesis will be tested at the 0.10 significance level. The rejection region consists of values of λ that are less than or equal to -1.645 ($= z_{0.05}$) or greater than or equal to 1.645.

Applying the formula from Chapter 3 to the data in Table 8.4, we find a correlation coefficient, r, to be -0.243. From this value z' and λ can be calculated:

$$z' = 0.5\left(\ln\left(\frac{1-0.243}{1+0.243}\right)\right) = -0.248 \text{ and}$$

$$\lambda = -0.248\sqrt{21-3} = -1.052.$$

Since λ is not in the rejection regions, we fail to reject the null hypothesis. The p-value of this test is 0.29 (twice of probability $[z \leq -1.052]$). The correlation is not significantly different from zero. Let us check whether we can draw the same conclusion from the corresponding confidence interval. Using the method discussed in Chapter 7, a 90 percent confidence interval is $(-0.562, 0.139)$ in which zero is included. The computer can be used to perform this test (see **Program Note 8.3** on the website), and most programs give correlation coefficients and their associated p-values.

This procedure can be used with the Spearman correlation coefficient for sample sizes greater than or equal to 10.

8.6 Testing Hypotheses about the Difference of Two Means

When we test hypotheses about the difference of two means, we need to first check whether the two means come from independent samples or from a single sample. When the two means are calculated from the sample, they are dependent. The test procedures used are different depending on whether the means are independent or dependent.

8.6.1 Difference of Two Independent Means

We begin with the consideration of independent means under various assumptions. The first test assumes that the variances are known, followed by the assumption that the variances are unknown but equal and then unknown and unequal. After these sections, we consider the difference of two dependent means.

Known Variances: The null hypothesis of interest for the difference of two independent means is

$$H_0: \mu_1 - \mu_2 = \Delta_0$$

where Δ_0 is the hypothesized difference of the two means. Usually Δ_0 is zero — that is, we are testing that the means have the same value. The alternative hypothesis could be either

$$H_a: \mu_1 - \mu_2 \neq \Delta_0$$

or that the difference is greater (less) than Δ_0. Regardless of the alternative hypothesis, when the variances are known, the test statistic is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

The rejection region for the two-sided alternative includes values of z less than or equal to $z_{\alpha/2}$ or greater than or equal to $z_{1-\alpha/2}$. The rejection region for the greater than alternative includes values of z greater than or equal to $z_{1-\alpha}$ and the rejection region for the less than alternative includes values of z less than or equal to z_α.

Example 8.8

We return to the Ramipril example from Chapter 7 and test the hypothesis that μ_1, the mean decrease in diastolic blood pressure associated with the 1.25 mg dose, is the same as μ_2, the mean decrease for the 5 mg dose. In practice, we should not initially focus on only two of the three doses; all three doses should be considered together at the start of the analysis. However, at this stage, we do not know how to analyze three means at one time — the topic of the next chapter. Therefore, we are temporarily ignoring the existence of the third dose (2.5 mg) of Ramipril that was used in the actual experiment.

As we expect that the higher dose of medication will have the greater effect, the null and alternative hypotheses are

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{and} \quad H_a: \mu_1 - \mu_2 < 0.$$

We perform the test at the 0.05 significance level; thus, if the test statistic is less than -1.645 ($= z_{0.05}$), we shall reject the null hypothesis in favor of the alternative hypothesis. The sample mean decreases, $\bar{x}_1$ and $\bar{x}_2$, are 10.6 and 14.9 mm Hg, respectively, and both sample means are based on 53 observations. Both σ_1 and σ_2 are assumed to be 9 mmHg. Therefore, the value of z, the test statistic, is

$$z = \frac{(10.6 - 14.9) - 0}{\sqrt{81/53 + 81/53}} = -2.46.$$

Since the test statistic is less than -1.645, we reject the null hypothesis in favor of the alternative hypothesis. There appears to be a difference in the effects of the two doses of Ramipril with the higher dose being associated with the greater mean decrease in diastolic blood pressure at the 0.05 significance level.

Unknown but Equal Population Variances: The null and alternative hypotheses are the same as in the preceding section. However, the test statistic for the difference of two independent means, when the variances are unknown, changes to

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s_p \sqrt{1/n_1 + 1/n_2}}.$$

For a two-sided alternative hypothesis, the rejection region includes values of t less than or equal to $t_{n-2, \alpha/2}$ or greater than or equal to $t_{n-2, 1-\alpha/2}$, where n is the sum of n_1 and n_2.

