CHAPTER 7

Hypothesis Tests Applied to Means

Objectives

To introduce the *t* test as a procedure for testing hypotheses with measurement data, and to show how it can be used with several different designs. To describe ways of estimating the magnitude of any differences that do appear.

Contents

7.1 Sampling Distribution of the Mean
7.2 Testing Hypotheses About Means—σ Known
7.3 Testing a Sample Mean When σ Is Unknown—The One-Sample *t* Test
7.4 Hypothesis Tests Applied to Means—Two Matched Samples
7.5 Hypothesis Tests Applied to Means—Two Independent Samples
7.6 A Final Worked Example
7.7 Heterogeneity of Variance: The Behrens–Fisher Problem
7.8 Hypothesis Testing Revisited

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, 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 we found 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 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. Beyond telling us what the mean and variance of the sampling distribution of the mean must be for any given sample size, the theorem states that as n increases, the shape of this sampling distribution approaches normal, *whatever* the shape of the parent population. 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.

[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.

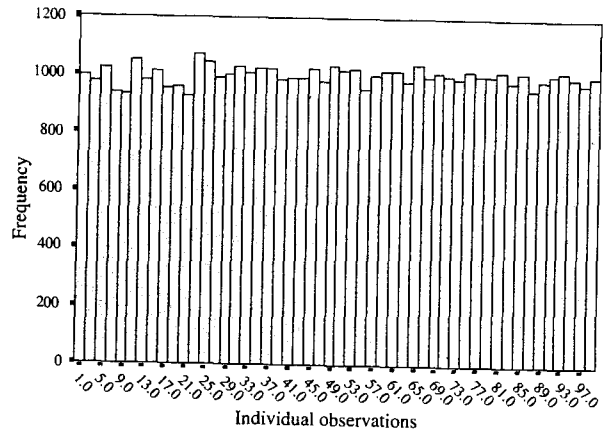


Figure 7.1 50,000 observations from a uniform distribution

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 the square root of 12), and the variance (σ^2) is thus 833.33.

Now suppose we drew 5,000 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. (Actually, the superimposed normal distribution fits the data quite 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, only this time drawing 5,000 samples of 30 observations each. The results for these samples are plotted in Figure 7.2b. Here you see that just as the central limit theorem predicted, the distribution is approximately normal, the mean is again at $\mu = 50$, and the standard deviation has been reduced to approximately $28.87/\sqrt{30} = 5.27$.

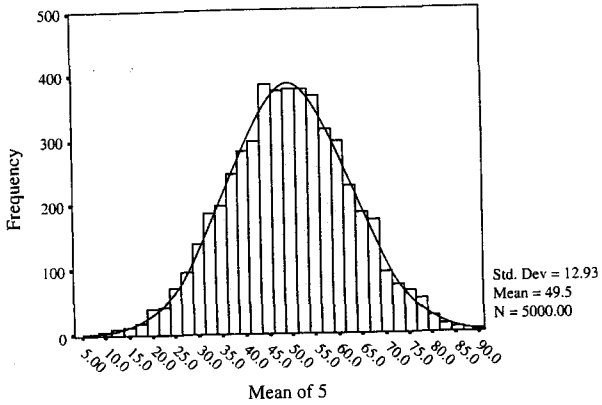


Figure 7.2a Sampling distribution of the mean when $n = 5$

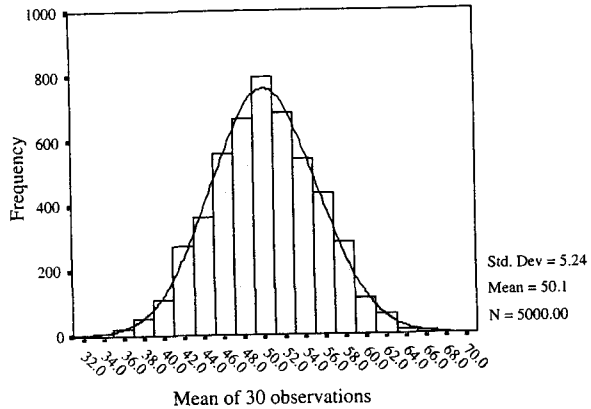


Figure 7.2b Sampling distribution of the mean when $n = 30$

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 it might be 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

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

and then using Appendix z to find the area below $z = -1.5$.[2] 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. This probability was not less than our preselected significance level of $\alpha = .05$, so 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 with a distribution of observations, we will compare a mean with a distribution of means (the sampling distribution 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 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**)[3] 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

[2] 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 $z = +1.5$.

[3] 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 $\sigma_{\bar{X}}$), 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 $\sigma_{\bar{X}_1 - \bar{X}_2}$. 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.

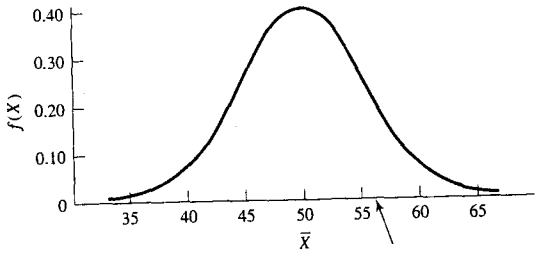


Figure 7.3 Sampling distribution of the mean for $n = 5$ drawn from a population with $\mu = 50$ and $\sigma = 10$

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 greater than two standard deviations above the mean.

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

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

For our data, this becomes

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

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 re-expressed 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 .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 $2(.0901) = .1802$. 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

.05 level of significance, we would not reject H_0 because the obtained probability is greater than .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. 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. 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

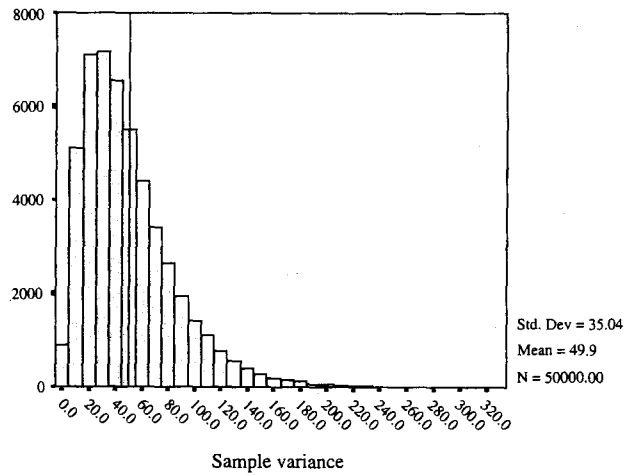


Figure 7.4 Sampling distribution of the sample variance

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.

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

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

We know that for any particular sample, s^2 is more likely than not to be smaller than the appropriate value of σ^2, so 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 .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.

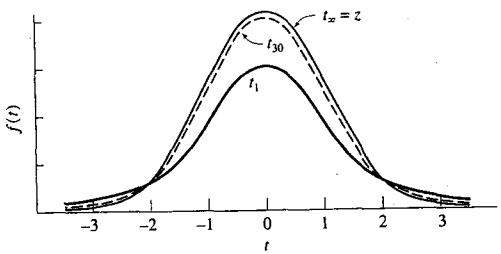


Figure 7.5 t distribution for 1, 30, and ∞ degrees of freedom

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. As $n \Rightarrow \infty$, $p(s^2 < \sigma^2) \Rightarrow p(s^2 > \sigma^2)$. (The symbol $\Rightarrow$ is read "approaches.") The skewness of the sampling distribution of s^2 disappears as the number of degrees of freedom increases, so 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 Gosset derived the t distribution 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 with 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 $\left[\sum (X - \overline{X})\right]$ 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 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

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

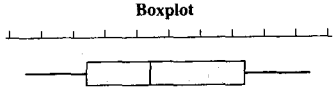
Raw Data

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

Stem-and-Leaf Display

Stem	Leaf
8*	3
8.	66999999
9*	222222
9.	5666688
10*	0000222444
10.	88888
11*	222
11.	667
12*	000000444
12.	567

Boxplot



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

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 LBW infants in general are 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. Included in this figure are 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 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 how raw scores are converted to scale scores, but it is worth checking. (Actually, 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 104.125. 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 $\mu = 100$.

