In Chapters 5 and 6, we considered tests dealing with frequency (categorical) data. In those situations, the results of any experiment can usually be represented by a few subtotals—the frequency of occurrence of each category of response. In this and subsequent chapters, we will deal with a different type of data, that which I have previously termed measurement or quantitative data.

In analyzing measurement data, our interest can focus either on differences between groups of subjects or on the relationship between two or more variables. The question of relationships between variables will be postponed until Chapters 9, 10, 15, and 16. This chapter will be concerned with the question of differences, and the statistic we will be most interested in will be the sample mean.

Low-birthweight (LBW) infants (who are often premature) are considered to be at risk for a variety of developmental difficulties. As part of an example we will return to later, suppose we took 25 LBW infants in an experimental group and 31 LBW infants in a control group, provided training to the parents of those in the experimental group on how to recognize the needs of LBW infants, and, when these children were 2 years old, obtained a measure of cognitive ability. Suppose that the LBW infants in the experimental group had a mean score of 117.2, whereas those in the control group had a mean score of 106.7. Is the observed mean difference sufficient evidence for us to conclude that 2-year-old LBW children in the experimental group generally score higher, on average, than do 2-year-old LBW control children? We will answer this particular question later; I mention the problem here to illustrate the kind of question we will discuss in this chapter.

7.1 Sampling Distribution of the Mean

sampling distribution of the mean

central limit theorem

As you should recall from Chapter 4, the sampling distribution of a statistic is the distribution of values we would expect to obtain for that statistic if we drew an infinite number of samples from the population in question and calculated the statistic on each sample. Because we are concerned in this chapter with sample *means,* we need to know something about the **sampling distribution of the mean.** Fortunately, all the important information about the sampling distribution of the mean can be summed up in one very important theorem: the central limit theorem. The **central limit theorem** is a factual statement about the distribution of means. In an extended form it states:

> Given a population with mean μ and variance σ^2, the sampling distribution of the mean (the distribution of sample means) will have a mean equal to μ (i.e., $\mu_{\overline{X}} = \mu$), a variance ($\sigma^2_{\overline{X}}$) equal to σ^2/n, and a standard deviation ($\sigma_{\overline{X}}$) equal to $\sigma/\sqrt{n}$. The distribution will approach the normal distribution as n, the *sample size,* increases.[1]

This is one of the most important theorems in statistics. It not only tells us what the mean and variance of the sampling distribution of the mean must be for any given sample size, but also states that as n increases, the shape of this sampling distribution

[1]The central limit theorem can be found stated in a variety of forms. The simplest form merely says that the sampling distribution of the mean approaches normal as n increases. The more extended form given here includes all the important information about the sampling distribution of the mean.

In Chapters 5 and 6, we considered tests dealing with frequency (categorical) data. In those situations, the results of any experiment can usually be represented by a few subtotals—the frequency of occurrence of each category of response. In this and subsequent chapters, we will deal with a different type of data, that which I have previously termed measurement or quantitative data.

In analyzing measurement data, our interest can focus either on differences between groups of subjects or on the relationship between two or more variables. The question of relationships between variables will be postponed until Chapters 9, 10, 15, and 16. This chapter will be concerned with the question of differences, and the statistic we will be most interested in will be the sample mean.

Low-birthweight (LBW) infants (who are often premature) are considered to be at risk for a variety of developmental difficulties. As part of an example we will return to later, suppose we took 25 LBW infants in an experimental group and 31 LBW infants in a control group, provided training to the parents of those in the experimental group on how to recognize the needs of LBW infants, and, when these children were 2 years old, obtained a measure of cognitive ability. Suppose that the LBW infants in the experimental group had a mean score of 117.2, whereas those in the control group had a mean score of 106.7. Is the observed mean difference sufficient evidence for us to conclude that 2-year-old LBW children in the experimental group generally score higher, on average, than do 2-year-old LBW control children? We will answer this particular question later; I mention the problem here to illustrate the kind of question we will discuss in this chapter.

7.1 Sampling Distribution of the Mean

sampling distribution of the mean

central limit theorem

As you should recall from Chapter 4, the sampling distribution of a statistic is the distribution of values we would expect to obtain for that statistic if we drew an infinite number of samples from the population in question and calculated the statistic on each sample. Because we are concerned in this chapter with sample *means,* we need to know something about the **sampling distribution of the mean.** Fortunately, all the important information about the sampling distribution of the mean can be summed up in one very important theorem: the central limit theorem. The **central limit theorem** is a factual statement about the distribution of means. In an extended form it states:

> Given a population with mean μ and variance σ^2, the sampling distribution of the mean (the distribution of sample means) will have a mean equal to μ (i.e., $\mu_{\bar{X}} = \mu$), a variance ($\sigma^2_{\bar{X}}$) equal to σ^2/n, and a standard deviation ($\sigma_{\bar{X}}$) equal to $\sigma/\sqrt{n}$. The distribution will approach the normal distribution as n, the *sample size,* increases.[1]

This is one of the most important theorems in statistics. It not only tells us what the mean and variance of the sampling distribution of the mean must be for any given sample size, but also states that as n increases, the shape of this sampling distribution

[1]The central limit theorem can be found stated in a variety of forms. The simplest form merely says that the sampling distribution of the mean approaches normal as n increases. The more extended form given here includes all the important information about the sampling distribution of the mean.

approaches normal, *whatever* the shape of the parent population.[2] The importance of these facts will become clear shortly.

The rate at which the sampling distribution of the mean approaches normal as n increases is a function of the shape of the parent population. If the population is itself normal, the sampling distribution of the mean will be normal regardless of n. If the population is symmetric but nonnormal, the sampling distribution of the mean will be nearly normal even for small sample sizes, especially if the population is unimodal. If the population is markedly skewed, sample sizes of 30 or more may be required before the means closely approximate a normal distribution.

uniform distribution

To illustrate the central limit theorem, suppose we have an infinitely large population of random numbers evenly distributed between 0 and 100. This population will have what is called a **uniform distribution**—every value between 0 and 100 will be equally likely. The distribution of 50,000 observations drawn from this population is shown in Figure 7.1. You can see that the distribution is very flat, as would be expected. For uniform distributions the mean (μ) is known to be equal to one-half of the range (50), the standard deviation (σ) is known to be equal to 28.87 (the range divided by $\sqrt{12}$), and the variance (σ^2) is thus 833.33.

Now suppose we drew 5000 samples of size 5 ($n = 5$) from this population and plotted the resulting sample *means.* Such sampling can be easily accomplished with a simple computer program; the results of just such a procedure are presented in Figure 7.2a, with a normal distribution superimposed. It is apparent that the distribution of means, although not exactly normal, is at least peaked in the center and trails off toward the extremes. (In fact, the superimposed normal distribution fits the data quite

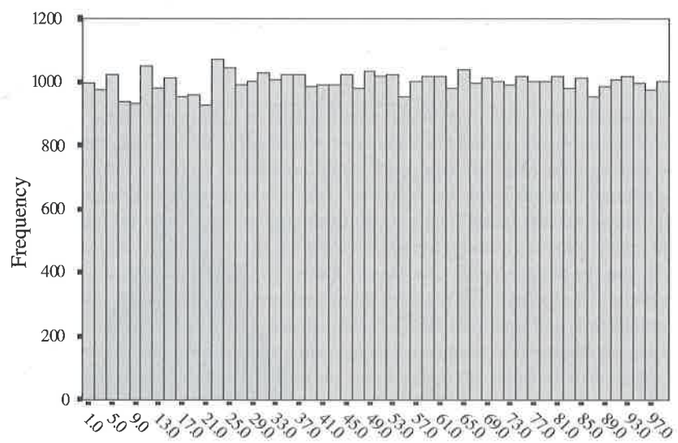


FIGURE 7.1 *50,000 observations from a uniform distribution*

[2]We traditionally let lowercase n stand for the number of observations in a single group or sample, and reserve uppercase N for the total number of cases over all groups. I am adopting that notation here.

well.) The mean and standard deviation of this distribution are shown, and they are extremely close to $\mu = 50$ and $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 28.87/\sqrt{5} = 12.91$. Any discrepancy between the actual values and those predicted by the central limit theorem is attributable to rounding error and to the fact that we did not draw an infinite number of samples.

Now suppose we repeated the entire procedure, but this time draw 5000 samples of 30 observations each. Again I have actually done this, and the results are plotted in Figure 7.2b. Here you see that just as the central limit theorem predicted, the

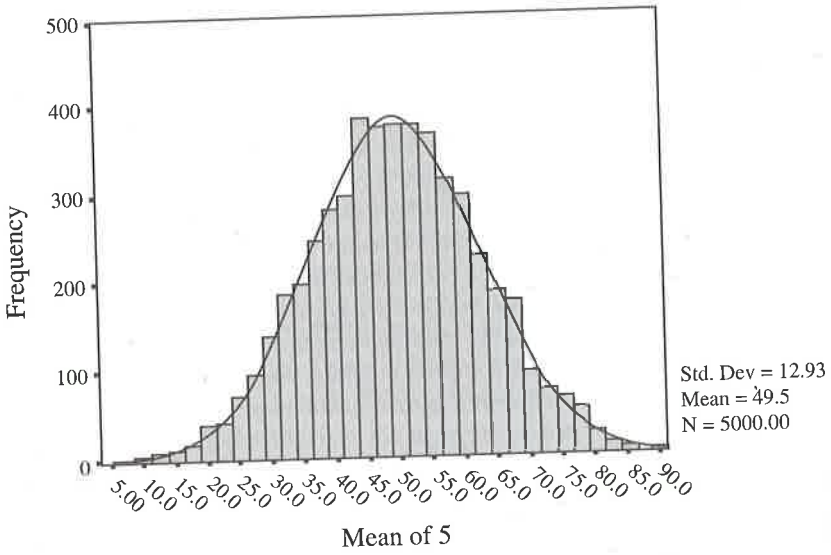


FIGURE 7.2a *Sampling distribution of the mean when $n = 5$*

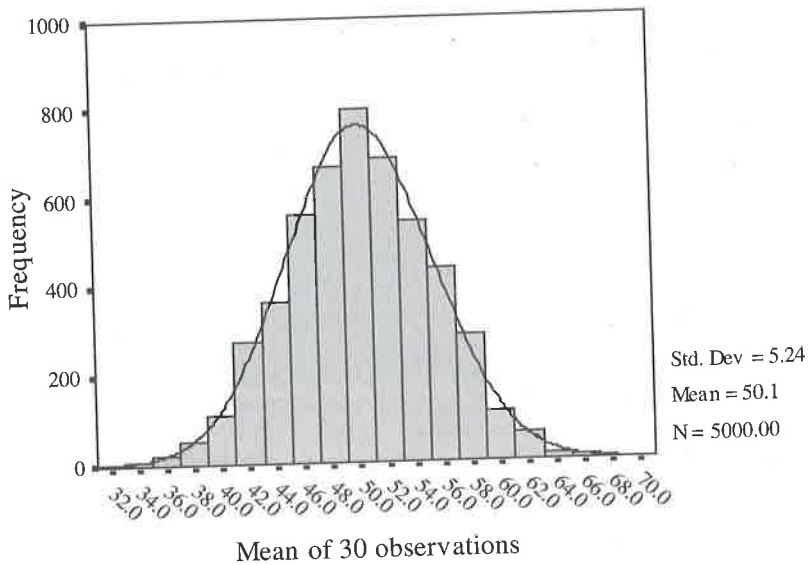


FIGURE 7.2b *Sampling distribution of the mean when $n = 30$*

well.) The mean and standard deviation of this distribution are shown, and they are extremely close to $\mu = 50$ and $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 28.87/\sqrt{5} = 12.91$. Any discrepancy between the actual values and those predicted by the central limit theorem is attributable to rounding error and to the fact that we did not draw an infinite number of samples.

Now suppose we repeated the entire procedure, but this time draw 5000 samples of 30 observations each. Again I have actually done this, and the results are plotted in Figure 7.2b. Here you see that just as the central limit theorem predicted, the

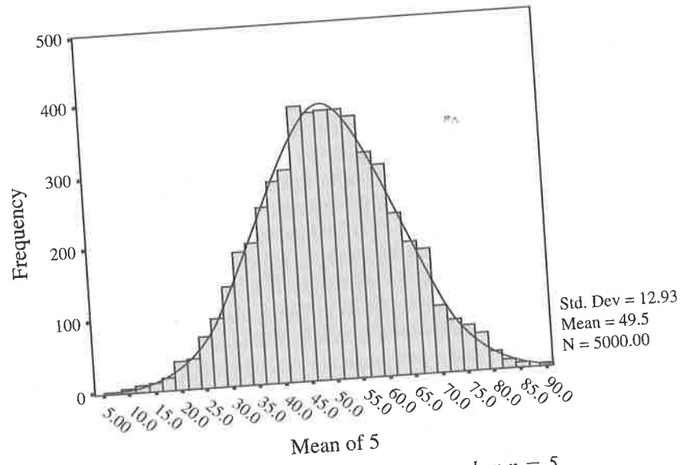


FIGURE 7.2a *Sampling distribution of the mean when $n = 5$*

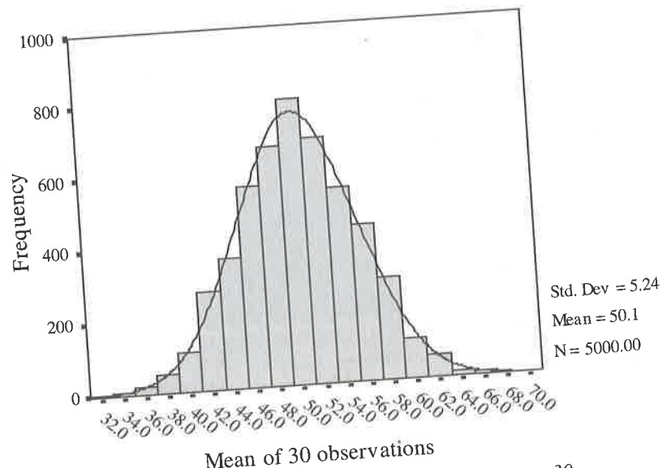


FIGURE 7.2b *Sampling distribution of the mean when $n = 30$*

distribution is approximately normal, the mean is again at $\mu = 50$, and the standard deviation has been reduced to approximately $28.87/\sqrt{30} = 5.27$.

