Chapter

7

Comparison of Two Independent Samples

Objectives

In this chapter we continue our study of comparisons of two independent samples by introducing hypothesis testing. We will

- explore how randomization can be used to form the basis of a statistical inference.
- demonstrate how to a conduct a two-sample *t* test to compare sample means and explain how this test relates to the confidence interval for the difference of two means.
- discuss the interpretation of *P*-values.
- take a closer look at how confounding and spurious association can limit the utility of a study.
- compare causal versus associative inferences and their relationships to experiments and observational studies.
- discuss the concepts of significance level, effect size, Type I and II errors, and power.
- distinguish between directional and nondirectional tests and examine how the *P*-values of these tests compare.
- consider the conditions under which the use of a *t* test is valid.
- show how to compare distributions using the Wilcoxon-Mann-Whitney test.

7.1 Hypothesis Testing: The Randomization Test

Consider taking a sample from a population and then randomly dividing the sample into two parts. We would expect the two parts of the sample to look similar, but not exactly alike. Now suppose that we have samples from two populations. If the two samples look quite similar to each other, we might infer that the two populations are identical; if the samples look quite different, we would infer that the populations differ. The question is, "How different do two samples have to be in order for us to infer that the populations that generated them are actually different?"

One way to approach this question is to compare the two sample means and to see how much they differ in comparison to the amount of difference we would expect to see due to chance.* The randomization test gives us a way to measure the variability in the difference of two sample means.

Example 7.1.1

Flexibility A researcher studied the flexibility of each of seven women, four of whom were in an aerobics class and three of whom were dancers. One measure she recorded was the "trunk flexion"—how far forward each of the women could

*One could compare the two sample medians rather than the means. We compare means so that we have a process similar to the *t* test, which is introduced in the next section and is based on means.

Table 7.1.1

Aerobics	Dance
38	48
45	59
58	61
64	
mean 51.25	56.00

stretch while seated on the floor.* The measures (in centimeters) are shown in Table 7.1.1.[1]

Do the data provide evidence that the flexibility is associated with being a dancer?

If being a dancer has no effect on flexibility, then one could argue that the seven data points in the study came from a common population: Some women have greater trunk flexion than others, but this has nothing to do with being a dancer.

Another way of saying this is

> **Claim:** The seven trunk flexion measures came from a single population; the labels "aerobics" and "dance" are arbitrary and have nothing to do with flexibility (as measured by trunk flexion). ■

If the claim stated in Example 7.1.1 is true, then any rearrangement of the seven observations into two groups, with four "aerobics" and three "dance" women, is as likely as any other rearrangement. Indeed, we could imagine writing the seven observations onto seven cards, shuffling the cards, and then drawing four of them to be the observations for the "aerobics" group, with the other three being the observations for the "dance" group.

Example 7.1.2

Flexibility There are 35 possible ways to divide the trunk flexion measures of the seven observations into two groups, of sizes 4 and 3. Table 7.1.2 lists each of the 35 possibilities, along with the difference in sample means for each. (We report the means to three decimal places, since we will be using these values in future calculations.) The two samples obtained in the study are listed first, followed by the other 34 ways that the samples might have turned out.

Table 7.1.2

Sample 1 ("aerobics")	Sample 2 ("dance")	Mean of sample 1	Mean of sample 2	Difference in means
38 45 58 64	48 59 61	51.25	56.00	**−4.75**
38 45 58 48	64 59 61	47.25	61.33	**−14.08**
38 45 58 59	64 48 61	50.00	57.67	**−7.67**
38 45 58 61	64 48 59	50.50	57.00	**−6.50**
38 45 64 48	58 59 61	48.75	59.33	**−10.58**
38 45 64 59	58 48 61	51.50	55.67	**−4.17**
38 45 64 61	58 48 59	52.00	55.00	**−3.00**
38 45 48 59	58 64 61	47.50	61.00	**−13.50**
38 45 48 61	58 64 59	48.00	60.33	**−12.33**
38 45 59 61	58 64 48	50.75	56.67	**−5.92**
38 58 64 48	45 59 61	52.00	55.00	−3.00
38 58 64 59	45 48 61	54.75	51.33	3.42
38 58 64 61	45 48 59	55.25	50.67	4.58
38 58 48 59	45 64 61	50.75	56.67	**−5.92**
38 58 48 61	45 64 59	51.25	56.00	**−4.75**

(*Continues on next page*)

*These data are part of a larger study—we are working with a subset of the full study in order to simplify matters.

Table 7.1.2 (Continued)

Sample 1 ("aerobics")	Sample 2 ("dance")	Mean of sample 1	Mean of sample 2	Difference in means
38 58 59 61	45 64 48	54.00	52.33	1.67
38 64 48 59	45 58 61	52.25	54.67	−2.42
38 64 48 61	45 58 59	52.75	54.00	−1.25
38 64 59 61	45 58 48	55.50	50.33	**5.17**
38 48 59 61	45 58 64	51.50	55.67	−4.17
45 58 64 48	38 59 61	53.75	52.67	1.08
45 58 64 59	38 48 61	56.50	49.00	**7.50**
45 58 64 61	38 48 59	57.00	48.33	**8.67**
45 58 48 59	38 64 61	52.50	54.33	−1.83
45 58 48 61	38 64 59	53.00	53.67	−0.67
45 58 59 61	38 64 48	55.75	50.00	**5.75**
45 64 48 59	38 58 61	54.00	52.33	1.67
45 64 48 61	38 58 59	54.50	51.67	2.83
45 64 59 61	38 58 48	57.25	48.00	**9.25**
45 48 59 61	38 58 64	53.25	53.33	−0.08
58 64 48 59	38 45 61	57.25	48.00	**9.25**
58 64 48 61	38 45 59	57.75	47.33	**10.42**
58 64 59 61	38 45 48	60.50	43.67	**16.83**
58 48 59 61	38 45 64	56.50	49.00	**7.50**
64 48 59 61	38 45 58	58.00	47.00	**11.00**

Figure 7.1.1 gives a visual display of these 35 possible values. The observed result of −4.75, which is highlighted, falls not far from the middle of the distribution.

Suppose that the labels "aerobics" and "dance" are, in fact, arbitrary and have nothing to do with trunk flexion. Then each of the 35 outcomes listed in Table 7.1.2, and shown in Figure 7.1.1, is equally likely. This means that the differences, shown in the last column of the table, are equally likely. Of the 35 differences, 20 of them are at least as large in magnitude as the –4.75 obtained in the study; these are shown in bold type in the table and filled in black or gray in the figure. Thus, if the claim is true (that the labels "aerobics" and "dance" are arbitrary), there is a 20/35 chance of obtaining a difference in sample means as large, in magnitude, as the difference that was observed.

The fraction 20/35 is approximately equal to 0.57, which is rather large. Thus, the observed data are consistent with the claim that the labels "aerobics" and "dance" are arbitrary and have nothing to do with flexibility. If the claim is true, we would expect to see a difference in sample means of 4.75 or more over half of the time, just due to chance alone. Therefore, this data provides little evidence that flexibility is associated with dancing. ■

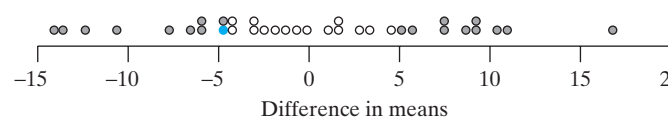


Figure 7.1.1 Distribution of "Difference in means" values, with the observed result of −4.75 colored blue, and values with observed results as or more extreme (in magnitude) than 4.75 colored gray

The process shown in Example 7.1.2 is called the **randomization test**.* In a randomization test one randomly divides the observed data into groups in order to see how likely it is that the observed difference is to arise due to chance alone.

Note: In Section 7.2 we will introduce a procedure known as the t test, which often provides a good approximation to the randomization test. The value of 20/35 (0.57) computed in Example 7.1.2 is called a P-value. (We have seen this term used earlier for the decision making in the context of the Shapiro–Wilk test for normality in Section 4.4. The general use of this term, and others, will be explained more fully in Section 7.2.) For the data in Example 7.1.1 the t test yields a P-value of 0.54. We can think of the 0.54 P-value from the t test as an approximation to the 0.57 P-value found with the randomization test.

Larger Samples

When we are dealing with small samples, such as in Example 7.1.1, we can list all of the possible outcomes from randomly assigning observations to groups. The following example shows how to handle large samples, where no such listing is possible.

Example 7.1.3

Leaf Area A plant physiologist investigated the effect of mechanical stress on the growth of soybean plants. Individually potted seedlings were divided into two groups. Those in the first group were stressed by shaking for 20 minutes twice daily, while those in the second group (the control group) were not shaken. After 16 days of growth the plants were harvested and total leaf area (cm^2) was measured for each plant. The data are given in Table 7.1.3 and are graphed in Figure 7.1.2.[2]

Table 7.1.3

Control	Stressed
314	283
320	312
310	291
340	259
299	216
268	201
345	267
271	326
285	241
mean 305.8	266.2

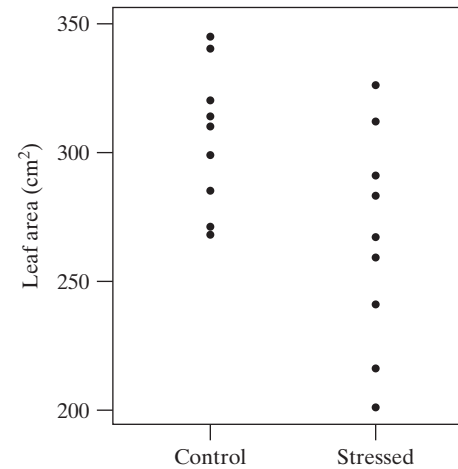


Figure 7.1.2 Parallel dotplots of leaf areas

The mean for the stressed plants is lower than for the control plants and Figure 7.1.2 provides some visual evidence of a difference between the two groups. On the other hand, the dotplots overlap quite a bit. Perhaps stressing the seedlings by shaking them has no actual effect on leaf area and the difference observed in this experiment (305.8 − 266.2 = 39.6) was simply due to chance. That is, it might be

*Many people would call this a permutation test, since it involves listing all possible permutations of the data.

that the "control" and "stressed" conditions have nothing to do with leaf area. If this is the case, then we can think of the 18 seedlings as having come from one population, with the division into "control" and "stressed" groups being arbitrary.

In Example 7.1.2 we could list all of the possible ways that the two groups could have been formed. However, in the current example there are 48,620 possible ways to select 9 of the 18 seedlings as the control group (and the other 9 as the stressed group). Thus, it is not feasible to create a table similar to Table 7.1.2 and list all the possibilities. What we can do, however, is to randomly sample from the 48,620 possibilities. One way to do this would be to (1) write the 18 observations on each of 18 cards; (2) shuffle the cards; (3) randomly deal out 9 of them as the control group, with the other 9 being the stress group; (4) calculate the difference in sample means; (5) record whether the magnitude of the difference in sample means is at least 39.6; (6) repeat steps (1)—(5) many times.

Consider the fraction of times that the magnitude of the difference in sample means is at least as large as the value of 39.6 obtained in the experiment. This is a measure of the evidence against the claim that "Stressing the seedlings by shaking them has no actual effect on leaf area."

Rather than use 18 cards, we could use a computer simulation to accomplish the same thing. In one simulation with 1,000 trials there were only 36 trials that gave a difference in sample means as large in magnitude as 39.6.* This indicates that the observed difference of 39.6 is unlikely to arise by chance—the chance is only 3.6%—so we have evidence that stressing the plants has an effect. Indeed, it appears that shaking the seedlings led to a reduction in average leaf area. ■

Note: The t test procedure (to be introduced in Section 7.2) yields a P-value of 0.033, which is a good approximation to the 0.036 P-value given by the randomization test.

Exercises 7.1.1–7.1.3

7.1.1 Suppose we have samples of five men and of five women and have conducted a randomization test to compare the sexes on the variable Y = pulse. Further, suppose we have found that in 120 out of the 252 possible outcomes under randomization the difference in means is at least as large as the difference in the two observed sample means. Does the randomization test provide evidence that the sexes differ with regard to pulse? Justify your answer using the randomization results.

7.1.2 In an investigation of the possible influence of dietary chromium on diabetic symptoms, some rats were fed a low-chromium diet and others were fed a normal diet. One response variable was activity of the liver enzyme GITH, which was measured using a radioactively labeled molecule. The accompanying table shows the results, expressed as thousands of counts per minute per gram of liver.[3] The sample means are 49.17 for the low-chromium diet and 51.90 for the normal diet; thus the difference in sample means is −2.73. There are 10 possible randomizations of the five observations into two groups, of sizes three and two.

(a) Create a list of these 10 randomizations (one of which is the original assignment of observations to the two groups) and for each case calculate the low-chromium diet mean minus the normal diet mean.

(b) How many of the 10 randomizations yield a difference in sample means as far from zero as −2.73, the difference in sample means for our observed samples?

*In this instance, we could also use a computer to consider the difference in means for each of the 48,620 possibilities and note how many of these yield differences larger than 39.6 in magnitude. However, as samples grow larger, listing all possibilities can be computationally expensive (even with fast computers) and only marginally more accurate than conducting simulations as we have described.

(c) Is there evidence that dietary chromium affects GITH liver enzyme activity? Justify your answer using the randomization results.

LOW-CHROMIUM DIET	NORMAL DIET
42.3	53.1
51.5	50.7
53.7	

7.1.3 The following table shows the number of bacteria colonies present in each of several petri dishes, after *E. coli* bacteria were added to the dishes and they were incubated for 24 hours. The "soap" dishes contained a solution prepared from ordinary soap; the "control" dishes contained a solution of sterile water. (These data are a subset of the larger data set seen in Exercise 6.6.9.) The sample means are 44 for the control group and 39.7 for the soap group; thus the difference in sample means is 4.3, with the control mean being larger, as would be expected if the soap were effective. There are 20 possible randomizations of the six observations into two groups, each of size three.

(a) Create a list of these 20 randomizations (one of which is the original assignment of observations to the two groups) and for each case calculate the control mean minus the soap mean.

(b) How many of the 20 randomizations produce a difference in means at least as large as 4.3?

(c) Is there evidence that the soap inhibits *E. coli* growth? Justify your answer using the randomization results.

CONTROL	SOAP
30	76
36	27
66	16

7.2 Hypothesis Testing: The *t* Test

In Chapter 6 we saw that two means can be compared by using a confidence interval for the difference $(\mu_1 - \mu_2)$. Now we will explore another approach to the comparison of means: the procedure known as *hypothesis testing*. The general idea is to formulate as a hypothesis the statement that μ_1 and μ_2 differ and then to see whether the data provide sufficient evidence in support of that hypothesis.

The Null and Alternative Hypotheses

The hypothesis that μ_1 and μ_2 are *not* equal is called an **alternative hypothesis** (or a research hypothesis) and is abbreviated H_A. It can be written as

$$H_A: \mu_1 \neq \mu_2$$

Its antithesis is the **null hypothesis**,

$$H_0: \mu_1 = \mu_2$$

which asserts that μ_1 and μ_2 are equal. A researcher would usually express these hypotheses more informally, as in the following example.

Example 7.2.1

Toluene and the Brain Abuse of substances containing toluene (for example, glue) can produce various neurological symptoms. In an investigation of the mechanism of these toxic effects, researchers measured the concentrations of various chemicals in the brains of rats that had been exposed to a toluene-laden atmosphere, and also in unexposed control rats. The concentrations of the brain chemical norepinephrine (NE) in the medulla region of the brain, for six toluene-exposed rats and five control rats, are given in Table 7.2.1 and displayed in Figure 7.2.1.[4]

The observed mean NE in the toluene group ($\bar{y}_1 = 540.8$ ng/gm) is substantially higher than the mean in the control group ($\bar{y}_2 = 444.2$ ng/gm). One might ask whether this observed difference indicates a real biological phenomenon—the effect of toluene—or whether the truth might be that toluene has no effect and that

Table 7.2.1 NE concentration (ng/gm)

	Toluene (Group 1)	Control (Group 2)
	543	535
	523	385
	431	502
	635	412
	564	387
	549	
n	6	5
$\bar{y}$	540.8	444.2
s	66.1	69.6
SE	27	31

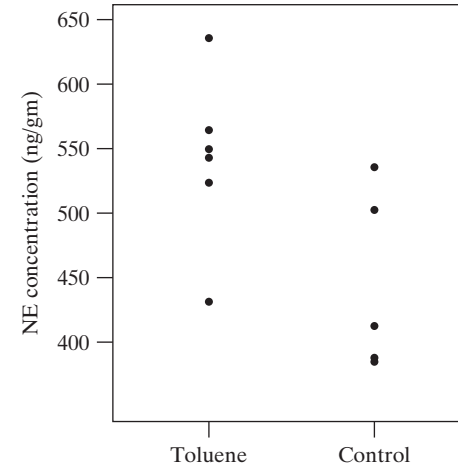


Figure 7.2.1 Parallel dotplots of NE concentration

the observed difference between $\bar{y}_1$ and $\bar{y}_2$ reflects only chance variation. Corresponding hypotheses, informally stated, would be

H_0^*: Toluene has no effect on NE concentration in rat medulla.

H_A^*: Toluene has some effect on NE concentration in rat medulla. ■

We denote the informal statements by different symbols (H_0^* and H_A^* rather than H_0 and H_A) because they make different assertions. In Example 7.2.1 the informal alternative hypothesis makes a very strong claim—not only that there is a difference, but that the difference is *caused* by toluene.*

A statistical **test of hypothesis** is a procedure for assessing the strength of evidence present in the data in support of H_A. The data are considered to demonstrate evidence for H_A if any discrepancies from H_0 (the opposite of H_A) could not be readily attributed to chance (that is, to sampling error).

The *t* Statistic

We consider the problem of testing the null hypothesis

$$H_0: \mu_1 = \mu_2$$

against the alternative hypothesis

$$H_A: \mu_1 \neq \mu_2$$

Note that the null hypothesis says that the two population means are equal, which is the same as saying that the difference between them is zero:

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

The alternative hypothesis asserts that the difference is not zero:

$$H_A: \mu_1 \neq \mu_2 \leftrightarrow H_A: \mu_1 - \mu_2 \neq 0$$

The ***t* test** is a standard method of choosing between these two hypotheses. To carry out the t test, the first step is to compute the **test statistic**, which for a t test is defined as

$$t_s = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}}$$

*Of course, our statements of H_0^* and H_A^* are abbreviated. Complete statements would include all relevant conditions of the experiment—adult male rats, toluene 1,000 ppm atmosphere for 8 hours, and so on. Our use of abbreviated statements should not cause any confusion.

Note that we subtract zero from $\bar{y}_1 - \bar{y}_2$ because H_0 states that $\mu_1 - \mu_2$ equals zero; writing "$(\bar{y}_1 - \bar{y}_2) - 0$" reminds us of what we are testing. The subscript "s" on t_s serves as a reminder that this value is calculated from the data ("s" for "sample"). The quantity t_s is the test statistic for the t test; that is, t_s provides the data summary that is the basis for the test procedure. Notice the structure of t_s: It is a measure of how far the difference between the sample means ($\bar{y}$'s) is from the difference we would expect to see if H_0 were true (zero difference), expressed in relation to the SE of the difference—the amount of variation we expect to see in differences of means from random samples. We illustrate with an example.