Example 8.9

Let us test, at the 0.05 significance level, the hypothesis that there is no difference in the population mean proportions of total calories coming from fat for fifth- and sixth-grade boys and seventh- and eighth-grade boys. The alternative hypothesis is that there is a difference — that is, that Δ_0 is not zero. The rejection region includes values of t less than or equal to -2.04 ($= t_{31,0.025}$) or greater than or equal to 2.04.

From Chapter 7, we know that $\bar{x}_1$, the sample mean proportion for the 14 fifth- and sixth-grade boys, is 0.329, and the corresponding value, $\bar{x}_2$, for the 19 seventh- and eighth-grade boys is 0.353. The value of s_p is 0.094. Therefore, the test statistic's value is

$$t = \frac{(0.329 - 0.353) - 0}{0.094\sqrt{1/14 + 1/19}} = -0.727.$$

Since t is not in the rejection region, we fail to reject the null hypothesis. There does not appear to be a difference in the proportion of calories coming from fat at the 0.01 significance level.

The computer can be used to perform this test (see **Program Note 8.4** on the website).

Unknown and Unequal Population Variances: The test statistic for testing the null hypothesis of a specified difference in the population means — that is,

$$H_0: \mu_1 - \mu_2 = \Delta_0$$

assuming that the population variances are unequal, is given by

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

The statistic t' approximately follows the t distribution with degrees of freedom, df, given by

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

For a two-sided alternative, if t' is less than or equal to $t_{df, \alpha/2}$ or greater than or equal to $t_{df, 1-\alpha/2}$, we reject the null hypothesis in favor of the alternative hypothesis. If the alternative hypothesis is

$$H_a: \mu_1 - \mu_2 < \Delta_0,$$

the rejection region consists of values for t' of less than or equal to $t_{df, \alpha}$. If the alternative hypothesis is

$$H_a: \mu_1 - \mu_2 > \Delta_0,$$

the rejection region consists of values for t' of greater than or equal to $t_{df, 1-\alpha}$.

Example 8.10

In Chapter 7, we examined the mean ages of the AML and ALL patients. Suppose we will consider that no difference in the population mean ages exists if the mean age of AML patients minus the mean age of ALL patients is less than or equal to 5 years. Thus, the null and alternative hypotheses are

$$H_0: \mu_1 - \mu_2 = 5 \text{ and } H_a: \mu_1 - \mu_2 > 5.$$

We shall perform this test at the 0.01 significance level, which means that we shall reject the null hypothesis in favor of the alternative hypothesis if t' is greater than or equal to 2.446 ($= t_{33, 0.99}$). The degrees of freedom of 33 for the t value is obtained by

$$df = \frac{\left(16.51^2/51 + 17.85^2/20\right)^2}{\left(\dfrac{\left(16.51^2/51\right)^2}{51 - 1} + \dfrac{\left(17.85^2/20\right)^2}{20 - 1}\right)} = 32.501.$$

Using the values for the sample means, standard deviations, and sample sizes from Chapter 7, we calculate t' to be

$$t' = \frac{(49.86 - 36.65) - 5}{\sqrt{16.51^2/51 + 17.85^2/20}} = 1.781.$$

Since t' is less than 2.446, we fail to reject the null hypothesis. There is not sufficient evidence to conclude that the difference in ages is greater than 5 years. Usually one would test the hypothesis of no difference instead of a difference of 5 years. However, by testing the difference of 5 years, we were able to demonstrate the calculations for a nonzero Δ_0. The computer can be used to perform this test (see **Program Note 8.5** on the website).

As we emphasized in Chapter 7, we seldom know much about the magnitude of the two variances. Therefore, in those situations in which we know little about the variances and have no reason to believe that they are equal, we recommend that the unequal variances assumption should be used.

8.6.2   Difference of Two Dependent Means

The test to be used in this section is the paired t test, one of the more well-known and widely used tests in statistics. The null hypothesis to be tested is that the mean difference of the paired observations has a specified value — that is,

$$H_0: \mu_d = \mu_{d0}$$

where μ_{d0} is usually zero. The test statistic is

$$t_d = \frac{\bar{x}_d - \mu_{d0}}{s_d \sqrt{n}}.$$

The rejection region for a two-sided alternative hypothesis includes values of t_d less than or equal to $t_{df,\alpha/2}$ or greater than or equal to $t_{df,1-\alpha/2}$. The rejection region for the alternative of less than includes values of t_d that are less than or equal to $t_{df,\alpha}$, and the rejection region for the alternative of greater than includes values of t_d that are greater than or equal to $t_{df,1-\alpha}$.