7.2 Testing Hypotheses about Means—σ Known

From the central limit theorem, we know all the important characteristics of the sampling distribution of the mean. (We know its shape, its mean, and its standard deviation.) On the basis of this information, we are in a position to begin testing hypotheses about means. But first we might do well to go back to something we discussed with respect to the normal distribution. In Chapter 4 we saw that we could test a hypothesis about the population from which a single score (in that case, a finger-tapping score) was drawn by calculating

$$z = \frac{X - \mu}{\sigma}$$

and then, if the population is normally distributed, by obtaining the probability of a value of z as low as the one obtained by using the tables of the standard normal distribution. We ran a one-tailed test on the null hypothesis that the tapping rate (70) of a single individual was drawn at random from a normally distributed population of healthy subjects' tapping rates with a mean of 100 and a standard deviation of 20. We did this by calculating

$$z = \frac{X - \mu}{\sigma}$$
$$= \frac{70 - 100}{20} = \frac{-30}{20}$$
$$= -1.5$$

and then using Appendix z to find the area below $z = -1.5$.[3] This value is 0.0668. Thus, approximately 7% of the time we would expect a score as low as this if we were sampling from a healthy population. Since this probability was not less than our preselected significance level of $\alpha = 0.05$, we could not reject the null hypothesis. The tapping rate for the person we examined was not an unusual rate for healthy individuals. Although in this example we were testing a hypothesis about a single observation, the same logic applies to testing hypotheses about sample means. The only difference is that instead of comparing an observation to a distribution of observations, we will compare a mean to a distribution of means (the sampling distribution of the mean), and instead of dividing by the standard deviation of observations, we will divide by the standard deviation of means (to be called the standard error of the mean).

In most situations in which we test a hypothesis about a population mean, we don't have any knowledge about the variance of that population. (This is the main reason we have *t* tests, which are the main focus of this chapter.) However, in a limited number of situations we do know σ. A discussion of testing a hypothesis when σ

[3]Recall that the normal distribution is symmetric, and thus there are no entries for negative values of z. The "smaller portion" for $z = -1.5$ is the same as the "smaller portion" for $z = +1.5$.

is known provides a good transition from what we already know about the normal distribution to what we want to know about *t* tests. An example of behavior problem scores on the Achenbach Child Behavior Checklist (CBCL) (Achenbach, 1991a) is a useful example for this purpose, because we know both the mean and the standard deviation for the population of Total Behavior Problems scores ($\mu = 50$ and $\sigma = 10$). Assume that a random sample of five children under stress had a mean score of 56.0. We want to test the null hypothesis that these five children are a random sample from a population of normal children (i.e., normal with respect to their general level of behavior problems). In other words, we want to test H_0: $\mu = 50$ against the alternative H_1: $\mu \neq 50$.

Because we know the mean and standard deviation of the population of general behavior problem scores, we can use the central limit theorem to obtain the sampling distribution when the null hypothesis is true. The central limit theorem states that if we obtain the sampling distribution of the mean from this population, it will have a mean of $\mu = 50$, a variance of $\sigma^2/n = 10^2/5 = 100/5 = 20$, and a standard deviation (usually referred to as the **standard error**)[4] of $\sigma/\sqrt{n} = 4.47$. This distribution is diagrammed in Figure 7.3. The arrow in Figure 7.3 represents the location of the sample mean.

standard error

Because we know that the sampling distribution is normally distributed with a mean of 50 and a standard error of 4.47, we can find areas under the distribution by referring to tables of the standard normal distribution. Thus, for example, because two standard errors is $2(4.47) = 8.94$, the area to the right of $\overline{X} = 58.94$ is simply the area under the normal distribution farther than two standard deviations above the mean.

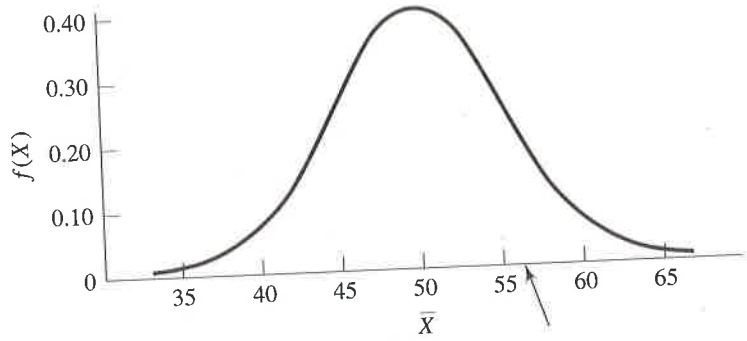


FIGURE 7 · 3 *Sampling distribution of the mean for $n = 5$ drawn from a population with $\mu = 50$ and $\sigma = 10$*

[4]The standard deviation of any sampling distribution is normally referred to as the *standard error* of that distribution. Thus, the standard deviation of means is called the standard error of the mean (symbolized by $\sigma_{\overline{X}}$), whereas the standard deviation of differences between means, which will be discussed shortly, is called the standard error of differences between means and is symbolized by $\sigma_{\overline{X}_1 - \overline{X}_2}$. Minor changes in terminology, such as calling a standard deviation a standard error, are not really designed to confuse students, though they probably have that effect.

is known provides a good transition from what we already know about the normal distribution to what we want to know about *t* tests. An example of behavior problem scores on the Achenbach Child Behavior Checklist (CBCL) (Achenbach, 1991a) is a useful example for this purpose, because we know both the mean and the standard deviation for the population of Total Behavior Problems scores ($\mu = 50$ and $\sigma = 10$). Assume that a random sample of five children under stress had a mean score of 56.0. We want to test the null hypothesis that these five children are a random sample from a population of normal children (i.e., normal with respect to their general level of behavior problems). In other words, we want to test H_0: $\mu = 50$ against the alternative H_1: $\mu \neq 50$.

Because we know the mean and standard deviation of the population of general behavior problem scores, we can use the central limit theorem to obtain the sampling distribution when the null hypothesis is true. The central limit theorem states that if we obtain the sampling distribution of the mean from this population, it will have a mean of $\mu = 50$, a variance of $\sigma^2/n = 10^2/5 = 100/5 = 20$, and a standard deviation (usually referred to as the **standard error**)[4] of $\sigma/\sqrt{n} = 4.47$. This distribution is diagrammed in Figure 7.3. The arrow in Figure 7.3 represents the location of the sample mean.

standard error

Because we know that the sampling distribution is normally distributed with a mean of 50 and a standard error of 4.47, we can find areas under the distribution by referring to tables of the standard normal distribution. Thus, for example, because two standard errors is 2(4.47) = 8.94, the area to the right of $\overline{X} = 58.94$ is simply the area under the normal distribution farther than two standard deviations above the mean.

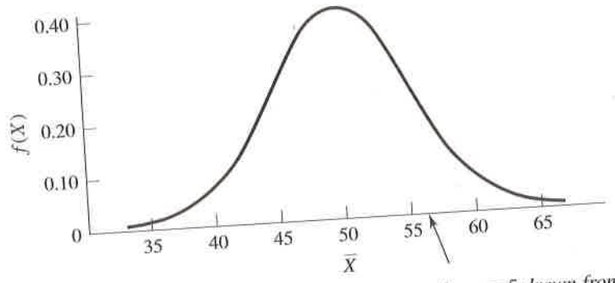


FIGURE 7·3 *Sampling distribution of the mean for n = 5 drawn from a population with μ = 50 and σ = 10*

[4]The standard deviation of any sampling distribution is normally referred to as the *standard error* of that distribution. Thus, the standard deviation of means is called the standard error of the mean (symbolized by $\sigma_{\overline{X}}$), whereas the standard deviation of differences between means, which will be discussed shortly, is called the standard error of differences between means and is symbolized by $\sigma_{\overline{X}_1 - \overline{X}_2}$. Minor changes in terminology, such as calling a standard deviation a standard error, are not really designed to confuse students, though they probably have that effect.

For our particular situation, we first need to know the probability of a sample mean greater than or equal to 56, and thus we need to find the area above $\overline{X} = 56$. We can calculate this in the same way we did with individual observations, with only a minor change in the formula for *z*:

$$z = \frac{X - \mu}{\sigma} \quad \text{becomes} \quad z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$$

which can also be written as

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

For our data this becomes

$$\frac{56 - 50}{4.47} = \frac{6}{4.47} = 1.34$$

Notice that the equation for *z* used here is in the same form as our earlier formula for *z*. The only differences are that *X* has been replaced by $\overline{X}$ and σ has been replaced by $\sigma_{\overline{X}}$. These differences occur because we are now dealing with a distribution of means, and thus the data points are now means, and the standard deviation in question is now the standard error of the mean (the standard deviation of means). The formula for *z* continues to represent (1) a point on a distribution, minus (2) the mean of that distribution, all divided by (3) the standard deviation of the distribution. Now rather than being concerned specifically with the distribution of $\overline{X}$, we have reexpressed the sample mean in terms of *z* scores and can now answer the question with regard to the standard normal distribution.

From Appendix *z* we find that the probability of a *z* as large as 1.34 is 0.0901. Because we want a two-tailed test of H_0, we need to double the probability to obtain the probability of a deviation as large as 1.34 standard errors *in either direction* from the mean. This is 2(0.0901) = 0.1802. Thus, with a two-tailed test (that stressed children have a mean behavior problem score that is different in either direction from that of normal children) at the 0.05 level of significance, we would not reject H_0 because the obtained probability is greater than 0.05. We would conclude that we have no evidence that stressed children show more or fewer behavior problems than other children.

7·3 Testing a Sample Mean When σ Is Unknown—The One-Sample *t* Test

The preceding example was chosen deliberately from among a fairly limited number of situations in which the population standard deviation (σ) is known. In the general case, we rarely know the value of σ and usually have to estimate it by way of the *sample* standard deviation (*s*). When we replace σ with *s* in the formula, however, the nature of the test changes. We can no longer declare the answer to be a *z* score and evaluate it using tables of *z*. Instead, we will denote the answer as *t* and evaluate it using tables of *t*, which are different from tables of *z*. The reasoning behind the switch

from z to t is really rather simple, although many texts ignore it. The basic problem that requires this change to t is related to the sampling distribution of the sample variance.

The Sampling Distribution of s^2

Because the t test uses s^2 as an estimate of σ^2, it is important that we first look at the sampling distribution of s^2. This sampling distribution gives us some insight into the problems we are going to encounter. We saw in Chapter 2 that s^2 is an *unbiased* estimate of σ^2, meaning that with repeated sampling the average value of s^2 will equal σ^2. Although an unbiased estimator is a nice thing, it is not everything. The problem is that the shape of the sampling distribution of s^2 is positively skewed, especially for small samples. (In fact, it is related to the chi-square distribution, as we saw in Chapter 6.) I drew 50,000 samples of $n = 5$ from a population with $\mu = 5$ and $\sigma^2 = 50$. I calculated the variance for each sample, and have plotted those 50,000 variances in Figure 7.4. Notice that the mean of this distribution is almost exactly 50, reflecting the unbiased nature of s^2 as an estimate of σ^2. However, the distribution is very positively skewed. Because of the skewness of this distribution, an individual value of s^2 is more likely to underestimate σ^2 than to overestimate it, especially for small samples. Also because of this skewness, the resulting value of t is likely to be larger than the value of z that we would have obtained had σ been known and used.

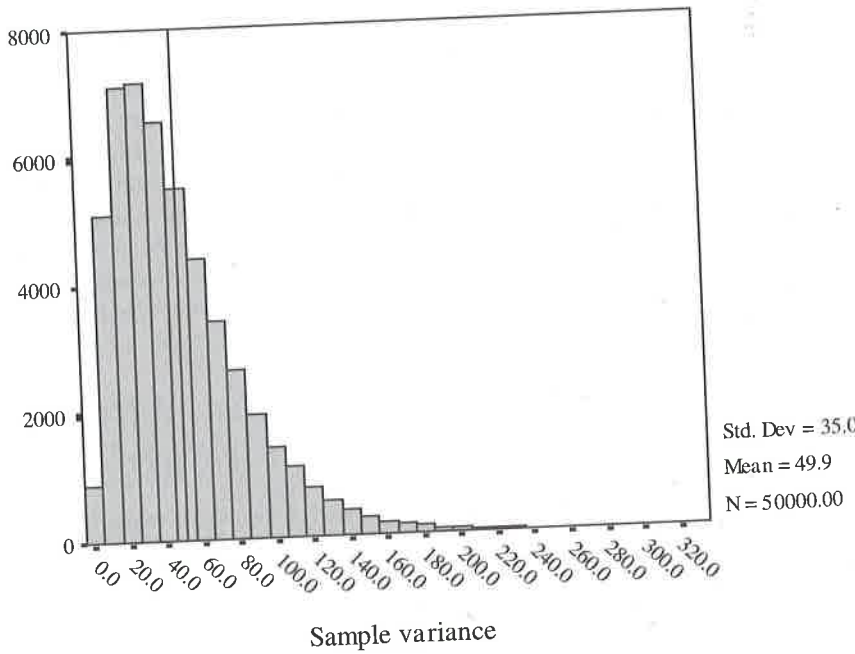


FIGURE 7.4 *Sampling distribution of the sample variance*

from z to t is really rather simple, although many texts ignore it. The basic problem that requires this change to t is related to the sampling distribution of the sample variance.

The Sampling Distribution of s^2

Because the t test uses s^2 as an estimate of σ^2, it is important that we first look at the sampling distribution of s^2. This sampling distribution gives us some insight into the problems we are going to encounter. We saw in Chapter 2 that s^2 is an *unbiased* estimate of σ^2, meaning that with repeated sampling the average value of s^2 will equal σ^2. Although an unbiased estimator is a nice thing, it is not everything. The problem is that the shape of the sampling distribution of s^2 is positively skewed, especially for small samples. (In fact, it is related to the chi-square distribution, as we saw in Chapter 6.) I drew 50,000 samples of $n = 5$ from a population with $\mu = 5$ and $\sigma^2 = 50$. I calculated the variance for each sample, and have plotted those 50,000 variances in Figure 7.4. Notice that the mean of this distribution is almost exactly 50, reflecting the unbiased nature of s^2 as an estimate of σ^2. However, the distribution is very positively skewed. Because of the skewness of this distribution, an individual value of s^2 is more likely to underestimate σ^2 than to overestimate it, especially for small samples. Also because of this skewness, the resulting value of t is likely to be larger than the value of z that we would have obtained had σ been known and used.