Example 7.2.2

Toluene and the Brain For the brain NE data of Example 7.2.1, the SE for $(\bar{Y}_1 - \bar{Y}_2)$ is

$$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\frac{66.1^2}{6} + \frac{69.6^2}{5}} = 41.195$$

and the value of t_s is

$$t_s = \frac{(540.8 - 444.2) - 0}{41.195} = 2.34$$

The t statistic shows that the difference between $\bar{y}_1$ and $\bar{y}_2$ is about 2.3 SEs from zero, the difference we'd expect to see if toluene had no effect on NE. ■

How shall we judge whether our data are sufficient evidence for H_A? A complete lack of evidence (*perfect* agreement with H_0) would be expressed by sample means that were identical and a resulting t statistic equal to zero ($t_s = 0$). But, even if the null hypothesis H_0 were true, we would not expect t_s to be exactly zero; we expect the sample means to differ from one another because of sampling variability (measured via $\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}$). Fortunately, we know what to expect regarding this sampling variability; in fact, the chance difference in the $\bar{Y}$'s is not likely to exceed a couple of standard errors when the null hypothesis is true. To put this more precisely, it can be shown mathematically that

> If H_0 is true, then the sampling distribution of t_s is well approximated by a Student's t distribution with degrees of freedom given by formula (6.7.1).*

The preceding statement is true if certain conditions are met. Briefly: We require independent random samples from normally distributed populations. These conditions will be considered in detail in Section 7.9.

The essence of the t test procedure is to identify where the observed value t_s falls in the Student's t distribution, as indicated in Figure 7.2.2. If t_s is near the center, as in Figure 7.2.2(a), then the data are regarded as compatible with H_0 because the observed difference between $(\bar{Y}_1 - \bar{Y}_2)$ and the null difference of zero can readily be attributed to chance variation caused by sampling error. (H_0 predicts that the sample means will be equal, since H_0 says that the population means are equal.)

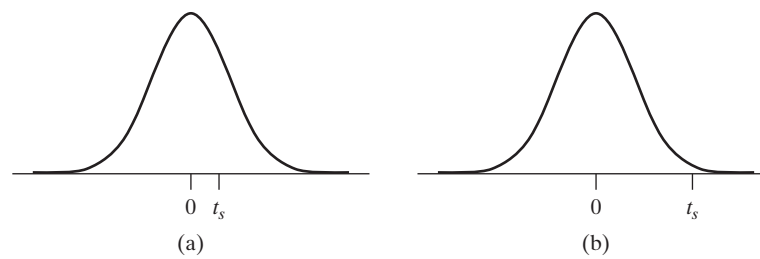


Figure 7.2.2 Essence of the t test. (a) Data compatible with H_0 (and thus a lack of significant evidence for H_A); (b) data incompatible with H_0 (and thus significant evidence for H_A).

*As we stated in Section 6.8, a conservative approximation to formula (6.7.1) is to use degrees of freedom given by the smaller of $n_1 - 1$ and $n_2 - 1$.

If, on the other hand, t_s falls in the far tail of the t distribution, as in Figure 7.2.2(b), then the data are regarded as evidence for H_A, because the observed deviation cannot be readily explained as being due to chance variation. To put this another way, if H_0 is true, then it is unlikely that t_s would fall in the far tails of the t distribution.

The *P*-Value

To judge whether an observed value t_s is "far" in the tail of the t distribution, we need a quantitative yardstick for locating t_s within the distribution. This yardstick is provided by the P-value, which can be defined (in the present context) as follows:

> The ***P*-value** of the test is the area under Student's t curve in the double tails beyond $-t_s$ and $+t_s$.

Thus, the P-value, which is sometimes abbreviated as simply "P," is the shaded area in Figure 7.2.3. Note that we have defined the P-value as the total area in *both* tails; this is sometimes called the "two-tailed" P-value.

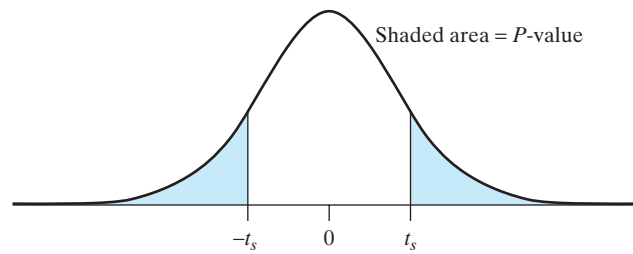


Figure 7.2.3 The two-tailed P-value for the t test

Example 7.2.3

Toluene and the Brain For the brain NE data of Example 7.2.1, the value of t_s is 2.34. We can ask, "If H_0 were true so that one would expect $\overline{Y}_1 - \overline{Y}_2 = 0$, on average, what is the probability that $\overline{Y}_1 - \overline{Y}_2$ would differ from zero by as many as 2.34 SEs?" The P-value answers this question. Formula (6.7.1) yields 8.47 degrees of freedom for these data. Thus, the P-value is the area under the t curve (with 8.47 degrees of freedom) beyond ± 2.34. This area, which was found using a computer, is shown in Figure 7.2.4 to be 0.0454. ■

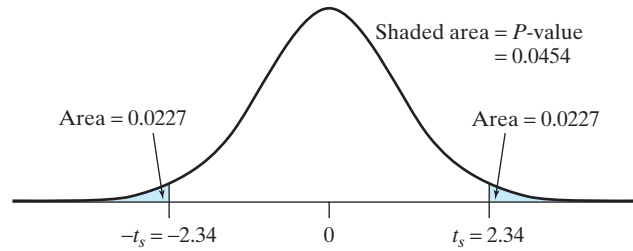


Figure 7.2.4 The two-tailed P-value for the toluene data

Definition The ***P*-value** for a hypothesis test is the probability, computed under the condition that the null hypothesis is true, of the test statistic being at least as extreme as the value of the test statistic that was actually obtained.

From the definition of P-value, it follows that the ***P*-value is a measure of compatibility between the data and H_0** and thus measures the **evidence for H_A**: A large P-value (close to 1) indicates a value of t_s near the center of the t distribution (lack of evidence for H_A), whereas a small P-value (close to 0) indicates a value of t_s in the far tails of the t distribution (evidence for H_A).

Drawing Conclusions from a *t* Test

The P-value is a measure of the evidence in the data for H_A, but where does one draw the line in determining how much evidence is sufficient? Most people would agree that P-value $= 0.0001$ indicates very strong evidence, and that P-value $= 0.80$ indicates a lack of evidence, but what about intermediate values? For example, should P-value $= 0.10$ be regarded as sufficient evidence for H_A? The answer is not intuitively obvious.

In much scientific research, it is not necessary to draw a sharp line. However, in many situations a *decision* must be reached. For example, the Food and Drug Administration (FDA) must decide whether the data submitted by a pharmaceutical manufacturer are sufficient to justify approval of a medication. As another example, a fertilizer manufacturer must decide whether the evidence favoring a new fertilizer is sufficient to justify the expense of further research.

Making a decision requires drawing a definite line between sufficient and insufficient evidence. The threshold value, on the P-value scale, is called the **significance level** of the test and is denoted by the Greek letter α (alpha). The value of α is chosen by whoever is making the decision. Common choices are $\alpha = 0.10, 0.05$, and 0.01. *If the P-value of the data is less than or equal to α, the data are judged to provide statistically significant evidence in favor of* H_A; *we also may say that* H_0 *is* ***rejected***. If the P-value of the data is greater than α, we say that the data provide insufficient evidence to claim that H_A is true, and thus H_0 is **not rejected**.

The following example illustrates the use of the t test to make a decision.

Example 7.2.4

Toluene and the Brain For the brain NE experiment of Example 7.2.1, the data are summarized in Table 7.2.2. Suppose we choose to make a decision at the 5% significance level, $\alpha = 0.05$. In Example 7.2.3 we found that the P-value of these data is 0.0454. This means that one of two things happened: Either (1) H_0 is true and we got a strange set of data just by chance or (2) H_0 is false. If H_0 is true, the kind of discrepancy we observed between $\bar{y}_1$ and $\bar{y}_2$ would happen only about 4.5% of the time. Because the P-value, 0.0454, is less than 0.05, we reject H_0 and conclude that the data provide statistically significant evidence in favor of H_A. The strength of the evidence is expressed by the statement that the P-value is 0.0454.

Table 7.2.2 NE concentration (ng/gm)

	Toluene	Control
n	6	5
$\bar{y}$	540.8	444.2
s	66.1	69.6

Conclusion: The data provide sufficient evidence at the 0.05 level of significance (P-value $= 0.0454$) that toluene increases NE concentration.* ■

The next example illustrates a t test in which there is a lack of sufficient evidence at the 0.05 level of significance for H_A.

Example 7.2.5

Fast Plants In Example 6.7.1 we saw that the mean height of fast plants was smaller when ancy was used than when water (the control) was used. Table 7.2.3 summarizes

*Because the alternative hypothesis was H_A: $\mu_1 \neq \mu_2$, some authors would say, "We conclude that toluene affects NE concentration," rather than saying that toluene increases NE concentration.

Table 7.2.3 Fourteen-day height of control and of ancy plants

	Control	Ancy
n	8	7
$\bar{y}$	15.9	11.0
s	4.8	4.7

the data. The difference between the sample means is $15.9 - 11.0 = 4.9$. The SE for the difference is

$$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\frac{4.8^2}{8} + \frac{4.7^2}{7}} = 2.46$$

Suppose we choose to use $\alpha = 0.05$ in testing

$$H_0: \mu_1 = \mu_2 \,(\text{i.e., } \mu_1 - \mu_2 = 0)$$

against the alternative hypothesis

$$H_A: \mu_1 \neq \mu_2 \,(\text{i.e., } \mu_1 - \mu_2 \neq 0)$$

The value of the test statistic is

$$t_s = \frac{(15.9 - 11.0) - 0}{2.46} = 1.99$$

Formula (6.7.1) gives 12.8 degrees of freedom for the t distribution. The P-value for the test is the probability of getting a t statistic that is at least as far away from zero as 1.99. Figure 7.2.5 shows that this probability is 0.0678. (This 4-digit P-value was found using a computer.) Because the P-value is greater than α, we have insufficient evidence for H_A; thus, we do not reject H_0. That is, these data do not provide sufficient evidence to conclude that μ_1 and μ_2 differ; the difference we observed between $\bar{y}_1$ and $\bar{y}_2$ could easily have happened by chance.

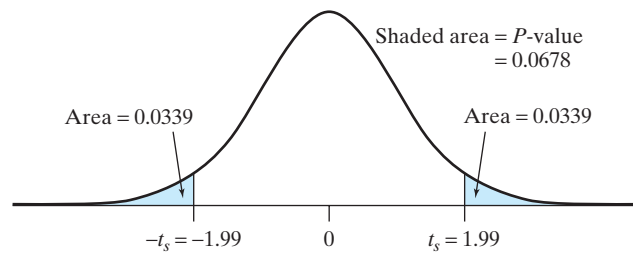


Figure 7.2.5 The two-sided P-value for the ancy data

Conclusion: The data do *not* provide sufficient evidence (P-value $= 0.0678$) at the 0.05 level of significance to conclude that ancy and water differ in their effects on fast plant growth (under the conditions of the experiment that was conducted). ■

Note carefully the phrasing of the conclusion in Example 7.2.5. We do *not* say that there is evidence *for* the null hypothesis, but only that there is insufficient evidence *against* it. When we do not reject H_0, this indicates a lack of evidence that H_0 is false, which is *not* the same thing as evidence that H_0 is true. The astronomer Carl Sagan (in another context) summed up this principle of evidence in this succinct statement:[5]

Absence of evidence is not evidence of absence.

In other words, nonrejection of H_0 is *not* the same as *acceptance* of H_0. (To avoid confusion, it may be best not to use the phrase "accept H_0" at all.)

Nonrejection of H_0 indicates that the data are compatible with H_0, but the data may *also* be quite compatible with H_A. For instance, in Example 7.2.5 we found that the observed difference between the sample means could be due to sampling variation, but this finding does not rule out the possibility that the observed difference is actually due to a real effect caused by ancy. (Methods for such ruling out of possible alternatives will be discussed in Section 7.7 and optional Section 7.8.)

In testing a hypothesis, the researcher starts out with the assumption that H_0 is true and then asks whether the data contradict that assumption. This logic can make sense even if the researcher regards the null hypothesis as implausible. For instance, in Example 7.2.5 it could be argued that there is almost certainly *some* difference (perhaps very small) between using ancy and not using ancy. The fact that we did not reject H_0 does not mean that we accept H_0.

Using Tables versus Using Technology

In analyzing data, how do we determine the *P*-value of a test? Statistical computer software, and some calculators, will provide exact *P*-values. If such technology is not available, then we can use formula (6.7.1) to find the degrees of freedom but round down to make the value an integer. A conservative alternative to using formula (6.7.1) is to use the smaller of $n_1 - 1$ and $n_2 - 1$ as the degrees of freedom for the test. A liberal approach is to use $n_1 + n_2 - 2$ as the degrees of freedom. (Formula (6.7.1) will always give degrees of freedom between the conservative value of the smaller of $n_1 - 1$ and $n_2 - 1$ and the liberal value of $n_1 + n_2 - 2$.) We can rely on the limited information in Table 4 to *bracket* the *P*-value, rather than to determine it exactly. The *P*-value found using the conservative approach will be somewhat larger than the exact *P*-value; the *P*-value found using the liberal approach will be somewhat smaller than the exact *P*-value. The following example illustrates the bracketing process.

Example 7.2.6

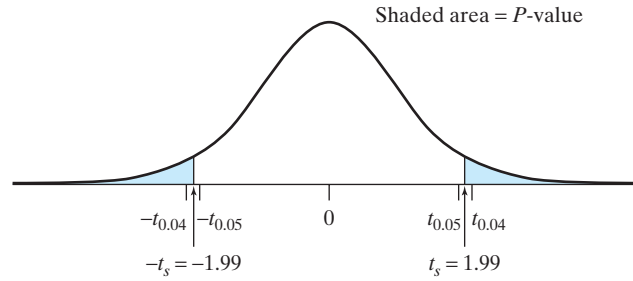
Fast Plants For the fast plant growth data, the value of the *t* statistic (as determined in Example 7.2.5) is $t_s = 1.99$. The smaller of $n_1 - 1$ and $n_2 - 1$ is $7 - 1 = 6$, so the conservative degrees of freedom are 6. The liberal degrees of freedom are $8 + 7 - 2 = 13$. Here is a copy of part of Table 4, with key numbers highlighted.

	Upper Tail Probability		
df	**.05**	**.04**	**.03**
6	**1.943**	**2.104**	2.313
7	1.895	2.046	2.241
8	1.860	2.004	2.189
9	1.833	1.973	2.150
10	1.812	1.948	2.120
11	1.796	1.928	2.096
12	1.782	1.912	2.076
13	1.771	**1.899**	**2.060**

We begin with the conservative degrees of freedom, 6. From the preceding table (or from Table 4) we find $t_{6,\,0.05} = 1.943$ and $t_{6,\,0.04} = 2.104$. The corresponding conservative *P*-value, based on a *t* distribution with 6 degrees of freedom, is shaded in

Figure 7.2.6. Because t_s is between the 0.04 and 0.05 critical values, the upper tail area must be between 0.04 and 0.05; thus, the conservative P-value must be between 0.08 and 0.10.

Figure 7.2.6 Conservative P-value for Example 7.2.6



The liberal degrees of freedom are $8 + 7 - 2 = 13$. From the preceding table (or from Table 4) we find $t_{13,\,0.04} = 1.899$ and $t_{13,\,0.03} = 2.060$. Because t_s is between these 0.03 and 0.04 critical values, the upper tail area must be between 0.06 and 0.08; thus, the liberal P-value must be between 0.06 and 0.08.

Putting these two together, we have

$$0.06 < P\text{-value} < 0.10$$ ■

If the observed t_s is not within the boundaries of Table 4, then the P-value is bracketed on only one side. For example, if t_s is greater than $t_{0.0005}$, then the two-sided P-value is bracketed as

$$P\text{-value} < 0.001$$

Reporting the Results of a *t* Test

In reporting the results of a t test, a researcher may choose to make a definite decision (to claim there is significant evidence for H_A or not significant evidence to support H_A) at a specified significance level α, or the researcher may choose simply to describe the results in phrases such as "There is very strong evidence that . . ." or "The evidence suggests that . . ." or "There is virtually no evidence that . . .". In writing a report for publication, it is very desirable to state the P-value so that the reader can make a decision on his or her own.

The term *significant* is often used in reporting results. For instance, an observed difference is said to be "statistically significant at the 5% level" if it is large enough to justify significant evidence for H_A at $\alpha = 0.05$. In Example 7.2.4 we saw that the observed difference between the two sample means in the toluene data is statistically significant at the 5% level, since the P-value is 0.0454, which is less than 0.05. In contrast, the fast plant data of Example 7.2.5 do not show a statistically significant difference at the 5% level, since the P-value for the fast plant data is 0.0678. However, the difference in sample means in the fast plant data *is* statistically significant at the $\alpha = 0.10$ level, since the P-value is less than 0.10. When α is not specified, it is usually understood to be 0.05; we should emphasize, however, that *α is an arbitrarily chosen value and there is nothing "official" about 0.05*. Unfortunately, the term "significant" is easily misunderstood and should be used with care; we will return to this point in Section 7.7.

Note: In this section we have considered tests of the form H_0: $\mu_1 = \mu_2$ (i.e., $\mu_1 - \mu_2 = 0$) versus H_A: $\mu_1 \neq \mu_2$ (i.e., $\mu_1 - \mu_2 \neq 0$); this is the most common pair of hypotheses. However, it may be that we wish to test that μ_1 is greater than μ_2

by some specific, nonzero amount, say c. To test H_0: $\mu_1 - \mu_2 = c$ versus H_A: $\mu_1 - \mu_2 \neq c$ we use the t test with test statistic given by

$$t_s = \frac{(\bar{y}_1 - \bar{y}_2) - c}{\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}}$$

From this point on, the test proceeds as before (i.e., as for the case when $c = 0$).

Exercises 7.2.1–7.2.17

[*Note*: Answers to hypothesis testing questions should include a statement of the conclusion in the context of the setting. (See Examples 7.2.4 and 7.2.5.)]

7.2.1 For each of the following data sets, use Table 4 to bracket the two-tailed P-value of the data as analyzed by the t test.

(a)

	SAMPLE 1	SAMPLE 2
n	4	3
$\bar{y}$	735	854

$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = 38$ with df $= 4$

(b)

	SAMPLE 1	SAMPLE 2
n	7	7
$\bar{y}$	5.3	5.0

$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = 0.24$ with df $= 12$

(c)

	SAMPLE 1	SAMPLE 2
n	15	20
$\bar{y}$	36	30

$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = 1.3$ with df $= 30$

7.2.2 For each of the following data sets, use Table 4 to bracket the two-tailed P-value of the data as analyzed by the t test.

(a)

	SAMPLE 1	SAMPLE 2
n	8	5
$\bar{y}$	100.2	106.8

$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = 5.7$ with df $= 10$

(b)

	SAMPLE 1	SAMPLE 2
n	8	8
$\bar{y}$	49.8	44.3

$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = 1.9$ with df $= 13$

(c)

	SAMPLE 1	SAMPLE 2
n	10	15
$\bar{y}$	3.58	3.00

$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = 0.12$ with df $= 19$

7.2.3 For each of the following situations, suppose H_0: $\mu_1 = \mu_2$ is being tested against H_A: $\mu_1 \neq \mu_2$. State whether or not there is significant evidence for H_A.