Example 8.11

We use this method to examine the effect of the 1.25 mg level of Ramipril. We shall analyze the first six weeks of observation of the subjects — four weeks of run-in followed by two weeks of treatment. The null hypothesis is that the mean difference in diastolic blood pressure between the value at the end of the run-in period and the value at the end of the first treatment period is zero. The alternative hypothesis of interest is that there is an effect — that is, that the mean difference is greater than zero. In symbols, the hypotheses are

$$H_0: \mu_d = 0 \text{ and } H_a: \mu_d > 0.$$

We perform the test at the 0.10 significance level. Thus, the rejection region includes values of t_d that are greater than or equal to 1.298 ($= t_{52,0.90}$), using 52 degrees of freedom because there were 53 pairs of observations that are being analyzed.

From Chapter 7, we find that the sample mean difference in diastolic blood pressure after the two weeks of treatment was 10.6 mmHg for the 53 subjects. The sample

standard deviation of the differences was 8.5 mmHg. Based on these data, we can calculate the value of t_d, and it is

$$t_d = \frac{(10.6 - 0)}{8.5\sqrt{53}} = 9.08.$$

Since t_d is greater than 1.298, we reject the null hypothesis in favor of the alternative hypothesis. It appears that there is a difference between the value of diastolic blood pressure at the end of the run-in period and the treatment period with the blood pressure at the end of the treatment period being significantly less than that at the end of the run-in period. Note that we only said that there was a difference, but we did not attribute the difference to the medication.

As we have discussed in Chapters 6 and 7, drawing any conclusion from this type of study design is very difficult. There are two concerns — the presence of extraneous factors and reversion to the mean — associated with this design. Without some control group, it is difficult to attribute any effects that are observed in the study group to the intervention because of the possibility of extraneous factors. In a tightly controlled experiment, the researcher may be able to remove all extraneous factors, but it is difficult. The presence of a control group is also useful in providing an estimate of the reversion-to-the-mean effect if such an effect exists. Thus, we are suggesting that the paired t test should be used with great caution — that is, in only those situations for which we believe that there are no extraneous factors and no reversion-to-the-mean effect. In other cases, we would randomly assign study subjects either to the control group or to the intervention group and compare the differences of the pre- and post-measures for both groups.

If we are comfortable with the use of the paired t test, it can easily be performed by the computer (see **Program Note 8.6** on the website).

8.7 Testing Hypotheses about the Difference of Two Proportions

As in the comparison of two means, when we test hypotheses about the difference of two proportions, we need to first check whether the two proportions come from independent samples or from a single sample. When two proportions come from a single sample or paired observations, they are dependent. The test procedures used are different depending on whether the proportions are independent or dependent.

8.7.1 Difference of Two Independent Proportions

As in Chapter 7, we are considering the case of two independent proportions. The null hypothesis is

$$H_0: \pi_1 - \pi_2 = \Delta_0$$

where Δ_0 usually is taken to be zero. The test statistic for this hypothesis, assuming that the sample sizes are large enough for the use of the normal approximation to the binomial to be appropriate, is

$$z_{\pi d} = \frac{(p_1 - p_2) - \Delta_0}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}.$$

The rejection region for a two-sided alternative includes values of $z_{\pi d}$ that are less than or equal to $z_{\alpha/2}$ or greater than or equal to $z_{1-\alpha/2}$. If the alternative is less than, the rejection region consists of values of $z_{\pi d}$ that are less than or equal to z_α; if the alternative is greater than, the rejection region consists of values of $z_{\pi d}$ that are greater than or equal to $z_{1-\alpha}$.

Example 8.12

We test the hypothesis, at the 0.01 significance level, that there is no difference in the proportions of milk that contain 80 to 120 percent of the amount of vitamin D stated on the label between the eastern and southwestern milk producers. The alternative hypothesis is that there is a difference. From Chapter 7, we find the values of p_1 and p_2 are 0.286 and 0.420, respectively. Thus, the test statistic is

$$z_{\pi d} = \frac{(0.286 - 0.420) - 0}{\sqrt{0.286(1-0.286)/42 + 0.420(1-0.420)/50}} = -1.358.$$

Since $z_{\pi d}$ is not in the rejection region, we fail to reject the null hypothesis. The computer can be used to perform this test (see **Program Note 8.7** on the website).