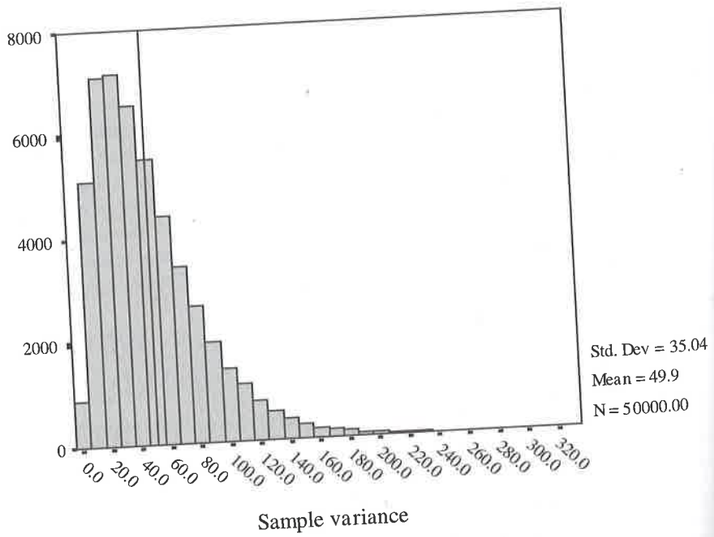


FIGURE 7.4 *Sampling distribution of the sample variance*

The t Statistic

We are going to take the formula that we just developed for z,

$$z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

and substitute s for σ to give

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\overline{X} - \mu}{\sqrt{\frac{s^2}{n}}}$$

Since we know that for any particular sample, s^2 is more likely than not to be smaller than the appropriate value of σ^2, we can see that the t formula is more likely than not to produce a larger answer (in absolute terms) than we would have obtained if we had solved for t using the true but unknown value of σ^2 itself. (You can see this in Figure 7.4, where more than half of the observations fall to the left of σ^2.) As a result, it would not be fair to treat the answer as a z score and use the table of z. To do so would give us too many "significant" results—that is, we would make more than 5% Type I errors. (For example, when we were calculating z, we rejected H_0 at the 0.05 level of significance whenever z exceeded ± 1.96. If we create a situation in which H_0 is true, repeatedly draw samples of $n = 5$, and use s^2 in place of σ^2, we will obtain a value of ± 1.96 or greater more than 10% of the time. The t cutoff in this case is 2.776.)

Student's *t* distribution

The solution to our problem was supplied in 1908 by William Gosset, who worked for the Guinness Brewing Company and wrote under the pseudonym of Student supposedly because the brewery would not allow him to publish under his own name. Gosset showed that if the data are sampled from a normal distribution, using s^2 in place of σ^2 would lead to a particular sampling distribution, now generally known as **Student's *t* distribution.** As a result of Gosset's work, all we have to do is substitute s^2, denote the answer as t, and evaluate t with respect to its own distribution, much as we evaluated z with respect to the normal distribution. The t distribution is tabled in Appendix t and examples of the actual distribution of t for various sample sizes are shown graphically in Figure 7.5.

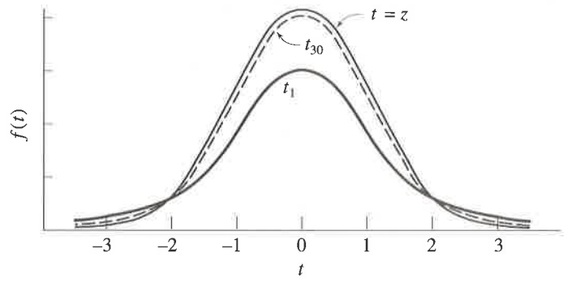


FIGURE 7.5 *t distribution for 1, 30, and ∞ degrees of freedom*

As you can see from Figure 7.5, the distribution of t varies as a function of the degrees of freedom, which for the moment we will define as one less than the number of observations in the sample. From what we know about χ^2, this is to be expected, since the distribution of χ^2, and thus the distribution of s^2, is also a function of the degrees of freedom. As $n \Rightarrow \infty$, $p(s^2 < \sigma^2) \Rightarrow p(s^2 > \sigma^2)$. Since the skewness of the sampling distribution of s^2 disappears as the number of degrees of freedom increases, the tendency for s to underestimate σ will also disappear. Thus, for an infinitely large number of degrees of freedom, t will be normally distributed and equivalent to z.

The test of one sample mean against a known population mean, which we have just performed, is based on the assumption that the sample was drawn from a normally distributed population. This assumption is required primarily because the t was derived assuming that the mean and variance are independent, which they are with a normal distribution. In practice, however, our t statistic can reasonably be compared to the t distribution whenever the sample size is sufficiently large to produce a normal sampling distribution of the mean. Most people would suggest that an n of 25 or 30 is "sufficiently large" for most situations, and for many situations it can be considerably smaller than that.

On the other hand, Wuensch (1993, personal communication) has argued convincingly that, at least with *very* skewed distributions, the fact that n is large enough to lead to a sampling distribution of the mean that appears to be normal does not guarantee that the resulting sampling distribution of t follows Student's t distribution. The derivation of t makes assumptions both about the distribution of means (which is under the control of the central limit theorem), and the variance, which is not controlled by that theorem.

Degrees of Freedom

I have mentioned that the t distribution is a function of the degrees of freedom (df). For the one-sample case, $df = n - 1$; the one degree of freedom has been lost because we used the sample mean in calculating s^2. To be more precise, we obtained the variance (s^2) by calculating the deviations of the observations from their own mean, $(X - \overline{X})$, rather than from the population mean $(X - \mu)$. Because the sum of the deviations about the mean $[\sum (X - \overline{X})]$ is always zero, only $n - 1$ of the deviations are free to vary (the nth deviation is determined if the sum of the deviations is to be zero).

Psychomotor Abilities of Low-Birthweight Infants

An example drawn from an actual study of low-birthweight (LBW) infants will be useful at this point because that same general study can serve to illustrate both this particular t test and other t tests to be discussed later in the chapter. Nurcombe et al. (1984) reported on an intervention program for the mothers of LBW infants. These infants present special problems for their parents because they are (superficially) unresponsive and unpredictable, in addition to being at risk for physical and developmental problems. The intervention program was designed to make mothers more aware of their infants' signals and more responsive to their needs, with the expectation that th

As you can see from Figure 7.5, the distribution of t varies as a function of the degrees of freedom, which for the moment we will define as one less than the number of observations in the sample. From what we know about χ^2, this is to be expected, since the distribution of χ^2, and thus the distribution of s^2, is also a function of the degrees of freedom. As $n \Rightarrow \infty$, $p(s^2 < \sigma^2) \Rightarrow p(s^2 > \sigma^2)$. Since the skewness of the sampling distribution of s^2 disappears as the number of degrees of freedom increases, the tendency for s to underestimate σ will also disappear. Thus, for an infinitely large number of degrees of freedom, t will be normally distributed and equivalent to z.

The test of one sample mean against a known population mean, which we have just performed, is based on the assumption that the sample was drawn from a normally distributed population. This assumption is required primarily because the t was derived assuming that the mean and variance are independent, which they are with a normal distribution. In practice, however, our t statistic can reasonably be compared to the t distribution whenever the sample size is sufficiently large to produce a normal sampling distribution of the mean. Most people would suggest that an n of 25 or 30 is "sufficiently large" for most situations, and for many situations it can be considerably smaller than that.

On the other hand, Wuensch (1993, personal communication) has argued convincingly that, at least with *very* skewed distributions, the fact that n is large enough to lead to a sampling distribution of the mean that appears to be normal does not guarantee that the resulting sampling distribution of t follows Student's t distribution. The derivation of t makes assumptions both about the distribution of means (which is under the control of the central limit theorem), and the variance, which is not controlled by that theorem.

Degrees of Freedom

I have mentioned that the t distribution is a function of the degrees of freedom (*df*). For the one-sample case, $df = n - 1$; the one degree of freedom has been lost because we used the sample mean in calculating s^2. To be more precise, we obtained the variance (s^2) by calculating the deviations of the observations from their own mean $(X - \overline{X})$, rather than from the population mean $(X - \mu)$. Because the sum of the deviations about the mean $[\sum (X - \overline{X})]$ is always zero, only $n - 1$ of the deviations are free to vary (the nth deviation is determined if the sum of the deviations is to be zero).

Psychomotor Abilities of Low-Birthweight Infants

An example drawn from an actual study of low-birthweight (LBW) infants will be useful at this point because that same general study can serve to illustrate both this particular t test and other t tests to be discussed later in the chapter. Nurcombe et al. (1984) reported on an intervention program for the mothers of LBW infants. These infants present special problems for their parents because they are (superficially) unresponsive and unpredictable, in addition to being at risk for physical and developmental problems. The intervention program was designed to make mothers more aware of their infants' signals and more responsive to their needs, with the expectation that this

would decrease later developmental difficulties often encountered with LBW infants. The study included three groups of infants: an LBW experimental group, an LBW control group, and a normal-birthweight (NBW) group. Mothers of infants in the last two groups did not receive the intervention treatment.

One of the dependent variables used in this study was the Psychomotor Development Index (PDI) of the Bayley Scales of Infant Development. This scale was first administered to all infants in the study when they were 6 months old. Because we would not expect to see differences in psychomotor development between the two LBW groups as early as 6 months, it makes some sense to combine the data from the two groups and ask whether the mean for LBW infants in general is significantly different from the normative population mean of 100 usually found with this index.

The data for the LBW infants on the PDI are presented in Table 7.1, which also shows a stem-and-leaf display and a boxplot. These two displays are important for examining the general nature of the distribution of the data and for searching for the presence of outliers.

From the stem-and-leaf display, we can see that the data, although not exactly normally distributed, at least are not badly skewed. Given our sample size (56), it is reasonable to assume that the sampling distribution of the mean would be reasonably normal. One interesting and unexpected finding that is apparent from the stem-and-leaf display is the prevalence of certain scores. For example, there are five scores of 108, but no other scores between 104 and 112. Similarly, there are six scores of 120, but no other scores between 117 and 124. Notice also that, with the exception of six scores of 89, there is a relative absence of odd numbers. A complete analysis of the

TABLE 7.1 Data for LBW infants on Psychomotor Development Index (PDI)

Raw Data

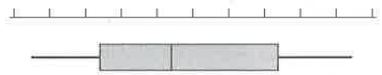
96	120	112	100
125	96	86	124
89	104	116	89
127	89	89	124
102	104	120	102
112	92	92	102
120	124	83	116
108	96	108	96
92	108	108	95
120	86	92	100
104	100	120	120
89	92	102	98
92	98	100	108
89	117	112	126

Mean = 104.125
S.D. = 12.584
N = 56

Stem-and-Leaf Display

Stem	Leaf
8*	3
8.	6 6 9 9 9 9 9 9
9*	2 2 2 2 2 2
9.	5 6 6 6 6 8 8
10*	0 0 0 0 2 2 2 2 4 4 4
10.	8 8 8 8 8
11*	2 2 2
11.	6 6 7
12*	0 0 0 0 0 0 4 4 4
12.	5 6 7

Boxplot



data requires that we at least notice these oddities and try to track down their source. It would be worthwhile to examine the scoring process to see whether there is a reason why scores often tended to fall in bunches. It is probably an artifact of the way raw scores are converted to scale scores, but it is worth checking. (In fact, if you check the scoring manual, you will find that these peculiarities are to be expected.) The fact that Tukey's exploratory data analysis (EDA) procedures lead us to notice these peculiarities is one of the great virtues of these methods. Finally, from the boxplot we can see that there are no serious outliers we need to worry about, which makes our task noticeably easier.

From the data in Table 7.1, we can see that the mean PDI score for our LBW infants is approximately 104. The norms for the PDI indicate that the population mean should be 100. Given the data, a reasonable first question concerns whether the mean of our LBW sample departs significantly from a population mean of 100. The t test is designed to answer this question.

From our formula for t and from the data, we have

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$$= \frac{104.125 - 100}{\frac{12.584}{\sqrt{56}}} = \frac{4.125}{1.682}$$

$$= 2.45$$

This value will be a member of the t distribution on $56 - 1 = 55$ df if the null hypothesis is true—that is, if the data were sampled from a population with $\mu = 100$.

A t value of 2.45 in and of itself is not particularly meaningful unless we can evaluate it against the sampling distribution of t. For this purpose, the critical values of t are presented in Appendix t. This table differs in form from the table of the normal distribution (z) because instead of giving the area above and below each specific value of t, which would require too much space, the table instead gives those values of t that cut off particular critical areas—for example, the 0.05 and 0.01 levels of significance. We saw a similar situation with respect to the χ^2 distribution. Also, in contrast to z, a different t distribution is defined for each possible number of degrees of freedom. Since we want to work at the two-tailed 0.05 level, we will want to know the value of t that cuts off $5/2 = 2.5\%$ in each tail. These critical values are generally denoted $t_{\alpha/2}$ or, in this case, $t_{0.025}$. From the table of the t distribution in Appendix t, an abbreviated version of which is shown in Table 7.2, we find that the critical value $t_{0.025}$ (rounding to 50 df for purposes of the table) is 2.009. (This is sometimes written as $t_{0.025}(50) = 2.009$ to indicate the degrees of freedom.) Because the obtained value of t, written t_{obt}, is greater than $t_{0.025}$, we will reject H_0 at $\alpha = 0.05$, two-tailed, that our sample came from a population of observations with $\mu = 100$. Instead, we will conclude that our sample of LBW children differed from the general population of children on the PDI. In fact, their mean was significantly *above* the normative population mean. This points out the advantage of using two-tailed tests, since we would have expected this group to score below the normative mean. (This might also sug

data requires that we at least notice these oddities and try to track down their source. It would be worthwhile to examine the scoring process to see whether there is a reason why scores often tended to fall in bunches. It is probably an artifact of the way raw scores are converted to scale scores, but it is worth checking. (In fact, if you check the scoring manual, you will find that these peculiarities are to be expected.) The fact that Tukey's exploratory data analysis (EDA) procedures lead us to notice these peculiarities is one of the great virtues of these methods. Finally, from the boxplot we can see that there are no serious outliers we need to worry about, which makes our task noticeably easier.

From the data in Table 7.1, we can see that the mean PDI score for our LBW infants is approximately 104. The norms for the PDI indicate that the population mean should be 100. Given the data, a reasonable first question concerns whether the mean of our LBW sample departs significantly from a population mean of 100. The t test is designed to answer this question.

From our formula for t and from the data, we have

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$$= \frac{104.125 - 100}{\frac{12.584}{\sqrt{56}}} = \frac{4.125}{1.682}$$

$$= 2.45$$

This value will be a member of the t distribution on $56 - 1 = 55$ df if the null hypothesis is true—that is, if the data were sampled from a population with $\mu = 100$.