(a) P-value $= 0.085$, $\alpha = 0.10$.

(b) P-value $= 0.065$, $\alpha = 0.05$.

(c) $t_s = 3.75$ with 19 degrees of freedom, $\alpha = 0.01$.

(d) $t_s = 1.85$ with 12 degrees of freedom, $\alpha = 0.05$.

7.2.4 For each of the following situations, suppose H_0: $\mu_1 = \mu_2$ is being tested against H_A: $\mu_1 \neq \mu_2$. State whether or not there is significant evidence for H_A.

(a) P-value $= 0.046$, $\alpha = 0.02$.

(b) P-value $= 0.033$, $\alpha = 0.05$.

(c) $t_s = 2.26$ with 5 degrees of freedom, $\alpha = 0.10$.

(d) $t_s = 1.94$ with 16 degrees of freedom, $\alpha = 0.05$.

7.2.5 In a study of the nutritional requirements of cattle, researchers measured the weight gains of cows during a 78-day period. For two breeds of cows, Hereford (HH) and Brown Swiss/Hereford (SH), the results are summarized in the following table.[6] [*Note*: Formula (6.7.1) yields 71.9 df.]

	HH	SH
n	33	51
$\bar{y}$	18.3	13.9
s	17.8	19.1

Use a t test to compare the means. Use $\alpha = 0.10$.

7.2.6 Backfat thickness is a variable used in evaluating the meat quality of pigs. An animal scientist measured backfat thickness (cm) in pigs raised on two different diets, with the results given in the table.[7]

	DIET 1	DIET 2
$\bar{y}$	3.49	3.05
s	0.40	0.40

Consider using the t test to compare the diets. Bracket the P-value, assuming that the number of pigs on each diet was

(a) 5

(b) 10

(c) 15

Use $n_1 + n_2 - 2$ as the approximate degrees of freedom.

7.2.7 Heart disease patients often experience spasms of the coronary arteries. Because biological amines may play a role in these spasms, a research team measured amine levels in coronary arteries that were obtained postmortem from patients who had died of heart disease and also from a control group of patients who had died from other causes. The accompanying table summarizes the concentration of the amine serotonin.[8]

	SEROTONIN (NG/GM)	
	HEART DISEASE	CONTROLS
n	8	12
$\bar{y}$	3,840	5,310
SE	850	640

(a) For these data, the SE of $(\overline{Y}_1 - \overline{Y}_2)$ is 1,064 and df = 14.3 (which can be rounded to 14). Use a t test to compare the means at the 5% significance level.

(b) Verify the value of $SE_{(\overline{Y}_1 - \overline{Y}_2)}$ given in part (a).

7.2.8 In a study of the periodical cicada *(Magicicada septendecim),* researchers measured the hind tibia lengths of the shed skins of 110 individuals. Results for males and females are shown in the accompanying table.[9]

		TIBIA LENGTH (μm)	
GROUP	n	MEAN	SD
Males	60	78.42	2.87
Females	50	80.44	3.52

(a) Use a t test to investigate the association of tibia length on gender in this species. Use the 5% significance level. [*Note*: Formula (6.7.1) yields 94.3 df.]

(b) Given the preceding data, if you were told the tibia length of an individual of this species, could you make a fairly confident prediction of its sex? Why or why not?

(c) Repeat the t test of part (a), assuming that the means and standard deviations were as given in the table, but that they were based on only one-tenth as many individuals (6 males and 5 females). [*Note*: Formula (6.7.1) yields 7.8 df.]

7.2.9 Myocardial blood flow (MBF) was measured for two groups of subjects after five minutes of bicycle exercise. The normoxia ("normal oxygen") group was provided normal air to breathe whereas the hypoxia group was provided with a gas mixture with reduced oxygen, to simulate high altitude. The results (ml/min/g) are shown in the table.[10] [*Note*: Formula (6.7.1) yields 12.2 df.]

	NORMOXIA	HYPOXIA
	3.45	6.37
	3.09	5.69
	3.09	5.58
	2.65	5.27
	2.49	5.11
	2.33	4.88
	2.28	4.68
	2.24	3.50
	2.17	
	1.34	
n	10	8
$\bar{y}$	2.51	5.14
s	0.60	0.84

Use a t test to investigate the effect of hypoxia on MBF. Use $\alpha = 0.05$.

7.2.10 In a study of the development of the thymus gland, researchers weighed the glands of 10 chick embryos. Five of the embryos had been incubated 14 days and 5 had been incubated 15 days. The thymus weights were as shown in the table.[11] [*Note*: Formula (6.7.1) yields 7.7 df.]

	THYMUS WEIGHT (MG)	
	14 DAYS	15 DAYS
	29.6	32.7
	21.5	40.3
	28.0	23.7
	34.6	25.2
	44.9	24.2
n	5	5
$\bar{y}$	31.72	29.22
s	8.73	7.19

(a) Use a t test to compare the means at $\alpha = 0.10$.

(b) Note that the chicks that were incubated longer had a smaller mean thymus weight. Is this "backward" result surprising, or could it easily be attributed to chance? Explain.

7.2.11 As part of an experiment on root metabolism, a plant physiologist grew birch tree seedlings in the greenhouse. He flooded four seedlings with water for one day and kept four others as controls. He then harvested the seedlings and analyzed the roots for ATP content. The results (nmol ATP per mg tissue) are shown in the table.[12] [*Note*: Formula (6.7.1) yields 5.6 df.]

	FLOODED	CONTROL
	1.45	1.70
	1.19	2.04
	1.05	1.49
	1.07	1.91
n	4	4
$\bar{y}$	1.190	1.785
s	0.184	0.241

Use a t test to investigate the effect of flooding. Use $\alpha = 0.05$.

7.2.12 After surgery a patient's blood volume is often depleted. In one study, the total circulating volume of blood plasma was measured for each patient immediately after surgery. After infusion of a "plasma expander" into the bloodstream, the plasma volume was measured again and the increase in plasma volume (ml) was calculated. Two of the plasma expanders used were albumin (25 patients) and polygelatin (14 patients). The accompanying table reports the increase in plasma volume.[13] [*Note*: Formula (6.7.1) yields 33.6 df.]

Use a t test to compare the mean increase in plasma volume under the two treatments. Let $\alpha = 0.01$.

	ALBUMIN	POLYGELATIN
n	25	14
mean increase	490	240
SE	60	30

7.2.13 Nutritional researchers conducted an investigation of two high-fiber diets intended to reduce serum cholesterol level. Twenty men with high serum cholesterol were randomly allocated to receive an "oat" diet or a "bean" diet for 21 days. The table summarizes the fall (before minus after) in serum cholesterol levels.[14] Use a t test to compare the diets at the 5% significance level. [*Note*: Formula (6.7.1) yields 17.9 df.]

		FALL IN CHOLESTEROL (MG/DL)	
DIET	n	MEAN	SD
Oat	10	53.6	31.1
Bean	10	55.5	29.4

7.2.14 Suppose we have conducted a t test, with $\alpha = 0.05$, and the P-value is 0.03. For each of the following statements, say whether the statement is true or false and explain why.

(a) We reject H_0 with $\alpha = 0.05$.

(b) We have significant evidence for H_A with $\alpha = 0.05$.

(c) We would reject H_0 if α were 0.10.

(d) We do not have significant evidence for H_A with $\alpha = 0.10$.

(e) If H_0 is true, the probability of getting a test statistic at least as extreme as the value of the t_s that was actually obtained is 3%.

(f) There is a 3% probability that H_0 is true.

7.2.15 Suppose we have conducted a t test, with $\alpha = 0.10$, and the P-value is 0.07. For each of the following statements, say whether the statement is true or false and explain why.

(a) We reject H_0 with $\alpha = 0.10$.

(b) We have significant evidence for H_A with $\alpha = 0.10$.

(c) We would reject H_0 if α were 0.05.

(d) We do not have significant evidence for H_A with $\alpha = 0.05$.

(e) The probability that $\overline{Y}_1$ is greater than $\overline{Y}_2$ is 0.07.

7.2.16 The following table shows the number of bacteria colonies present in each of several petri dishes, after *E. coli* bacteria were added to the dishes and they were incubated for 24 hours. The "soap" dishes contained a solution prepared from ordinary soap; the "control" dishes contained a solution of sterile water. (These data were seen in Exercise 6.6.9.)

	CONTROL	SOAP
	30	76
	36	27
	66	16
	21	30
	63	26
	38	46
	35	6
	45	
n	8	7
$\bar{y}$	41.8	32.4
s	15.6	22.8
SE	5.5	8.6

Use a t test to investigate whether soap affects the number of bacteria colonies that form. Use $\alpha = 0.10$. [*Note*: Formula (6.7.1) yields 10.4 degrees of freedom for these data.]

7.2.17 Researchers studied the effect of a houseplant fertilizer on radish sprout growth. They randomly selected some radish seeds to serve as controls, while others were planted in aluminum planters to which fertilizer sticks were added. Other conditions were held constant between the two groups. The following table shows data on the heights of plants (in cm) two weeks after germination.[15]

Use a t test to investigate whether the fertilizer has an effect on average radish sprout growth. Use $\alpha = 0.05$. [*Note*: Formula (6.7.1) yields 53.5 degrees of freedom for these data.]

CONTROL		FERTILIZED	
3.4	1.6	2.8	1.9
4.4	2.9	1.9	2.7
3.5	2.3	3.6	2.3
2.9	2.8	1.2	1.8
2.7	2.5	2.4	2.7
2.6	2.3	2.2	2.6
3.7	1.6	3.6	1.3
2.7	1.6	1.2	3.0
2.3	3.0	0.9	1.4
2.0	2.3	1.5	1.2
1.8	3.2	2.4	2.6
2.3	2.0	1.7	1.8
2.4	2.6	1.4	1.7
2.5	2.4	1.8	1.5
n	28		28
$\bar{y}$	2.58		2.04
s	0.65		0.72

7.3 Further Discussion of the t Test

In this section we discuss more fully the method and interpretation of the t test.

Relationship between Test and Confidence Interval

There is a close connection between the confidence interval approach and the hypothesis testing approach to the comparison of μ_1 and μ_2. Consider, for example, a 95% confidence interval for $(\mu_1 - \mu_2)$ and its relationship to the t test at the 5% significance level. The t test and the confidence interval use the same three quantities—$(\overline{Y}_1 - \overline{Y}_2)$, $\mathrm{SE}_{(\overline{Y}_1-\overline{Y}_2)}$ and $t_{0.025}$—but manipulate them in different ways.

In the t test, when $\alpha = 0.05$, we have significant evidence for H_A (and so we reject H_0) if the P-value is less than or equal to 0.05. This happens if and only if the test statistic, t_s, is in the tail of the t distribution, at or beyond $\pm t_{0.025}$. If the magnitude of t_s (symbolized as $|t_s|$) is greater than or equal to $t_{0.025}$, then the P-value is less than or equal to 0.05 and we have significant evidence for H_A; if $|t_s|$ is less than $t_{0.025}$, then the P-value is greater than 0.05 and we do *not* have significant evidence for H_A. Figure 7.3.1 shows this relationship.

Thus, we lack significant evidence for H_A: $\mu_1 - \mu_2 \neq 0$ if and only if $|t_s| < t_{0.025}$. That is, we lack significant evidence for H_A when

$$\frac{|\bar{y}_1 - \bar{y}_2|}{\mathrm{SE}_{(\overline{Y}_1-\overline{Y}_2)}} < t_{0.025}$$

This is equivalent to

$$|\bar{y}_1 - \bar{y}_2| < t_{0.025}\,\mathrm{SE}_{(\overline{Y}_1-\overline{Y}_2)}$$

or

$$-t_{0.025}\,\mathrm{SE}_{(\overline{Y}_1-\overline{Y}_2)} < (\bar{y}_1 - \bar{y}_2) < t_{0.025}\,\mathrm{SE}_{(\overline{Y}_1-\overline{Y}_2)}$$

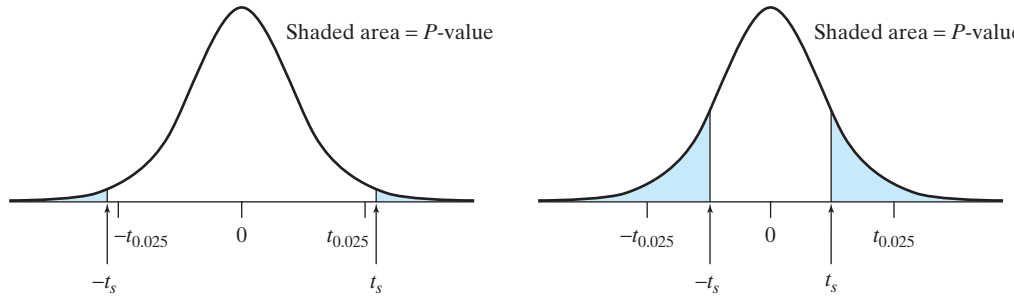


Figure 7.3.1 Possible outcomes of the *t* test at $\alpha = 0.05$. (a) If $|t_s| \geq t_{0.025}$ then P-value ≤ 0.05 and there is significant evidence for H_A (so H_0 is rejected). (b) If $|t_s| < t_{0.025}$, then P-value > 0.05 and there is a lack of significant evidence for H_A.

which is equivalent to

$$-(\bar{y}_1 - \bar{y}_2) - t_{0.025}\,\mathrm{SE}_{(\bar{Y}_1-\bar{Y}_2)} < 0 < -(\bar{y}_1 - \bar{y}_2) + t_{0.025}\,\mathrm{SE}_{(\bar{Y}_1-\bar{Y}_2)}$$

or

$$(\bar{y}_1 - \bar{y}_2) + t_{0.025}\,\mathrm{SE}_{(\bar{Y}_1-\bar{Y}_2)} > 0 > (\bar{y}_1 - \bar{y}_2) - t_{0.025}\,\mathrm{SE}_{(\bar{Y}_1-\bar{Y}_2)}$$

or

$$(\bar{y}_1 - \bar{y}_2) - t_{0.025}\,\mathrm{SE}_{(\bar{Y}_1-\bar{Y}_2)} < 0 < (\bar{y}_1 - \bar{y}_2) + t_{0.025}\,\mathrm{SE}_{(\bar{Y}_1-\bar{Y}_2)}$$

Thus, we have shown that we lack significant evidence for H_A: $\mu_1 - \mu_2 \neq 0$ if and only if the confidence interval for $(\mu_1 - \mu_2)$ includes zero. Conversely, if the 95% confidence interval for $(\mu_1 - \mu_2)$ does not cover zero, then we have significant evidence for H_A: $\mu_1 - \mu_2 \neq 0$ when $\alpha = 0.05$. (The same relationship holds between the 90% confidence interval and the test at $\alpha = 0.10$, and so on.) We illustrate with an example.

Example 7.3.1

Crawfish Lengths Biologists took samples of the crawfish species *Orconectes sanborii* from two rivers in central Ohio, the Upper Cuyahoga River (CUY) and East Fork of Pine Creek (EFP), and measured the length (mm) of each crawfish captured.[16] Table 7.3.1 shows the summary statistics; Figure 7.3.2 shows parallel boxplots of the data. The EFP sample distribution is shifted down from the CUY distribution; both distributions are reasonably symmetric.

Table 7.3.1 Crawfish data: length (mm)

	CUY	EFP
n	30	30
$\bar{y}$	22.91	21.97
s	3.78	2.90

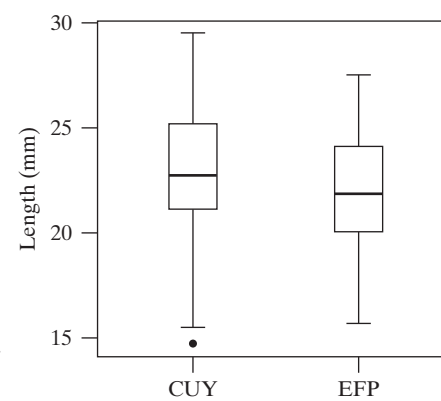


Figure 7.3.2 Boxplots of the crawfish data

For these data the two SEs are $3.78/\sqrt{30} = 0.69$ and $2.90/\sqrt{30} = 0.53$ for CUY and EFP, respectively. The degrees of freedom are

$$\mathrm{df} = \frac{(0.69^2 + 0.53^2)^2}{0.69^4/30 + 0.53^4/30} = 56.3$$

The quantities needed for a t test with $\alpha = 0.05$ are

$$\bar{y}_1 - \bar{y}_2 = 22.91 - 21.97 = 0.94$$

and

$$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{0.69^2 + 0.53^2} = 0.87$$

The test statistic is

$$t_s = \frac{(22.91 - 21.97) - 0}{0.87} = \frac{0.94}{0.87} = 1.08$$

The P-value for this test (found using a computer) is 0.2850, which is greater than 0.05, so we do not reject H_0. (A quick look at Table 4, using df = 50, shows that the P-value is between 0.20 and 0.40.)

If we construct a 95% confidence interval for $(\mu_1 - \mu_2)$ we get

$$0.94 \pm 2.006 \times 0.87$$

or $(-2.68, 0.81)$.*

The confidence interval includes zero, which is consistent with not having significant evidence for H_A: $\mu_1 - \mu_2 \neq 0$ in the t test. Note that this equivalence between the test and the confidence interval makes common sense; according to the confidence interval, μ_1 may be as much as 2.68 less, or as much as 0.81 more, than μ_2; it is natural, then, to say that we are uncertain as to whether μ_1 is greater than (or less than, or equal to) μ_2. ■

In the context of the Student's t method, the confidence interval approach and hypothesis testing approach are different ways of using the same basic information. The confidence interval has the advantage that it indicates the magnitude of the difference between μ_1 and μ_2. The testing approach has the advantage that the P-value describes on a continuous scale the strength of the evidence that μ_1 and μ_2 are really different. In Section 7.7 we will explore further the use of a confidence interval to supplement the interpretation of a t test. In later chapters we will encounter other hypothesis tests that cannot so readily be supplemented by a confidence interval.

Interpretation of α

In analyzing data or making a decision based on data, you will often need to choose a significance level α. How do you know whether to choose $\alpha = 0.05$ or $\alpha = 0.01$ or some other value? To make this judgment, it is helpful to have an *operational* interpretation of α. We now give such an interpretation.

Recall from Section 7.2 that the sampling distribution of t_s, if H_0 is true, is a Student's t distribution. Let us assume for definiteness that df = 60 and that α is chosen equal to 0.05. The critical value (from Table 4) is $t_{0.025} = 2.000$. Figure 7.3.3

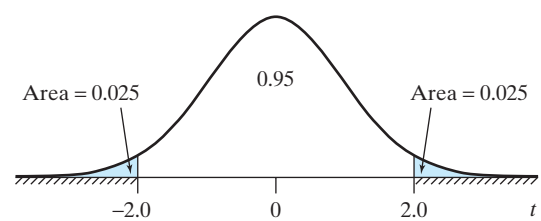


Figure 7.3.3 A t test at $\alpha = 0.05$. There is significant evidence for H_A if t_s falls in the hatched region

*The value of $t_{0.025} = 2.006$ is based on 56.3 degrees of freedom. If we were to use 50 degrees of freedom (i.e., if we had to rely on Table 4, rather than a computer) the t multiplier would be 2.009. This makes almost no difference in the resulting confidence interval.

shows the Student's t distribution and the values ± 2.000. The total shaded area in the figure is 0.05; it is split into two equal parts of area 0.025 each. We can think of Figure 7.3.3 as a formal guide for deciding whether the evidence is strong enough to significantly support H_A: If the observed value of t_s falls in the hatched regions of the t_s axis, then there is significant evidence for H_A. But the chance of this happening is 5%, if H_0 is true. Thus, we can say that

$$\Pr\{\text{data provide significant evidence for } H_A\} = 0.05 \text{ if } H_0 \text{ is true}$$

This probability has meaning in the context of a meta-study (depicted in Figure 7.3.4) in which we repeatedly sample from two populations and calculate a value of t_s. It is important to realize that the probability refers to a situation in which H_0 is true. In order to concretely picture such a situation, you are invited to suspend disbelief for a moment and come on an imaginary trip in Example 7.3.2.