8.7.2 Difference of Two Dependent Proportions

In Chapter 7 we discussed the confidence interval for the difference between two dependent proportions, π_d. Recall that the proportions of a particular attribute at two time points for the same individuals are not independent and the sample data for these situations is arranged as follows:

Attribute at Time		Number of Subjects
1	2	
Present	Present	a
Present	Absent	b
Absent	Present	c
Absent	Absent	d
Total		n

The estimated proportion of subjects with the attribute at time 1 is $p_1 = (a + b)/n$, and the estimated proportion with the attribute at time 2 is $p_2 = (a + c)/n$. The difference between the two estimated proportions is

$$p_d = p_1 - p_2 = (a + b)/n - (a + c)/n = (b - c)/n.$$

Here we want to test the difference of two dependent proportions. The null and alternative hypotheses in this situation are

$$H_o: \pi_d = \Delta_0 \text{ versus } H_a: \pi_d \neq \Delta_0.$$

The test statistic, assuming the sample size is large enough for the normal approximation to the binomial to be appropriate, is

$$z_{\pi d} = \frac{p_d - \Delta_0 - 1/n}{Estimated\ SE(p_d)}$$

where $1/n$ is the continuity correction suggested by Edwards (1948) and Δ_0 is zero in most situations. Expression of the estimated standard error for the difference of two dependent proportions was given in Chapter 7, and we repeat it here:

Then the test statistic becomes

$$z_{\pi d} = \frac{(b-c)/n - 1/n}{(1/n)\sqrt{(b+c) - (b-c)^2/n}} = \frac{b-c-1}{\sqrt{(b+c) - (b-c)^2/n}}.$$

This test is valid when the average of the discordant cell frequencies ($[b + c]/2$) is 5 or more. When it is less than 5, a binomial test is recommended instead of the z-test. The binomial test can be done by restricting our attention to the $(b + c)$ pairs. Under the null hypothesis the difference of the proportions conveniently follows a binomial distribution with $\pi = 0.5$ and sample size of $(b + c)$.

Example 8.13

We use the same data from the biostatistics and epidemiology test in Example 7.9 and the data were tabulated in a 2 by 2 table.

	Epidemiology		
Biostatistics	**Failed**	**Passed**	**Total**
Failed	**9** (a)	9 (b)	**18**
Passed	**1** (c)	81 (d)	82
Total	**10**	90	**100** (n)

Let us test the null hypothesis that there is no difference between the failure rates in biostatistics (18 percent) and epidemiology (10 percent) against the two-sided alternative hypothesis that they are different. We use a significance level of 0.05. The test statistic, $z_{\pi d}$, is

$$z_{\pi d} = \frac{9-1-1}{\sqrt{(9+1) - (9-1)^2/100}} = 2.289.$$

Since 2.289 is larger than 1.96, we reject the null hypothesis, suggesting that the two failure rates are significantly different. This result is consistent with the conclusion based on a 95 percent confidence interval shown in Example 7.9. The p-value of this test is 0.022. If we conducted this test at the 0.01 significance level, we could not reject the null hypothesis.

8.8   Tests of Hypotheses and Sample Size

We considered the sample size issue in the context of confidence interval in Chapter 7. We now consider the sample size in the context of hypothesis testing. As seen in testing hypotheses, the decision rule is based on the probabilities of Type I and Type II errors and increasing the sample size is one way of decreasing the error probabilities. Thus,

specification of Type I and Type II errors leads to the determination of required sample size. We consider the sample size issue for three situations below.

Testing a Single Mean: The z-values specifying α for a μ_0 (null hypothesis and β for a μ_1 (alternative hypothesis) leads to the determination of n. The z value specifying the upper α percentage point of the normal distribution is

$$z_\alpha = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

and the z value specifying the lower β percentage point is

$$z_\beta = \frac{\bar{x} - \mu_1}{\sigma/\sqrt{n}}.$$

Solving these two equations for n, eliminating $\bar{x}$, we get

$$n = \left[\frac{(z_\alpha - z_\beta)\sigma}{\mu_1 - \mu_0} \right]^2.$$

Note that the above formula includes the population standard deviation σ. Thus, in addition to specification of two error levels, we must have some idea of the underlying variability of a variable under consideration. The formula suggests that a larger sample size is required when the alternative hypothesis (μ_1) is closer to the null hypothesis (μ_0). While z_α is different depending on whether the test is specified as one-sided or two-sided, z_β always refers to one side of the normal curve. Also note that z_β carries the opposite sign of z_α.