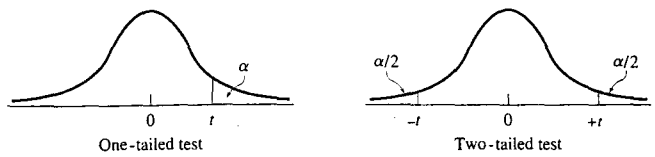
A *t* value of 2.45 in and of itself is not particularly meaningful unless we can evaluate it against the sampling distribution of *t*. For this purpose, the critical values of *t* are presented in Appendix *t*. This table differs in form from the table of the normal distribution (*z*) because instead 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 .05 and .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. We want to work at the two-tailed .05 level, so 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, $t_{.025}$. 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 $t_{.025}$ (rounding to 50 *df* for purposes of the table) = 2.009. (This is sometimes written as $t_{.025}(50) = 2.009$ to indicate the degrees of freedom.) Because the obtained value of *t*, written t_{obt}, is greater than $t_{.025}$, we will reject H_0 at $\alpha = .05$, 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 statistically significantly *above* the normative population mean. This points out the advantage of using two-tailed tests because 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. Actually, however, a number of other studies using the PDI have reported similarly high means.)

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.[4] The moon illusion has fascinated psychologists for years and refers to the fact that when we see the moon near the horizon, it appears to be considerably larger than when we see it high in the sky. Kaufman and Rock concluded that this illusion could be explained on the basis of the greater *apparent*

[4] A more 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.*

Table 7.2 Percentage points of the t distribution



	Level of Significance for One-Tailed Test								
	.25	.20	.15	.10	.05	.025	.01	.005	.0005
	Level of Significance for Two-Tailed Test								
df	.50	.40	.30	.20	.10	.05	.02	.01	.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.

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 to match the size of a standard "moon" that appeared at its zenith, or vice versa. (In these measurements, they used an artifical moon 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_0: $\mu \neq 1.00$.

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 $\alpha = .05$, the critical value of $t_{.025}(9) = \pm 2.262$. The obtained value of t was 4.29. Because $4.29 > 2.262$, we can reject H_0 at $\alpha = .05$ 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.)

Confidence Interval on μ

point estimate

confidence limits

confidence interval

Confidence intervals are a useful way to convey the meaning of an experimental result that goes beyond the simple hypothesis test. The 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 one 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.**[5] In Chapter 3, we saw 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 is to ask 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. But for a whole range of values for μ, 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 we know $\overline{X} = 1.463$, $s = 0.341$, $n = 10$. We also know that the critical two-tailed value for t at $\alpha = .05$ is $t_{.025}(9) = \pm 2.262$. We will substitute

[5] We often speak of "confidence limits" and "confidence interval" as if they were synonymous. The pretty much are, except that the limits are the end points of the interval. Don't be confused when you see them used interchangeably.

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

$$\mu_{upper} = +0.244 + 1.463 = 1.707$$
$$\mu_{lower} = -0.244 + 1.463 = 1.219$$

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

$$CI_{.95} = 1.219 \leq \mu \leq 1.707$$

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

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

We have a 95% confidence interval because we used the two-tailed critical value of t at $\alpha = .05$. For the 99% limits we would take $t_{.01/2} = t_{.005} = \pm 3.250$. Then the 99% confidence interval is

$$CI_{.99} = \overline{X} \pm t_{.01/2}(s_{\overline{X}}) = 1.463 \pm 3.250(0.108) = 1.12 \leq \mu \leq 1.814$$

We can now say that the probability is .95 that intervals calculated as we have calculated the 95% interval earlier include the true mean ratio for the moon illusion. It is very tempting to say that the probability is .95 that the interval 1.219 to 1.707 includes the true mean ratio for the moon illusion, and the probability is .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,*

$$\overline{X} \pm t_{.025}(s_{\overline{X}})$$

has a probability of .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 ($p = .00$). Put in slightly different form,

$$\overline{X} \pm t_{.025}(s_{\overline{X}})$$

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,* rather than in the *interval.* Many would maintain that it is perfectly reasonable to say that my confidence is .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 .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 intervals that I computed includes the value of 1.00, which represents no illusion. We already knew this for the 95% confidence

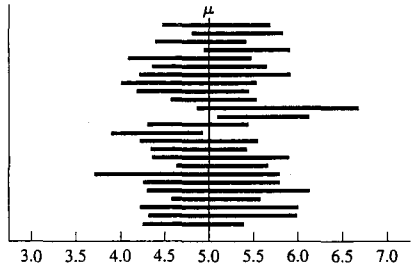


Figure 7.6 Confidence intervals computed on 25 samples from a population with $\mu = 5$

interval because we had rejected that null hypothesis when we ran our t test at that significance level.

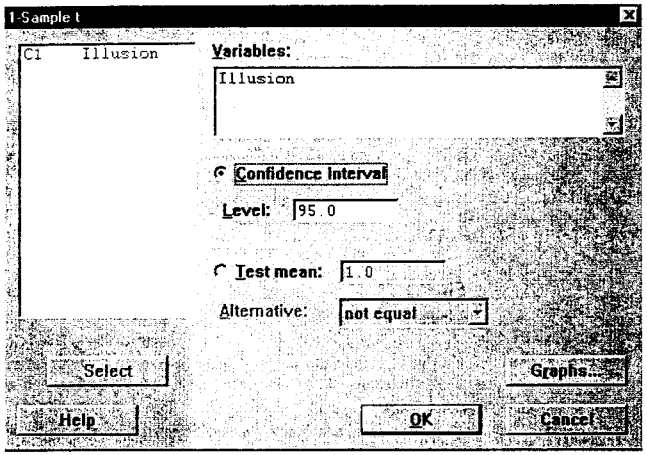
I should add another way of looking at the interpretation of confidence limits. Statements of the form $p(1.219 < \mu < 1.707) = .95$ are not interpreted in the usual way. (Actually, I probably shouldn't use p in that equation.) The parameter μ is not a variable—it does not jump around from experiment 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 7.6. 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. In this case, we know that the value of μ equals 5, so 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, 2 misses out of 25 seems reasonable. Notice also that the confidence intervals vary in width. This variability is because the width of an interval is a function of the standard deviation of the sample, and some samples have larger standard deviations than others.

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. 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 Minitab computes the exact probability of a Type I error (the ***p* level**), rather than comparing t with a tabled value. Thus, whereas we concluded that the probability of a Type I error was *less than* .05, Minitab reveals that the actual probability is .0020. Most computer programs operate in this way.

***p* level**



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

7.4 Hypothesis Tests Applied to Means—Two Matched Samples

matched samples

repeated measures

related samples

matched-sample t test

In Section 7.3, we considered the situation in which we had one sample mean ($\overline{X}$) and wanted 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) 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 want 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.**

Table 7.3 Data from Everitt on weight gain

ID	1	2	3	4	5	6	7	8	9	10
Before	83.8	83.3	86.0	82.5	86.7	79.6	76.9	94.2	73.4	80.5
After	95.2	94.3	91.5	91.9	100.3	76.7	76.8	101.6	94.9	75.2
Diff	11.4	11.0	5.5	9.4	13.6	−2.9	−0.1	7.4	21.5	−5.3

ID	11	12	13	14	15	16	17	Mean	St. Dev
Before	81.6	82.1	77.6	83.5	89.9	86.0	87.3	83.23	5.02
After	77.8	95.5	90.7	92.5	93.8	91.7	98.0	90.49	8.48
Diff	−3.8	13.4	13.1	9.0	3.9	5.7	10.7	7.26	7.16

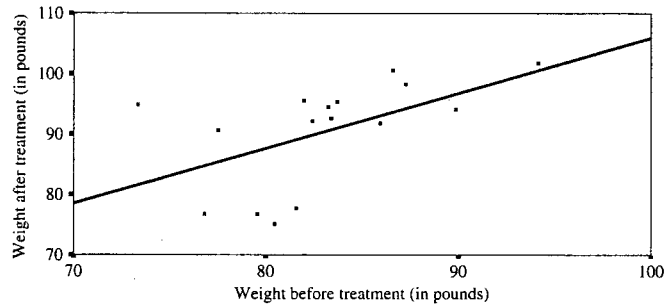


Figure 7.7 Relationship of weight before and after family therapy, for a group of 17 anorexic girls

Treatment of Anorexia

Everitt, in Hand, Daly, Lunn, McConway, and Ostrowski (1994), reported on family therapy as a treatment for anorexia. There were 17 girls in this experiment, and they were weighed before and after treatment. The weights of the girls, in pounds,[6] are given in Table 7.3. The row of difference scores was obtained by subtracting the Before score from the After score, so that a negative difference represents weight *loss,* and a positive difference represents a *gain.*

One of the first things we should probably do, although it takes us away from *t* tests for a moment, is to plot the relationship between Before Treatment and After Treatment weights, looking to see if there is, in fact, a relationship, and how linear that relationship is. Such a plot is given in Figure 7.7. Notice that the relationship is basically linear, with a slope quite near 1.0. Such a slope suggests that how much the girl weighed at the beginning of

[6] Everitt reported that these weights were in kilograms, but if so he has a collection of anorexic young girls whose mean weight is about 185 pounds, and that just doesn't sound reasonable. The example is completely unaffected by the units in which we record weight.

therapy did not seriously influence how much weight she gained or lost by the end of therapy. (We will discuss regression lines and slopes further in Chapter 9.)