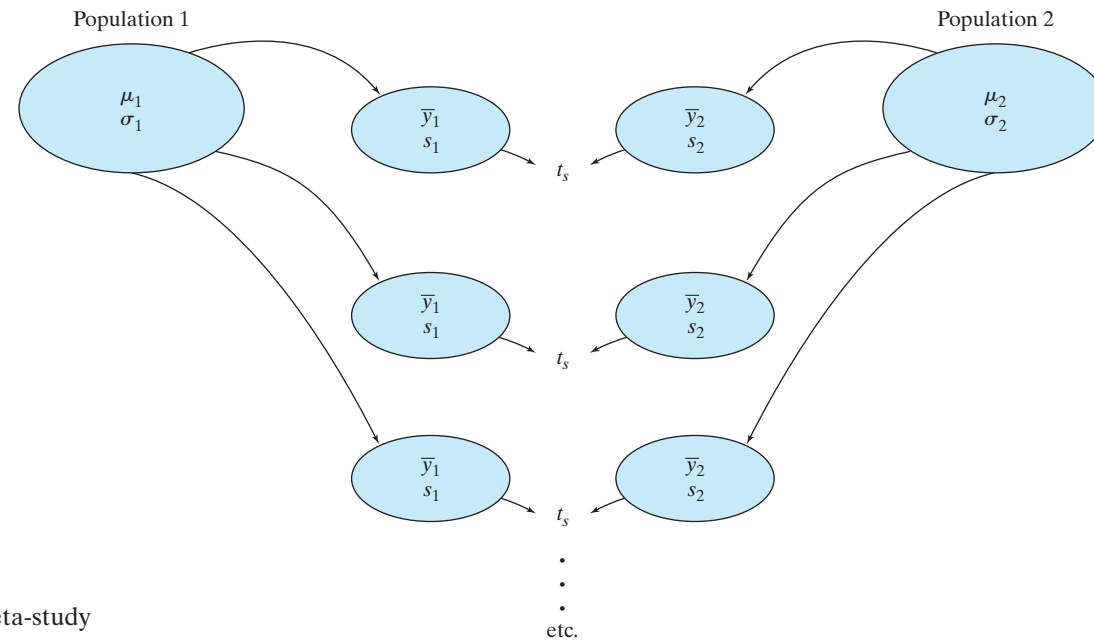


Figure 7.3.4 Meta-study for the t test

Example 7.3.2

Music and Marigolds* Imagine that the scientific community has developed great interest in the influence of music on the growth of marigolds. One school of investigation centers on whether music written by Bach or Mozart produces taller plants. Plants are randomly allocated to listen to Bach (treatment 1) or Mozart (treatment 2) and, after a suitable period of listening, data are collected on plant height. The null hypothesis is

H_0: Marigolds respond equally well to Bach or Mozart.

or

$$H_0: \mu_1 = \mu_2$$

where

μ_1 = Mean height of marigolds if exposed to Bach

μ_2 = Mean height of marigolds if exposed to Mozart

*This example is intentionally fanciful.

Assume for the sake of argument that H_0 is in fact true. Imagine now that many investigators perform the Bach versus Mozart experiment, and that each experiment results in data with 60 degrees of freedom. Suppose each investigator analyzes his or her data with a t test at $\alpha = 0.05$. What conclusions will the investigators reach? In the meta-study of Figure 7.3.4, suppose each pair of samples represents a different investigator. Since we are assuming that μ_1 and μ_2 are actually equal, the values of t_s will deviate from 0 only because of chance sampling error. If all the investigators were to get together and make a frequency distribution of their t_s values, that distribution would follow a Student's t curve with 60 degrees of freedom. The investigators would make their decisions as indicated by Figure 7.3.3, so we would expect them to have the following experiences:

95% of them would (correctly) not find significant evidence for H_A;

2.5% of them would find significant evidence for H_A and conclude (incorrectly) that the plants prefer Bach.

2.5% of them would find significant evidence for H_A and conclude (incorrectly) that the plants prefer Mozart.

Thus, a total of 5% of the investigators would find significant evidence for the alternative hypothesis. ■

Example 7.3.2 provides an image for interpreting α. Of course, in analyzing data, we are not dealing with a meta-study but rather with a single experiment. When we perform a t test at the 5% significance level, we are playing the role of one of the investigators in Example 7.3.2, and the others are imaginary. If we find significant evidence for H_A, there are two possibilities:

1. H_A is in fact true; or
2. H_0 is in fact true, but we are one of the unlucky 5% who obtained data that provided significant evidence for H_A anyway. In this case, we can think of the significant evidence for H_A as "setting off a false alarm."

We feel "confident" in claiming our evidence for H_A is significant because the second possibility is unlikely (assuming that we regard 5% as a small percentage). Of course, we never know (unless someone replicates the experiment) whether or not we are one of the unlucky 5%.

Significance Level versus *P*-Value Students sometimes find it hard to distinguish between significance level (α) and P-value.* For the t test, both α and the P-value are tail areas under Student's t curve. But α is an arbitrary prespecified value; it can be (and should be) chosen before looking at the data. By contrast, the P-value is determined from the data; indeed, giving the P-value is a way of describing the data. You may find it helpful at this point to compare Figure 7.2.3 with Figure 7.3.3. The shaded area represents P-value in the former and α in the latter figure.

Type I and Type II Errors

We have seen that α can be interpreted as a probability:

$$\alpha = \Pr\{\text{finding significant evidence for } H_A\} \text{ if } H_0 \text{ is true}$$

*Unfortunately, the term "significance level" is not used consistently by all people who write about statistics. A few authors use the terms "significance level" or "significance probability" where we have used "P-value."

Claiming that data provide evidence that significantly supports H_A when H_0 is true is called a **Type I error**. In choosing α, we are choosing our level of protection against Type I error. Many researchers regard 5% as an acceptably small risk. If we do not regard 5% as small enough, we might choose to use a more conservative value of α such as $\alpha = 0.01$; in this case the percentage of true null hypotheses that we reject would be not 5% but 1%.

In practice, the choice of α may depend on the context of the particular experiment. For example, a regulatory agency might demand more exacting proof of efficacy for a toxic drug than for a relatively innocuous one. Also, a person's choice of α may be influenced by his or her prior opinion about the phenomenon under study. For instance, suppose an agronomist is skeptical of claims for a certain soil treatment; in evaluating a new study of the treatment, he might express his skepticism by choosing a very conservative significance level (say, $\alpha = 0.001$), thus indicating that it would take a lot of evidence to convince him that the treatment is effective. For this reason, written reports of an investigation should include a P-value, so that each reader is free to choose his or her own value of α in evaluating the reported results.

If H_A is true, but we do not observe sufficient evidence to support H_A, then we have made a **Type II error**. Table 7.3.2 displays the situations in which Type I and Type II errors can occur. For example, if we find significant evidence for H_A, then we eliminate the possibility of a Type II error, but by rejecting H_0 we may have made a Type I error.

Table 7.3.2 Possible outcomes of testing H_0

		True situation	
		H_0 true	H_A true
OUR DECISION	Lack of significant evidence for H_A	Correct	Type II error
	Significant evidence for H_A	Type I error	Correct

The consequences of Type I and Type II errors can be very different. The following two examples show some of the variety of these consequences.

Example 7.3.3

Marijuana and the Pituitary Cannabinoids, which are substances contained in marijuana, can be transmitted from mother to young through the placenta and through the milk. Suppose we conduct the following experiment on pregnant mice: We give one group of mice a dose of cannabinoids and keep another group as controls. We then evaluate the function of the pituitary gland in the offspring. The hypotheses would be

H_0: Cannabinoids do not affect pituitary of offspring.

H_A: Cannabinoids do affect pituitary of offspring.

If in fact cannabinoids do not affect the pituitary of the offspring, but we conclude that our data provide significant evidence for H_A, we would be making a Type I error; the consequence might be unnecessary alarm if the conclusion were made public. On the other hand, if cannabinoids do affect the pituitary of the offspring, but our t test results in a lack of significant evidence for H_A, this would be a Type II error; one consequence might be unjustifiable complacency on the part of marijuana-smoking mothers. ■

Example 7.3.4

Immunotherapy Chemotherapy is standard treatment for a certain cancer. Suppose we conduct a clinical trial to study the efficacy of supplementing the chemotherapy with immunotherapy (stimulation of the immune system). Patients are given either chemotherapy or chemotherapy plus immunotherapy. The hypotheses would be

H_0: Immunotherapy is not effective in enhancing survival.

H_A: Immunotherapy does affect survival.

If immunotherapy is actually not effective, but we conclude that our data provide significant evidence for H_A and thus conclude that immunotherapy is effective, then we have made a Type I error. The consequence, if this conclusion is acted on by the medical community, might be the widespread use of unpleasant, dangerous, and worthless immunotherapy. If, on the other hand, immunotherapy is actually effective, but our data do not enable us to detect that fact (perhaps because our sample sizes are too small), then we have made a Type II error, with consequences quite different from those of a Type I error: The standard treatment will continue to be used until someone provides convincing evidence that supplementary immunotherapy is effective. If we still "believe" in immunotherapy, we can conduct another trial (perhaps with larger samples) to try again to establish its effectiveness. ■

As the foregoing examples illustrate, the consequences of a Type I error are usually quite different from those of a Type II error. The likelihoods of the two types of error may be very different, also. The significance level α is the probability of obtaining significant evidence for H_A if H_0 is true. Because α is chosen at will, the hypothesis testing procedure "protects" you against Type I error by giving you control over the risk of such an error. This control is independent of the sample size and other factors. The chance of a Type II error, by contrast, depends on many factors, and may be large or small. In particular, an experiment with small sample sizes often has a high risk of Type II error.

We are now in a position to reexamine Carl Sagan's aphorism that "Absence of evidence is not evidence of absence." Because the risk of Type I error is controlled and that of Type II error is not, our state of knowledge is much stronger after rejection of a null hypothesis than after nonrejection. For example, suppose we are testing whether a certain soil additive is effective in increasing the yield of field corn. If we find significant evidence for H_A and claim the additive is effective, then either (1) we are right; or (2) we have made a Type I error. Since the risk of a Type I error is controlled, we can be relatively confident of our conclusion that the additive is effective (although not necessarily very effective). Suppose, on the other hand, that the data are such that there is a lack of evidence for the additive's effectiveness—we do not have evidence for H_A. Then either (1) we are right (that is, H_0 is true), or (2) we have made a Type II error. Since the risk of a Type II error may be quite high, we cannot say confidently that the additive is ineffective. In order to justify a claim that the additive is ineffective, we would need to supplement our test of hypothesis with further analysis, such as a confidence interval or an analysis of the chance of Type II error. We will consider this in more detail in Sections 7.6 and 7.7.

Power

As we have seen, Type II error is an important concept. The probability of making a Type II error is denoted by β:

$$\beta = \Pr\{\text{lack of significant evidence for } H_A\} \text{ if } H_A \text{ is true}$$

The chance of not making a Type II error when H_A is true—that is, the chance of having significant evidence for H_A when H_A is true—is called the **power** of a statistical test:

$$\text{Power} = 1 - \beta = \Pr\{\text{significant evidence for } H_A\} \text{ if } H_A \text{ is true}$$

Thus, the power of a *t* test is a measure of the sensitivity of the test, or the ability of the test procedure to detect a difference between μ_1 and μ_2 when such a difference really *does* exist. In this way the power is analogous to the resolving power of a microscope.

The power of a statistical test depends on many factors in an investigation, including the sample sizes, the inherent variability of the observations, and the magnitude of the difference between μ_1 and μ_2. All other things being equal, using larger samples gives more information and thereby increases power. In addition, we will see that some statistical tests can be more powerful than others, and that some study designs can be more powerful than others.

The planning of a scientific investigation should always take power into consideration. No one wants to emerge from lengthy and perhaps expensive labor in the lab or the field, only to discover upon analyzing the data that the sample sizes were insufficient or the experimental material too variable, so that experimental effects that were considered important were not detected. Two techniques are available to aid the researcher in planning for adequate sample sizes. One technique is to decide how small each standard error ought to be and choose n using an analysis such as that of Section 6.4. A second technique is a quantitative analysis of the power of the statistical test. Such an analysis for the *t* test is discussed in Section 7.7.

Exercises 7.3.1–7.3.8

7.3.1 (Sampling exercise) Refer to the collection of 100 ellipses shown with Exercise 3.1.1, which can be thought of as representing a natural population of the organism *C. ellipticus.* Use random digits (from Table 1 or your calculator) to choose two random samples of five ellipses each. Use a metric ruler to measure the body length of each ellipse; measurements to the nearest millimeter will be adequate.

(a) Compare the means of your two samples, using a *t* test at $\alpha = 0.05$.

(b) Did the analysis of part (a) lead you to a Type I error, a Type II error, or no error?

7.3.2 (Sampling exercise) Simulate choosing random samples from two different populations, as follows. First, proceed as in Exercise 7.3.1 to choose two random samples of five ellipses each and measure their lengths. Then add 6 mm to *each* measurement in one of the samples.

(a) Compare the means of your two samples, using a *t* test at $\alpha = 0.05$.

(b) Did the analysis of part (a) lead you to a Type I error, a Type II error, or no error?

7.3.3 (Sampling exercise) Prepare simulated data as follows. First, proceed as in Exercise 7.3.1 to choose two random samples of five ellipses each and measure their lengths. Then, toss a coin. If the coin falls heads, add 6 mm to *each* measurement in one of the samples. If the coin falls tails, do not modify either sample.

(a) Prepare two copies of the simulated data. On the Student Copy, show the data only; on the Instructor Copy, indicate also which sample (if any) was modified.

(b) Give your Instructor Copy to the instructor and trade your Student Copy with another student when you are told to do so.

(c) After you have received another student's paper, compare the means of his or her two samples using a two-tailed *t* test at $\alpha = 0.05$. If you reject H_0, decide which sample was modified.

7.3.4 Suppose a new drug is being considered for approval by the Food and Drug Administration. The null hypothesis is that the drug is not effective. If the FDA approves the drug, what type of error, Type I or Type II, could not possibly have been made?

7.3.5 In Example 7.3.1, the null hypothesis was not rejected. What type of error, Type I or Type II, might have been made in that *t* test?

7.3.6 Suppose that a 95% confidence interval for $(\mu_1 - \mu_2)$ is calculated to be (1.4, 6.7). If we test H_0: $\mu_1 - \mu_2 = 0$ versus H_A: $\mu_1 - \mu_2 \neq 0$ using $\alpha = 0.05$, will we reject H_0? Why or why not?

7.3.7 Suppose that a 95% confidence interval for $(\mu_1 - \mu_2)$ is calculated to be $(-7.4, -2.3)$. If we test H_0: $\mu_1 = \mu_2$ versus H_A: $\mu_1 \neq \mu_2$ using $\alpha = 0.10$, will we reject H_0? Why or why not?

7.3.8 A dairy reasearcher has developed a new technique for culturing cheese that is purported to age cheese in substantially less time than traditional methods without affecting other properties of the cheese. Retrofitting cheese manufacturing plants with this new technology will initially cost millions of dollars, but if it indeed reduces aging time—even marginally—it will lead to higher company profits in the long run. If, on the other hand, the new method is no better than the old, the retrofit would be a financial mistake. Before making the decision to retrofit, an experiment will be performed to compare culture times of the new and old methods.

(a) In plain English, what are the null and alternative hypotheses for this experiment?

(b) In the context of the problem, what would be the consequence of a Type I error?

(c) In the context of the problem, what would be the consequence of a Type II error?

(d) In your opinion, which type of error would be more serious? Justify your answer. (It is possible to argue both sides.)

7.4 Association and Causation

When we are comparing two populations we often focus on the nature of the relationship between a **response variable**, Y—a variable that measures an outcome of interest—and an **explanatory variable** X—a variable used to explain or predict an outcome. As we will explore next, with data collected from an **experiment** we can assess whether or not there is evidence that X *affects* the mean value of Y. That is, we can ask, Do changes in X *cause* changes in Y? (For example, does toluene affect the mean amount of norepinephrine in the brain?) With **observational studies** our conclusions are more limited—we are not able to make causal claims, but rather only conclusions regarding association between X and Y. For example, we can ask, Are changes in X associated with changes in the mean value of Y? Or, Is there evidence that the mean values of Y differ for two populations? (For example, do crawfish captured from two different locations have different mean lengths?)

Thus, our ability to investigate such questions depends on how the data were collected: experimentally or with an observational study. Below are examples of each type of study as they pertain to comparing the means of two samples, followed by a more formal discussion of these study types.

Example 7.4.1

Hematocrit in Males and Females Hematocrit level is a measure of the concentration of red cells in blood. Table 7.4.1 gives the sample means and standard deviations of hematocrit values for two samples of 17-year-old American youths—489 males and 469 females.[17] ■

Table 7.4.1 Hematocrit (percent)

	Males	Females
Mean	45.8	40.6
SD	2.8	2.9

Example 7.4.2

Pargyline and Sucrose Consumption A study was conducted to determine the effect of the psychoactive drug Pargyline on feeding behavior in the black blowfly *Phormia regina*. The response variable was the amount of sucrose (sugar) solution a fly would drink in 30 minutes. The experimenters used two separate groups of flies: a group injected with Pargyline (905 flies) and a control group injected with saline (900 flies). Comparing the responses of the two groups provides an indirect assessment of the effect of Pargyline. (One might propose that a more *direct* way to determine

the effect of the drug would be to measure each fly twice—on one occasion after injecting Pargyline and on another occasion after injecting saline. However, this direct method is not practical because the measurement procedure disturbs the fly so much that each fly can be measured only once.) Table 7.4.2 shows the means and standard deviations for the two groups.[18] ■

Table 7.4.2 Sucrose consumption (mg)

	Control	Pargyline
Mean	14.9	46.5
SD	5.4	11.7

Examples 7.4.1 and 7.4.2 both involve two-sample comparisons, but notice that the two studies differ in a fundamental way. In Example 7.4.1 the samples come from populations that occur naturally; the investigator is merely an observer:

Population 1: Hematocrit values of 17-year-old U.S. males
Population 2: Hematocrit values of 17-year-old U.S. females

By contrast, the two populations in Example 7.4.2 do not actually exist but rather are defined in terms of specific experimental conditions; in a sense, the populations are created by experimental intervention:

Population 1: Sucrose consumptions of blowflies when injected with saline
Population 2: Sucrose consumptions of blowflies when injected with Pargyline

These two types of two-sample comparisons—the observational and the experimental—are both widely used in research. The formal methods of analysis are often the same for the two types, but the interpretation of the results is often somewhat different. For instance, in Example 7.4.2 it might be reasonable to say that Pargyline *causes* the increase in sucrose consumption, while no such notion applies in Example 7.4.1.