Example 8.14

Let us consider an example. A researcher wants to determine a required sample size for a study to test whether male patients who do not exercise have elevated serum uric acid vales. The serum uric acid levels of males are known to be distributed normally with mean = 5.4 mg/100 mL and standard deviation of 1. The investigator wants to perform a one-sided test at the 5 percent significance level. The null hypothesis is that the population mean is 5.4. This indicates $\mu_0 = 5.4$ and $z_\alpha = 1.645$. He further specifies that if the true difference is as much as 0.4 mg/100 mL, he wishes to risk 10 percent chance of failing to reject the null hypothesis. This indicates that $\mu_1 = 5.8$ and $z_\beta = -0.28$. Then the required sample is

$$n = \left[\frac{(1.645 + 1.282)(1)}{5.8 - 5.4} \right]^2 = 53.5 \approx 54.$$

Testing a Single Proportion: As in the case of proportion, the specification of α for a π_0 and β for a π_1 would lead to the following two equations, ignoring the continuity correction:

$$z_\alpha = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \quad \text{and} \quad z_\beta = \frac{p - \pi_1}{\sqrt{\pi_1(1-\pi_1)/n}}.$$

Solving for n, eliminating p, we get

$$n = \left[\frac{z_\alpha \sqrt{\pi_0 (1-\pi_0)} - z_\beta \sqrt{\pi_1 (1-\pi_1)}}{\pi_1 - \pi_0} \right]^2 .$$

Example 8.15

Consider the planning of a survey to find out how smoking behavior changed while students were in college. A comprehensive survey four years ago found that 30 percent of freshmen smoked. The investigator wants to know how many seniors to be sampled this year. He wants to perform a two-tailed test at the 0.05 level. This suggests that $z_\alpha = 1.96$. The null hypothesis is $\pi_0 = 0.3$. He also states that if the proportion is changed as much as 5 percentage points, then he wishes to risk 10 percent chance of failing to reject the null hypothesis. This indicates $\pi_1 = 0.35$ and $z_\beta = -1.28$ (as mentioned earlier, one-tailed z value is used for z_β). Then the required sample size is

$$n = \left[\frac{1.96\sqrt{(.3)(.7)} - (-1.282)\sqrt{(.35)(.65)}}{.35 - .3} \right]^2 = 910.48 \approx 911.$$

Testing the Difference of Two Means: The required sample size for testing the difference of two independent means can also be determined in a similar manner. We assume equal variance in two groups ($\sigma_1^2 = \sigma_2^2$) and an equal division of the sample size between the two groups ($n_1 = n_2$). Specifying α error for the null hypothesis ($\mu_1 - \mu_2 = \Delta_0$), we have

$$z_\alpha = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{2\sigma_1^2/n_1}} .$$

Specifying β error for the alternative hypothesis ($\mu_1 - \mu_2 = \Delta_1$), we have

$$z_\beta = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_1}{\sqrt{2\sigma_1^2/n_1}} .$$

By solving these two equations for n_1, eliminating $(\bar{x}_1 - \bar{x}_2)$, we get

$$n_1 = 2\left[\frac{(z_\alpha - z_\beta)\sigma_1}{\Delta_1 - \Delta_0} \right]^2 .$$

Since Δ_0 is zero in most application, the denominator is usually Δ_1, the value specified in the alternative hypothesis. Note that n_1 is the sample size in each group and the total sample size for the study is $2n_1$.

Example 8.16

Let us consider a case of designing a clinical nutritional study of special diet regimen to lower blood pressure among hypertensive adult males (diastolic blood pressure over 90 mmHg). The investigator expects to demonstrate that the new diet would reduce diastolic blood pressure by 4 mmHg in three months. He is willing to risk a

Type I error of 5 percent and a Type II error of 10 percent for a one-sided test. NHANES III data show that the mean and standard deviation of diastolic blood pressure among hypertensive males are 95.4 mmHg and 5.6 mmHg. The required sample size in each group can be calculated by

$$n_1 = 2\left(\frac{5.6(1.645+1.282)}{4}\right)^2 = 33.5 \approx 34.$$

The proposed study would need a total of 68 subjects, allocated randomly and equally between the treatment and the control group.

As discussed in Chapter 7, the determination of sample size for a study is not as simple as the preceding example may suggest. In practice, we seldom know the population standard deviation, and we need to obtain an estimate of its value from the literature or from a pilot study. Setting the error levels low may lead to a very large sample size that can not possibly be carried out. The balancing of the error levels against availability of resources may require an iterative process until a satisfactory solution is found.

8.9   Statistical and Practical Significance

We must not confuse statistical significance with practical significance. For example, in the diet study discussed earlier, if we had a large enough sample, an observed value for p of 0.51 could be significantly different from the null hypothesis value of 0.50. However, this finding would be of little practical use. For a result to be important, it should be both statistically and practically significant. The test determines statistical significance, but the investigator must determine whether or not the observed difference is large enough to be practically significant.

When reporting the results of a study, many researchers have simply indicated whether or not the result was statistically significant and/or given only the p-value associated with the test statistic. This is useful information, but it is more informative to include the confidence interval for the parameter as well.

In conclusion, while hypothesis tests are useful to check whether observed results are attributable to chance, estimates of effects should also be considered before conclusions are drawn.