The primary question we want is ask is whether subjects gained weight as a function of the therapy sessions. We have an experimental problem here because it is possible that weight gain resulted merely from the passage of time, and that therapy had nothing to do with it. However, I know from other data in that experiment that a group that did not receive therapy did not gain weight over the same period, which strongly suggests that the simple passage of time was not an important variable. If you were to calculate the weight of these girls before and after therapy, the means would be 83.23 and 90.49 lbs, respectively, which translates to a gain of a little over 7 pounds. However, we still need to test to see whether this difference is likely to represent a true difference in population means, or a chance difference. By this, I mean that we need to test the null hypothesis that the mean *in the population* of Before scores is equal to the mean *in the population* of After scores. In other words, we are testing H_0: $\mu_A = \mu_B$.

Difference Scores

difference scores

gain scores

Although it would seem obvious to view the data as representing two samples of scores, one set obtained before the therapy program and one after, 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,** or **gain scores,** and are shown in the row labeled "Diff" in Table 7.3. They represent the degree of weight gain between one measurement session and the next—presumably as a result of our intervention. If the therapy program had *no* effect (i.e., if H_0 is true), the average weight would not change from session to session. By chance, some participants would happen to have a higher weight on X_2 than on X_1, and some would have a lower weight, but *on the average* there would be no difference.

If we now think of our data as being the set of difference scores, the null hypothesis becomes the hypothesis that the mean of a population of difference scores (denoted μ_D) equals 0. Because 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.

The *t* Statistic

We are now at precisely the same place we were in the previous section 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 the 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 and $\overline{D}$ and s_D are the mean and the standard deviation of the difference scores and N is the number of difference scores (i.e., the number of *pairs,* not the number of raw scores). From Table 7.3, we see that the mean difference score was 7.26, and the standard deviation of the differences was 7.16. For our data

$$t = \frac{\overline{D} - 0}{s_{\overline{D}}} = \frac{\overline{D} - 0}{\frac{s_D}{\sqrt{N}}} = \frac{7.26 - 0}{\frac{7.16}{\sqrt{17}}} = \frac{7.26}{1.74} = 4.18$$

Degrees of Freedom

The degrees of freedom for the matched-sample case are exactly the same as they were for the one-sample case. Because 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). Because 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 because this sample variance is obtained using the sample mean ($\overline{D}$), we will lose one *df* to the mean and have $N - 1$ *df*. In other words, *df* = number of *pairs* minus 1.

We have 17 difference scores in this example, so we will have 16 degrees of freedom. From Appendix *t*, we find that for a two-tailed test at the .05 level of significance, $t_{.05}(16) = \pm 2.12$. Our obtained value of $t(4.18)$ exceeds 2.12, so we will reject H_0 and conclude that the difference scores were not sampled from a population of difference scores where $\mu_D = 0$. In practical terms, this means that the subjects weighed significantly more after the intervention program than before it. Although we would like to think that this means that the program was successful, keep in mind the possibility that this could just be normal growth. The fact remains, however, that for whatever reason, the weights were sufficiently higher on the second occasion to allow us to reject H_0: $\mu_D = \mu_1 - \mu_2 = 0$.

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 because the observer looked straight at the moon with eyes level when it was on the horizon, whereas when the moon 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.

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 $\alpha = .05$, the alternative hypothesis is H_1: $\mu_D \neq 0$.

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

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

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

$$= 0.44$$

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 Appendix t, we find that $t_{.025}(9) = \pm 2.262$. Because $t_{obt} = 0.44$ 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.[7] (These data also include a second test of Holway and Boring's hypothesis because 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 Section 7.3, and the illusion is in fact statistically significant.)

Confidence Limits on Matched Samples

We can calculate confidence limits on matched samples in the same way we did for the one-sample case because in matched samples the data come down to a single column of difference scores. Returning to Everitt's data on anorexia, we have

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

and thus

$$CI_{.95} = \overline{D} \pm t_{.05/2}(s_{\overline{D}}) = \overline{D} \pm t_{.025}\frac{s_D}{\sqrt{n}}$$

$$CI_{.95} = 7.26 \pm 2.12(1.74)$$

$$CI_{.95} = 7.26 \pm 3.69$$

$$= 3.57 \leq \mu \leq 10.95$$

[7] A glance at Appendix t will reveal that a t less than 1.96 (the critical value for z) will never be significant at $\alpha = .05$, 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.

Notice that this confidence interval does not include $\mu_D = 0.0$, which is consistent with the fact that we rejected the null hypothesis.

Effect Size

In Chapter 6, we looked at effect size measures as a way of understanding the magnitude of the effect that we see in an experiment—rather than simply the statistical significance. When we are looking at the difference between two related measures we can, and should, also compute effect sizes. In this case, there is a slight complication as we will see shortly.

d-Family of Measures

A number of different effect sizes measures are often recommended, and for a complete coverage of this topic I suggest the reference by Kline (2004). As I did in Chapter 6, I am going to distinguish between measures based on differences between groups (the *d*-family) and measures based on correlations between variables (the *r*-family). However, in this chapter I am not going to discuss the *r*-family measures, partly because I find them less informative and partly because they are more easily and logically discussed in Chapter 11 when we come to the analysis of variance.

Cohen's *d*

There is considerable confusion in the naming of measures, and for clarification on that score I refer the reader to Kline (2004). Here I will use the more common approach, which Kline points out is not quite technically correct, and refer to my measure as **Cohen's *d*.** Measures proposed by Hedges and by Glass are very similar and are often named almost interchangeably.

The data on treatment of anorexia offer a good example of a situation in which it is relatively easy to report on the difference in ways that people will understand. All of us step onto a scale occasionally, and we have some general idea what it means to gain or lose 5 or 10 pounds. So for Everitt's data, we could simply report that the difference was significant ($t = 4.18$, $p < .05$) and that girls gained an average of 7.26 pounds. For girls who started out weighing, on average, 83 pounds, that is a substantial gain. In fact, it might make sense to convert pounds gained to a percentage and say that the girls increased their weight by $7.26/83.23 = 9\%$.

An alternative measure would be to report the gain in standard deviation units. This idea goes back to Cohen, who originally formulated the problem in terms of a statistic (d), where

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

In this equation, the numerator is the difference between two population means, and the denominator is the standard deviation of either population. In our case, we can modify that slightly to let the numerator be the mean gain ($\mu_{After} - \mu_{Before}$), and let the denominator be the population standard deviation *of the pretreatment weights*. To put this in terms of statistics, rather than parameters, we can substitute sample means and standard deviations instead of population values. This leaves us with

$$\hat{d} = \frac{\overline{X}_1 - \overline{X}_2}{s_{X_1}} = \frac{90.49 - 83.23}{5.02} = \frac{7.26}{5.02} = 1.45$$

I have put a "hat" over the d to indicate that we are calculating an estimate of d, and I have put the standard deviation of the pretreatment scores in the denominator. Our estimate tells us that, on average, the girls involved in family therapy gained nearly one and a half standard deviations of pretreatment weights over the course of therapy.

In this particular example, I find it easier to deal with the mean weight gain, rather than d, simply because I know something meaningful about weight. However, if this experiment

had measured the girls' self-esteem, rather than weight, I would not know what to think if you said that they gained 7.26 self-esteem points because that scale means nothing to me. I would be impressed, however, if you said that they gained nearly one and a half standard deviation units in self-esteem.

The issue is not quite as simple as I have made it out to be because there are alternative ways of approaching the problem. One way would be to use the average of the pre- and post-score standard deviations, rather than just the standard deviation of the pre-scores. However, when we are measuring gain, it makes sense to me to measure it in the metric of the original weights. You may come across situations where you would think that it makes more sense to use the average standard deviation. In addition, it would be perfectly possible to use the standard deviation of the difference scores in the denominator for d. Kline (2004) discusses this approach and concludes, "If our natural reference for thinking about scores on (some) measure is their original standard deviation, it makes most sense to report standardized mean change (using that standard deviation)." However, the important point here is to keep in mind that such decisions often depend on substantive considerations in the particular research field, and no one measure is uniformly best.