A t value of 2.45 in and of itself is not particularly meaningful unless we can evaluate it against the sampling distribution of t. For this purpose, the critical values of t are presented in Appendix t. This table differs in form from the table of the normal distribution (z) because instead of giving the area above and below each specific value of t, which would require too much space, the table instead gives those values of t that cut off particular critical areas—for example, the 0.05 and 0.01 levels of significance. We saw a similar situation with respect to the χ^2 distribution. Also, in contrast to z, a different t distribution is defined for each possible number of degrees of freedom. Since we want to work at the two-tailed 0.05 level, we will want to know the value of t that cuts off $5/2 = 2.5\%$ in each tail. These critical values are generally denoted $t_{\alpha/2}$ or, in this case, $t_{0.025}$. From the table of the t distribution in Appendix t, an abbreviated version of which is shown in Table 7.2, we find that the critical value of $t_{0.025}$ (rounding to 50 df for purposes of the table) is 2.009. (This is sometimes written as $t_{0.025}(50) = 2.009$ to indicate the degrees of freedom.) Because the obtained value of t, written t_{obt}, is greater than $t_{0.025}$, we will reject H_0 at $\alpha = 0.05$, two-tailed, that our sample came from a population of observations with $\mu = 100$. Instead, we will conclude that our sample of LBW children differed from the general population of children on the PDI. In fact, their mean was significantly *above* the normative population mean. This points out the advantage of using two-tailed tests, since we would have expected this group to score below the normative mean. (This might also suggest that we check our scoring procedures to make sure we are not systematically overscoring our subjects. In fact, however, a number of other studies using the Bayley have reported similarly high means.)

TABLE 7.2 Percentage points of the t distribution

	Level of Significance for One-Tailed Test								
	0.25	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0005**
	Level of Significance for Two-Tailed Test								
df	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.001**
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.62
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
...	...	...	...	...	...	...	...	...	...
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
50	0.679	0.849	1.047	1.299	1.676	**2.009**	2.403	2.678	3.496
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.390
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

Source: The entries in this table were computed by the author.

The Moon Illusion

It will be useful to consider a second example, this one taken from a classic paper by Kaufman and Rock (1962) on the moon illusion.[5] Kaufman and Rock concluded that

[5]A very recent paper on this topic by Lloyd Kaufman and his son James Kaufman was published in the January 2000 issue of the *Proceedings of the National Academy of Sciences.*

the commonly observed fact that the moon near the horizon appears larger than does the moon at its zenith (highest point overhead) could be explained on the basis of the greater *apparent* distance of the moon when it is at the horizon. As part of a very complete series of experiments, the authors initially sought to estimate the moon illusion by asking subjects to adjust a variable "moon" that appeared to be on the horizon so as to match the size of a standard "moon" that appeared at its zenith, or vice versa. (In these measurements, they used not the actual moon but an artificial one created with special apparatus.) One of the first questions we might ask is whether there really is a moon illusion—that is, whether a larger setting is required to match a horizon moon or a zenith moon. The following data for 10 subjects are taken from Kaufman and Rock's paper and present the ratio of the diameter of the variable and standard moons. A ratio of 1.00 would indicate no illusion, whereas a ratio other than 1.00 would represent an illusion. (For example, a ratio of 1.50 would mean that the horizon moon appeared to have a diameter 1.50 times the diameter of the zenith moon.) Evidence in support of an illusion would require that we reject H_0: $\mu = 1.00$ in favor of H_1: $\mu \neq 1.00$.

Obtained ratio:	1.73	1.06	2.03	1.40	0.95
	1.13	1.41	1.73	1.63	1.56

For these data, $n = 10$, $\overline{X} = 1.463$, and $s = 0.341$. A t test on H_0: $\mu = 1.00$ is given by

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$$= \frac{1.463 - 1.000}{\frac{0.341}{\sqrt{10}}} = \frac{0.463}{0.108}$$

$$= 4.29$$

From Appendix t, with $10 - 1 = 9$ df for a two-tailed test at $\alpha = 0.05$, the critical value of $t_{0.025}(9)$ is ± 2.262. The obtained value of t was 4.29. Since $4.29 > 2.262$, we can reject H_0 at $\alpha = 0.05$ and conclude that the true mean ratio under these conditions is not equal to 1.00. In fact, it is greater than 1.00, which is what we would expect on the basis of our experience. (It is always comforting to see science confirm what we have all known since childhood, but in this case the results also indicate that Kaufman and Rock's experimental apparatus performed as it should.)

Using Minitab to Run One-Sample t Tests

With a large data set, it is often convenient to use a program such as Minitab to compute t values. Exhibit 7.1 shows how Minitab can be used to obtain a one-sample t test and confidence limits for the moon-illusion data. (Confidence limits will be discussed in Section 7.7; for now you can ignore them.) To get both the t test and the confidence limits, you have to specify separate analyses by clicking on different radio buttons. These buttons are shown in the first part of Exhibit 7.1. Notice that Minitab's results agree, within rounding error, with those we obtained by hand. Notice also that

the commonly observed fact that the moon near the horizon appears larger than does the moon at its zenith (highest point overhead) could be explained on the basis of the greater *apparent* distance of the moon when it is at the horizon. As part of a very complete series of experiments, the authors initially sought to estimate the moon illusion by asking subjects to adjust a variable "moon" that appeared to be on the horizon so as to match the size of a standard "moon" that appeared at its zenith, or vice versa. (In these measurements, they used not the actual moon but an artificial one created with special apparatus.) One of the first questions we might ask is whether there really is a moon illusion—that is, whether a larger setting is required to match a horizon moon or a zenith moon. The following data for 10 subjects are taken from Kaufman and Rock's paper and present the ratio of the diameter of the variable and standard moons. A ratio of 1.00 would indicate no illusion, whereas a ratio other than 1.00 would represent an illusion. (For example, a ratio of 1.50 would mean that the horizon moon appeared to have a diameter 1.50 times the diameter of the zenith moon.) Evidence in support of an illusion would require that we reject H_0: $\mu = 1.00$ in favor of H_1: $\mu \neq 1.00$.

Obtained ratio:	1.73	1.06	2.03	1.40	0.95
	1.13	1.41	1.73	1.63	1.56

For these data, $n = 10$, $\overline{X} = 1.463$, and $s = 0.341$. A t test on H_0: $\mu = 1.00$ is given by

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

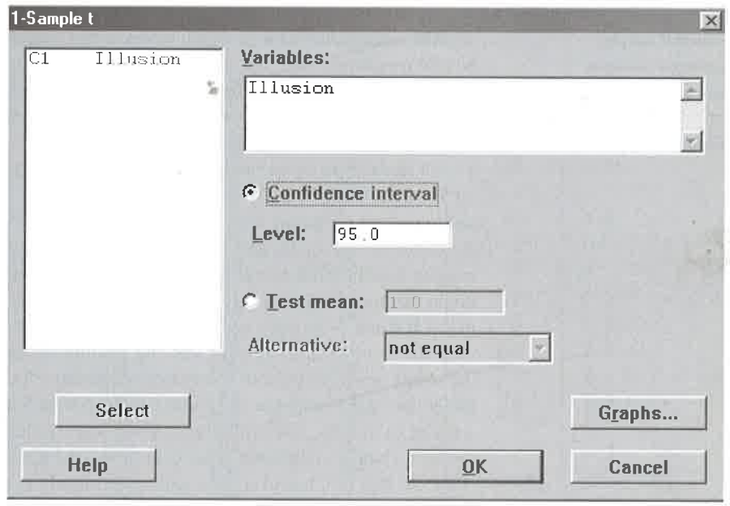
$$= \frac{1.463 - 1.000}{\frac{0.341}{\sqrt{10}}} = \frac{0.463}{0.108}$$

$$= 4.29$$

From Appendix t, with $10 - 1 = 9$ df for a two-tailed test at $\alpha = 0.05$, the critical value of $t_{0.025}(9)$ is ± 2.262. The obtained value of t was 4.29. Since $4.29 > 2.262$, we can reject H_0 at $\alpha = 0.05$ and conclude that the true mean ratio under these conditions is not equal to 1.00. In fact, it is greater than 1.00, which is what we would expect on the basis of our experience. (It is always comforting to see science confirm what we have all known since childhood, but in this case the results also indicate that Kaufman and Rock's experimental apparatus performed as it should.)

Using Minitab to Run One-Sample t Tests

With a large data set, it is often convenient to use a program such as Minitab to compute t values. Exhibit 7.1 shows how Minitab can be used to obtain a one-sample t test and confidence limits for the moon-illusion data. (Confidence limits will be discussed in Section 7.7; for now you can ignore them.) To get both the t test and the confidence limits, you have to specify separate analyses by clicking on different radio buttons. These buttons are shown in the first part of Exhibit 7.1. Notice that Minitab's results agree, within rounding error, with those we obtained by hand. Notice also that



T Confidence Intervals

Variable	N	Mean	StDev	SE Mean	95.0 % CI
Illusion	10	1.463	0.341	0.108	(1.219, 1.707)

T-Test of the Mean

Test of mu = 1.000 vs mu not = 1.000

Variable	N	Mean	StDev	SE Mean	T	P
Illusion	10	1.463	0.341	0.108	4.30	0.0020

EXHIBIT 7.1 Minitab for one-sample t test and confidence limits

p level

Minitab computes the exact probability of a Type I error (the ***p* level**), rather than comparing t to a tabled value. Thus, whereas we concluded that the probability of a Type I error was *less than* 0.05, Minitab reveals that the actual probability is 0.0020. Most computer programs operate in this way.

7.4 Hypothesis Tests Applied to Means—Two Matched Samples

In Section 7.3 we considered the situation in which we had one sample mean ($\overline{X}$) and wished to test to see whether it was reasonable to believe that such a sample mean would have occurred if we had been sampling from a population with some specified mean (often denoted μ_0). Another way of phrasing this is to say that we were testing to determine whether the mean of the population from which we sampled (call it μ_1)

matched samples
repeated measures
related samples

matched-sample *t* test

was equal to some particular value given by the null hypothesis (μ_0). In this section we will consider the case in which we have two **matched samples** (often called **repeated measures,** when the same subjects respond on two occasions, or **related samples,** correlated samples, paired samples, or dependent samples) and wish to perform a test on the difference between their two means. In this case we want what is sometimes called the **matched-sample *t* test.**

As an example of the analysis of matched (related) samples, we will consider the intervention study of LBW infants referred to earlier. Part of the study on LBW infants involved collecting data on the Bayley Mental Development Index (MDI) when the children were 6, 12, and 24 months old. Previous work on LBW children would suggest that the MDI scores for the control group of nonintervention LBW children might decline noticeably between 6 and 24 months of age. The data for the control group at 6 and 24 months, and the differences between them, are shown in Table 7.3, with the stem-and-leaf display and the boxplot on the difference scores. Because these data represent pairs of scores from each child, the two sets of scores are related rather than independent (a child with a high score at 6 months is likely to have a high score at 24 months, and similarly for a low-scoring child). Therefore, the appropriate test for a change in the mean score over time is the matched-sample *t* test. Before carrying out that test, however, let's look more closely at the data.

We want a *t* test on the difference between the 6- and 24-month means for these matched samples. We will designate the first set of scores as X_1 and the second set as X_2. The null hypothesis we want to test is the hypothesis that the mean (μ_1) of the population of scores from which the first set of data was drawn is equal to the mean (μ_2) of the population from which the second set of data was drawn. In other words we want to test

$$H_0: \mu_1 = \mu_2$$

or, equivalently,

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

Note that we have no interest in the values of μ_1 and μ_2 themselves, but only in whether they are equal. Note also that we will be testing this null hypothesis using data from matched samples.

Difference Scores

difference scores

Although it would seem most obvious to view the data as representing two samples of scores, one set obtained at 6 months and the other at 24 months, it is also possible, and very profitable, to transform the data into one set of scores—the set of differences between X_1 and X_2 for each subject. These differences are called **difference scores** and are indicated in the third column of Table 7.3. These scores are often represented by D (for difference) and can be thought of as the degree of improvement or decrement between the two testing times. If in fact the MDI scores of LBW children do not in general decrease over time (i.e., if H_0 is true), the average score would not change from session to session. By chance, some subjects would happen to have

19

matched samples
repeated measures
related samples

matched-sample *t* test

was equal to some particular value given by the null hypothesis (μ_0). In this section we will consider the case in which we have two **matched samples** (often called **repeated measures,** when the same subjects respond on two occasions, or **related samples,** correlated samples, paired samples, or dependent samples) and wish to perform a test on the difference between their two means. In this case we want what is sometimes called the **matched-sample *t* test.**

As an example of the analysis of matched (related) samples, we will consider the intervention study of LBW infants referred to earlier. Part of the study on LBW infants involved collecting data on the Bayley Mental Development Index (MDI) when the children were 6, 12, and 24 months old. Previous work on LBW children would suggest that the MDI scores for the control group of nonintervention LBW children might decline noticeably between 6 and 24 months of age. The data for the control group at 6 and 24 months, and the differences between them, are shown in Table 7.3, with the stem-and-leaf display and the boxplot on the difference scores. Because these data represent pairs of scores from each child, the two sets of scores are related rather than independent (a child with a high score at 6 months is likely to have a high score at 24 months, and similarly for a low-scoring child). Therefore, the appropriate test for a change in the mean score over time is the matched-sample *t* test. Before carrying out that test, however, let's look more closely at the data.

We want a *t* test on the difference between the 6- and 24-month means for these matched samples. We will designate the first set of scores as X_1 and the second set as X_2. The null hypothesis we want to test is the hypothesis that the mean (μ_1) of the population of scores from which the first set of data was drawn is equal to the mean (μ_2) of the population from which the second set of data was drawn. In other words, we want to test

$$H_0: \mu_1 = \mu_2$$

or, equivalently,

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

Note that we have no interest in the values of μ_1 and μ_2 themselves, but only in whether they are equal. Note also that we will be testing this null hypothesis using data from matched samples.

Difference Scores

difference scores

Although it would seem most obvious to view the data as representing two samples of scores, one set obtained at 6 months and the other at 24 months, it is also possible, and very profitable, to transform the data into one set of scores—the set of differences between X_1 and X_2 for each subject. These differences are called **difference scores** and are indicated in the third column of Table 7.3. These scores are often represented by D (for difference) and can be thought of as the degree of improvement or decrement between the two testing times. If in fact the MDI scores of LBW children do not in general decrease over time (i.e., if H_0 is true), the average score would not change from session to session. By chance, some subjects would happen to have a

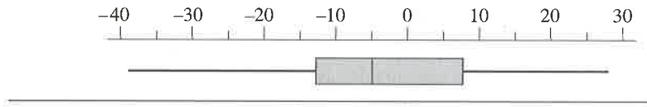
TABLE 7.3 Data and difference scores on Mental Development Index (MDI) for the LBW control group at 6 and 24 months of age

MDI-6 Months	MDI-24 Months	Difference (D)
124	114	−10
94	88	−6
115	102	−13
110	127	17
116	104	−12
139	104	−35
116	91	−25
110	96	−14
129	104	−25
120	106	−14
105	91	−14
88	102	14
120	104	−16
120	100	−20
116	114	−2
105	109	4
100	109	9
91	119	28
129	91	−38
84	81	−3
91	114	23
116	119	3
100	102	2
113	111	−2
89	80	−9
102	119	17
110	119	9
116	123	7
124	119	−5
126	114	−12
123	132	9
Mean	111.0	106.71	−4.29
S.D.	13.85	12.95	16.04
n	31	31	31

Stem	Leaf
−3.	5 8
−3*	
−2.	5 5
−2*	0
−1.	6
−1*	0 2 2 3 4 4 4
−0.	5 6 9
−0*	2 2 3
0*	2 3 4
0.	7 9 9 9
1*	4
1.	7 7
2*	3
2.	8

Code |2*|3 = 23

Boxplot



higher score on X_2 than on X_1 and some would have a lower score, but *on the average* there would be no difference.