Conclusion

In this chapter we have introduced hypothesis testing and the associated terminology. A key point is that the calculation of the probabilities of errors should be conducted before the study is performed. By doing this, we can determine whether or not the study, as designed, can deliver answers to the question of interest with reasonable error levels. We also added to the material on confidence intervals that was presented in Chapter 7, demonstrating the equivalence of the confidence intervals to the test of hypothesis. We showed how to test hypotheses about the more common parameters used with normally distributed data and how to calculate power for a test of hypothesis about the mean when the population variance was known. In addition, we presented statistics to be used in

the tests of hypotheses about the difference of two means and two proportions. This latter material prepares us for the analysis of variance, in which we extend the test of hypothesis to comparing two or more means. Finally, we pointed out that statistical significance must not be confused with practical significance.

EXERCISES

8.1 In the diet study with a sample size of 20 pairs, suppose that we used a rejection region of 0 to 4 and 16 to 20. The null and alternative hypotheses are the same as in the chapter, and we are still interested in the specific alternative that π is 0.80. What are the values of α and β based on this decision rule? What is the power of the test for this specific alternative? We again observed 13 pairs favoring diet 1. What is the p-value for this result?

8.2 Suppose that the null and alternative hypotheses in the diet study were

$$H_0: \pi = 0.50 \text{ versus } H_a: \pi > 0.50.$$

Conduct the test at the 0.05 significance level. What is the decision rule that you will use? What are the probabilities of Type I and Type II errors for a sample size of 20 pairs and the specific alternative that π is 0.80?

8.3 What specific alternative value for π do you think indicates an important difference in the diet study? Provide an example of another study for which the binomial distribution could be used. What value would you use for the specific alternative for π in your study? What is the rationale for your choice for π in this new study?

8.4 Complete Table 8.2 by providing the values of power for π ranging from 0.05 to 0.50 in increments of 0.05. Graph the values of the power function versus the values of π for π ranging from 0.05 to 0.95. This graph is the power curve of the binomial test using the critical region of 0 to 2 and 10 to 12. What is the value of power when π is 0.50? Is there a specific name for this value? Describe the shape of the power curve. Discuss why the power curve, when the null hypothesis is π is equal to 0.50, must have this shape.

8.5 Frickhofen et al. (1991) performed a study on the effect of using cyclosporine in addition to antilymphocyte globulin and methylprednisolone in the treatment of aplastic anemia patients. There was a sample of 43 patients that received the cyclosporine in addition to the other treatment. Assume that the use of antilymphocyte globulin and methylprednisolone without cyclosporine results in complete or partial remission in 40 percent of aplastic anemia patients at the end of three months of treatment. We wish to determine if the use of cyclosporine can increase significantly the percentage of patients with complete or partial remission. What are the appropriate null and alternative hypotheses? Assume that the test is to be performed at the 0.01 significance level. What is the decision rule to be used? What is the probability of a Type II error based on the sample size of 43 and your decision rule? Twenty-eight patients achieved complete or partial remission at the end of three months. Is this a statistically significant result at the 0.01 level? What is the p-value of the test?

8.6 In a recent study, Hall (1989) examined the pulmonary functioning of 135 male Caucasian asbestos product workers. An earlier study had suggested that the development of clinical manifestations of the exposure to asbestos required a

minimum of 20 years. Therefore, Hall partitioned his data set into two groups, one with less than 20 years of exposure to asbestos and the other with 20 or more years of exposure. Two of the variables used to examine pulmonary function are the forced vital capacity (FVC) measured in liters and the percent of the predicted FVC value where the prediction is based on age, height, sex, and race. Age is a particularly important variable to consider, since there is a strong positive correlation between FVC and age. The sample means and standard deviations of FVC and percent of the predicted FVC for each of the two groups are as follows:

	Length of Exposure			
	<20 Years ($n = 66$)		20 Years ($n = 69$)	
Variable	Mean	S.D.	Mean	S.D.
FVC (L)	5.19	0.78	4.27	0.63
% Predicted. FVC	104	9.7	45	12.8

Choose the more appropriate of these two variables to use in a test of whether or not there is a difference in the means of the two exposure groups. Perform the test at the 0.05 significance level. Explain your choice for which variable to use and also your choice of a one- or two-sided alternative hypothesis. What assumption did you make about the population variances? Does this study support the idea that there is a difference between those with less than 20 years of exposure and those with 20 or more years of exposure? What is the p-value of the test? What, if any, other variable should be taken into account in the analysis?