Confidence Limits on *d*

Just as we were able to establish confidence limits on our estimate of the population mean (μ), we can establish confidence limits on d. However, it is not a simple process to do so, and I refer the reader to Kline (2004) or Cumming and Finch (2001). The latter provide a very inexpensive computer program to make these calculations.

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 disabled 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 sets 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 disabled whereas others are only mildly disabled. 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, thus causing their scores to be related.

Many experimental designs involve related samples. They all have one thing in common, and that is 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 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 using a compute 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. This is very similar to what we have already done, so I will not repeat that here.

Writing Up the Results of a Dependent *t*

Suppose that we want to write up the results of Everitt's study of family therapy for anorexia. We would want to be sure to include the relevant sample statistics ($\bar{X}$, s^2, and N), as well as the test of statistical significance. But we would also want to include confidence

limits on the mean weight gain following therapy, and our effect size estimate (d). We might write,

> Everitt ran a study on the effect of family therapy on weight gain in girls suffering from anorexia. He collected weight data on 17 girls before therapy, provided family therapy to the girls and their families, and then collected data on the girls' weight at the end of therapy.
>
> The mean weight gain for the $N = 17$ girls was 7.26 pounds, with a standard deviation of 7.16. A two-tailed t-test on weight gain was statistically significant ($t(16) = 4.18$, $p < .05$), revealing that on average the girls did gain weight over the course of therapy. A 95% confidence interval on mean weight gain was 3.57–10.95, which is a notable weight gain even at the low end of the interval. Cohen's $d = 1.45$, indicating that the girls' weight gain was nearly 1.5 standard deviations relative to their original pretest weights. It would appear that family therapy has made an important contribution to the treatment of anorexia in this experiment.

7.5 Hypothesis Tests Applied to Means—Two Independent Samples

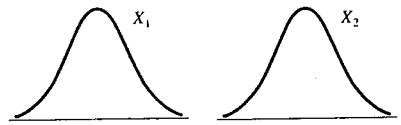
One of the most common uses of the t test involves testing the difference between the means of two independent groups. We might want to compare the mean number of trials needed to reach criterion on a simple visual discrimination task for two groups of rats—one raised under normal conditions and one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



	$\overline{X}_{11}$	$\overline{X}_{21}$	$\overline{X}_{11} - \overline{X}_{21}$
	$\overline{X}_{12}$	$\overline{X}_{22}$	$\overline{X}_{12} - \overline{X}_{22}$
	$\overline{X}_{13}$	$\overline{X}_{23}$	$\overline{X}_{13} - \overline{X}_{23}$
	...	...	...
	$\overline{X}_{1\infty}$	$\overline{X}_{2\infty}$	$\overline{X}_{1\infty} - \overline{X}_{2\infty}$
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 7.8 Schematic set of means and mean differences when sampling from two populations

are presented schematically in Figure 7.8. 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 ($\overline{X}_1 - \overline{X}_2$). We are most interested in this third column because 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.[8]

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. Because 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 7.9.

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 distributed variables is itself normally distributed. Because Figure 7.9 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 7.9 must itself be at least approximately normal.

[8] 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.)

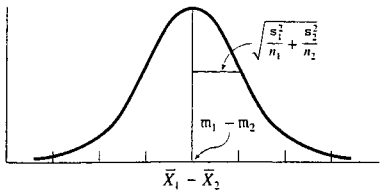


Figure 7.9 Sampling distribution of mean differences

The *t* Statistic

standard error of differences between means

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**[9] is

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

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 $\alpha = .05$ 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.

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

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

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

The null hypothesis is generally the hypothesis that $\mu_1 - \mu_2 = 0$, so 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.) Because 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

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

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

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

pooled variance estimate

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

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

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 how 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.

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 their corresponding sample means, 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, & 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. The researchers 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 is the degree of arousal at the end of the 4-minute video, with larger values indicating greater arousal.

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			

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 let $\alpha = .05$. It is important to keep in mind, however, that there is nothing particularly sacred about any of these decisions. (Think about how Jones and Tukey [2000] would have written this paragraph. Where would they have differed from what is here, and why might their approach be clearer?)

[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.

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 heterogeneity 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, $t_{.025}(62) \cong \pm 2.003$ (with linear interpolation). Because the value of t_{obt} far exceeds $t_{\alpha/2}$, we will reject H_0 (at $\alpha = .05$) 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. (How would the conclusions of Jones and Tukey (2000) compare with the one given here?)

Confidence Limits on $\mu_1 - \mu_2$

In addition to testing a null hypothesis about population means (i.e., testing H_0: $\mu_1 - \mu_2 = 0$), it is useful to set confidence limits and effect sizes on the difference between μ_1 and μ_2. The logic for setting confidence limits is exactly the same as it was for the one-sample case. The calculations are also exactly 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

$$CI_{95} = (\overline{X}_1 - \overline{X}_2) \pm t_{.05} s_{\overline{X}_1 - \overline{X}_2}$$

For the homophobia study, we have

$$CI_{95} = (\overline{X}_1 - \overline{X}_2) \pm t_{.05} s_{\overline{X}_1 - \overline{X}_2} = (24.00 - 16.5) \pm 2.00\sqrt{\frac{144.48}{35} + \frac{144.48}{29}}$$

$$= 7.50 \pm 2.00(3.018) = 7.5 \pm 6.04$$

$$1.46 \leq (\mu_1 - \mu_2) \leq 13.54$$

The probability is .95 that an interval computed as we computed this interval encloses the difference in arousal to homosexual videos between homophobic and nonhomophobic participants. Although the interval is wide, it does not include 0. This is consistent with our rejection

of the null hypothesis and allows us to state that homophobic individuals are, in fact, more sexually aroused by homosexual videos than are nonhomophobic individuals. However, I think that we would be remiss if we simply ignored the width of this interval. Although the difference between groups is statistically significant, there is still considerable uncertainty about how large the difference is. In addition, keep in mind that the dependent variable is the "degree of sexual arousal" on an arbitrary scale. Even if your confidence interval were quite narrow, it is difficult to know what to make of the result in absolute 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, as we will in the next section, it is often more informative, as I think it is here.

Effect Size

The confidence interval that we just calculated has shown us that we still have considerable uncertainty about the difference in sexual arousal between groups, even though our statistically significant difference tells us that the homophobic group actually shows more arousal than the nonhomophobic group does. Again, we come to the issue of finding ways to present information to our readers that conveys the magnitude of the difference between our groups. We will use an effect size measure based on Cohen's d. It is very similar to the one that we used in the case of two dependent samples, where we divided the difference between the means by a standard deviation. We will again call this statistic (d). In this case, however, our standard deviation will be the estimated standard deviation of either population. More specifically, we will pool the two variances and take the square root of the result, and that will give us our best estimate of the standard deviation of the populations from which the numbers were drawn.[11] (If we had noticeably different variances, we would most likely use the standard deviation of one sample and note to the reader that this is what we had done.)

For our data on homophobia, we have

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

This result expresses the difference between the two groups in standard deviation units and tells us that the mean arousal for homophobic participants was nearly 2/3 of a standard deviation higher than the arousal of nonhomophobic participants. That strikes me as a big difference. (Using the software by Cumming and Finch [2001], we find that the confidence intervals on d are 0.1155 and 1.125, which is also rather wide. At the same time, even the lower limit on the confidence interval is meaningfully large.)

A word of caution: In the example of homophobia, the units of measurement were largely arbitrary, and a 7.5 difference had no intrinsic meaning to us. Thus, it made more sense to express it in terms of standard deviations because we have at least some understanding of what that means. However, there are many cases wherein the original units are meaningful, and in those cases it may not make much sense to standardize the measure (i.e., report it in standard deviation units). We might prefer to specify the difference between means, or the ratio of means, or some similar statistic. The earlier example of the moon illusion is a case in point. There, it is far more meaningful to speak of the horizon moon appearing approximately half-again as large as the zenith moon, and I see no advantage, and some obfuscation, in converting to standardized units. The important goal is to give the reader an appreciation of the size of a difference, and you should choose that measure that

[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.

best expresses this difference. In one case, a standardized measure such as d is best, and in other cases, other measures, such as the distance between the means, is better.

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 $d = .20$ as an effect that is small, but probably meaningful, an effect size of $d = .50$ 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 $d = .80$ as large. We should not make too much of Cohen's levels, but they are helpful as a rough guide.