From the stem-and-leaf display in Table 7.3, we can see that the distribution is slightly, but not markedly, skewed. This display highlights the large variability in the difference scores. Two children showed a decrease (negative difference score) of over 30 points, whereas two other children increased their scores by at least 20 points. This substantial variability does not invalidate the t test, but it does suggest that the question of whether the MDI *means* decrease may not be the only one we should be asking. In fact, although we will not do so here, it may be more important to try to explain the large variability in difference scores—why do some children show so much improvement while others show so much deterioration? (For example, using the techniques to be discussed in Chapter 9, you might plot the degree of change against initial score level.) The boxplot does not indicate the presence of any outliers. Although large positive and negative scores occur here, they do not qualify as outliers because they are not uniquely large—the large standard deviation of the difference scores has already been discussed.

If we now think of our data as being the column of difference scores, the null hypothesis becomes the hypothesis that the mean of a population of difference scores (denoted μ_D) equals zero. Since it can be shown that $\mu_D = \mu_1 - \mu_2$, we can write H_0: $\mu_D = \mu_1 - \mu_2 = 0$. But now we can see that we are testing a hypothesis using *one* sample of data (the sample of difference scores), and we already know how to do that from Section 7.3.

The t Statistic

We are now at precisely the same place we were with the one-sample t test when we had a sample of data and a null hypothesis ($\mu = 0$). The only difference is that in this case the data are difference scores, and the mean and standard deviation are based on the differences. Recall that t was defined as the difference between a sample mean and a population mean, divided by the standard error of the mean. Then we have

$$t = \frac{\overline{D} - 0}{s_{\overline{D}}} = \frac{\overline{D} - 0}{\frac{s_D}{\sqrt{n}}}$$

where $\overline{D}$ and $s_{\overline{D}}$ are the mean and standard deviation of the difference scores and n is the number of difference scores (the number of *pairs*, not the number of raw scores). For our data

$$t = \frac{\overline{D} - 0}{s_{\overline{D}}} = \frac{-4.29 - 0}{\frac{16.04}{\sqrt{31}}} = \frac{-4.29}{2.88} = -1.49$$

Degrees of Freedom

The degrees of freedom for the matched-sample case are the same as they were for the one-sample case. Since we are working with the difference scores, n will be equal

higher score on X_2 than on X_1 and some would have a lower score, but *on the average* there would be no difference.

From the stem-and-leaf display in Table 7.3, we can see that the distribution is slightly, but not markedly, skewed. This display highlights the large variability in the difference scores. Two children showed a decrease (negative difference score) of over 30 points, whereas two other children increased their scores by at least 20 points. This substantial variability does not invalidate the t test, but it does suggest that the question of whether the MDI *means* decrease may not be the only one we should be asking. In fact, although we will not do so here, it may be more important to try to explain the large variability in difference scores—why do some children show so much improvement while others show so much deterioration? (For example, using the techniques to be discussed in Chapter 9, you might plot the degree of change against initial score level.) The boxplot does not indicate the presence of any outliers. Although large positive and negative scores occur here, they do not qualify as outliers because they are not uniquely large—the large standard deviation of the difference scores has already been discussed.

If we now think of our data as being the column of difference scores, the null hypothesis becomes the hypothesis that the mean of a population of difference scores (denoted μ_D) equals zero. Since it can be shown that $\mu_D = \mu_1 - \mu_2$, we can write H_0: $\mu_D = \mu_1 - \mu_2 = 0$. But now we can see that we are testing a hypothesis using *one* sample of data (the sample of difference scores), and we already know how to do that from Section 7.3.

The t Statistic

We are now at precisely the same place we were with the one-sample t test when we had a sample of data and a null hypothesis ($\mu = 0$). The only difference is that in this case the data are difference scores, and the mean and standard deviation are based on the differences. Recall that t was defined as the difference between a sample mean and a population mean, divided by the standard error of the mean. Then we have

$$t = \frac{\overline{D} - 0}{s_{\overline{D}}} = \frac{\overline{D} - 0}{\frac{s_D}{\sqrt{n}}}$$

where $\overline{D}$ and $s_{\overline{D}}$ are the mean and standard deviation of the difference scores and n is the number of difference scores (the number of *pairs*, not the number of raw scores). For our data

$$t = \frac{\overline{D} - 0}{s_{\overline{D}}} = \frac{-4.29 - 0}{\frac{16.04}{\sqrt{31}}} = \frac{-4.29}{2.88} = -1.49$$

Degrees of Freedom

The degrees of freedom for the matched-sample case are the same as they were for the one-sample case. Since we are working with the difference scores, n will be equal

to the number of differences (or the number of *pairs* of observations, or the number of *independent* observations, all of which amount to the same thing). Since the variance of these difference scores (s_D^2) is used as an estimate of the variance of a population of difference scores (σ_D^2), and since this sample variance is obtained using the sample mean ($\overline{D}$), we will lose one df to the mean and have $n - 1$ df.

Because we have 31 difference scores in this example, we will have 30 degrees of freedom. From Appendix t we find that for a two-tailed test at the 0.05 level of significance, $t_{0.025}(30) = \pm 2.042$. Our obtained value of t (-1.49) is smaller than the critical value; thus we will not reject H_0. We have no reason to doubt that our difference scores were sampled from a population of difference scores where $\mu_D = 0$. In practical terms, this means that we have not shown that the control group of LBW infants exhibited a decreasing level of performance over time. Although this finding is very positive from the point of view of the children involved (who are, after all, the important people!), it makes it more difficult to show that the intervention program is useful. An intervention program looks best when you can show that people who do *not* receive it *deteriorate* over time, whereas those who *do* receive it stay the same or even *improve.*

The Moon Illusion Revisited

As a second example, we will return to the work by Kaufman and Rock (1962) on the moon illusion. An important hypothesis about the source of the moon illusion was put forth by Holway and Boring (1940), who suggested that the illusion was due to the fact that when the moon was on the horizon, the observer looked straight at it with eyes level, whereas when it was at its zenith, the observer had to elevate his eyes as well as his head. Holway and Boring proposed that this difference in the elevation of the eyes was the cause of the illusion. Kaufman and Rock thought differently. To test Holway and Boring's hypothesis, Kaufman and Rock devised an apparatus that allowed them to present two artificial moons (one at the horizon and one at the zenith) and to control whether the subjects elevated their eyes to see the zenith moon. In one case, the subject was forced to put his head in such a position as to be able to see the zenith moon with eyes level. In the other case, the subject was forced to see the zenith moon with eyes raised. (The horizon moon was always viewed with eyes level.) In both cases, the dependent variable was the ratio of the perceived size of the horizon moon to the perceived size of the zenith moon (a ratio of 1.00 would represent no illusion). If Holway and Boring were correct, there should have been a greater illusion (larger ratio) in the eyes-elevated condition than in the eyes-level condition, although the moon was always perceived to be in the same place, the zenith. The actual data for this experiment are given in Table 7.4 on page 196.

In this example, we want to test the *null hypothesis* that the means are equal under the two viewing conditions. Because we are dealing with related observations (each subject served under both conditions), we will work with the difference scores and test H_0: $\mu_D = 0$. Using a two-tailed test at $\alpha = 0.05$, the alternative hypothesis is H_1: $\mu_D \neq 0$.

TABLE 7.4 Magnitude of the moon illusion when zenith moon is viewed with eyes level and with eyes elevated

Observer	Eyes Elevated	Eyes Level	Difference (D)
1	1.65	1.73	–0.08
2	1.00	1.06	–0.06
3	2.03	2.03	0.00
4	1.25	1.40	–0.15
5	1.05	0.95	0.10
6	1.02	1.13	–0.11
7	1.67	1.41	0.26
8	1.86	1.73	0.13
9	1.56	1.63	–0.07
10	1.73	1.56	0.17
			$\overline{D} = 0.019$
			$s_D = 0.137$
			$s_{\overline{D}} = 0.043$

From the formula for a t test on related samples, we have

$$t = \frac{\overline{D} - 0}{s_{\overline{D}}} = \frac{\overline{D} - 0}{\frac{s_D}{\sqrt{n}}}$$

$$= \frac{0.019 - 0}{\frac{0.137}{\sqrt{10}}} = \frac{0.019}{0.043}$$

$$= 0.44$$

From Appendix t, we find that $t_{0.025}(9) = \pm 2.262$. Since $t_{obt} = 0.44$ is less th 2.262, we will fail to reject H_0 and will decide that we have no evidence to sugg that the illusion is affected by the elevation of the eyes.[6] (In fact, these data also clude a second test of Holway and Boring's hypothesis since they would have p dicted that there would not be an illusion if subjects viewed the zenith moon with e level. On the contrary, the data reveal a considerable illusion under this condition test of the significance of the illusion with eyes level can be obtained by the meth discussed in the previous section, and the illusion is in fact significant.)

[6] A glance at Appendix t will reveal that a t less than 1.96 (the critical value for z) will never be signific $\alpha = 0.05$, regardless of the number of degrees of freedom. Moreover, unless you have at least 50 degree freedom, t values less than 2.00 will not be significant, often making it unnecessary for you even to bot looking at the table of t.

TABLE 7·4 Magnitude of the moon illusion when zenith moon is viewed with eyes level and with eyes elevated

Observer	Eyes Elevated	Eyes Level	Difference (D)
1	1.65	1.73	–0.08
2	1.00	1.06	–0.06
3	2.03	2.03	0.00
4	1.25	1.40	–0.15
5	1.05	0.95	0.10
6	1.02	1.13	–0.11
7	1.67	1.41	0.26
8	1.86	1.73	0.13
9	1.56	1.63	–0.07
10	1.73	1.56	0.17
			$\overline{D} = 0.019$
			$s_D = 0.137$
			$s_{\overline{D}} = 0.043$

From the formula for a t test on related samples, we have

$$t = \frac{\overline{D} - 0}{s_{\overline{D}}} = \frac{\overline{D} - 0}{\frac{s_D}{\sqrt{n}}}$$

$$= \frac{0.019 - 0}{\frac{0.137}{\sqrt{10}}} = \frac{0.019}{0.043}$$

$$= 0.44$$

From Appendix t, we find that $t_{0.025}(9) = \pm 2.262$. Since $t_{obt} = 0.44$ is less than 2.262, we will fail to reject H_0 and will decide that we have no evidence to suggest that the illusion is affected by the elevation of the eyes.[6] (In fact, these data also include a second test of Holway and Boring's hypothesis since they would have predicted that there would not be an illusion if subjects viewed the zenith moon with eyes level. On the contrary, the data reveal a considerable illusion under this condition. A test of the significance of the illusion with eyes level can be obtained by the methods discussed in the previous section, and the illusion is in fact significant.)

[6]A glance at Appendix t will reveal that a t less than 1.96 (the critical value for z) will never be significant at $\alpha = 0.05$, regardless of the number of degrees of freedom. Moreover, unless you have at least 50 degrees of freedom, t values less than 2.00 will not be significant, often making it unnecessary for you even to bother looking at the table of t.

Matched Samples

In many, but certainly not all, situations in which we will use the matched-sample t test, we will have two sets of data from the same subjects. For example, we might ask each of 20 people to rate their level of anxiety before and after donating blood. Or we might record ratings of level of disability made using two different scoring systems for each of 20 handicapped individuals in an attempt to see whether one scoring system leads to generally lower assessments than does the other. In both examples, we would have 20 pairs of numbers, two numbers for each person, and would expect these two sets of numbers to be related (or, in the terminology we will later adopt, to be correlated). Consider the blood-donation example. People differ widely in level of anxiety. Some seem to be anxious all of the time no matter what happens, and others just take things as they come and do not worry about anything. Thus, there should be a relationship between an individual's anxiety level before donating blood and her anxiety level after donating blood. In other words, if we know what a person's anxiety score was before donation, we can make a reasonable guess what it was after donation. Similarly, some people are severely handicapped whereas others are only mildly handicapped. If we know that a particular person received a high assessment using one scoring system, it is likely that he also received a relatively high assessment using the other system. The relationship between data sets does not have to be perfect—it probably never will be. The fact that we can make better-than-chance predictions is sufficient to classify two sets of data as matched or related.

In the two preceding examples, I chose situations in which each person in the study contributed two scores. Although this is the most common way of obtaining related samples, it is not the only way. For example, a study of marital relationships might involve asking husbands and wives to rate their satisfaction with their marriage, with the goal of testing to see whether wives are, on average, more or less satisfied than husbands. (You will see an example of just such a study in the exercises for this chapter.) Here each individual would contribute only one score, but the couple *as a unit* would contribute a pair of scores. It is reasonable to assume that if the husband is very dissatisfied with the marriage, his wife is probably also dissatisfied, and vice versa, and thus their scores are related.

Many experimental designs involve related samples. They all have one thing in common, and that is the fact that knowing one member of a pair of scores tells you something—maybe not much, but something—about the other member. Whenever this is the case, we say that the samples are matched.

Missing Data

Ideally, with matched samples we have a score on each variable for each case or pair of cases. If a subject participates in the pretest, she also participates in the posttest. If one member of a couple provides data, so does the other member. When we are finished collecting data, we have a complete set of paired scores. Unfortunately, experiments do not usually work out as cleanly as we would like.

Suppose, for example, that we want to compare scores on a checklist of children's behavior problems completed by mothers and fathers, with the expectation that mothers are more sensitive to their children's problems than are fathers, and thus will produce higher scores. Most of the time both parents will complete the form. But there might be 10 cases where the mother sent in her form but the father did not, and 5 cases where we have a form from the father but not from the mother. The normal procedure in this situation is to eliminate the 15 pairs of parents where we do not have complete data, and then run a matched-sample t test on the data that remain. This is the way almost everyone would analyze the data. There is an alternative, however, that allows us to use all of the data if we are willing to assume that data are missing at random and not systematically. (By this I mean that we have to assume that we are not more likely to be missing Dad's data when the child is reported by Mom to have very few problems, nor are we less likely to be missing Dad's data for a very behaviorally disordered child.)