8.7 Kirklin et al. (1981) performed a study of infants less than 3 months old who underwent open heart surgery. There were 175 infants in their study based on data from 1967 to 1980. It was suggested that the survival probabilities improved over time. To examine this, the data were broken into two time periods. Test the hypothesis that there is a difference in the survival probabilities over these two time periods versus the alternative hypothesis of no difference over time at the 0.01 significance level. Use the following hypothetical data, based on data presented in the study.

Date	Probability of Survival	Sample Size
Jan. 1967 to Dec. 1973	0.46	66
Jan. 1974 to July 1980	0.64	109

Provide possible reasons why there might be a difference in the survival probabilities over time.

8.8 Data from the National Institute of Occupational Safety and Health for the 1980–88 period were used to obtain estimates of the annual workplace fatality rates by state (PCHRG 1992). The average annual state rates over the nine-year period are given in Exercise 7.4. There is tremendous variability in the rates, ranging from a low of 1.9 to a high of 33.1 deaths per 100,000 workers. Provide some possible reasons for this variability. For the state of your residence, test the hypothesis of no difference in the crude workplace fatality rate and the

national average of 7.2 per 100,000 workers. Exercise 7.7 gives the population total for your state. Perform this test against a two-sided alternative at the 0.05 significance level. What is the p-value of the test? Provide possible reasons why there is or is not evidence of a difference between your state and the national average.

8.9 In the study by Reisin et al. (1978) one of the goals was to observe the effect of weight loss without salt restriction on blood pressure. We shall focus on one of the intervention groups, the group that was on a weight reduction program and given no medication. The program consisted of a strict diet with caloric intake reduced to about 50 percent of the usual adult intake for a two-month period. Before examining the data for an effect on blood pressure, it is necessary to determine whether or not the diet worked. The summary weight data for the sample of 24 patients was a mean reduction of 8.8 kilograms, and the standard deviation of the weight changes was 4.3 kilograms. This is a paired-t test situation for this single group. However, there was also a control group that was not part of the weight reduction effort. During the period when the study group lost an average of 8.8 kilograms, the control group showed an average decrease of only 0.7 kilograms. The results from the control group increase our confidence in the use of the paired-t test here. Test the null hypothesis of no weight reduction versus the appropriate one-sided alternative hypothesis at the 0.01 significance level. Did the weight reduction program work?

8.10 There have been a number of drug recalls during 1993 because of the failure of the drugs to meet dissolution specifications, content uniformity specifications, or because of subpotency (PCHRG 1993). Three products from the Parke-Davis Division of the Warner-Lambert Company were recalled. One of the products, Tedral, did not meet either the dissolution or content uniformity specifications.

  Suppose that the content uniformity specification as expressed in terms of the variance. For example, say that the variance of the amount of phenobarbital in tablets was supposed to be less than or equal to 0.015 grams^2. We selected a sample of 30 tablets and found the sample standard deviation of phenobarbital to be 0.14 grams. Test the appropriate hypothesis to determine, at the 0.10 level, whether there is compliance with the content uniformity specification for the amount of phenobarbital in the tablets.

8.11 In Chapter 7, using data from Table 7.6, we saw that there was a statistically significant (at the 0.01 level) difference in the mean ages of the AML and ALL patients. The difference in ages is important, particularly if the length of survival is strongly related to age. Calculate the sample Pearson correlation coefficient between age and length of survival based on all the patients in Table 7.6. Then test the null hypothesis, at the 0.05 level, that the population correlation coefficient is -0.30 versus the alternative hypothesis that the correlation is less (more negative) than -0.30. Here we are using -0.30 or more negative values to indicate a strong inverse correlation. Based on your analysis, is it necessary to control for the effect of age in the comparison of the length of survival of the AML and ALL patients?

8.12 In Exercise 7.10, we examined progress towards the Surgeon General's goal of reducing the proportion of 12- to 18-year-old adolescents who smoked to below 6 percent for a hypothetical community. We found that in 1990, of the 12- to

18-year-olds in the sample, 11 of 85 admitted that they smoked. Test the hypothesis that the hypothetical community has already attained the Surgeon General's goal at the 0.05 significance level. Should you use a one- or two-sided alternative hypothesis? Explain your reasoning.

8.13 Opponents of a national health system argue that it will lead to rationing of services, something that is viewed as being unacceptable to people in the United States. To determine how people in the United States really felt about rationing of services, the American Board of Family Practice had a survey conducted and some of the results are reported by Potter and Porter (1989). One question asked whether or not people would approve of rationing medical attention in the case of a terminal illness. Suppose that we have decided that there is substantial support for rationing if the proportion of the population who would approve of rationing in this case is 40 percent. In the sample of 1007 Americans, 34 percent supported rationing in the case of terminal illness. Test the hypothesis that the population proportion equals 40 percent versus the alternative hypothesis that it is less than 40 percent. Use the 0.01 significance level. It is interesting to note that 43 percent of the physicians surveyed supported rationing in this situation.