Reporting Results

Reporting results for a t test on two independent samples is basically similar to reporting results for the case of dependent samples. In Adams's et al. study of homophobia, two groups of participants were involved—one group scoring high on a scale of homophobia and the other scoring low. When presented with sexually explicit homosexual videos, the homophobic group actually showed a higher level of sexual arousal (the mean difference = 7.50 units). A t test of the difference between means produced a statistically significant result ($p < .05$), and Cohen's $d = .62$ showed that the two groups differed by nearly 2/3 of a standard deviation. However, the confidence limits on the population mean difference were rather wide ($1.46 \leq \mu_1 - \mu_2 \leq 13.54$), suggesting that we do not have a tight handle on the size of our difference.

SPSS Analysis

The SPSS analysis of the Adams et al. (1996) data is given in Exhibit 7.2. 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 ($p = .534$), 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.

Group Statistics

GROUP		N	Mean	Std. Deviation	Std. Error Mean
Arousal	Homophobic	35	24.0000	12.2013	2.0624
	Nonhomophobic	29	16.5034	11.7966	2.1906

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
								95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Equal variances assumed	.391	.534	2.484	62	.016	7.4966	3.0183	1.4630	13.5301
Equal variances not assumed			2.492	60.495	.015	7.4966	3.0087	1.4794	13.5138

Exhibit 7.2 SPSS analyses of Adams et al. (1996) data

Next note that the t supplied by SPSS is the same as we calculated, and that the probability associated with this value of t (.016) is less than $\alpha = .05$, 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.

7.6 A Final Worked Example

Joshua Aronson has done extensive work on what he refers to as "stereotype threat," which refers to the fact that "members of stereotyped groups often feel extra pressure in situations where their behavior can confirm the negative reputation that their group lacks a valued ability." (Aronson, Lustina, Good, Keough, Steele, & Brown, 1998) This feeling of stereotype threat is then hypothesized to affect their performance, generally by lowering it from what it would have been had they not felt threatened. Considerable work has been done with ethnic groups who are stereotypically reputed to do poorly in some area, but Aronson et al. went a step further to ask if stereotype threat could actually lower the performance of white males—a group not normally associated with stereotype threat.

Aronson et al. (1998) used two independent groups of college students who were known to excel in mathematics, and for whom doing well in math was considered important. The researchers assigned 11 students to a control group that was simply asked to complete a difficult mathematics exam. They assigned 12 students to a threat condition, in which they were told that Asian students typically did better than other students in math tests, and that the purpose of the exam was to help the experimenter to understand why this difference exists. Aronson reasoned that simply telling white students that Asians did better on math tests would arouse feelings of stereotype threat and diminish the students' performance.

The data in Table 7.6 have been constructed to have nearly the same means and standard deviations as Aronson's data. The dependent variable is the number of items correctly solved.

First, we need to specify the null hypothesis, the significance level, and whether we will use a one- or a two-tailed test. We want to test the null hypothesis that the two conditions perform equally well on the test, so we have H_0: $\mu_1 = \mu_2$. We will set alpha at $\alpha = .05$, in line with what we have been using. Finally, we will choose to use a two-tailed test because it is reasonably possible for either group to show superior math performance.

Next, we need to calculate the pooled variance estimate:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{10(3.17^2) + 11(3.03^2)}{11 + 12 - 2}$$

$$= \frac{10(10.0489) + 11(9.1809)}{21} = \frac{201.4789}{21} = 9.5942$$

Table 7.6 Data from Aronson et al. (1998)

Control Subjects				Threat Subjects			
4	9	12	8	7	8	7	2
9	13	12	13	6	9	7	10
13	7	6		5	0	10	8
Mean = 9.64				Mean = 6.58			
st. dev. = 3.17				st. dev. = 3.03			
$n_1 = 11$				$n_2 = 12$			

Finally, we can calculate t using the pooled variance estimate:

$$t = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{(9.64 - 6.58)}{\sqrt{\frac{9.5942}{11} + \frac{9.5942}{12}}} = \frac{3.06}{\sqrt{1.6717}} = \frac{3.06}{1.2929} = 2.37$$

For this example, we have $n_1 + n_2 - 2 = 21$ degrees of freedom. From Appendix t, we find $t_{.025} = 2.080$. Because $2.37 > 2.080$, we will reject H_0 and conclude that the two population means are not equal.

If you were writing up the results of this experiment, you might write something like the following:

> This experiment tested the hypothesis that stereotype threat will disrupt the performance even of a group that is not usually thought of as having a negative stereotype with respect to performance on math tests. Aronson et al. (1998) asked two groups of participants to take a difficult math exam. These were white male college students who reported that they typically performed well in math and that good math performance was important to them. One group of students ($n = 11$) was simply given the math test and asked to do as well as they could. A second, randomly assigned group ($n = 12$) was informed that Asian males often outperformed white males and that the test was intended to help to explain the difference in performance. The test itself was the same for all participants. The results showed that the Control subjects answered a mean of 9.64 problems correctly, whereas the subjects in the Threat group completed only a mean of 6.58 problems. The standard deviations were 3.17 and 3.03, respectively. This represents an effect size (d) of .99, meaning that the two groups differed in the number of items correctly completed by nearly one standard deviation.
>
> Student's t test was used to compare the groups. The resulting $t(21)$ was 2.37, and was significant at $p < .05$, showing that stereotype threat significantly reduced the performance of those subjects to whom it was applied. The 95% confidence interval on the difference in means is $0.3712 \leq \mu_1 - \mu_2 \leq 5.7488$. This is quite a wide interval, but keep in mind that the two sample sizes were 11 and 12. An alternative way of comparing groups is to note that the Threat group answered 32% fewer items correctly than did the Control group.

7.7 Heterogeneity of Variance: The Behrens–Fisher Problem

homogeneity of variance

We have already seen that one of the assumptions underlying the t test for two independent samples is the assumption of **homogeneity of variance** ($\sigma_1^2 = \sigma_2^2 = \sigma^2$). To be more specific, we can say that *when* H_0 is true and *when* we have homogeneity of variance, then, pooling the variances, the ratio

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

is distributed as t on $n_1 + n_2 - 2$ df. If we can assume homogeneity of variance, there is no difficulty, and the techniques discussed in this section are not needed. When we do not have homogeneity of variance, however, this ratio is not, strictly speaking, distributed as t. This leaves us with a problem, but fortunately a solution (or a number of competing solutions) exists.

heterogeneous variances

First, unless $\sigma_1^2 = \sigma_2^2 = \sigma^2$, it makes no sense to pool (average) variances because the reason we were pooling variances in the first place was that we assumed them to be estimating the same quantity. For the case of **heterogeneous variances,** we will first dispense with pooling procedures and define

$$t' = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where s_1^2 and s_2^2 are taken to be heterogeneous variances. As noted earlier the expression that I have just denoted as t' is *not* necessarily distributed as t on $n_1 + n_2 - 2\,df$. If we knew what the sampling distribution of t' actually looked like, there would be no problem. We would just evaluate t' against that sampling distribution. Fortunately, although there is no universal agreement, we know at least the approximate distribution of t'.

The Sampling Distribution of t'

Behrens–Fisher problem

One of the first attempts to find the exact sampling distribution of t' was begun by Behrens and extended by Fisher, and the general problem of heterogeneity of variance has come to be known as the **Behrens–Fisher problem.** Based on this work, the Behrens–Fisher distribution of t' was derived and is presented in a table in Fisher and Yates (1953). However, because this table covers only a few degrees of freedom, it is not particularly useful for most purposes.

Welch–Satterthwaite solution

An alternative solution was developed apparently independently by Welch (1938) and by Satterthwaite (1946). The **Welch–Satterthwaite solution** is particularly important because we will refer back to it when we discuss the analysis of variance. Using this method, t' is viewed as a legitimate member of the t distribution, but for an unknown number of degrees of freedom. The problem then becomes one of solving for the appropriate df, denoted df':

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

The degrees of freedom (df') are then taken to the nearest integer.[12] The advantage of this approach is that df' is bounded by the smaller of $n_1 - 1$ and $n_2 - 1$ at one extreme and $n_1 + n_2 - 2\,df$ at the other. More specifically, $Min(n_1 - 1, n_2 - 1) \le df' \le n_1 + n_2 - 2$. Because the critical value of t decreases as df increases, we can first evaluate t' as if df' were at its minimum. If the difference is significant, it will certainly be significant for the true df'.

[12] Welch (1947) later suggested that letting

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

might be a more accurate solution, although the difference is negligible.