Bohj (1978) proposed an ingenious test in which you basically compute a matched-sample t for those cases in which both scores are present, then compute an additional independent group t (to be discussed next) between the scores of mothers without fathers and fathers without mothers, and finally combine the two t statistics. This combined t can then be evaluated against special tables. These tables are available in Wilcox (1986), and approximations to critical values of this combined statistic are discussed briefly in Wilcox (1987a). This test is sufficiently awkward that you would not use it simply because you are missing two or three observations. But it can be extremely useful when many pieces of data are missing. For a more extensive discussion, see Wilcox (1987b).

Using Computer Software for t Tests on Matched Samples

The use of almost any computer software to analyze matched samples *can* involve nothing more than a command to create a variable that is the difference between the two scores we are comparing. We then run a simple one-sample t test to test the null hypothesis that those difference scores came from a population with a mean of 0. Alternatively, some software, such as SPSS, allows you to specify that you want a t on two related samples, and then to specify the two variables that represent those samples. Since this is very similar to what we have already done, I will not repeat that here.

7·5 Hypothesis Tests Applied to Means—Two Independent Samples

One of the most common uses of the t test involves testing the difference between the means of two independent groups. We might wish to compare the mean number of trials needed to reach criterion on a simple visual discrimination task for two groups of rats—one raised under normal conditions and one raised under conditions of sensory deprivation. Or we might wish to compare the mean levels of retention of a group

Suppose, for example, that we want to compare scores on a checklist of children's behavior problems completed by mothers and fathers, with the expectation that mothers are more sensitive to their children's problems than are fathers, and thus will produce higher scores. Most of the time both parents will complete the form. But there might be 10 cases where the mother sent in her form but the father did not, and 5 cases where we have a form from the father but not from the mother. The normal procedure in this situation is to eliminate the 15 pairs of parents where we do not have complete data, and then run a matched-sample t test on the data that remain. This is the way almost everyone would analyze the data. There is an alternative, however, that allows us to use all of the data if we are willing to assume that data are missing at random and not systematically. (By this I mean that we have to assume that we are not more likely to be missing Dad's data when the child is reported by Mom to have very few problems, nor are we less likely to be missing Dad's data for a very behaviorally disordered child.)

Bohj (1978) proposed an ingenious test in which you basically compute a matched-sample t for those cases in which both scores are present, then compute an additional independent group t (to be discussed next) between the scores of mothers without fathers and fathers without mothers, and finally combine the two t statistics. This combined t can then be evaluated against special tables. These tables are available in Wilcox (1986), and approximations to critical values of this combined statistic are discussed briefly in Wilcox (1987a). This test is sufficiently awkward that you would not use it simply because you are missing two or three observations. But it can be extremely useful when many pieces of data are missing. For a more extensive discussion, see Wilcox (1987b).

Using Computer Software for t Tests on Matched Samples

The use of almost any computer software to analyze matched samples *can* involve nothing more than a command to create a variable that is the difference between the two scores we are comparing. We then run a simple one-sample t test to test the null hypothesis that those difference scores came from a population with a mean of 0. Alternatively, some software, such as SPSS, allows you to specify that you want a t on two related samples, and then to specify the two variables that represent those samples. Since this is very similar to what we have already done, I will not repeat that here.

7.5 Hypothesis Tests Applied to Means—Two Independent Samples

One of the most common uses of the t test involves testing the difference between the means of two independent groups. We might wish to compare the mean number of trials needed to reach criterion on a simple visual discrimination task for two groups of rats—one raised under normal conditions and one raised under conditions of sensory deprivation. Or we might wish to compare the mean levels of retention of a group of college students asked to recall active declarative sentences and a group asked to recall passive negative sentences. Or we might place subjects in a situation in which another person needed help; we could compare the latency of helping behavior when subjects were tested alone and when they were tested in groups.

In conducting any experiment with two independent groups, we would most likely find that the two sample means differed by some amount. The important question, however, is whether this difference is sufficiently large to justify the conclusion that the two samples were drawn from different populations—that is, using the example of helping behavior, is the mean of the population of latencies from singly tested subjects different from the mean of the population of latencies from group-tested subjects? Before we consider a specific example, however, we will need to examine the sampling distribution of differences between means and the t test that results from it.

Distribution of Differences Between Means

sampling distribution of differences between means

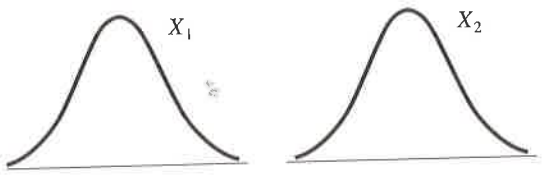
When we are interested in testing for a difference between the mean of one population (μ_1) and the mean of a second population (μ_2), we will be testing a null hypothesis of the form H_0: $\mu_1 - \mu_2 = 0$ or, equivalently, $\mu_1 = \mu_2$. Because the test of this null hypothesis involves the difference between independent sample means, it is important that we digress for a moment and examine the **sampling distribution of differences between means.**

Suppose that we have two populations labeled X_1 and X_2 with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. We now draw pairs of samples of size n_1 from population X_1 and of size n_2 from population X_2, and record the means and the difference between the means for each pair of samples. Because we are sampling independently from each population, the sample means will be independent. (Means are paired only in the trivial and presumably irrelevant sense of being drawn at the same time.) The results of an infinite number of replications of this procedure are presented schematically in Figure 7.6 on page 200. In the lower portion of this figure, the first two columns represent the sampling distributions of $\overline{X}_1$ and $\overline{X}_2$, and the third column represents the sampling distribution of mean differences ($\overline{X}_1 - \overline{X}_2$). It is this third column we are most interested in, since we are concerned with testing differences between means. The mean of this distribution can be shown to equal $\mu_1 - \mu_2$. The variance of this distribution of differences is given by what is commonly called the **variance sum law,** a limited form of which states

variance sum law

> The variance of a sum or difference of two *independent* variables is equal to the sum of their variances.[7]

[7]The complete form of the law omits the restriction that the variables must be independent and states that the variance of their sum or difference is $\sigma^2_{X_1 \pm X_2} = \sigma_1^2 + \sigma_2^2 \pm 2\rho\sigma_1\sigma_2$ where the notation $\pm$ is interpreted as plus when we are speaking of their sum and as minus when we are speaking of their difference. The term ρ (rho) in this equation is the correlation between the two variables (to be discussed in Chapter 9) and is equal to zero when the variables are independent. (The fact that $\rho \neq 0$ when the variables are not independent was what forced us to treat the related-sample case separately.)



	$\overline{X}_{11}$	$\overline{X}_{21}$	$\overline{X}_{11} - \overline{X}_{21}$
	$\overline{X}_{12}$	$\overline{X}_{22}$	$\overline{X}_{12} - \overline{X}_{22}$
	$\overline{X}_{13}$	$\overline{X}_{23}$	$\overline{X}_{13} - \overline{X}_{23}$
	...	...	...
	$\overline{X}_{1\infty}$	$\overline{X}_{2\infty}$	$\overline{X}_{1\infty} - \overline{X}_{2\infty}$
Mean	μ_1	μ_2	$\mu_1 - \mu_2$
Variance	$\frac{\sigma_1^2}{n_1}$	$\frac{\sigma_2^2}{n_2}$	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
S.D.	$\frac{\sigma_1}{\sqrt{n_1}}$	$\frac{\sigma_2}{\sqrt{n_2}}$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

FIGURE 7.6 *Schematic set of means and mean differences when sampling from two populations*

We know from the central limit theorem that the variance of the distribution of $\overline{X}_1$ is σ_1^2/n_1 and the variance of the distribution of $\overline{X}_2$ is σ_2^2/n_2. Since the variables (sample means) are independent, the variance of the difference of these two variables is the sum of their variances. Thus

$$\sigma^2_{\overline{X}_1 - \overline{X}_2} = \sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Having found the mean and the variance of a set of differences between means, we know most of what we need to know. The general form of the sampling distribution of mean differences is presented in Figure 7.7.

The final point to be made about this distribution concerns its shape. An important theorem in statistics states that the sum or difference of two independent normally

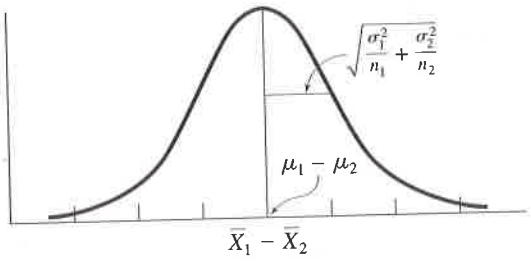
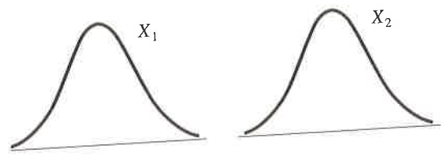


FIGURE 7.7 *Sampling distribution of mean differences*



	$\overline{X}_{11}$	$\overline{X}_{21}$	$\overline{X}_{11} - \overline{X}_{21}$
	$\overline{X}_{12}$	$\overline{X}_{22}$	$\overline{X}_{12} - \overline{X}_{22}$
	$\overline{X}_{13}$	$\overline{X}_{23}$	$\overline{X}_{13} - \overline{X}_{23}$
	...	...	...
	$\overline{X}_{1\infty}$	$\overline{X}_{2\infty}$	$\overline{X}_{1\infty} - \overline{X}_{2\infty}$
Mean	μ_1	μ_2	$\mu_1 - \mu_2$
Variance	$\frac{\sigma_1^2}{n_1}$	$\frac{\sigma_2^2}{n_2}$	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
S.D.	$\frac{\sigma_1}{\sqrt{n_1}}$	$\frac{\sigma_2}{\sqrt{n_2}}$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

FIGURE 7.6 *Schematic set of means and mean differences when sampling from two populations*

We know from the central limit theorem that the variance of the distribution of $\overline{X}_1$ is σ_1^2/n_1 and the variance of the distribution of $\overline{X}_2$ is σ_2^2/n_2. Since the variables (sample means) are independent, the variance of the difference of these two variables is the sum of their variances. Thus

$$\sigma^2_{\overline{X}_1-\overline{X}_2} = \sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Having found the mean and the variance of a set of differences between means, we know most of what we need to know. The general form of the sampling distribution of mean differences is presented in Figure 7.7.

The final point to be made about this distribution concerns its shape. An important theorem in statistics states that the sum or difference of two independent normally

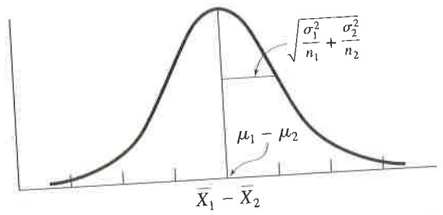


FIGURE 7.7 *Sampling distribution of mean differences*

distributed variables is itself normally distributed. Because Figure 7.7 represents the difference between two sampling distributions of the mean, and because we know that the sampling distribution of means is at least approximately normal for reasonable sample sizes, the distribution in Figure 7.7 must itself be at least approximately normal.

The *t* Statistic

Given the information we now have about the sampling distribution of mean differences, we can proceed to develop the appropriate test procedure. Assume for the moment that knowledge of the population variances (σ_i^2) is not a problem. We have earlier defined z as a statistic (a point on the distribution) minus the mean of the distribution, divided by the standard error of the distribution. Our statistic in the present case is $(\overline{X}_1 - \overline{X}_2)$, the observed difference between the sample means. The mean of the sampling distribution is $(\mu_1 - \mu_2)$, and, as we saw, the **standard error of differences between means**[8] is

standard error of differences between means

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

Thus we can write

$$z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\overline{X}_1-\overline{X}_2}}$$

$$= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The critical value for $\alpha = 0.05$ is $z = \pm 1.96$ (two-tailed), as it was for the one-sample tests discussed earlier.

The preceding formula is not particularly useful except for the purpose of showing the origin of the appropriate t test, because we rarely know the necessary population variances. (Such knowledge is so rare that it is not even worth imagining cases in which we would have it, although a few do exist.) We can circumvent this problem just as we did in the one-sample case, by using the sample variances as estimates of the population variances. This, for the same reasons discussed earlier for the one-sample t, means that the result will be distributed as t rather than z.

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_{\overline{X}_1-\overline{X}_2}}$$

$$= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

[8]Remember that the standard deviation of any sampling distribution is called the standard error of that distribution.

Since the null hypothesis is generally the hypothesis that $\mu_1 - \mu_2 = 0$, we will drop that term from the equation and write

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Pooling Variances

Although the equation for t that we have just developed is appropriate when the sample sizes are equal, it requires some modification when the sample sizes are unequal. This modification is designed to improve the estimate of the population variance. One of the assumptions required in the use of t for two independent samples is that $\sigma_1^2 = \sigma_2^2$ (i.e., the samples come from populations with equal variances, regardless of the truth or falsity of H_0). The assumption is required regardless of whether n_1 and n_2 are equal. Such an assumption is often reasonable. We frequently begin an experiment with two groups of subjects who are equivalent and then do something to one (or both) group(s) that will raise or lower the scores by an amount equal to the effect of the experimental treatment. In such a case, it often makes sense to assume that the variances will remain unaffected. (Recall that adding or subtracting a constant—here, the treatment effect—to or from a set of scores has no effect on its variance.) Since the population variances are assumed to be equal, this common variance can be represented by the symbol σ^2, without a subscript.

In our data we have two estimates of σ^2, namely s_1^2 and s_2^2. It seems appropriate to obtain some sort of an average of s_1^2 and s_2^2 on the grounds that this average should be a better estimate of σ^2 than either of the two separate estimates. We do not want to take the simple arithmetic mean, however, because doing so would give equal weight to the two estimates, even if one were based on considerably more observations. What we want is a **weighted average,** in which the sample variances are weighted by their degrees of freedom $(n_i - 1)$. If we call this new estimate s_p^2 then

weighted average

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

The numerator represents the sum of the variances, each weighted by its degrees of freedom, and the denominator represents the sum of the weights or, equivalently, the degrees of freedom for s_p^2.