8.14 Anderson et al. (1990) performed a study on the effects of oat bran on serum cholesterol for males with high or borderline high values of serum cholesterol. High values of serum cholesterol are greater than or equal to 240 mg/dL (6.20 mmol/L). We wish to use the data from the study to determine whether or not there is a linear relation between body mass index and serum cholesterol. The body mass index is defined as weight (in kilograms) divided by the square of height (in meters). The data are

Body Mass Index	Serum Cholesterol	Body Mass Index	Serum Cholesterol
29.0	7.29	26.3	8.04
21.6	8.43	21.8	7.96
27.2	5.43	24.8	5.77
25.2	6.96	24.5	6.23
25.1	6.65	23.5	6.26
27.9	8.20	24.8	6.21
31.9	5.92	24.4	5.92

Test the hypothesis of no correlation between body mass index and serum cholesterol at the 0.10 level. Explain your choice of a one- or two-sided alternative hypothesis. What is the *p*-value of the test?

8.15 Exercise 7.6 shows 15 hypothetical serum cholesterol values. For these data, test the hypothesis that the population variance equals 100 $(mg/dL)^2$ versus the alternative hypothesis that the population variance is greater than 100 $(mg/dL)^2$. Perform the test at the 0.025 level. Discuss the results of this test in relation to the confidence interval obtained in Exercise 7.6. Recall that this test requires that the cholesterol values follow a normal distribution. Examine the assumption of normality of the cholesterol values.

8.16 For the same data from Exercise 7.6, test the hypothesis that the measuring process works — that is, test the hypothesis that the population mean of the values measured by this process equals 190 versus the alternative hypothesis

that the population mean is not equal to 190 mg/dL. Perform the test at the 0.02 significance level.

REFERENCES

Anderson, J. W., D. B. Spencer, C. C. Hamilton, et al. "Oat-Bran Cereal Lowers Serum Total and LDL Cholesterol in Hypercholesterolemic Men." *American Journal of Clinical Nutrition* 52:495–499, 1990.

Edwards, A. L. Note on the "Correction for Continuity" in Testing the Significance of the Difference between Correlated Proportions. *Psychometrika,* 13:185–187, 1948.

Frickhofen, N., J. P. Kaltwasser, H. Schrezenmeier, et al. "Treatment of Aplastic Anemia with Antilymphocyte Globulin and Methylprednisolone with or without Cyclosporine." *The New England Journal of Medicine* 324:1297–1304, 1991.

Hall, S. K. "Pulmonary Health Risk." *Journal of Environmental Health* 52:165–167, 1989.

Kirklin, J. K., E. H. Blackstone, J. W. Kirklin, et al. "Intracardiac Surgery under 3 Months of Age: Incremental Risk Factors for Hospitality Mortality." *The American Journal of Cardiology* 48:500–506, 1981.

Lee, E. S., and R. N. Forthofer. *Analyzing Complex Survey Data*, Second Edition. Thousand Oaks, CA: Sage Publications, 2006. Chapter 6, data processed from National Center for Health Statistics. The Third National Health and Nutrition Examination Survey, Adult sample of Phase I, 1999–2000.

National Center for Health Statistics. *Health, United States, 1990.* Hyattsville, MD: Public Health Service. DHHS Pub. No. 91-1232, 1991.

National Center for Health Statistics. *Health, United States, 1991 and Prevention Profile.* Hyattsville, MD: Public Health Service. DHHS Pub. No. 92-1232, 1992.

National Center for Health Statistics. *Health, United States, 2004 with Chartbook on Trends in the Health of Americans.* Hyattsville, MD: DHHS Pub. No. 2004-1232, 2004, Table 35.

Potter, C., and J. Porter. "American Perceptions of the British National Health Service: Five Myths." *Journal of Health Politics, Policy and Law* 14:341–365, 1989.

Public Citizen Health Research Group (PCHRG). "Work-Related Injuries Reached Record Level Last Year." *Public Citizen Health Research Group Health Letter* 8(12):1–3, 9, 1992.

Public Citizen Health Research Group (PCHRG). "Drug Recalls March 9–June 7, 1993." *Public Citizen Health Research Group Health Letter* 9(7):9–10, 1993.

Reisin, E., R. Abel, M. Modan, et al. "Effect of Weight Loss without Salt Restriction on the Reduction of Blood Pressure in Overweight Hypertensive Patients." *The New England Journal of Medicine* 298:1–6, 1978.