If the difference is not significant, we can then evaluate t' at its maximum $n_1 + n_2 - 2$. If it is not significant at this point, no reduction in the degrees of freedom (by more accurate calculation of df') would cause it to be significant. Thus, the only time we actually need to calculate df' is when the value of t' would not be significant for $Min(n_1 - 1, n_2 - 1)\,df$ but would be significant for $n_1 + n_2 - 2\,df$.

In this book, we will rely primarily on the Welch–Satterthwaite approximation. It has the distinct advantage of applying easily to problems that arise in the analysis of variance, and it is not noticeably more awkward than the other solutions.

Testing for Heterogeneity of Variance

How do we know whether we even have heterogeneity of variance to begin with? We do not know σ_1^2 and σ_2^2 (if we did we would not be solving for t), so we must in some way test their difference by using our two sample variances (s_1^2 and s_2^2).

A number of solutions have been put forth for testing for heterogeneity of variance. One of the simpler ones was advocated by Levene (1960), who suggested replacing each value of X either by its absolute deviation from the group mean—$d_{ij} = |X_{ij} - \overline{X}_j|$—or by its squared deviation—$d_{ij} = (X_{ij} - \overline{X}_j)^2$—where i and j represent the ith subject in the jth group. He then proposed running a standard two-sample t test on the d_{ij}s. This test makes intuitive sense, because if there is greater variability in one group, the absolute, or squared, values of the deviations will be greater. If t is significant, we would then declare the two groups to differ in their variances. Alternative approaches have been proposed—see, for example, O'Brien (1981)—but they are rarely implemented in standard software, and I will not elaborate on them here.

The procedures just described are suggested as replacements for the more traditional F test, which is a ratio of the larger sample variance to the smaller. This F has been shown by many people to be severely affected by nonnormality of the data and should not be used. The F test is still computed and printed by many of the large computer packages, but I do not recommend using it.

The Robustness of t with Heterogeneous Variances

robust

I mentioned that the t test is what is described as **robust,** meaning that it is more or less unaffected by moderate departures from the underlying assumptions. For the t test for two independent samples, we have two major assumptions and one side condition that must be considered. The two assumptions are those of normality of the sampling distribution of differences between means and homogeneity of variance. The side condition is the condition of equal sample sizes versus unequal sample sizes. Although we have just seen how the problem of heterogeneity of variance can be handled by special procedures, it is still relevant to ask what happens if we use the standard approach even with heterogeneous variances.

Box (1953), Norton (1953), Boneau (1960), and many others have investigated the effects of violating, both independently and jointly, the underlying assumptions of t. The general conclusion to be drawn from these studies is that for equal sample sizes, violating the assumption of homogeneity of variance produces very small effects—the nominal value of $\alpha = .05$ is most likely within ± 0.02 of the true value of α. By this we mean that if you set up a situation with unequal variances *but with* H_0 *true* and proceed to draw (and compute t on) a large number of pairs of samples, you will find that somewhere between 3% and 7% of the sample t values actually exceed $\pm t_{.025}$. This level of inaccuracy is not intolerable. The same kind of statement applies to violations of the assumption of normality, provided that the true populations are roughly the same shape or else both are symmetric. If the

distributions are markedly skewed (especially in opposite directions), serious problems arise unless their variances are fairly equal.

With unequal sample sizes, however, the results are more difficult to interpret. In Boneau's study, for example, sample variances were pooled in all cases because this is probably the most common procedure in practice (although it is incorrect for heterogeneous variances). Boneau found that when there was heterogeneity of variance *and* unequal sample sizes, the actual and normative probability values differed considerably. Keep in mind, however, that Boneau was pooling variances and evaluating t on $n_1 + n_2 - 2$ *df*. We do not know what would have happened had he solved for t' and then evaluated t' on df' degrees of freedom. We do know, however, that had he done so, it would be reasonable to expect that the test would have proven to be robust because the Welch–Satterthwaite solution does not require the homogeneity assumption.

The investigator who has collected data that she thinks may violate one or more of the underlying assumptions should refer to the article by Boneau (1960). This article may be old, but it is quite readable and contains an excellent list of references to other work in the area. A good summary of alternative procedures can be found in Games, Keselman, and Rogan (1981).

Wilcox (1992) has argued persuasively for the use of trimmed samples for comparing group means with heavy-tailed distributions. (Interestingly, statisticians seem to have a fondness for trimmed samples, whereas psychologists and other social science practitioners seem not to have heard of trimming.) He provides results showing dramatic increases in power when compared with more standard approaches. Alternative nonparametric approaches, including "resampling statistics," are discussed in Chapter 18 of this book. These can be very powerful techniques that do not require unreasonable assumptions about the populations from which you have sampled. I suspect that resampling statistics and related procedures will be in the mainstream of statistical analysis in the not-too-distant future.

7.8 Hypothesis Testing Revisited

In Chapter 4, we spent quite a bit of time on examining the process of hypothesis testing. I pointed out that the traditional approach involves setting up a null hypothesis, and then generating a statistic that tells us how likely we are to find the obtained results if, in fact, the null hypothesis is true. In other words, we calculate the probability of the data given the null, and if that probability is very low, we reject the null.

In that chapter, we also looked briefly at a proposal by Jones and Tukey (2000) in which they approached the problem slightly differently. Now that we have several examples, this is a good point to go back and look at their proposal. In discussing the Adams et al. study of homophobia, I suggested that you think about how Jones and Tukey would have approached the issue. I am not going to repeat the traditional approach because that is laid out in each of the examples of how to write up our results.

The study by Adams et al. (1996) makes a good example. I imagine that all of us would be willing to agree that the null hypothesis of equal population means in the two conditions is highly unlikely to be true. Even laying aside the argument about differences in the 10th decimal place, it just seems unlikely that people who differ appreciably in amount of homophobia would show exactly the same mean level of arousal to erotic videos. We don't know which group will show the greater arousal, but one population mean is certain to be larger than the other. So we can rule out the null hypothesis (H_0: $\mu_H - \mu_N = 0$) as a viable possibility. That leaves us with three possible conclusions we could draw as a result of our test. The first is that $\mu_H < \mu_N$, the second is that $\mu_H > \mu_N$, and the third is that we do not have sufficient evidence to draw a conclusion.

Now let's look at the possibilities of error. It could actually be that $\mu_H < \mu_N$, but that we draw the opposite conclusion by deciding that the nonhomophobic participants are more aroused. This is what Jones and Tukey call a "reversal," and the probability of making this error if we use a *one-tailed* test at $\alpha = .05$ is .05. Alternatively, it could be that $\mu_H > \mu_N$ but that we make the error of concluding that the nonhomophobic participants are less aroused. Again with a one-tailed test, the probability of making this error is .05. It is not possible for us to make both of these errors because one of the hypotheses is true, so using a *one-tailed* test (in both directions) at $\alpha = .05$ gives us a 5% error rate. In our particular example, the critical value for a one-tailed test on 62 *df* is approximately 1.68. Because our obtained value of t was 2.48, we will conclude that homophobic participants are more aroused, on average, than nonhomophobic participants were. Notice that in writing this paragraph I have not used the phrase "Type I error" because that refers to rejecting a true null, and I have already said that the null can't possibly be true. Notice that my conclusion did not contain the phrase "rejecting the hypothesis." Instead, I referred to "drawing a conclusion." These are subtle differences, but I hope this example clarifies the position taken by Jones and Tukey.

Key Terms

Sampling distribution of the mean (7.1)
Central limit theorem (7.1)
Uniform distribution (7.1)
Standard error (7.2)
Student's t distribution (7.3)
Point estimate (7.3)
Confidence limits (7.3)
Confidence interval (7.3)
p level (7.3)
Matched samples (7.4)
Repeated measures (7.4)
Related samples (7.4)
Matched-sample t test (7.4)
Difference scores (7.4)
Gain scores (7.4)
Cohen's d (7.4)
Sampling distribution of differences between means (7.5)
Variance sum law (7.5)
Standard error of differences between means (7.5)
Weighted average (7.5)
Pooled variance estimate (7.5)
Homogeneity of variance (7.7)
Heterogeneous variances (7.7)
Behrens–Fisher problem (7.7)
Welch–Satterthwaite solution (7.7)
Robust (7.7)

Exercises

7.1 The following numbers represent 100 random numbers drawn from a rectangular population with a mean of 4.5 and a standard deviation of 2.7. Plot the distribution of these digits.