Observational versus Experimental Studies

A major consideration in interpreting the results of a biological study is whether the study was observational or experimental. In an **experiment**, the researcher intervenes in or manipulates the experimental conditions.* In an **observational study**, the researcher merely observes an existing situation, as in the following example.

Example 7.4.3

Cigarette Smoking In studies of the effects of smoking cigarettes, both experimental and observational approaches have been used. Effects in animals can be studied experimentally, because animals (for instance, dogs) can be allocated to treatment groups and the groups can be given various doses of cigarette smoke. Effects in humans are usually studied observationally. In one study, for example, pregnant women were questioned about their smoking habits, dietary habits, and so on.[19] When the babies were born, their physical and mental development was followed.

*The conditions being manipulated must be those defining the populations being compared. For example, if five men and five women are given the same drug and then the sexes are compared, the comparison of men to women is observational, not experimental.

One striking finding related to the babies' birthweights: The smokers tended to have smaller babies than the nonsmokers. The difference was not attributable to chance (the P-value was less than 10^{-5}). Nevertheless, it was far from clear that the difference was *caused* by smoking, because the women who smoked differed from the nonsmokers in many other aspects of their lifestyle besides smoking—for instance, they had very different dietary habits. ■

As Example 7.4.3 illustrates, it can be difficult to determine the exact nature of a cause–effect relationship in an observational study. In an experiment, on the other hand, a cause–effect relationship may be easy to see, based on the way in which the researcher manipulated the experimental conditions. To help fix the ideas, consider studying cholesterol level. Suppose a group of patients with high cholesterol levels enrolls in a clinical trial—that is, in a medical experiment—in which some of the patients are randomly chosen to receive a new drug and others are given a standard drug that has shown only modest effects in the past. If a two-sample t test shows that the mean cholesterol level decreased more for those on the new drug than for those on the standard drug, then the researcher can conclude that the new drug *caused* the superior outcome and is better than the standard drug.

Now consider a two-sample t test to compare average cholesterol level in a random sample of 50-year-olds to average cholesterol level in a random sample of 25-year-olds. Suppose a two-sample t test gives a small P-value, with the 50-year-olds having higher cholesterol than the 25-year-olds. We could be fairly confident that cholesterol level tends to increase with age. However, it would be *possible* that some other explanation were at work. For example, maybe diets have changed over time and the 25-year-olds are eating foods that the 50-year-olds don't eat, causing the 25-year-olds to have low cholesterol; perhaps if the 25-year-olds keep the same diet until they are 50, they will still have low cholesterol at age 50.

As a third example, consider comparing a random sample of home owners to a random sample of renters. Suppose a two-sample t test shows a significantly higher mean cholesterol level among the home owners than among the renters. We should not conclude that buying a home causes one's cholesterol level to rise. Rather, we should consider that people who own homes tend to be older than are renters. It might very well be the case that age is the causal factor, which explains why the home owners have higher cholesterol than do the renters.

All three of these cases might involve a two-sample t test and the rejection of H_0. Indeed, we might get the same P-value in each test. However, the conclusions we can draw from the three situations are quite different. The scope of the inference we can draw depends on the way in which the data are collected. Experiments allow us to infer cause–effect relationships that can only be guessed at in observational studies. Sometimes an observational study will leave us feeling reasonably confident that we understand the causal mechanism at work; however, we will see that drawing such conclusions is fraught with danger. For this reason, researchers interested in drawing causal conclusions should make great efforts to conduct controlled experiments rather than observational studies.

More on Observational Studies

The difficulties in interpreting observational studies arise from two primary sources:

Nonrandom selection from populations
Uncontrolled extraneous variables

The following example illustrates both of these.

Example 7.4.4

Race and Brain Size In the nineteenth century, much effort was expended in the attempt to show "scientifically" that certain human races were inferior to others. A leading researcher on this subject was the American physician S. G. Morton, who won widespread admiration for his studies of human brain size. Throughout his life, Morton collected human skulls from various sources, and he carefully measured the cranial capacities of hundreds of these skulls. His data appeared to suggest that (as he suspected) the "inferior" races had smaller cranial capacities. Table 7.4.3 gives a summary of Morton's data comparing Caucasian skulls to those of Native Americans.[20] According to a t test, the difference between these two samples is "statistically significant" (P-value < 0.001). But is it *meaningful*?

Table 7.4.3 Cranial capacity (in^3)

	Caucasian	Native American
Mean	87	82
SD	8	10
n	52	144

In the first place, the notion that cranial capacity is a measure of intelligence is no longer taken seriously. Leaving that question aside, one can still ask whether it is true that the mean cranial capacity of Native Americans is less than that of Caucasians. Such an inference beyond the actual data requires that the data be viewed as random samples from their respective populations. Of course, in actuality, Morton's data are not random samples but "samples of convenience," because Morton measured those skulls that he happened to obtain. But might the data be viewed "as if" they were generated by random sampling? One way to approach this question is to look for sources of bias. In 1977, the noted biologist Stephen Jay Gould reexamined Morton's data with this goal in mind, and indeed Gould found several sources of bias. For instance, the 144 Native American skulls represent many different groups of Native Americans; as it happens, 25% of the skulls (that is, 36 of them) were from Inca Peruvians, who were a small-boned people with small skulls, while relatively few were from large-skulled tribes such as the Iroquois. Clearly a comparison between Native Americans and Caucasians is meaningless unless somehow adjusted for such imbalances. When Gould made such an adjustment, he found that the difference between Native Americans and Caucasians vanished. ■

Even though the story of Morton's skulls is more than 100 years old, it can still serve to alert us to the pitfalls of inference. Morton was a conscientious researcher and took great care to make accurate measurements; Gould's reexamination did not reveal any suggestion of conscious fraud on Morton's part. Morton may have overlooked the biases in his data because they were *invisible* biases; that is, they related to aspects of the selection process rather than aspects of the measurements themselves.

When we look at a set of observational data, we can sometimes become so hypnotized by its apparent *solidity* and *objectivity* that we forget to ask how the observational units—the persons or things that were observed—were selected. The question should always be asked. If the selection was haphazard rather than truly random, the results can be severely distorted.

Confounding

Many observational studies are aimed at discovering some kind of causal relationship. Such discovery can be very difficult because of extraneous variables that enter in an uncontrolled (and perhaps unknown) way. The investigator must be guided by the maxim:

Association is not causation.

For instance, it is known that some populations whose diets are high in fiber enjoy a reduced incidence of colon cancer. But this observation does not in itself show that it is the high-fiber diet, rather than some other factor, that provides the protection against colon cancer.

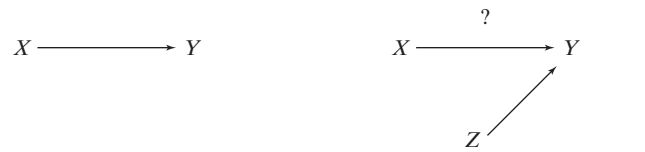
The following example shows how uncontrolled extraneous variables can cloud an observational study, and what kinds of steps can be taken to clarify the picture.

Example 7.4.5

Smoking and Birthweight In a large observational study of pregnant women, it was found that the women who smoked cigarettes tended to have smaller babies than the nonsmokers.[19] (This study was mentioned in Example 7.4.3.) It is plausible that smoking could cause a reduction in birthweight, for instance, by interfering with the flow of oxygen and nutrients across the placenta. But of course plausibility is not proof. In fact, the investigators found that the smokers differed from the nonsmokers with respect to many other variables. For instance, the smokers drank more whiskey than the nonsmokers. Alcohol consumption might plausibly be linked to a deficit in growth. ■

In Example 7.4.5 three variables are presented; let us refer to these as X = smoking, Y = birthweight, and Z = alcohol consumption. There is an association between X and Y, but is there a *causal* link between them? Or is there a causal link between Z and Y? Figure 7.4.1 gives a schematic representation of the situation. Changes in X are associated with changes in Y. However, changes in Z are also associated with changes in Y. We say that the effect that X has on Y is **confounded** with the effect that Z has on Y. In the context of Example 7.4.5, we say that the effect that smoking has on birthweight is confounded with the effect that alcohol consumption has on birthweight. In observational studies, confounding of effects is a common problem.

Figure 7.4.1 Schematic representation of causation (a) and of confounding (b)



(a) (b) The effect of X on Y is confounded with the effect of Z on Y

Example 7.4.6

Smoking and Birthweight The study presented in Example 7.4.5 uncovered many confounding variables. For example, the smokers drank more coffee than the nonsmokers. In addition—and this is especially puzzling—it was found that the smokers began to menstruate at younger ages than the nonsmokers. This phenomenon (early onset of menstruation) could not possibly have been *caused* by smoking, because it occurred (in almost all instances) *before* the woman began to smoke. One interpretation that has been proposed is that the two populations—women who choose to smoke and those who do not—are different in some biological way; thus, it has been suggested that the reduced birthweight is due "to the *smoker,* not the *smoking*."[21]

A number of more recent studies have attempted to shed some light on the relationship between maternal smoking and infant development. Researchers in one study observed, in addition to smoking habits, about 50 extraneous variables, including the mother's age, weight, height, blood type, upper arm circumference, religion, education, income, and so on.[22] After applying complex statistical methods of adjustment, they concluded that birthweight varies with smoking even when these extraneous factors are held constant. This says that there quite likely is a link between X = smoking and Y = birthweight as shown in Figure 7.4.1, although several other variables also affect birthweight. The point is that the presence of confounding doesn't mean that a link does not exist between X and Y, only that it is tangled up with other effects, so that we have to be cautious when interpreting the findings of an observational study.

In another study of pregnant women, researchers measured various quantities related to the functioning of the placenta.[23] They found that, compared to nonsmokers, women who smoked had more abnormalities of the placenta, and that their infants had very much higher blood levels of cotinine, a substance derived from nicotine. They also found evidence that, in the women who smoked, the circulation of blood in the placenta was notably improved by abstaining from smoking for three hours.

A third study used a matched design to try to isolate the effect of smoking behavior. The investigators identified 159 women who had smoked during one pregnancy but quit smoking before the next pregnancy.[24] These women were individually matched with 159 women who smoked during two consecutive pregnancies; pairs were matched with respect to the birthweight of the first child, amount of smoking during the first pregnancy, and several other factors. Thus, the members of a pair were believed to have identical "reproductive potential." The researchers then considered the birthweight of the second child; they found that the women who had quit smoking gave birth to infants who weighed more than the infants of their matched controls who continued to smoke. Of course, we cannot rule out the possibility that the women who quit smoking also quit other harmful habits, such as drinking too much alcohol, and that the increased birthweight was not really caused by giving up smoking. ■

Example 7.4.6 shows that observational studies can provide information about causality but must be interpreted cautiously. Researchers generally agree that a causal interpretation of an observed association requires extra support—for instance, that the association be observed consistently in observational studies conducted under various conditions and taking various extraneous factors into account, and also, ideally, that the causal link be supported by experimental evidence. We do not mean to say that an observed association *cannot* be causally interpreted, but only that such interpretation requires particular caution.

Spurious Association

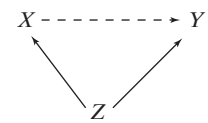
Example 7.4.7

Ultrasound It is quite common for a physician to use ultrasound examination of the fetus of a pregnant woman. However, when ultrasound technology was first used, there were concerns that the procedure might be harmful to the baby. An early study seemed to bear this out: On average, babies exposed to ultrasound in the womb were lighter at birth than were babies not exposed to ultrasound.[25] Later, a study was done in which some women were randomly chosen to have ultrasounds and others were not given ultrasounds. This study found no difference in birthweight between the two groups.[26] It seems that the reason a difference appeared in the first

study was that ultrasound was being used mostly for women who were experiencing problem pregnancies. The complications with the pregnancy were leading to low birthweight, not the use of ultrasound. ■

Figure 7.4.2 gives a schematic representation of the situation in Example 7.4.7. Changes in X (having an ultrasound examination) are associated with changes in Y (lower birthweight). However, X and Y are both dependent on a third variable Z (whether or not there are problems with the pregnancy), which is the variable that is driving the relationship. Changes in X and changes in Y are a common response to the third variable Z. We say that the association between X and Y is **spurious**: When we control for the "lurking variable" Z, the link between X and Y disappears. In the case of Example 7.4.7, it was not having an ultrasound that influenced birthweight; what mattered was whether or not there were problems with the pregnancy.

Figure 7.4.2 Schematic representation of spurious association



The association between X and Y is spurious; controlling for the lurking variable Z eliminates the X–Y link.

More on Experiments

An experiment is a study in which the researcher intervenes and imposes treatment conditions. The following is a simple example.

Example 7.4.8

Headache Pain Suppose a researcher gives ibuprofen to some people who have headaches and aspirin to others and then measures how long it takes for each person's headache to disappear. In this case, there are two treatments: ibuprofen and aspirin. By assigning people to treatment groups—ibuprofen and aspirin—the researcher is conducting an experiment. ■

When we are discussing an experiment, we refer to the units to which the treatments are assigned as **experimental units**. In an agricultural experiment, an experimental unit might be a plot of land. In general, an experimental unit is the smallest unit to which a treatment is applied in an experiment. Thus, in Example 7.4.8 the experimental units are individual people, since treatment is assigned on a person-by-person basis.

If treatments are assigned at random, for example, by tossing a coin and letting heads mean the person gets ibuprofen, while tails means the person gets aspirin, then the experiment is a *randomized* experiment. Sometimes an experiment is conducted in which one group is given a treatment and a second—the control group—is given nothing. For example, one could investigate the effectiveness of ibuprofen in treating headache pain by giving it to some people, while giving no painkiller to others. In contrast, the experiment in which some people are given ibuprofen and others are given aspirin is said to have an "active" control—the aspirin group.

Randomization Distributions

In Section 5.2 we developed the concept of a sampling distribution for the sample mean, $\overline{Y}$, by considering how $\overline{Y}$ varies from one random sample to another. Strictly

speaking, this provides the foundation for inference when analyzing an observational study, but not when the data arise from an experiment—in which treatments are assigned to experimental units, rather than a random sample being taken from a population. However, the concepts of Section 5.2 can be extended in a natural way to develop the **randomization distribution** of $\overline{Y}$, which is the distribution that $\overline{Y}$ takes on under all possible random assignments within an experiment. Randomization distributions then form the foundation for inference for experiments.

Only Statistical?

The term "statistical" is sometimes used—or, rather, misused—as an epithet. For instance, some people say that the evidence linking dietary cholesterol and heart disease is "only statistical." What they really mean is "only observational." Statistical evidence can be very strong indeed, if it flows from a randomized experiment rather than an observational study. As we have seen in the preceding examples, statistical evidence from an observational study must be interpreted with great care, because of potential distortions caused by extraneous variables.

Exercises 7.4.1–7.4.9

7.4.1 In 2005, 5.3% of the deaths in the United States were caused by chronic lower respiratory diseases (e.g., asthma and emphysema). In Arizona, 6.2% of deaths were due to chronic lower respiratory diseases.[27] Does this mean that living in Arizona exacerbates respiratory problems? If not, how can we explain the Arizona rate being above the national rate?

7.4.2 It has been hypothesized that silicone breast implants cause illness. In one study it was found that women with implants were more likely to smoke, to be heavy drinkers, to use hair dye, and to have had an abortion than were women in a comparison group who did not have implants.[28] Use the language of statistics to explain why this study casts doubt on the claim that implants cause illness.

7.4.3 Consider the setting of Exercise 7.4.2.

(a) What is the explanatory variable?

(b) What is the response variable?

(c) What are the observational units?

7.4.4 In a study of 1,040 subjects, researchers found that the prevalence of coronary heart disease increased as the number of cups of coffee consumed per day increased.[29]

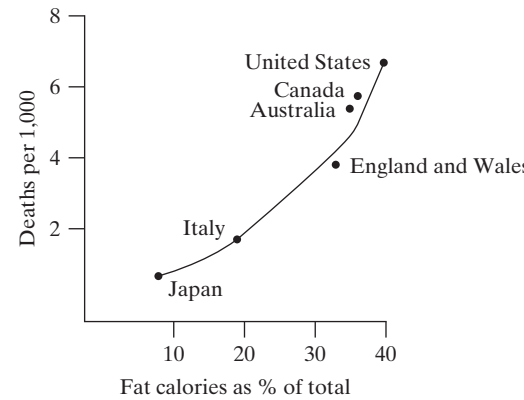
(a) What is the explanatory variable?

(b) What is the response variable?

(c) What are the observational units?

7.4.5 For an early study of the relationship between diet and heart disease, the investigator obtained data on heart disease mortality in various countries and on national average dietary compositions in the same countries. The accompanying graph shows, for six countries, the 1948–1949 death rate from degenerative heart disease (among men aged 55–59 years) plotted against the amount of fat in the diet.[30]

In what ways might this graph be misleading? Which extraneous variables might be relevant here? Discuss.



7.4.6 Shortly before Valentine's Day in 1999, a newspaper article was printed with the headline "Marriage makes for healthier, longer life, studies show." The headline was based on studies that showed that married persons live longer and have lower rates of cancer, heart disease, and stroke than do those who never marry.[31] Use the language of statistics to discuss the headline. Use a schematic diagram similar to Figure 7.4.1 or Figure 7.4.2 to support your explanation of the situation.

7.4.7 In June 2009, the *New York Times* published an article entitled "Alcohol's Good for You? Some Scientists Doubt It." The author wrote, "Study after study suggests that alcohol in moderation may promote heart health and even ward off diabetes and dementia. The evidence is so plentiful that some experts consider moderate drinking—about one drink a day for women, about two for men—a central component of a healthy lifestyle." Later in the article, the author wrote, "For some scientists, the question will not go away. No study, these critics say, has ever proved a causal relationship between moderate drinking and lower risk of death." Explain using the language of statistics and a schematic diagram similar to Figure 7.4.1 or Figure 7.4.2 why the critics say no study has ever proved a causal relationship.

7.4.8 In a study of the relationship between birthweight and race, birth records of babies born in Illinois were examined. The researchers found that the percentage of low birthweight babies among babies born to U.S.-born white women was much lower than the percentage of low birthweight babies among babies born to U.S.-born black women. This suggests that race plays an important role in determining the chance that a baby will have a low birthweight. However, the percentage of low birthweight babies among babies born to African-born black women was roughly equal to the percentage among babies born to U.S.-born white women.[32] Use the language of statistics to discuss what these data say about the relationships between low birthweight, race, and mother's birthplace. Use a schematic diagram similar to Figure 7.4.1 or Figure 7.4.2 to support your explanation.

7.4.9 Does the release of a Harry Potter book lead children to spend more time reading and thus reduce the number of accidents they have? Doctors in England compared the number of emergency room visits due to musculoskeletal injuries to children aged 7 to 15 during two types of weekends: (1) following the release dates of two books in the Harry Potter series and (2) during 24 "control" weekends, for one hospital. The following table shows the data, with the "Harry Potter weekends" in italics.[33]

WEEKEND	INJURIES	WEEKEND	INJURIES
6/7/03	63	7/10/04	57
6/14/03	77	7/17/04	66
6/21/03	*36**	7/24/04	62
6/28/03	63	6/4/05	51
7/5/03	75	6/11/05	83
7/12/03	71	6/18/05	60
7/19/03	60	6/25/05	66
7/26/03	52	7/2/05	74
6/5/04	78	7/9/05	75
6/12/04	84	*7/16/05*	*37**
6/19/04	70	7/23/05	46
6/26/04	75	7/30/05	68
7/3/04	81	8/6/05	60

(a) Given the nature of the data, can we make an inference about the release of Harry Potter books *causing* a change in accidents? Why or why not?