The weighted average of the two sample variances is usually referred to as a **pooled variance estimate** (a rather inelegant name, but reasonably descriptive). Having defined the pooled estimate (s_p^2), we can now write

pooled variance estimate

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Notice that both this formula for t and the one we have just been using involve dividing the difference between the sample means by an estimate of the standard error of the difference between means. The only difference concerns the way in which this standard error is estimated. When the sample sizes are equal, it makes absolutely no

Since the null hypothesis is generally the hypothesis that $\mu_1 - \mu_2 = 0$, we will drop that term from the equation and write

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Pooling Variances

Although the equation for t that we have just developed is appropriate when the sample sizes are equal, it requires some modification when the sample sizes are unequal. This modification is designed to improve the estimate of the population variance. One of the assumptions required in the use of t for two independent samples is that $\sigma_1^2 = \sigma_2^2$ (i.e., the samples come from populations with equal variances, regardless of the truth or falsity of H_0). The assumption is required regardless of whether n_1 and n_2 are equal. Such an assumption is often reasonable. We frequently begin an experiment with two groups of subjects who are equivalent and then do something to one (or both) group(s) that will raise or lower the scores by an amount equal to the effect of the experimental treatment. In such a case, it often makes sense to assume that the variances will remain unaffected. (Recall that adding or subtracting a constant—here, the treatment effect—to or from a set of scores has no effect on its variance.) Since the population variances are assumed to be equal, this common variance can be represented by the symbol σ^2, without a subscript.

In our data we have two estimates of σ^2, namely s_1^2 and s_2^2. It seems appropriate to obtain some sort of an average of s_1^2 and s_2^2 on the grounds that this average should be a better estimate of σ^2 than either of the two separate estimates. We do not want to take the simple arithmetic mean, however, because doing so would give equal weight to the two estimates, even if one were based on considerably more observations. What we want is a **weighted average,** in which the sample variances are weighted by their degrees of freedom $(n_i - 1)$. If we call this new estimate s_p^2 then

weighted average

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

The numerator represents the sum of the variances, each weighted by its degrees of freedom, and the denominator represents the sum of the weights or, equivalently, the degrees of freedom for s_p^2.

The weighted average of the two sample variances is usually referred to as a **pooled variance estimate** (a rather inelegant name, but reasonably descriptive). Having defined the pooled estimate (s_p^2), we can now write

pooled variance estimate

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Notice that both this formula for t and the one we have just been using involve dividing the difference between the sample means by an estimate of the standard error of the difference between means. The only difference concerns the way in which this standard error is estimated. When the sample sizes are equal, it makes absolutely no difference whether or not you pool variances; the answer will be the same. When the sample sizes are unequal, however, pooling can make quite a difference.[9]

Degrees of Freedom

Two sample variances (s_1^2 and s_2^2) have gone into calculating t. Each of these variances is based on squared deviations about its corresponding sample mean, and therefore each sample variance has $n_i - 1$ df. Across the two samples, therefore, we will have $(n_1 - 1) + (n_2 - 1) = (n_1 + n_2 - 2)$ df. Thus, the t for two independent samples will be based on $n_1 + n_2 - 2$ degrees of freedom.

Homophobia and Sexual Arousal

Adams, Wright, and Lohr (1996) were interested in some basic psychoanalytic theories that homophobia may be unconsciously related to the anxiety of being or becoming homosexual. They administered the Index of Homophobia to 64 heterosexual males, and classed them as homophobic or nonhomophobic on the basis of their score. They then exposed homophobic and nonhomophobic heterosexual men to videotapes of sexually explicit erotic stimuli portraying heterosexual and homosexual behavior, and recorded their level of sexual arousal. Adams et al. reasoned that if homophobia were unconsciously related to anxiety about one's own sexuality, homophobic individuals would show greater arousal to the homosexual videos than would nonhomophobic individuals.

In this example we will examine only the data from the homosexual video. (There were no group differences for the heterosexual and lesbian videos.) The data in Table 7.5 were created to have the same means and pooled variance as the data that Adams et al. collected, so our conclusions will be the same as theirs.[10] The dependent variable

TABLE 7.5 Data from Adams et al. on level of sexual arousal in homophobic and nonhomophobic heterosexual males

Homophobic						Nonhomophobic					
39.1	38.0	14.9	20.7	19.5	32.2	24.0	17.0	35.8	18.0	–1.7	11.1
11.0	20.7	26.4	35.7	26.4	28.8	10.1	16.1	–0.7	14.1	25.9	23.0
33.4	13.7	46.1	13.7	23.0	20.7	20.0	14.1	–1.7	19.0	20.0	30.9
19.5	11.4	24.1	17.2	38.0	10.3	30.9	22.0	6.2	27.9	14.1	33.8
35.7	41.5	18.4	36.8	54.1	11.4	26.9	5.2	13.1	19.0	–15.5	
8.7	23.0	14.3	5.3	6.3							
Mean		24.00				**Mean**		16.50			
Variance		148.87				**Variance**		139.16			
n		35				***n***		29			

[9]Notice the way in which the denominator is written in the last equation. This same form of expression will occur often in our discussions of the analysis of variance and multiple comparison procedures.

[10]I actually added 12 points to each mean, largely to avoid many negative scores, but it doesn't change the results or the calculations in the slightest.

is degree of arousal at the end of the 4-minute video, with larger values indicating greater arousal.

Before we consider any statistical test, and ideally even before the data are collected, we must specify several features of the test. First we must specify the null and alternative hypotheses:

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

The alternative hypothesis is bidirectional (we will reject H_0 if $\mu_1 < \mu_2$ or if $\mu_1 > \mu_2$), and thus we will use a two-tailed test. For the sake of consistency with other examples in this book, we will use $\alpha = 0.05$. It is important to keep in mind, however, that there is nothing particularly sacred about any of these decisions. Given the null hypothesis as stated, we can now calculate t:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Because we are testing H_0: $\mu_1 - \mu_2 = 0$, the $\mu_1 - \mu_2$ term has been dropped from the equation. We should pool our sample variances because they are so similar that we do not have to worry about homogeneity of variance. Doing so, we obtain

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$
$$= \frac{34(148.87) + 28(139.16)}{35 + 29 - 2} = 144.48$$

Notice that the pooled variance is slightly closer in value to s_1^2 than to s_2^2 because of the greater weight given s_1^2 in the formula. Then

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{24.00 - 16.50}{\sqrt{\frac{144.48}{35} + \frac{144.48}{29}}} = \frac{7.50}{\sqrt{9.11}} = 2.48$$

For this example, we have $n_1 - 1 = 34$ *df* for the homophobic group and $n_2 - 1 =$ 28 *df* for the nonhomophobic group, making a total of $n_1 - 1 + n_2 - 1 = 62$ *df*. From the sampling distribution of t in Appendix t, $t_{0.025}(62) \cong \pm 2.003$ (with linear interpolation). Since the value of t_{obt} far exceeds $t_{\alpha/2}$, we will reject H_0 (at $\alpha = 0.05$) and conclude that there is a difference between the means of the populations from which our observations were drawn. In other words, we will conclude (statistically) that $\mu_1 \neq$ and (practically) that $\mu_1 > \mu_2$. In terms of the experimental variables, homophobic subjects show greater arousal to a homosexual video than do nonhomophobic subjects.

Effect Size

The fact that the groups have been shown to differ from one another with a standard null hypothesis test does not mean that we are finished with this example. Psychologists are beginning to insist on analyses that go beyond a simple test of the n

is degree of arousal at the end of the 4-minute video, with larger values indicating greater arousal.

Before we consider any statistical test, and ideally even before the data are collected, we must specify several features of the test. First we must specify the null and alternative hypotheses:

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

The alternative hypothesis is bidirectional (we will reject H_0 if $\mu_1 < \mu_2$ or if $\mu_1 > \mu_2$), and thus we will use a two-tailed test. For the sake of consistency with other examples in this book, we will use $\alpha = 0.05$. It is important to keep in mind, however, that there is nothing particularly sacred about any of these decisions. Given the null hypothesis as stated, we can now calculate t:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Because we are testing H_0: $\mu_1 - \mu_2 = 0$, the $\mu_1 - \mu_2$ term has been dropped from the equation. We should pool our sample variances because they are so similar that we do not have to worry about homogeneity of variance. Doing so, we obtain

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$
$$= \frac{34(148.87) + 28(139.16)}{35 + 29 - 2} = 144.48$$

Notice that the pooled variance is slightly closer in value to s_1^2 than to s_2^2 because of the greater weight given s_1^2 in the formula. Then

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{24.00 - 16.50}{\sqrt{\frac{144.48}{35} + \frac{144.48}{29}}} = \frac{7.50}{\sqrt{9.11}} = 2.48$$

For this example, we have $n_1 - 1 = 34$ df for the homophobic group and $n_2 - 1 = 28$ df for the nonhomophobic group, making a total of $n_1 - 1 + n_2 - 1 = 62$ df. From the sampling distribution of t in Appendix t, $t_{0.025}(62) \cong \pm 2.003$ (with linear interpolation). Since the value of t_{obt} far exceeds $t_{\alpha/2}$, we will reject H_0 (at $\alpha = 0.05$) and conclude that there is a difference between the means of the populations from which our observations were drawn. In other words, we will conclude (statistically) that $\mu_1 \neq \mu_2$ and (practically) that $\mu_1 > \mu_2$. In terms of the experimental variables, homophobic subjects show greater arousal to a homosexual video than do nonhomophobic subjects.

Effect Size

The fact that the groups have been shown to differ from one another with a standard null hypothesis test does not mean that we are finished with this example. Psychologists are beginning to insist on analyses that go beyond a simple test of the null

effect size

hypothesis. In particular, we have begun to ask about some measure of **effect size,** and some sort of confidence intervals when they are meaningful. We will discuss effect sizes in this section, and confidence intervals in the next.

Cohen's d

One of the people who crusaded early for measures that would indicate something about the magnitude of an experimental result was Jacob Cohen. Within the context of power analyses, to be discussed in the next chapter, he defined a statistic that he called **d**:

$$\mathbf{d} = \frac{\mu_1 - \mu_2}{\sigma}$$

where μ_1 and μ_2 are the means of the two populations in question, and σ is the common population variance. Because **d** as defined here involves only parameters, it is itself a parameter. That works well when talking about power, as Cohen was doing, but it does not work for our purposes. However, we can create a useful statistic that estimates **d** by simply redefining **d** in terms of sample statistics.[11] (We could put a hat (caret) on it if we wanted to emphasize that it is an estimator, but we seldom do.) Thus

$$\mathbf{d} = \frac{\overline{X}_1 - \overline{X}_2}{s_p}$$

where s_p is the square root of the pooled variance. For our example this becomes

$$\mathbf{d} = \frac{\overline{X}_1 - \overline{X}_2}{s_p} = \frac{24.00 - 16.50}{12.02} = \frac{7.5}{12.02} = 0.62$$

From the definition of **d** you can see that we have really expressed the difference between our two means in terms of the size of a standard deviation. In other words, we have standardized the difference. In our example this can be interpreted to mean that the homophobic and nonhomophobic group differed by about six-tenths of a standard deviation, which is actually quite a substantial difference.

One reason for standardizing our measure is to put it in more meaningful terms. To say that the groups differed by 7.5 units in arousal is not particularly informative. Is that a big difference or a little difference? We have no real way to know, because the units (mm of penile circumference) are not something that most of us have an intuitive feel for. But when we standardize the measure it is often more informative, as I think it is here.

The American Psychological Association has considered effect size measures in some detail, and has recommended that some measure be given in reporting experimental results. When units have no intrinsic meaning in themselves, standardized measures such as **d** are appropriate. When the units do have meaning, there is often less need to standardize. For example, a 5-point difference in IQ, or a 1.3-point difference in a four-year college grade point average is meaningful in its own right. A 12-point difference on the Howell Scale of Weird Personalities is not.

[11] Hedges (1982) was the one who first recommended stating this formula in terms of statistics with the pooled estimate of the standard deviation substituted for the population value. It is sometimes referred to as Hedges' **g**.

As you will see in the next chapter, Cohen laid out some very general guidelines for what he considered small, medium, and large effect sizes. He characterized **d** = 0.20 as an effect that is small, but probably meaningful, an effect size of **d** = 0.50 as a medium effect that most people would be able to notice (such as a half of a standard deviation difference in IQ), and an effect size of **d** = 0.80 as large. We should not make too much of Cohen's levels, but they are helpful as a rough guide.

The Squared Point-Biserial Correlation (r_{pb}^2)

In Chapter 10 we will discuss the point-biserial correlation and how it can be used as a measure of the magnitude of an effect. I cannot develop that statistic here, but I need to at least mention it because it is a common measure that can be used to elaborate on the magnitude of an effect beyond merely stating that we have rejected the null hypothesis.

In Chapter 10 we will define

$$r_{pb}^2 = \frac{t^2}{t^2 + df}$$

If we apply this equation to our results, we will find that

$$r_{pb}^2 = \frac{t^2}{t^2 + df} = \frac{2.48^2}{2.48^2 + 62} = \frac{6.15}{68.15} = 0.09$$

As you will see later, this result can be taken to mean that 9 percent of the variability in arousal scores is associated with differences between homophobia and nonhomophobia. We will attach a more meaningful interpretation when you encounter this later.

7.6 Confidence Intervals

Confidence intervals are another way to convey the meaning of an experimental result that goes beyond the simple hypothesis test. In the case of homophobia we would be particularly interested in confidence intervals on the *difference* between the two means, but life will be much simpler if we start with an example of a confidence interval around a single mean.

The earlier data on the moon illusion offer an excellent example of a case in which we are particularly interested in estimating the true value of μ—in this case the true ratio of the perceived size of the horizon moon to the perceived size of the zenith moon. The sample mean ($\overline{X}$), as you already know, is an unbiased estimate of μ. When we have a specific estimate of a parameter, we call this a **point estimate.** There are also **interval estimates,** which are attempts to set limits that have a high probability of encompassing the true (population) value of the mean [the mean (μ) of a whole population of observations]. What we want, here, are **confidence limits** on μ. These limits enclose what is called a **confidence interval.**[12] In Chapter 3, we

point estimate
interval estimates
confidence limits
confidence interval

[12] We often speak of "confidence limits" and "confidence interval" as if they were synonymous. They are essentially the same, except that the limits are the endpoints of the interval. Don't be confused when you see them used interchangeably.

As you will see in the next chapter, Cohen laid out some very general guidelines for what he considered small, medium, and large effect sizes. He characterized $\mathbf{d} = 0.20$ as an effect that is small, but probably meaningful, an effect size of $\mathbf{d} = 0.50$ as a medium effect that most people would be able to notice (such as a half of a standard deviation difference in IQ), and an effect size of $\mathbf{d} = 0.80$ as large. We should not make too much of Cohen's levels, but they are helpful as a rough guide.

The Squared Point-Biserial Correlation (r_{pb}^2)

In Chapter 10 we will discuss the point-biserial correlation and how it can be used as a measure of the magnitude of an effect. I cannot develop that statistic here, but I need to at least mention it because it is a common measure that can be used to elaborate on the magnitude of an effect beyond merely stating that we have rejected the null hypothesis.