6	4	8	7	8	7	0	8	2	8	5	7
4	8	2	6	9	0	2	6	4	9	0	4
9	3	4	2	8	2	0	4	1	4	7	4
1	7	4	2	4	1	4	2	8	7	9	7
3	7	4	7	3	1	6	7	1	8	7	2
7	6	2	1	8	6	2	3	3	6	5	4
1	7	2	1	0	2	6	0	8	3	2	4
3	8	4	5	7	0	8	4	2	8	6	3
7	3	5	1								

7.2 I drew 50 samples of 5 scores each from the same population that the data in Exercise 7.1 came from, and calculated the mean of each sample. The means are shown here. Plot the distribution of these means.

2.8	6.2	4.4	5.0	1.0	4.6	3.8	2.6	4.0	4.8
6.6	4.6	6.2	4.6	5.6	6.4	3.4	5.4	5.2	7.2
5.4	2.6	4.4	4.2	4.4	5.2	4.0	2.6	5.2	4.0
3.6	4.6	4.4	5.0	5.6	3.4	3.2	4.4	4.8	3.8
4.4	2.8	3.8	4.6	5.4	4.6	2.4	5.8	4.6	4.8

7.3 Compare the means and the standard deviations for the distribution of digits in Exercise 7.1 and the sampling distribution of the mean in Exercise 7.2.

a. What would the central limit theorem lead you to expect in this situation?

b. Do the data correspond to what you would predict?

7.4 How would the result in Exercise 7.2 differ if you had drawn more samples of size 5?

7.5 How would the result in Exercise 7.2 differ if you had drawn 50 samples of size 15?

7.6 In 1979, the 238 students from North Dakota who took the verbal portion of the SAT exam had a mean score of 525. The standard deviation was not reported.

a. Is this result consistent with the idea that the SAT has a mean of 500 and a standard deviation of 100?

b. Would you have rejected H_0 had you been looking for evidence that SAT scores in general have been declining over the years from the mean of 500?

c. If you rejected H_0 in part (a), you might draw some conclusions about North Dakota's students or our assumption about the general population of students. What are those possible conclusions?

7.7 Why do the data in Exercise 7.6 not really speak to the issue of whether American education in general is in a terrible state?

7.8 In 1979, the 2,345 students from Arizona who took the math portion of the SAT had a mean score of 524. Is this consistent with the notion of a population mean of 500 if we assume that $\sigma = 100$?

7.9 Why does the answer to Exercise 7.8 differ substantially from the answer to Exercise 7.6 even though the means are virtually the same?

7.10 Compute 95% confidence limits on μ for the data in Exercise 7.6.

7.11 Everitt, in Hand et al. (1994), reported on several different therapies as treatments for anorexia. There were 29 girls in a cognitive-behavior therapy condition, and they were weighed before and after treatment. The weight gains of the girls, in pounds, are given here. The scores was obtained by subtracting the Before score from the After score, so that a negative difference represents weight *loss*, and a positive difference represents a *gain*.

1.7	0.7	−0.1	−0.7	−3.5	14.9	3.5	17.1	−7.6	1.6	11.7
6.1	1.1	−4.0	20.9	−9.1	2.1	−1.4	1.4	−0.3	−3.7	−0.8
2.4	12.6	1.9	3.9	0.1	15.4	−0.7				

a. What does the distribution of these values look like?

b. Did the girls in this group gain a statistically significant amount of weight?

7.12 Compute 95% confidence limits on the weight gain in Exercise 7.11.

7.13 Katz, Lautenschlager, Blackburn, and Harris (1990) examined the performance of 28 students who answered multiple-choice items on the SAT without having read the passages to which the items referred. The mean score (out of 100) was 46.6, with a standard deviation of 6.8. Random guessing would have been expected to result in 20 correct answers.

a. Were these students responding at better-than-chance levels?

b. If performance is statistically significantly better than chance, does it mean that the SAT test is not a valid predictor of future college performance?

7.14 Compas and others (1994) were surprised to find that young children under stress actually report fewer symptoms of anxiety and depression than we would expect. But they also noticed that their scores on a Lie Scale (a measure of the tendency to give socially desirable answers)

were higher than expected. The population mean for the Lie scale on the Children's Manifest Anxiety Scale (Reynolds & Richmond, 1978) is known to be 3.87. For a sample of 36 children under stress, Compas et al. found a sample mean of 4.39, with a standard deviation of 2.61.

a. How would we test whether this group shows an increased tendency to give socially acceptable answers?

b. What would the null hypothesis and research hypothesis be?

c. What can you conclude from the data?

7.15 Calculate the 95% confidence limits for μ for the data in Exercise 7.14. Are these limits consistent with your conclusion in Exercise 7.14?

7.16 Hoaglin, Mosteller, and Tukey (1983) present data on blood levels of beta-endorphin as a function of stress. They took beta-endorphin levels for 19 patients 12 hours before surgery, and again 10 minutes before surgery. The data are presented here, in fmol/ml:

ID	1	2	3	4	5	6	7	8	9	10
12 hours	10.0	6.5	8.0	12.0	5.0	11.5	5.0	3.5	7.5	5.8
10 minutes	6.5	14.0	13.5	18.0	14.5	9.0	18.0	42.0	7.5	6.0
ID	11	12	13	14	15	16	17	18	19	
12 hours	4.7	8.0	7.0	17.0	8.8	17.0	15.0	4.4	2.0	
10 minutes	25.0	12.0	52.0	20.0	16.0	15.0	11.5	2.5	2.0	

Based on these data, what effect does increased stress have on beta-endorphin levels?

7.17 Why would you use a matched-sample t test in Exercise 7.16?

7.18 Construct 95% confidence limits on the true mean difference between beta-endorphin levels at the two times described in Exercise 7.16.

7.19 Hout, Duncan, and Sobel (1987) reported on the relative sexual satisfaction of married couples. They asked each member of 91 married couples to rate the degree to which they agreed with "Sex is fun for me and my partner" on a four-point scale ranging from "never or occasionally" to "almost always." The data appear below (I know it's a lot of data, but it's an interesting question):

Husband	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Wife	1	1	1	1	1	1	1	2	2	2	2	2	2	2	3
Husband	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2
Wife	3	4	4	4	1	1	2	2	2	2	2	2	2	2	3
Husband	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
Wife	3	3	4	4	4	4	4	4	4	1	2	2	2	2	2
Husband	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4
Wife	3	3	3	3	4	4	4	4	4	4	4	4	4	1	1
Husband	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Wife	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3
Husband	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Wife	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4

Start by running a matched-sample t test on these data. Why is a matched-sample test appropriate?

7.20 In the study referred to in Exercise 7.19, what, if anything does your answer to that question tell us about whether couples are sexually compatible? What do we know from this analysis, and what don't we know?

7.21 For the data in Exercise 7.19, create a scatterplot and calculate the correlation between husband's and wife's sexual satisfaction. How does this amplify what we have learned from the analysis in Exercise 7.19. (I do not discuss scatterplots and correlation until Chapter 9, but

a quick glance at Chapter 9 should suffice if you have difficulty. SPSS will easily do the calculation.)

7.22 Construct 95% confidence limits on the true mean difference between the Sexual Satisfaction scores in Exercise 7.19, and interpret them with respect to the data.

7.23 Some would object that the data in Exercise 7.19 are clearly discrete, if not ordinal, and that it is inappropriate to run a *t* test on them. Can you think what might be a counter argument? (This is not an easy question, and I really asked it mostly to make the point that there could be controversy here.)

7.24 Give an example of an experiment in which using related samples would be ill advised because taking one measurement might influence another measurement.

7.25 Everitt, in Hand et al. (1994), (see Exercise 7.11) reported on family therapy as a treatment for anorexia. There were 17 girls in this experiment, and they were weighed before and after treatment. The weights of the girls, in pounds, are given here. The row of difference scores was obtained by subtracting the Before score from the After score, so that a negative difference represents weight *loss*, and a positive difference represents a *gain*.

ID	1	2	3	4	5	6	7	8	9	10
Before	83.8	83.3	86.0	82.5	86.7	79.6	76.9	94.2	73.4	80.5
After	95.2	94.3	91.5	91.9	100.3	76.7	76.8	101.6	94.9	75.2
Diff	11.4	11.0	5.5	9.4	13.6	−2.9	−.1	7.4	21.5	−5.3

ID	11	12	13	14	15	16	17	Mean	St. Dev
Before	81.6	82.1	77.6	83.5	89.9	86.0	87.3	83.23	5.02
After	77.8	95.5	90.7	92.5	93.8	91.7	98.0	90.49	8.48
Diff	−3.8	13.4	13.1	9.0	3.9	5.7	10.7	7.26	7.16

a. What null hypothesis would these data lead you to want to test?

b. Run the appropriate *t* test and draw the appropriate conclusion.