(b) The average for the Harry Potter weekends is 36.5, with a standard deviation of 0.7. The corresponding numbers for the other (control) weekends are 67.4 and 10.4. Use a t test to investigate the claim that the small number of injuries during Harry Potter weekends is consistent with chance variation. Use $\alpha = 0.01$. [*Note*: Formula (6.7.1) yields 23.9 degrees of freedom for these data.]

7.5 One-Tailed t Tests

The t test described in the preceding sections is called a **two-tailed t test** or a **two-sided t test** because the null hypothesis is rejected if t_s falls in either tail of the Student's t distribution and the P-value of the data is a two-tailed area under Student's t curve. A two-tailed t test is used to test the null hypothesis

$$H_0: \mu_1 = \mu_2$$

against the alternative hypothesis

$$H_A: \mu_1 \neq \mu_2$$

This alternative H_A is called a **nondirectional alternative**.

Directional Alternative Hypotheses

In some studies it is apparent from the beginning—*before* the data are collected—that there is only one reasonable direction of deviation from H_0. In such situations it is appropriate to formulate a directional alternative hypothesis. The following is a directional alternative:

$$H_A: \mu_1 < \mu_2$$

Another directional alternative is

$$H_A: \mu_1 > \mu_2$$

The following two examples illustrate situations where directional alternatives are appropriate.

Example 7.5.1

Niacin Supplementation Consider a feeding experiment with lambs. The observation Y will be weight gain in a two-week trial. Ten animals will receive diet 1, and 10 animals will receive diet 2, where

$$\text{Diet 1} = \text{Standard ration} + \text{Niacin}$$
$$\text{Diet 2} = \text{Standard ration}$$

On biological grounds it is expected that niacin may increase weight gain; there is no reason to suspect that it could possibly decrease weight gain. An appropriate formulation would be

H_0: Niacin is not effective in increasing weight gain ($\mu_1 = \mu_2$).

H_A: Niacin is effective in increasing weight gain ($\mu_1 > \mu_2$). ■

Example 7.5.2

Hair Dye and Cancer Suppose a certain hair dye is to be tested to determine whether it is carcinogenic (cancer causing). The dye will be painted on the skins of 20 mice (group 1), and an inert substance will be painted on the skins of 20 mice (group 2) that will serve as controls. The observation Y will be the number of tumors appearing on each mouse. An appropriate formulation is

H_0: The dye is not carcinogenic ($\mu_1 = \mu_2$).

H_A: The dye is carcinogenic ($\mu_1 > \mu_2$). ■

Note: If H_A is directional, then some people would rewrite H_0 to include the "opposite direction." For example, if H_A is $H_A: \mu_1 > \mu_2$, then we could write H_0 as $H_0: \mu_1 \leq \mu_2$. Thus, the null hypothesis is stating that the mean of population 1 *is* not greater than the mean of population 2, whereas the alternative hypothesis asserts that the mean of population 1 *is* greater than the mean of population 2. Between these two hypotheses, all possibilities are covered.

The One-Tailed Test Procedure

When the alternative hypothesis is directional, the t test procedure must be modified. The modified procedure is called a **one-tailed *t* test** and is carried out in two steps as follows:

Step 1 Check directionality—see if the data deviate from H_0 in the direction specified by H_A:

(a) If not, the P-value is greater than 0.50.

(b) If so, proceed to step 2.

Step 2 The P-value of the data is the *one-tailed* area beyond t_s.

To conclude the test, one can make a decision at a prespecified significance level α: H_0 is rejected if P-value $\leq \alpha$.

The rationale of the two-step procedure is that the P-value measures deviation from H_0 in the direction specified by H_A. The one-tailed P-value is illustrated in Figure 7.5.1 for two cases in which the data deviate from H_0 in the direction specified by H_A. Figure 7.5.2 illustrates the P-value for (a) a case in which the data are consistent with H_A: $\mu_1 > \mu_2$ and (b) a case in which the data are inconsistent with H_A: $\mu_1 > \mu_2$. The two-step testing procedure is demonstrated in Example 7.5.3.

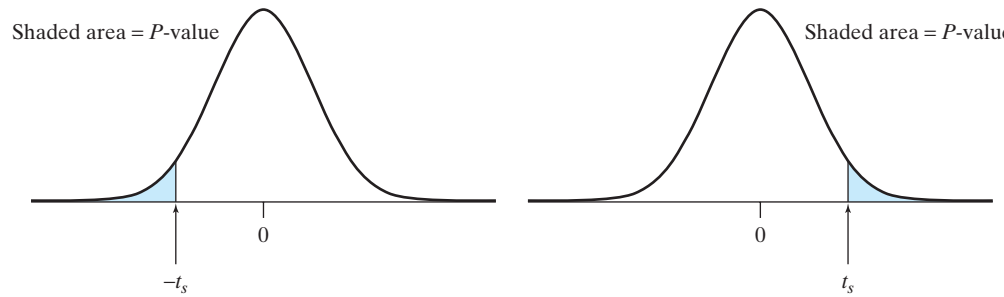


Figure 7.5.1 One-tailed P-value for a t test, (a) if the alternative is H_A: $\mu_1 < \mu_2$ and t_s is negative; (b) if the alternative is H_A: $\mu_1 > \mu_2$ and t_s is positive

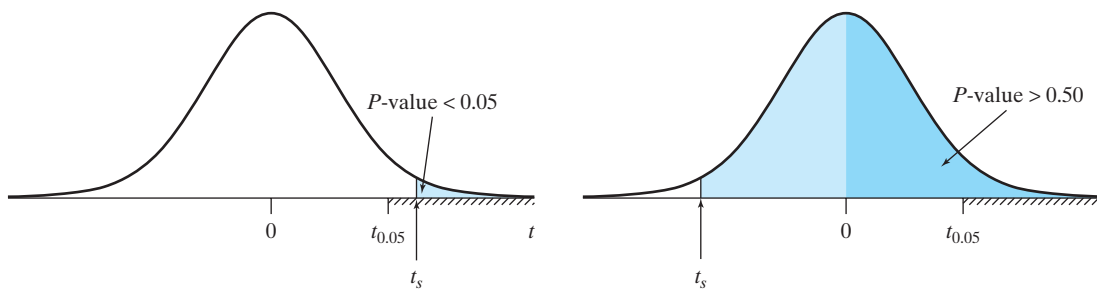


Figure 7.5.2 One-tailed P-value for a t test, (a) in which the data are consistent with H_A: $\mu_1 > \mu_2$; (b) in which the data are inconsistent with H_A: $\mu_1 > \mu_2$

Example 7.5.3 Niacin Supplementation Consider the lamb feeding experiment of Example 7.5.1. The alternative hypothesis is

$$H_A: \mu_1 > \mu_2$$

We will claim significant evidence for H_A if $\overline{Y}_1$ is sufficiently greater than $\overline{Y}_2$. Suppose formula (6.7.1) yields df = 18. The critical values from Table 4 are reproduced in Table 7.5.1.

Table 7.5.1 Critical values with df = 18

Tail area	0.20	0.10	0.05	0.04	0.03	0.025	0.02	0.01	0.005	0.0005
Critical value	0.862	1.330	1.734	1.855	2.007	2.101	2.214	2.552	2.878	3.922

To illustrate the one-tailed test procedure, suppose that we have[34]

$$\text{SE}_{(\overline{Y}_1-\overline{Y}_2)} = 2.2 \text{ lb}$$

and that we choose $\alpha = 0.05$. Let us consider various possibilities for the two sample means.

(a) Suppose the data give $\overline{y}_1 = 10$ lb and $\overline{y}_2 = 13$ lb. This deviation from H_0 is opposite to the assertion of H_A: We have $\overline{y}_1 < \overline{y}_2$, but H_A asserts that $\mu_1 > \mu_2$. Consequently, P-value > 0.50, so we would not find significant evidence for H_A at any significance level. (We would never use an α greater than 0.50.) We conclude that the data provide no evidence that niacin is effective in increasing weight gain.

(b) Suppose the data give $\overline{y}_1 = 14$ lb and $\overline{y}_2 = 10$ lb. This deviation from H_0 is in the direction of H_A (because $\overline{y}_1 > \overline{y}_2$), so we proceed to step 2. The value of t_s is

$$t_s = \frac{(14 - 10) - 0}{2.2} = 1.82$$

The (one-tailed) P-value for the test is the probability of getting a t statistic, with 18 degrees of freedom, that is as large or larger than 1.82. This upper tail probability (found with a computer) is 0.043, as shown in Figure 7.5.3.

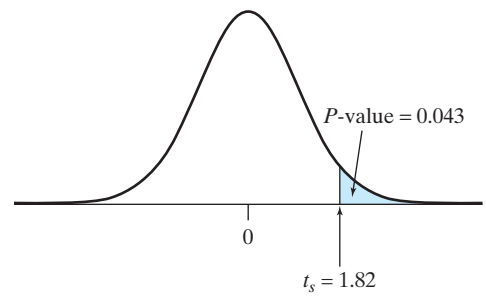


Figure 7.5.3 One-tailed P-value for the t test in Example 7.5.3

If we did not have a computer or graphing calculator available, we could use Table 4 to bracket the P-value. From Table 4, we see that the P-value would be bracketed as follows:

$0.04 <$ one-tailed P-value < 0.05

Since P-value $< \alpha$, we reject H_0 and conclude that there is some evidence that niacin is effective.

(c) Suppose the data give $\overline{y}_1 = 11$ lb and $\overline{y}_2 = 10$ lb. Then, proceeding as in part (b), we compute the test statistic as $t_s = 0.45$. The P-value is 0.329.

If we did not have a computer or graphing calculator available, we could use Table 4 to bracket the P-value as

$$P\text{-value} > 0.20$$

Since P-value $> \alpha$, we do not find significant evidence for H_A; we conclude that there is insufficient evidence to claim that niacin is effective. Thus, although these data deviate from H_0 in the direction of H_A, the amount of deviation is not great enough to justify significant evidence for H_A. ■

Notice that what distinguishes a one-tailed from a two-tailed t test is the way in which the P-value is determined, but not the directionality or nondirectionality of the conclusion. If we find significant evidence for H_A, our conclusion may be considered directional even if our H_A is nondirectional.* (For instance, in Example 7.2.4 we concluded that toluene increases NE concentration.)

Directional versus Nondirectional Alternatives

The same data will give a different P-value depending on whether the alternative hypothesis is directional or nondirectional. Indeed, if the data deviate from H_0 in the direction specified by H_A, the P-value for a directional alternative hypothesis will be 1/2 of the P-value for the test that uses a nondirectional alternative. It can happen that the same data will provide significant evidence for H_A using the one-tailed procedure but not using the two-tailed procedure, as Example 7.5.4 shows.

Example 7.5.4 Niacin Supplementation Consider part (b) of Example 7.5.3. In that example we chose $\alpha = 0.05$ and tested

$$H_0: \mu_1 = \mu_2$$

against the directional alternative hypothesis

$$H_A: \mu_1 > \mu_2$$

With $\bar{y}_1 = 14$ lb and $\bar{y}_2 = 10$ lb, the test statistic was $t_s = 1.82$ and the P-value was 0.043, as indicated in Figure 7.5.3. Our conclusion was to claim there is significant evidence for H_A.

However, suppose we had wished to test

$$H_0: \mu_1 = \mu_2$$

against the nondirectional alternative hypothesis

$$H_A: \mu_1 \neq \mu_2$$

With the same data of $\bar{y}_1 = 14$ lb and $\bar{y}_2 = 10$ lb, the test statistic is still $t_s = 1.82$. The P-value, however, is 0.086, as shown in Figure 7.5.4. Thus, P-value $> \alpha$ and we do not reject H_0.

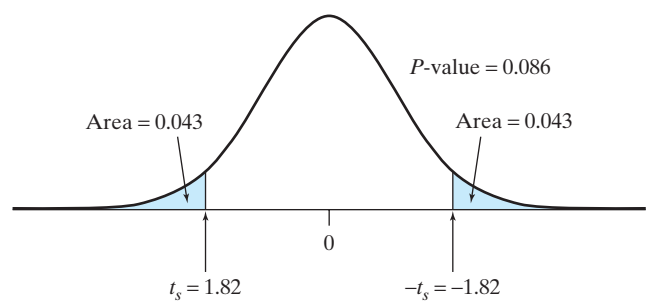


Figure 7.5.4 Two-tailed P-value for the t test in Example 7.5.4

*Some authors prefer not to draw a directional conclusion if H_A is nondirectional.

Hence, the one-tailed procedure finds significant evidence for H_A, but the two-tailed procedure does not. In this sense, it is "easier" to claim that the evidence significantly supports H_A with the one-tailed procedure than with the two-tailed procedure. ■

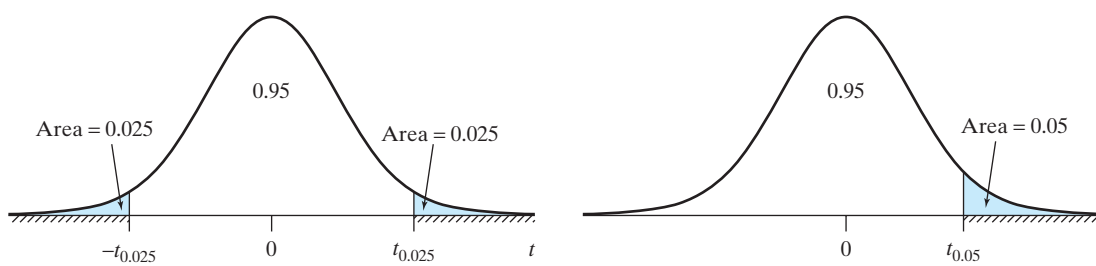
Why is the two-tailed P-value cut in half when the alternative hypothesis is directional? In Example 7.5.4, the researcher would conclude by saying, "The data suggest that niacin increases weight gain. But if niacin has no effect, then the kind of data I got in my experiment—having two sample means that differ by 1.82 SEs or more—would happen fairly often (P-value = 0.086). Sometimes the niacin diet would come out on top; sometimes the standard diet would come out on top. I cannot find significant evidence for H_A on the basis of what I have seen in these data." In Example 7.5.3(b), the researcher would conclude by saying, "*Before the experiment was run*, I suspected that niacin increases weight gain. The data provide evidence in support of this theory. If niacin has no effect, then the kind of data I got in my experiment—having the niacin diet sample mean exceed the standard diet that differ by 1.82 SEs or more—would rarely happen (P-value 0.043). (Before the experiment was run I dismissed the possibility that the niacin diet mean could be less than the standard diet mean.) Thus, I can claim my evidence significantly supports H_A." The researcher in Example 7.5.3(b) is using *two* sources of information to claim the significance of evidence for H_A: (1) what the data have to say (as measured by the tail area) and (2) previous expectations (which allow the researcher to ignore the lower tail area—the 0.043 area under the curve below −1.82 in Figure 7.5.4).

Note that the modification in procedure, when going from a two-tailed to a one-tailed test, preserves the interpretation of significance level α as given in Section 7.3, that is,

$$\alpha = \Pr\{\text{reject } H_0\} \text{ if } H_0 \text{ is true}$$

For instance, consider the case $\alpha = 0.05$. Figure 7.5.5 shows that the total shaded area—the probability of rejecting H_0—is equal to 0.05 in both a two-tailed test and a one-tailed test. This means that, if a great many investigators were to test a true H_0, then 5% of them would find significant evidence for H_A and commit a Type I error; this statement is true whether the alternative H_A is directional or nondirectional.

The crucial point in justification of the modified procedure for testing against a directional H_A is that *if* the direction of deviation of the data from H_0 is *not* as specified by H_A, then we will not claim that the evidence significantly supports H_A. For example, in the carcinogenesis experiment of Example 7.5.2, if the mice exposed to the hair dye had *fewer* tumors than the control group, we might (1) simply conclude



(a) Nondirectional H_A: $\mu_1 \neq \mu_2$ (b) Directional H_A: $\mu_1 > \mu_2$

Figure 7.5.5 Two-tailed and one-tailed t test with $\alpha = 0.05$. The data provide significant evidence for H_A if t_s falls in the hatched region of the t-axis

that the data do not indicate a carcinogenic effect, or (2) if the exposed group had *substantially* fewer tumors, so that the test statistic t_s was very far in the wrong tail of the t distribution, we might look for methodological errors in the experiment—for example, mistakes in lab technique or in recording the data, nonrandom allocation of the mice to the two groups, and so on—but we would not claim significant evidence for H_A.

A one-tailed t test is especially natural when only one direction of deviation from H_0 is believed to be plausible. However, one-tailed tests are also used in situations where deviation in both directions is possible, but only one direction is of interest. For instance, in the niacin experiment of Example 7.5.3, it is not necessary that the experimenter believe that it is *impossible* for niacin to reduce weight gain rather than increase it. Deviations in the wrong direction (less weight gain on niacin) would not lead to claiming there is significant evidence for H_A, and thus we would not make claims about the effect of niacin; this is the essential feature that distinguishes a directional from a nondirectional formulation.

Choosing the Form of H_A

When is it legitimate to use a directional H_A, and so to perform a one-tailed test? The answer to this question is linked to the directionality check—step 1 of the two-step test procedure given previously. Clearly such a check makes sense only if H_A was formulated before the data were inspected. (If we were to formulate a directional H_A that was "inspired" by the data, then of course the data would always deviate from H_0 in the "right" direction and the test procedure would always proceed to step 2.) This is the rationale for the following rule.

Rule for Directional Alternatives

It is legitimate to use a directional alternative H_A only if H_A is formulated before seeing the data and there is no scientific interest in results that deviate in a manner opposite to that specified by H_A.

In research, investigators often get more pleasure from finding significant evidence for an alternative hypothesis than not finding evidence. In fact, research reports often contain phrases such as "we are unable to find significant evidence for the alternative hypothesis" or "the results failed to reach statistical significance." Under these circumstances, one might wonder what the consequences would be if researchers succumbed to the natural temptation to ignore the preceding rule for using directional alternatives. After all, very often one can think of a rationale for an effect *ex post facto*—that is, after the effect has been observed. A return to the imaginary experiment on plants' musical tastes will illustrate this situation.

Example 7.5.5

Music and Marigolds Recall the imaginary experiment of Example 7.3.2, in which investigators measure the heights of marigolds exposed to Bach or Mozart. Suppose, as before, that the null hypothesis is true, that df $= 60$, and that the investigators all perform t tests at $\alpha = 0.05$. Now suppose in addition that all of the investigators violate the rule for use of directional alternatives, and that they formulate H_A after seeing the data. Half of the investigators would obtain data for which $\bar{y}_1 > \bar{y}_2$, and they would formulate the alternative

$$H_A: \mu_1 > \mu_2 \text{ (plants prefer Bach)}$$

The other half would obtain data for which $\bar{y}_1 < \bar{y}_2$, and they would formulate the alternative

$$H_A: \mu_1 < \mu_2 \text{ (plants prefer Mozart)}$$

Now envision what would happen. Since the investigators are using directional alternatives, they will all compute *P*-values using only one tail of the distribution. We would expect them to have the following experiences:

90% of them would get a t_s in the middle 90% of the distribution and would not find significant evidence for H_A.