In Chapter 10 we will define

$$r_{pb}^2 = \frac{t^2}{t^2 + df}$$

If we apply this equation to our results, we will find that

$$r_{pb}^2 = \frac{t^2}{t^2 + df} = \frac{2.48^2}{2.48^2 + 62} = \frac{6.15}{68.15} = 0.09$$

As you will see later, this result can be taken to mean that 9 percent of the variability in arousal scores is associated with differences between homophobia and nonhomophobia. We will attach a more meaningful interpretation when you encounter this later.

7.6 Confidence Intervals

Confidence intervals are another way to convey the meaning of an experimental result that goes beyond the simple hypothesis test. In the case of homophobia we would be particularly interested in confidence intervals on the *difference* between the two means, but life will be much simpler if we start with an example of a confidence interval around a single mean.

The earlier data on the moon illusion offer an excellent example of a case in which we are particularly interested in estimating the true value of μ—in this case, the true ratio of the perceived size of the horizon moon to the perceived size of the zenith moon. The sample mean ($\overline{X}$), as you already know, is an unbiased estimate of μ. When we have a specific estimate of a parameter, we call this a **point estimate.** There are also **interval estimates,** which are attempts to set limits that have a high probability of encompassing the true (population) value of the mean [the mean (μ) of a whole population of observations]. What we want, here, are **confidence limits** on μ. These limits enclose what is called a **confidence interval.**[12] In Chapter 3, we saw

point estimate
interval estimates

confidence limits
confidence interval

[12]We often speak of "confidence limits" and "confidence interval" as if they were synonymous. They are essentially the same, except that the limits are the endpoints of the interval. Don't be confused when you see them used interchangeably.

how to set "probable limits" on an observation. A similar line of reasoning will apply here, where we attempt to set confidence limits on a parameter.

If we want to set limits that are likely to include μ given the data at hand, what we really want to ask is how large, or small, the true value of μ could be without causing us to reject H_0 if we ran a t test on the obtained sample mean. In other words, if μ were quite small (or quite large), we would have been unlikely to obtain the sample data. For a whole range of values for μ, however, we would expect data like those we obtained. We want to calculate what those values of μ are.

An easy way to see what we are doing is to start with the formula for t for the one-sample case:

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

From the moon illusion data (see pp. 189–190), we know $\overline{X} = 1.463$, $s = 0.341$, $n = 10$. We also know that the critical two-tailed value for t at $\alpha = 0.05$ is $t_{0.025}(9) = \pm 2.262$. We will substitute these values in the formula for t, but this time we will solve for the μ associated with this value of t:

$$t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \qquad \pm 2.262 = \frac{1.463 - \mu}{\frac{0.341}{\sqrt{10}}} = \frac{1.463 - \mu}{0.108}$$

Rearranging to solve for μ, we have

$$\mu = \pm 2.262(0.108) + 1.463 = \pm 0.244 + 1.463$$

Using the $+0.244$ and -0.244 separately to obtain the upper and lower limits for μ, we have

$$\mu_{\text{upper}} = +0.244 + 1.463 = 1.707$$

$$\mu_{\text{lower}} = -0.244 + 1.463 = 1.219$$

and thus we can write the 95% confidence limits as 1.219 and 1.707 and the confidence interval as

$$\text{CI}_{0.95} = 1.219 \leq \mu \leq 1.707$$

Testing a null hypothesis about any value of μ outside these limits would lead to rejection of H_0, whereas testing a null hypothesis about any value of μ inside those limits would not lead to rejection. The general expression is

$$\text{CI}_{1-\alpha} = \overline{X} \pm t_{\alpha/2}(s_{\overline{X}}) = \overline{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

We have a 95% confidence interval because we used the two-tailed critical value of t at $\alpha = 0.05$. For the 99% limits we would take $t_{0.01/2} = t_{0.005} = \pm 3.250$. Then the 99% confidence interval is

$$\text{CI}_{0.99} = \overline{X} \pm t_{0.005}(s_{\overline{X}}) = 1.463 \pm 3.250(0.108) = 1.112 \leq \mu \leq 1.814$$

We can now say that the probability is 0.95 that intervals calculated as we have calculated the 95% interval above include the true mean ratio for the moon illusion. It is very tempting to say that the probability is 0.95 that the interval 1.219 to 1.707 includes the true mean ratio for the moon illusion, and the probability is 0.99 that the interval 1.112 to 1.814 includes μ. However, most statisticians would object to the statement of a confidence limit expressed in this way. They would argue that *before the experiment is run* and the calculations are made, an interval *of the form*

$$\overline{X} \pm t_{0.025}(s_{\overline{X}})$$

has a probability of 0.95 of encompassing μ. However, μ is a fixed (though unknown) quantity, and once the data are in, the specific interval 1.219 to 1.707 either includes the value of μ ($p = 1.00$) or it does not ($p = 0$). Put in slightly different form,

$$\overline{X} \pm t_{0.025}(s_{\overline{X}})$$

is a random variable (it will vary from one experiment to the next), but the specific interval 1.219 to 1.707 is not a random variable and therefore does not have a probability associated with it. Good (1999) has made the point that we place our confidence in the *method,* not in the *interval.* Many would maintain that it is perfectly reasonable to say that my confidence is 0.95 that if you were to tell me the true value of μ, it would be found to lie between 1.219 and 1.707. But there are many people just lying in wait for you to say that the *probability* is 0.95 that μ lies between 1.219 and 1.707. When you do, they will pounce!

Note that neither the 95% nor the 99% confidence interval that I computed includes the value of 1.00, which represents no illusion. We already knew this for the 95% confidence interval because we had rejected that null hypothesis when we ran our t test at that significance level.

I should add another way of looking at the interpretation of confidence limits. Statements of the form $p(1.219 < \mu < 1.707) = 0.95$ are not interpreted in the usual way. (In fact, I probably shouldn't use p in that equation.) The parameter μ is not a variable—it does not jump around from experiment to experiment. Rather, μ is a constant, and the interval is what varies from experiment to experiment. Thus, we can think of the parameter as a stake and the experimenter, in computing confidence limits, as tossing rings at it. Ninety-five percent of the time, a ring of specified width will encircle the parameter; 5% of the time, it will miss. A confidence statement is a statement of the probability that the ring has been on target; it is not a statement of the probability that the target (parameter) landed in the ring.

A graphic demonstration of confidence limits is shown in Figure 7.8. To generate this figure, I drew 25 samples of $n = 4$ from a population with a mean (μ) of 5. For every sample, a 95% confidence limit on μ was calculated and plotted. For example, the limits produced from the first sample (the top horizontal line) were approximately 4.46 and 5.72, whereas those for the second sample were 4.83 and 5.80. Since in this case we know that the value of μ equals 5, I have drawn a vertical line at that point. Notice that the limits for samples 12 and 14 do not include $\mu = 5$. We would expect that 95% confidence limits would encompass μ 95 times out of 100. Therefore, two misses out of 25 seems reasonable. Notice also that the confidence intervals vary in width. This variability is due to the fact that the width of an interval is

We can now say that the probability is 0.95 that intervals calculated as we have calculated the 95% interval above include the true mean ratio for the moon illusion. It is very tempting to say that the probability is 0.95 that the interval 1.219 to 1.707 includes the true mean ratio for the moon illusion, and the probability is 0.99 that the interval 1.112 to 1.814 includes μ. However, most statisticians would object to the statement of a confidence limit expressed in this way. They would argue that *before the experiment is run* and the calculations are made, an interval *of the form*

$$\overline{X} \pm t_{0.025}(s_{\overline{X}})$$

has a probability of 0.95 of encompassing μ. However, μ is a fixed (though unknown) quantity, and once the data are in, the specific interval 1.219 to 1.707 either includes the value of μ ($p = 1.00$) or it does not ($p = 0$). Put in slightly different form,

$$\overline{X} \pm t_{0.025}(s_{\overline{X}})$$

is a random variable (it will vary from one experiment to the next), but the specific interval 1.219 to 1.707 is not a random variable and therefore does not have a probability associated with it. Good (1999) has made the point that we place our confidence in the *method*, not in the *interval*. Many would maintain that it is perfectly reasonable to say that my confidence is 0.95 that if you were to tell me the true value of μ, it would be found to lie between 1.219 and 1.707. But there are many people just lying in wait for you to say that the *probability* is 0.95 that μ lies between 1.219 and 1.707. When you do, they will pounce!

Note that neither the 95% nor the 99% confidence interval that I computed includes the value of 1.00, which represents no illusion. We already knew this for the 95% confidence interval because we had rejected that null hypothesis when we ran our t test at that significance level.

I should add another way of looking at the interpretation of confidence limits. Statements of the form $p(1.219 < \mu < 1.707) = 0.95$ are not interpreted in the usual way. (In fact, I probably shouldn't use p in that equation.) The parameter μ is not a variable—it does not jump around from experiment to experiment. Rather, μ is a constant, and the interval is what varies from experiment to experiment. Thus, we can think of the parameter as a stake and the experimenter, in computing confidence limits, as tossing rings at it. Ninety-five percent of the time, a ring of specified width will encircle the parameter; 5% of the time, it will miss. A confidence statement is a statement of the probability that the ring has been on target; it is not a statement of the probability that the target (parameter) landed in the ring.

A graphic demonstration of confidence limits is shown in Figure 7.8. To generate this figure, I drew 25 samples of $n = 4$ from a population with a mean (μ) of 5. For every sample, a 95% confidence limit on μ was calculated and plotted. For example, the limits produced from the first sample (the top horizontal line) were approximately 4.46 and 5.72, whereas those for the second sample were 4.83 and 5.80. Since in this case we know that the value of μ equals 5, I have drawn a vertical line at that point. Notice that the limits for samples 12 and 14 do not include $\mu = 5$. We would expect that 95% confidence limits would encompass μ 95 times out of 100. Therefore, two misses out of 25 seems reasonable. Notice also that the confidence intervals vary in width. This variability is due to the fact that the width of an interval is

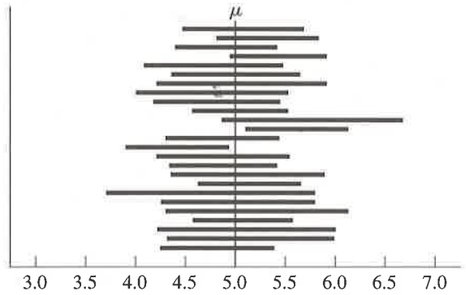


FIGURE 7.8 *Confidence intervals computed on 25 samples from a population with $\mu = 5$*

a function of the standard deviation of the sample, and some samples have larger standard deviations than others.

Confidence Limits on $\mu_1 - \mu_2$

Now we can get down to what we really care about here, which are the 95% confidence limits on the difference in arousal between homophobic and nonhomophobic males. The logic for setting these confidence limits is the same as for doing so in the one-sample case. The calculations are the same except that we use the difference between the means and the standard error of differences between means in place of the mean and the standard error of the mean. Thus, for the 95% confidence limits on $\mu_1 - \mu_2$, we have

$$\text{CI}_{0.95} = (\overline{X}_1 - \overline{X}_2) \pm t_{0.025} s_{\overline{X}_1 - \overline{X}_2}$$

For the comparison of the homophobic and nonhomophobic males (see Table 7.5), we have

$$\begin{aligned}\text{CI}_{0.95} &= (24.00 - 16.5) \pm 2.00\sqrt{\frac{144.48}{35} + \frac{144.48}{29}} \\ &= 7.5 \pm 2.00(3.018) \\ &= 7.5 \pm 6.04 \\ &= 1.46 \leq (\mu_1 - \mu_2) \leq 13.54\end{aligned}$$

The probability is 0.95 that an interval such as 1.46 to 13.54 encloses the true difference between the mean arousal scores for theoretical populations of homophobic and nonhomophobic males. Notice that our calculations and interpretation are the same as they were in the previous example. The only difference is that we have replaced $\overline{X}$ with $(\overline{X}_1 - \overline{X}_2)$ and $s_{\overline{X}}$ with $s_{\overline{X}_1 - \overline{X}_2}$.

Different measures tell us different things, and some are more informative than others in a particular situation. Earlier we saw that the standardized effect size measure (**d**) showed that the two sample means differed by 0.62 standard deviation, which

is a sizeable difference. I commented that the unstandardized difference (7.5 units) was not particularly informative because the dependent variable is not measured in units that have any intuitive meaning to us. Similarly, because the units do not have intuitive meaning, the confidence interval does not strike me as being particularly helpful in this situation. Most of us probably don't care what the true difference is, as long as it is greater than zero. In this particular case the fact that homophobic participants were more aroused than nonhomophobic participants is probably all that we need to know. Expressing the difference in terms of standard deviations is somewhat helpful because it tells us that we are talking about something substantial (0.62 standard deviation strikes me as being a very large difference). Precision in raw score units (or an interval of raw score units) doesn't give us anything we are likely to care about.

Confidence limits do have an important role in many situations, however. In the exercises at the end of this chapter you are asked to calculate a confidence interval on the difference is mean weight gain between two different kinds of therapy for anorexia. In that case the difference is measured in pounds, and you will recognize that we are talking about a substantial difference if the interval is 12–15 pounds, and that we are talking about something pretty trivial if the interval is 0.8 to 2.1 pounds. In each case the difference is significant because the interval does not include zero. (My apologies to non-U.S. users of this book, though I imagine that you will be less confused by English units than most Americans claim to be by metric units.)

The moon illusion example is another case where the confidence interval is meaningful. We know that a ratio of 1.0 would mean that the two moons appear equally large. A ratio of 1.25 means that the horizon moon is 25% larger than the zenith moon. Our interval was $1.219 \leq \mu \leq 1.707$. Thus we know that the horizon moon appears, very roughly, 25% to 75% larger. That is a meaningful statistic.

We generally calculate confidence intervals on means, and I have rarely seen intervals calculated on other statistics, such as the variance. It can be done, and if you really need an interval on the variance, see a previous edition of this book, but since no one seems interested in doing it, I have left other forms of confidence intervals out of this discussion.

SPSS Analysis

The SPSS analysis of the Adams et al. (1996) data is given in Table 7.6. Notice that SPSS first provides what it calls Levene's test for equality of variances. We will discuss this test shortly, but it is simply a test on our assumption of homogeneity of variance. We do not come close to rejecting the null hypothesis that the variances are homogeneous ($p = 0.534$), so we don't have to worry about that here. From now on we will assume equal variances, and will focus on the next-to-bottom row of the table.

Next note that the t supplied by SPSS is the same as we calculated, and that the probability associated with this value of t (0.016) is less than $\alpha = 0.05$, leading to rejection of the null hypothesis. Note also that SPSS prints the difference between the means and the standard error of that difference, both of which we have seen in our own calculations. Finally, SPSS prints the 95% confidence interval on the difference between means, and it agrees with ours.