7.26 What would happen in Exercise 7.25 if I subtracted the After score from the Before score, instead of the other way around?

7.27 Calculate a confidence interval on the weight gain for the girls in Everitt's study in Exercise 7.25.

7.28 Graph the relationship between the Before and After scores to evaluate the degree to which the two sets of scores are related. (I do not discuss scatterplots until Chapter 9, but you should be able to work this out for yourself.)

7.29 In the study referred to in Exercise 7.13, Katz et al. (1990) compared the performance on SAT items of a group of 17 students who were answering questions about a passage after having read the passage with the performance of a group of 28 students who had not seen the passage. The mean and standard deviation for the first group were 69.6 and 10.6, whereas for the second group they were 46.6 and 6.8.

a. What is the null hypothesis?

b. What is the alternative hypothesis?

c. Run the appropriate *t* test.

d. Interpret the results.

7.30 Many mothers experience a sense of depression shortly after the birth of a child. Design a study to examine postpartum depression and, from material in this chapter, tell how you would estimate the mean increase in depression.

7.31 In Exercise 7.25, we saw data from Everitt that showed that girls receiving family therapy gained weight over the course of that therapy. However, it is possible that they just gained weight because they got older. One way to control for this is to look at the amount of weight gained by the Family Therapy group ($n = 17$) in contrast with the amount gained by girls in a Control group ($n = 26$), who received no therapy. The data on weight gain for the two groups is shown below.

Control		Family Therapy	
−0.5	3.3	11.4	9.0
−9.3	11.3	11.0	3.9
−5.4	0.0	5.5	5.7
12.3	−1.0	9.4	10.7
−2.0	−10.6	13.6	
−10.2	−4.6	−2.9	
−12.2	−6.7	−0.1	
11.6	2.8	7.4	
−7.1	0.3	21.5	
6.2	1.8	−5.3	
−0.2	3.7	−3.8	
−9.2	15.9	13.4	
8.3	−10.2	13.1	
Mean	−0.45		7.26
St Dev.	7.99		7.16
Variance	63.82		51.23

Run the appropriate test to compare the group means. What would you conclude?

7.32 Calculate the confidence interval on $\mu_1 - \mu_2$ for the data in Exercise 7.31.

7.33 In Exercise 7.19, we saw pairs of observations on sexual satisfaction for husbands and wives. Suppose that those data had actually come from unrelated males and females, such that the data are no longer paired. What effect you expect this to have on the analysis?

7.34 Run the appropriate *t* test on the data in Exercise 7.19 assuming that the observations are independent. What would you conclude?

7.35 Why isn't the difference between the results in Exercises 7.34 and 7.19 greater than it is?

7.36 What is the role of random assignment in the Everitt's anorexia study referred to in Exercise 7.31, and under what conditions might we find it difficult to carry out random assignment?

7.37 The Thematic Apperception Test (TAT) presents subjects with ambiguous pictures and asks them to tell a story about them. These stories can be scored in any number of ways. Werner, Stabenau, and Pollin (1970) asked mothers of 20 Normal and 20 Schizophrenic children to complete the TAT and scored for the number of stories (out of 10) that exhibited a positive parent-child relationship. The data follow:

Normal	8	4	6	3	1	4	4	6	4	2
Schizophrenic	2	1	1	3	2	7	2	1	3	1
Normal	2	1	1	4	3	3	2	6	3	4
Schizophrenic	0	2	4	2	3	3	0	1	2	2

a. What would you assume to be the experimental hypothesis behind this study?

b. What would you conclude with respect to that hypothesis?

7.38 In Exercise 7.37, why might it be smart to look at the variances of the two groups?

7.39 In Exercise 7.37, a significant difference might lead someone to suggest that poor parent-child relationships are the cause of schizophrenia. Why might this be a troublesome conclusion?

7.40 Much has been made of the concept of experimenter bias, which refers to the fact that even the most conscientious experimenters tend to collect data that come out in the desired direction (they see what they want to see). Suppose we use students as experimenters. All the experimenters are told that subjects will be given caffeine before the experiment, but one-half of the experimenters are told that we expect caffeine to lead to good performance and

one-half are told that we expect it to lead to poor performance. The dependent variable is the number of simple arithmetic problems the subjects can solve in 2 minutes. The data obtained are as follows:

Expectation good:	19	15	22	13	18	15	20	25	22
Expectation poor:	14	18	17	12	21	21	24	14	

What can you conclude?

7.41 Calculate 95% confidence limits on $\mu_1 - \mu_2$ for the data in Exercise 7.40.

7.42 An experimenter examining decision making asked 10 children to solve as many problems as they could in 10 minutes. One group (5 subjects) was told that this was a test of their innate problem-solving ability; a second group (5 subjects) was told that this was just a time-filling task. The data follow:

Innate ability:	4	5	8	3	7
Time-filling task:	11	6	9	7	9

Does the mean number of problems solved vary with the experimental condition?

7.43 A second investigator repeated the experiment described in Exercise 7.42 and obtained the same results. However, she thought that it would be more appropriate to record the data in terms of minutes per problem (e.g., 4 problems in 10 minutes = 10/4 = 2.5 minutes/problem). Thus, her data were as follows:

Innate ability:	2.50	2.00	1.25	3.33	1.43
Time-filling task:	0.91	1.67	1.11	1.43	1.11

Analyze and interpret these data with the appropriate t test.

7.44 What does a comparison of Exercises 7.42 and 7.43 show you?

7.45 I stated earlier that Levene's test consists of calculating the absolute (or squared) differences between individual observations and their group's mean, and then running a t test on those differences. Using any computer software, it is simple to calculate those absolute and squared differences and then to run a t test on them. Calculate both and determine which approach SPSS is using in the example. (*Hint*, $F = t^2$ here, and the F value that SPSS actually calculated was 0.391148, to 6 decimal places.)

7.46 Research on clinical samples (i.e., people referred for diagnosis or treatment) has suggested that children who experience the death of a parent may be at risk for developing depression or anxiety in adulthood. Mireault (1990) collected data on 140 college students who had experienced the death of a parent, 182 students from two-parent families, and 59 students from divorced families. The data are found in the file Mireault.dat and are described in Appendix: Computer Exercises.

a. Use any statistical program to run t tests to compare the first two groups on the Depression, Anxiety, and Global Symptom Index t scores from the Brief Symptom Inventory (Derogatis, 1983).

b. Are these three t tests independent of one another? (*Hint:* To do this problem you will have to ignore or delete those cases in Group 3 [the Divorced group]. Your instructor or the appropriate manual will explain how to do this for the particular software that you are using.)

7.47 It is commonly reported that women show more symptoms of anxiety and depression than men. Would the data from Mireault's study support this hypothesis?

7.48 Now run separate t tests to compare Mireault's Group 1 versus Group 2, Group 1 versus Group 3, and Group 2 versus Group 3 on the Global Symptom Index. (This is not a good way to compare the three group means, but it is being done here because it leads to more appropriate analyses in Chapter 12.)

7.49 Present meaningful effect sizes estimate(s) for the matched pairs data in Exercise 7.25.

7.50 Present meaningful effect sizes estimate(s) for the two independent group data in Exercise 7.31.

Discussion Questions

7.51 In Chapter 6 (Exercise 6.38), we examined data presented by Hout et al. on the sexual satisfaction of married couples. We did that by setting up a contingency table and computing χ^2 on that table. We looked at those data again in a different way in Exercise 7.19, where we ran a t test comparing the means. Instead of asking subjects to rate their statement "Sex is fun for me and my partner" as "Never, Fairly Often, Very Often, or Almost Always," we converted their categorical responses to a four-point scale from 1 = "Never" to 4 = "Almost Always."

a. How does the "scale of measurement" issue relate to this analysis?

b. Even setting aside the fact that this exercise and Exercise 6.37 use different statistical tests, the two exercises are asking quite different questions of the data. What are those different questions?

c. What might you do if 15 wives refused to answer the question, although their husbands did, and 8 husbands refused to answer the question when their wives did?

d. How comfortable are you with the t test analysis, and what might you do instead?

7.52 Write a short paragraph containing the information necessary to describe the results of the experiment discussed in Exercise 7.31. This should be an abbreviated version of what you would write in a research article.