5% of them would get a t_s in the top 5% of the distribution and would conclude that the plants prefer Bach.

5% of them would get a t_s in the bottom 5% of the distribution and would conclude that the plants prefer Mozart.

Thus, a total of 10% of the investigators would claim there is significant evidence for H_A. Of course each investigator individually never realizes that the overall percentage of Type I errors is 10% rather than 5%. And the conclusions that plants prefer Bach or Mozart could be supported by *ex post facto* rationales that would be limited only by the imagination of the investigators. ■

As Example 7.5.5 illustrates, a researcher who uses a directional alternative when it is not justified pays the price of a doubled risk of Type I error. Moreover, those who read the researcher's report will not be aware of this doubling of risk, which is why some scientists advocate never using a directional alternative.

Exercises 7.5.1–7.5.13

7.5.1 For each of the following data sets, use Table 4 to bracket the one-tailed *P*-value of the data as analyzed by the *t* test, assuming that the alternative hypothesis is $H_A: \mu_1 > \mu_2$.

(a)

	SAMPLE 1	SAMPLE 2
n	10	10
$\bar{y}$	10.8	10.5

$\text{SE}_{(\bar{Y}_1-\bar{Y}_2)} = 0.23$ with df = 18

(b)

	SAMPLE 1	SAMPLE 2
n	100	100
$\bar{y}$	750	730

$\text{SE}_{(\bar{Y}_1-\bar{Y}_2)} = 11$ with df = 180

7.5.2 For each of the following data sets, use Table 4 to bracket the one-tailed *P*-value of the data as analyzed by the *t* test, assuming that the alternative hypothesis is $H_A: \mu_1 > \mu_2$.

(a)

	SAMPLE 1	SAMPLE 2
n	10	10
$\bar{y}$	3.24	3.00

$\text{SE}_{(\bar{Y}_1-\bar{Y}_2)} = 0.61$ with df = 17

(b)

	SAMPLE 1	SAMPLE 2
n	6	5
$\bar{y}$	560	500

$\text{SE}_{(\bar{Y}_1-\bar{Y}_2)} = 45$ with df = 8

(c)

	SAMPLE 1	SAMPLE 2
n	20	20
$\bar{y}$	73	79

$\text{SE}_{(\bar{Y}_1-\bar{Y}_2)} = 2.8$ with df = 35

7.5.3 For each of the following situations, suppose H_0: $\mu_1 = \mu_2$ is being tested against H_A: $\mu_1 > \mu_2$. State whether or not there is significant evidence for H_A.

(a) $t_s = 3.75$ with 19 degrees of freedom, $\alpha = 0.01$.

(b) $t_s = 2.6$ with 5 degrees of freedom, $\alpha = 0.10$.

(c) $t_s = 2.1$ with 7 degrees of freedom, $\alpha = 0.05$.

(d) $t_s = 1.8$ with 7 degrees of freedom, $\alpha = 0.05$.

7.5.4 For each of the following situations, suppose H_0: $\mu_1 = \mu_2$ is being tested against H_A: $\mu_1 < \mu_2$. State whether or not there is significant evidence for H_A.

(a) $t_s = -1.6$ with 23 degrees of freedom, $\alpha = 0.05$.

(b) $t_s = -2.3$ with 5 degrees of freedom, $\alpha = 0.10$.

(c) $t_s = 0.4$ with 16 degrees of freedom, $\alpha = 0.10$.

(d) $t_s = -2.8$ with 27 degrees of freedom, $\alpha = 0.01$.

7.5.5 Ecological researchers measured the concentration of red cells in the blood of 27 field-caught lizards (*Sceloporis occidetitalis*). In addition, they examined each lizard for infection by the malarial parasite *Plasmodium*. The red cell counts ($10^{-3} \times$ cells per mm^3) were as reported in the table.[35]

	INFECTED ANIMALS	NONINFECTED ANIMALS
n	12	15
$\bar{y}$	972.1	843.4
s	245.1	251.2

One might expect that malaria would reduce the red cell count, and in fact previous research with another lizard species had shown such an effect. Do the data support this expectation? Assume that the data are normally distributed. Test the null hypothesis of no difference against the alternative that the infected population has a lower red cell count. Use a t test at

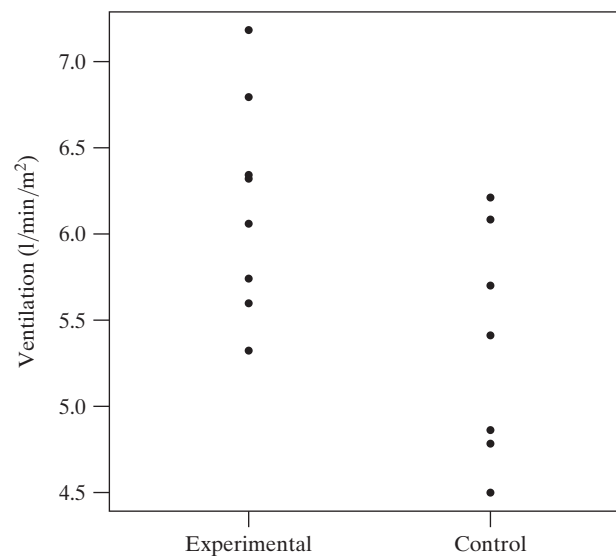
(a) $\alpha = 0.05$

(b) $\alpha = 0.10$

[*Note*: Formula (6.7.1) yields 24 df.]

7.5.6 A study was undertaken to compare the respiratory responses of hypnotized and nonhypnotized subjects to certain instructions. The 16 male volunteers were allocated at random to an experimental group to be hypnotized or to a control group. Baseline measurements were taken at the start of the experiment. In analyzing the data, the researchers noticed that the baseline breathing patterns of the two groups were different; this was surprising, since all the subjects had been treated the same up to that time. One explanation proposed for this unexpected difference was that the experimental group were more excited in anticipation of the experience of being hypnotized. The accompanying table presents a summary of the baseline measurements of total ventilation (liters of air per minute per square meter of body area). Parallel dotplots of the data are given in the following graph.[36] [*Note*: Formula (6.7.1) yields 14 df.]

	EXPERIMENTAL	CONTROL
	5.32	4.50
	5.60	4.78
	5.74	4.79
	6.06	4.86
	6.32	5.41
	6.34	5.70
	6.79	6.08
	7.18	6.21
n	8	8
$\bar{y}$	6.169	5.291
s	0.621	0.652



(a) Use a t test to test the hypothesis of no difference against a nondirectional alternative. Let $\alpha = 0.05$.

(b) Use a t test to test the hypothesis of no difference against the alternative that the experimental conditions produce a larger mean than the control conditions. Let $\alpha = 0.05$.

(c) Which of the two tests, that of part (a) or part (b), is more appropriate? Explain.

7.5.7 In a study of lettuce growth, 10 seedlings were randomly allocated to be grown in either standard nutrient solution or in a solution containing extra nitrogen. After 22 days of growth, the plants were harvested and weighed, with the results given in the table.[37] Are the data sufficient to conclude that the extra nitrogen

enhances plant growth under these conditions? Use a t test at $\alpha = 0.10$ against a directional alternative. (Assume that the data are normally distributed.) [*Note:* Formula (6.7.1) yields 7.7 df.]

	LEAF DRY WEIGHT (GM)		
NUTRIENT SOLUTION	n	MEAN	SD
Standard	5	3.62	0.54
Extra nitrogen	5	4.17	0.67

7.5.8 Research has shown that for mammals giving birth to a son versus a daughter places a greater strain on mothers. Does this affect the health of their next child? A study compared the birthweights of humans born after a male versus after a female. Summary statistics for the sample of size 76 are given in the following table; the data appeared to be normally distributed.[38] Use a t test, with $\alpha = 0.05$ and a directional alternative, to investigate the research hypothesis that birthweight is lower when the elder sibling is male. [*Note:* Formula (6.7.1) yields 69.5 df.]

	BIRTHWEIGHT (KG)		
SEX OF ELDER SIBLING	n	MEAN	SD
Male	33	3.32	0.62
Female	43	3.63	0.63

7.5.9 An entomologist conducted an experiment to see if wounding a tomato plant would induce changes that improve its defense against insect attack. She grew larvae of the tobacco hornworm (*Manduca sexta*) on wounded plants or control plants. The accompanying table shows the weights (mg) of the larvae after seven days of growth.[39] (Assume that the data are normally distributed.) How strongly do the data support the researcher's expectation? Use a t test at the 5% significance level. Let H_A be that wounding the plant tends to diminish larval growth. [*Note:* Formula (6.7.1) yields 31.8 df.]

	WOUNDED	CONTROL
n	16	18
$\bar{y}$	28.66	37.96
s	9.02	11.14

7.5.10 A pain-killing drug was tested for efficacy in 50 women who were experiencing uterine cramping pain following childbirth. Twenty-five of the women were randomly allocated to receive the drug, and the remaining 25 received a placebo (inert substance). Capsules of drug or placebo were given before breakfast and again at noon. A pain relief score, based on hourly questioning throughout the day, was computed for each woman. The possible pain relief scores ranged from 0 (no relief) to 56 (complete relief for 8 hours). Summary results are shown in the table.[40] [*Note:* Formula (6.7.1) yields 47.2 df.]

		PAIN RELIEF SCORE	
TREATMENT	n	MEAN	SD
Drug	25	31.96	12.05
Placebo	25	25.32	13.78

(a) Test for evidence of efficacy using a t test. Use a directional alternative and $\alpha = 0.05$.

(b) If the alternative hypothesis were nondirectional, how would the answer to part (a) change?

7.5.11 Postoperative ileus (POI) is a form of gastrointestinal dysfunction that commonly occurs after abdominal surgery and results in absent or delayed gastrointestinal motility. Does rocking in a chair after abdominal surgery reduce postoperative ileus (POI) duration? Sixty-six postoperative abdominal surgery patients were randomly divided into two groups. The experimental group ($n = 34$) received standard care plus the use of a rocking chair while the control group ($n = 32$) received only standard care. For each patient, the postoperative time until first flatus (days) (an indication that the POI has ended) was measured. The results are tabulated here.[41]

	TIME UNTIL FIRST FLATUS (DAYS)		
	n	MEAN (DAYS)	SD
Rocking	34	3.16	0.86
Control	32	3.88	0.80

(a) Is there evidence that use of the rocking chair reduces POI duration (i.e., the time until first flatus)? Use a t test with a directional alternative with $\alpha = 0.05$.

(b) While the researchers hypothesized that the use of a rocking chair could reduce POI duration, it is not unreasonable to hypothesize that the use of a rocking chair could increase POI duration. Based on this possibility, discuss the appropriateness of using a directional versus nondirectional test. (*Hint:* Consider what medical recommendations might be made based on this research.)

7.5.12 In Example 7.2.6 we considered testing H_0: $\mu_1 = \mu_2$ against the nondirectional alternative hypothesis H_A: $\mu_1 \neq \mu_2$ and found that the P-value could be

bracketed as $0.06 < P\text{-value} < 0.10$. Recall that the sample mean for the group 1 (the control group) was 15.9, which was less than the sample mean of 11.0 for group 2 (the group treated with Ancymidol). However, Ancymidol is considered to be a growth inhibitor, which means that one would expect the control group to have a larger mean than the treatment group if ancy has any effect on the type of plant being studied (in this case, the Wisconsin Fast Plant). Suppose the researcher had expected ancy to retard growth—before conducting the experiment—and had conducted a test of H_0: $\mu_1 = \mu_2$ against the nondirectional alternative hypothesis H_A: $\mu_1 > \mu_2$, using $\alpha = 0.05$. What would be the bounds on the P-value? Would H_0 be rejected? Why or why not? What would be the conclusion of the experiment? (*Note:* This problem requires almost no calculation.)

7.5.13 (Computer exercise) An ecologist studied the habitat of a marine reef fish, the six bar wrasse (*Thalassoma hardwicke*), near an island in French Polynesia that is surrounded by a barrier reef. He examined 48 patch reef settlements at each of two distances from the reef crest: 250 meters from the crest and 800 meters from the crest. For each patch reef, he calculated the "settler density," which is the number of settlers (juvenile fish) per unit of settlement habitat. Before collecting the data, he hypothesized that the settler density might decrease as distance from the reef crest increased, since the way that waves break over the reef crest causes resources (i.e., food) to tend to decrease as distance from the reef crest increases. Here are the data:[42]

250 METERS			800 METERS		
0.318	0.758	0.318	0.941	0.289	0.399
0.637	0.372	0.524	0.279	0.392	0.955
0.196	0.637	1.404	1.021	0.725	0.531
0.624	1.560	0.000	0.108	1.318	0.252
0.909	0.207	1.061	0.738	0.612	1.179
0.295	0.685	0.590	0.907	0.637	0.442
0.594	0.000	0.363	0.503	0.181	0.291
0.442	1.303	1.567	0.637	0.941	0.579
1.220	0.898	1.577	1.498	0.265	0.252
1.303	1.157	0.312	0.866	0.979	0.373
0.187	0.970	0.758	0.588	0.909	0.000
1.560	0.624	0.505	0.606	0.283	0.463
0.849	1.592	0.909	0.490	0.337	1.248
2.411	1.019	0.362	0.163	0.813	2.010
1.705	0.829	0.329	0.277	0.000	1.213
1.019	0.884	0.909	0.293	0.544	0.808

For 250 meters, the sample mean is 0.818 and the sample SD is 0.514. For 800 meters, the sample mean is 0.628 and the sample SD is 0.413. Do these data provide statistically significant evidence, at the 0.10 level, to support the ecologist's theory? Investigate with an appropriate graph and test.

7.6 More on Interpretation of Statistical Significance

Ideally, statistical analysis should aid the researcher by helping to clarify whatever message is contained in the data. For this purpose, it is not enough that the statistical calculations be correct; the results must also be correctly interpreted. In this section we explore some principles of interpretation that apply not only to the t test, but also to other statistical tests to be discussed later.

Significant Difference versus Important Difference

The term *significant* is often used in describing the results of a statistical analysis. For example, if an experiment to compare a drug against a placebo gave data with a very small P-value, then the conclusion might be stated as "The effect of the drug was highly significant." As another example, if two fertilizers for wheat gave a yield comparison with a large P-value, then the conclusion might be stated as "The wheat yields did not differ significantly between the two fertilizers" or "The difference between the fertilizers was not significant." As a third example, suppose a substance is tested for toxic effects by comparing exposed animals and control animals, and that the null hypothesis of no difference is not rejected. Then the conclusion might be stated as "No significant toxicity was found."

Clearly such phraseology using the term *significant* can be seriously misleading. After all, in ordinary English usage, the word significant connotes "substantial" or "important." In statistical jargon, however, the statement

> "The difference was significant"

means nothing more or less than

> "The null hypothesis of no difference was rejected."

This is to say, "We found sufficient evidence that the difference in sample means was not caused by chance error alone."

By the same token, the statement

> "The difference was not significant"

means

> "There was not sufficient evidence that the observed difference in means was due to anything other than chance variation."

It would perhaps be preferable if a different word were used in place of "significant," such as "discernible" (meaning that the test discerned a difference). Alas, the specialized usage of the word *significant* has become quite common in scientific writing and understandably is the source of much confusion.

It is essential to recognize that a statistical test provides information about only one question: Is the difference observed in the data large enough to infer that a difference in the same direction exists in the population? The question of whether a difference is *important,* as opposed to (statistically) significant, cannot be decided on the basis of the P-values alone but must also include an examination of the magnitude of the estimated population difference as well as specific expertise in the research area or practical situation. The following two examples illustrate this fact.

Example 7.6.1 Serum LD Lactate dehydrogenase (LD) is an enzyme that may show elevated activity following damage to the heart muscle or other tissues. A large study of serum LD levels in healthy young people yielded the results shown in Table 7.6.1.[43]

Table 7.6.1 Serum LD (U/l)

	Males	Females
n	270	264
$\bar{y}$	60	57
s	11	10

The difference between males and females is quite significant; in fact, $t_s = 3.3$, which gives a P-value ≈ 0.001. However, this does not imply that the difference $(60 - 57 = 3$ U/l) is large or important in any practical sense. ■

Example 7.6.2 Body Weight Imagine that we are studying the body weight of men and women, and we obtain the fictitious but realistic data shown in Table 7.6.2.[44]

Table 7.6.2 Body weight (lb)

	Males	Females
n	2	2
$\bar{y}$	175	143
s	35	34

For these data the t test gives $t_s = 0.93$ and a P-value ≈ 0.45. The observed difference between males and females is not small (it is $175 - 143 = 32$ lb), yet it is not statistically significant for any reasonable choice of α. The lack of statistical significance does not imply that the sex difference in body weight is small or unimportant. It means only that the data are inadequate to characterize the difference in the population means. A sample difference of 32 lb could easily happen by chance if the two populations are identical, especially with such small sample sizes. ■

Effect Size

The preceding examples show that the statistical significance or nonsignificance of a difference does not indicate whether the difference is important. Nevertheless, the question of "importance" can and should be addressed in most data analyses. To assess importance, one needs to consider the *magnitude* of the difference. In Example 7.6.1 the male versus female difference is "statistically significant," but this is largely due to the sample sizes being quite large. A t test uses the test statistic

$$t_s = \frac{(\bar{y}_1 - \bar{y}_2)}{\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}}$$

If n_1 and n_2 are large, then $\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}$ will be small and the test statistic will tend to be large even when the difference in observed means $(\bar{Y}_1 - \bar{Y}_2)$ is very small. Thus, one might find significant evidence for H_A due to the sample size being large, even if μ_1 and μ_2 are nearly equal. The sample size acts like a magnifying glass: *The larger the sample size, the smaller the difference that can be detected in a hypothesis test.*

The **effect size** in a study is the difference between μ_1 and μ_2, expressed relative to the standard deviation of one of the populations. If the two populations have the same standard deviation, σ, then the effect size is*

$$\text{Effect size} = \frac{|\mu_1 - \mu_2|}{\sigma}$$

Of course, when working with sample data we can only calculate an *estimated* effect size by using sample values in place of the unknown population values.

Example 7.6.3

Serum LD For the data given in Example 7.6.1 the difference in sample means, $60 - 57 = 3$, is less than one-third of a standard deviation. Using the larger sample SD we can calculate a sample effect size of

$$\text{Effect size} = \frac{|\bar{y}_1 - \bar{y}_2|}{s} = \frac{60 - 57}{11} = 0.27$$

*If the standard deviations are not equal, we can use the larger SD in defining the effect size.

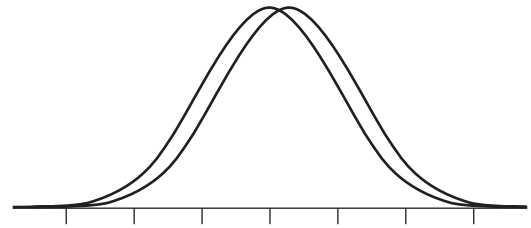


Figure 7.6.1 Overlap between two normally distributed populations when the effect size is 0.27

This indicates that there is a lot of overlap between the two groups. Figure 7.6.1 shows the extent of the overlap that occurs if two normally distributed populations differ on average by 0.27 SDs. ■

Example 7.6.4

Body Weight For the data given in Example 7.6.2 the difference in sample means, $175 - 143 = 32$, is roughly one standard deviation. The sample effect size is

$$\text{Effect size} = \frac{|\bar{y}_1 - \bar{y}_2|}{s} = \frac{175 - 143}{35} = 0.91$$

Figure 7.6.2 shows the extent of the overlap that occurs if two normally distributed populations differ on average by 0.91 SD. ■

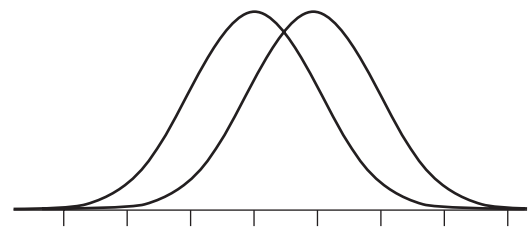


Figure 7.6.2 Overlap between two normally distributed populations when the effect size is 0.91

The definition of effect size that we are using is probably unfamiliar to the biologically oriented reader. It is more common in biology to "standardize" a difference of two quantities by expressing it as a percentage of one of them. For example, the weight difference given in Table 7.6.2 between males and females, expressed as a percentage of mean female weight, is

$$\frac{\bar{y}_1 - \bar{y}_2}{\bar{y}_2} = \frac{175 - 143}{143} = 0.22 \text{ or } 22\%$$

Thus, the males are about 22% heavier than the females. However, from a statistical viewpoint it is often more relevant that the average weights for males and females are 0.91 SD apart.

Confidence Intervals to Assess Importance

Calculating the effect size is one way to quantify how far apart two sample means are. Another reasonable approach is to use the observed difference $(\overline{Y}_1 - \overline{Y}_2)$ to construct a confidence interval for the population difference $(\mu_1 - \mu_2)$. In interpreting the confidence interval, the judgment of what is "important" is made on the basis of experience with the particular practical situation. The following three examples illustrate this use of confidence intervals.

Example 7.6.5

Serum LD For the LD data of Example 7.6.1, a 95% confidence interval for $(\mu_1 - \mu_2)$ is

$$3 \pm 1.8$$

or

$$(1.2, 4.8)$$

This interval implies (with 95% confidence) that the population difference in means between the sexes does not exceed 4.8 U/l. As an expert, a physician evaluating this information would know that typical day-to-day fluctuation in a person's LD level is around 6.5 U/l, which is higher than 4.8 U/l, the highest we estimate the mean sex difference to be, and therefore this difference is negligible from the medical standpoint. Consequently, the physician might conclude that it is unnecessary to differentiate between the sexes in establishing clinical thresholds for diagnosis of illness. In this case, the sex difference in LD may be said to be statistically significant but medically unimportant. To put this another way, the data suggest that men do in fact tend to have higher levels than women, but not higher in any clinically useful way. ■

Example 7.6.6

Body Weight For the body-weight data of Example 7.6.2, a 95% confidence interval for $(\mu_1 - \mu_2)$ is

$$32 \pm 149$$

or

$$(-117, 181)$$

From this confidence interval we cannot tell whether the true difference (between the population means) is large favoring females, is small, or is large favoring males. Because the confidence interval contains numbers of both small and large magnitude, it does not tell us whether the difference between the sexes is important or unimportant. With such a wide confidence interval a researcher would likely wish to conduct a larger study to better assess the importance of the difference. Suppose, for example, that the means and standard deviations were as given in Table 7.6.2, but that they were based on 2,000 rather than 2 people of each sex. Then the 95% confidence interval would be

$$32 \pm 2$$

or

$$(30, 34)$$

This interval would imply (with 95% confidence) that the difference is at least 30 lb, an amount that might reasonably be regarded as important, at least for some purposes. ■

Example 7.6.7

Yield of Tomatoes Suppose a horticulturist is comparing the yields of two varieties of tomatoes; yield is measured as pounds of tomatoes per plant. On the basis of practical considerations, the horticulturist has decided that a difference between the varieties is "important" only if it exceeds 1 pound per plant, on the average. That is, the difference is important if

$$|\mu_1 - \mu_2| > 1.0 \text{ lb}$$

Suppose the horticulturist's data give the following 95% confidence interval:

$$(0.2, 0.3)$$

Because the largest estimate for the population difference is only 0.3 lb (all values in the interval are less than 1.0 lb), the data support (with 95% confidence) the assertion that the difference is *not* important, using the horticulturist's criterion. ■

In many investigations, statistical significance and practical importance are both of interest. The following example shows how the relationship between these two concepts can be visualized using confidence intervals.

Example 7.6.8

Yield of Tomatoes Let us return to the tomato experiment of Example 7.6.7. The confidence interval was

$$(0.2, 0.3)$$

Recall from Section 7.3 that the confidence interval can be interpreted in terms of a t test. Because all values within the confidence interval are positive, a t test (two-tailed) at $\alpha = 0.05$ finds significant evidence for H_A. Thus, the difference between the two varieties is statistically significant, although it is not horticulturally important: The data indicate that variety 1 is better than variety 2, but also that it is not much better. The distinction between significance and importance for this example can be seen in Figure 7.6.3, which shows the confidence interval plotted on the $(\mu_1 - \mu_2)$-axis. Note that the confidence interval lies entirely to one side of zero and also entirely to one side of the "importance" threshold of 1.0.

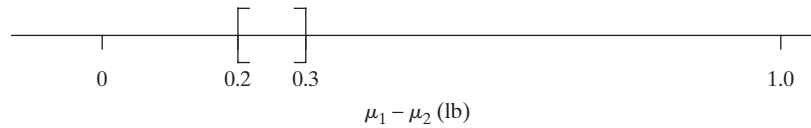


Figure 7.6.3 Confidence interval for Example 7.6.8

To further explore the relationship between significance and importance, let us consider other possible outcomes of the tomato experiment. Table 7.6.3 shows how the horticulturist would interpret various possible confidence intervals, still using the criterion that a difference must exceed 1.0 lb in order to be considered important.

Table 7.6.3 Interpretation of confidence intervals

95% confidence interval	Is the difference significant?	Is the difference important?
(0.2, 0.3)	Yes	No
(1.2, 1.3)	Yes	Yes
(0.2, 1.3)	Yes	Cannot tell
(−0.2, 0.3)	No	No
(−1.2, 1.3)	No	Cannot tell

Table 7.6.3 shows that a significant difference may or may not be important, and an important difference may or may not be significant. In practice, the assessment of importance using confidence intervals is a simple and extremely useful supplement to a test of hypothesis. ■

Exercises 7.6.1–7.6.8

7.6.1 A field trial was conducted to evaluate a new seed treatment that was supposed to increase soybean yield. When a statistician analyzed the data, the statistician found that the mean yield from the treated seeds was 40 lb/acre greater than that from control plots planted with untreated seeds. However, the statistician declared the difference to be "not (statistically) significant." Proponents of the treatment objected strenuously to the statistician's statement, pointing out that, at current market prices, 40 lb/acre would bring a tidy sum, which would be highly significant to the farmer. How would you answer this objection?[45]

7.6.2 In a clinical study of treatments for rheumatoid arthritis, patients were randomly allocated to receive either a standard medication or a newly designed medication. After a suitable period of observation, statistical analysis showed that there was no significant difference in the therapeutic response of the two groups, but that the incidence of undesirable side effects was significantly lower in the group receiving the new medication. The researchers concluded that the new medication should be regarded as clearly preferable to the standard medication, because it had been shown to be equally effective therapeutically and to produce fewer side effects. In what respect is the researchers' reasoning faulty? (Assume that the term "significant" refers to rejection of H_0 at $\alpha = 0.05$.)

7.6.3 There is an old folk belief that the sex of a baby can be guessed before birth on the basis of its heart rate. In an investigation to test this theory, fetal heart rates were observed for mothers admitted to a maternity ward. The results (in beats per minute) are summarized in the table.[46]

	HEART RATE (bpm)		
	n	Mean	SE
Males	250	137.21	0.62
Females	250	137.18	0.53

Construct a 95% confidence interval for the difference in population means. Does the confidence interval support the claim that the population mean sex difference (if any) in fetal heart rates is small and unimportant? (Use your own "expert" knowledge of heart rate to make a judgment of what is "unimportant.")

7.6.4 Coumaric acid is a compound that may play a role in disease resistance in corn. A botanist measured the concentration of coumaric acid in corn seedlings grown in the dark or in a light/dark photoperiod. The results (nmol acid per gm tissue) are given in the accompanying table.[47] [*Note:* Formula (6.7.1) yields 5.7 df.]

	DARK	PHOTOPERIOD
n	4	4
$\bar{y}$	106	102
s	21	27

Suppose the botanist considers the effect of lighting conditions to be "important" if the difference in means is 20%, that is, about 20 nmol/g. Based on a 95% confidence interval, do the preceding data indicate whether the true difference is "important"?

7.6.5 Repeat Exercise 7.6.4, assuming that the means and standard deviations are as given in the table, but that the sample sizes are 10 times as large (that is, $n = 40$ for "dark" and $n = 40$ for "photoperiod"). [*Note:* Formula (6.7.1) yields 73.5 df.]

7.6.6 Researchers measured the breadths, in mm, of the ankles of 460 youth (ages 11–16); the results are shown in the table.[48]

	MALES	FEMALES
n	244	216
$\bar{y}$	55.3	53.3
s	6.1	5.4

Calculate the sample effect size from these data.

7.6.7 As part of a large study of serum chemistry in healthy people, the following data were obtained for the serum concentration of uric acid in men and women aged 18–55 years.[49]

	SERUM URIC ACID (mmol/l)	
	MEN	WOMEN
n	530	420
$\bar{y}$	0.354	0.263
s	0.058	0.051

Construct a 95% confidence interval for the true difference in population means. Suppose the investigators feel that the difference in population means is "clinically

important" if it exceeds 0.08 mmol/l. Does the confidence interval indicate whether the difference is "clinically important"? [Note: Formula (6.7.1) yields 934 df.]

7.6.8 Repeat Exercise 7.6.7, assuming that the means and standard deviations are as given in the table, but that the sample sizes are only one-tenth as large (that is, 53 men and 42 women). [Note: Formula (6.7.1) yields 92 df.]

7.7 Planning for Adequate Power (Optional)

We have defined the power of a statistical test as

$$\text{Power} = \Pr\{\text{significant evidence for } H_A\} \text{ if } H_A \text{ is true}$$

To put this another way, the power of a test is the probability of obtaining data that provide statistically significant evidence for H_A when H_A is true.

Since the power is the probability of *not* making an error (of Type II), high power is desirable: If H_A is true, a researcher would like to find that out when conducting a study. But power comes at a price. All other things being equal, more observations (larger samples) bring more power, but observations cost time and money. In this section we explain how a researcher can rationally plan an experiment to have adequate power for the purposes of the research project and yet cost as little as possible.

Specifically, we will consider the power of the two-sample t test, conducted at significance level α. We will assume that the populations are normal with equal SDs, and we denote the common value of the SD by σ (that is, $\sigma_1 = \sigma_2 = \sigma$). It can be shown that in this case, for a given total sample size of $2n$, the power is maximized if the sample sizes are equal; thus we will assume that n_1 and n_2 are equal and denote the common value by n (that is, $n_1 = n_2 = n$).

Under the above conditions, the power of the t test depends on the following factors: (a) α; (b) σ; (c) n; and (d) $(\mu_1 - \mu_2)$. After briefly discussing each of these factors, we will address the all-important question of choosing the value of n.

Dependence of Power on α

In choosing α, one chooses a level of protection against Type I error. However, this protection is traded for vulnerability to Type II error. If, for example, one chooses $\alpha = 0.01$ rather than $\alpha = 0.05$, then one is requiring stronger evidence for H_A before choosing to claim there is significant evidence for H_A, and so is (perhaps unwittingly) also choosing to increase the risk of Type II error and reduce the power. Thus, there is an unavoidable trade-off between the risk of Type I error and the risk of Type II error.

Dependence on σ

The larger σ, the smaller the power (all other things being equal). Recall from Chapter 5 that the reliability of a sample mean is determined by the quantity

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$$

The larger σ is, the more variability there is in the sample mean. Thus, having a larger σ implies having samples that produce less reliable information about each

population mean, and so less power to discern a difference between them. In order to increase power, then, a researcher usually tries to design the investigation so as to have σ as small as possible. For example, a botanist will try to hold light conditions constant throughout a greenhouse area, a pharmacologist will use genetically identical experimental animals, and so on. Usually, however, σ cannot be reduced to zero; there is still considerable variation in the observations.

Dependence on n

The larger n, the higher the power (all other things being equal). If we increase n, we decrease $\sigma/\sqrt{n}$; this improves the precision of the sample means ($\overline{Y}_1$ and $\overline{Y}_2$). In addition, larger n gives more information about σ; this is reflected in a reduced critical value for the test (reduced because of more df). Thus, increasing n increases the power of the test in two ways.

Dependence on $(\mu_1 - \mu_2)$

In addition to the factors we have discussed, the power of the t test also depends on the actual difference between the population means, that is, on $(\mu_1 - \mu_2)$. This dependence is very natural, as illustrated by the following example.

Example 7.7.1

Heights of People In order to clearly illustrate the concepts, we consider a familiar variable, body height of people. Imagine what would happen if an investigator were to measure the heights of two random samples of eleven people each ($n = 11$), and then conduct a two-tailed t test at $\alpha = 0.05$.

(a) First, suppose that sample 1 consisted of 17-year-old males and sample 2 consisted of 17-year-old females. The two population means differ substantially; in fact, $(\mu_1 - \mu_2)$ is about 5 inches ($\mu_1 \approx 69.1$ and $\mu_2 \approx 64.1$ inches).[50] It can be shown (as we will see) that in this case the investigator has about a 99% chance of obtaining significant evidence for a difference (i.e., H_A) and correctly concluding that the males in the population of 17-year-olds are taller (on average) than the females.

(b) By contrast, suppose that sample 1 consisted of 17-year-old females and sample 2 consisted of 14-year-old females. The two population means differ, but by a modest amount; the difference is $(\mu_1 - \mu_2) = 0.6$ inches ($\mu_1 \approx 64.1$ and $\mu_2 \approx 63.5$ inches). It can be shown that in this case the investigator has less than a 10% chance of obtaining significant evidence of a difference (i.e, H_A); in other words, there is more than a 90% chance that the investigator will fail to detect the fact that 17-year-old girls are taller than 14-year-old girls. (In fact, it can be shown that there is a 29% chance that $\overline{Y}_1$ will be less than $\overline{Y}_2$—that is, there is a 29% chance that eleven 17-year-old girls chosen at random will be shorter on the average than eleven 14-year-old girls chosen at random!)

The contrast between cases (a) and (b) is not due to any change in the SDs; in fact, for each of the three populations the value of σ is about 2.5 inches. Rather, the contrast is due to the simple fact that, with a fixed n and σ, it is easier to detect a large difference than a small difference. ■

Planning a Study

Suppose an investigator is planning a study for which the t test will be appropriate. How shall she take into account all the factors that influence the power of the test? First consider the choice of significance level α. A simple approach is to begin by determining the cost of an adequately powerful study using a somewhat liberal choice (say, $\alpha = 0.05$ or 0.10). If that cost is not high, the investigator can consider reducing α (say, to 0.01) and see if an adequately powerful study is still affordable.

Suppose, then, that the investigator has chosen a working value of α. Suppose also that the experiment has been designed to reduce σ as far as practicable, and that the investigator has available an estimate or guess of the value of σ.

At this point, the investigator needs to ask herself about the magnitude of the difference she wants to detect. As we saw in Example 7.7.1, a given sample size may be adequate to detect a large difference in population means, but entirely inadequate to detect a small difference. As a more realistic example, an experiment using 5 rats in a treatment group and 5 rats in a control group might be large enough to detect a substantial treatment effect, while detection of a subtle treatment effect would require more rats (perhaps 30) in each group.

The preceding discussion suggests that choosing a sample size for adequate power is somewhat analogous to choosing a microscope: We need high resolving power if we want to see a very tiny structure; for large structures a hand lens will do. In order to proceed with planning the experiment, the investigator needs to decide how large an effect she is looking for.

Recall that in Section 7.7, we defined the effect size in a study as the difference between μ_1 and μ_2, expressed relative to the standard deviation of one of the populations. If, as we are assuming here, the two populations have the same standard deviation, σ, then the effect size is

$$\text{Effect size} = \frac{|\mu_1 - \mu_2|}{\sigma}$$

That is, the effect size is the difference in population means expressed relative to the common population SD. The effect size is a kind of "signal to noise ratio," where $(\mu_1 - \mu_2)$ represents the signal we want to detect and σ represents the background noise that tends to obscure the signal. Figure 7.7.1(a) shows two normal curves for which the effect size is 0.5; Figure 7.7.1(b) shows two normal curves for which the effect size is 4. Clearly, at a fixed sample size it is easier to detect the difference between the curves in graph (b) than it is in graph (a).

If α and the effect size have been specified, then the power of the t test depends only on the sample sizes (n). Table 5 at the end of the book shows the value of n

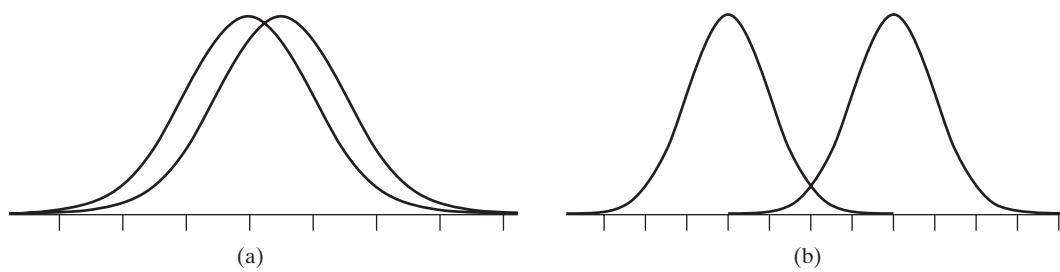


Figure 7.7.1 Normal distributions with an effect size (a) of 0.5 and (b) of 4

required in order to achieve a specified power against a specified effect size. Let us see how Table 5 applies to our familiar example of body height.

Example 7.7.2

Heights of People In Example 7.7.1, case (a), we considered samples of 17-year-old males and 17-year-old females. The effect size is

$$\frac{|\mu_1 - \mu_2|}{\sigma} = \frac{|69.1 - 64.1|}{2.5} = \frac{5}{2.5} = 2.0$$

For a two-tailed t test at $\alpha = 0.05$, Table 5 shows that the sample size required for a power of 0.99 is $n = 11$; this is the basis for the claim in Example 7.7.1 that the investigator has a 99% chance of detecting the difference between males and females. Figure 7.7.2 shows the two distributions being considered in Example 7.7.2. Suppose 100 researchers each conduct the following study. Take a random sample of eleven 17-year-old males and a random sample of eleven 17-year-old females, find the sample average heights of the two groups, and then conduct a two-tailed t test of H_0: $\mu_1 = \mu_2$ using $\alpha = 0.05$. We would expect 99 of the 100 researchers to find statistically significant evidence that the average heights of 17-year-old males and females differ (i.e., significant evidence for H_A). We would expect one of the 100 researchers to not find sufficient evidence for a difference, at the 0.05 level of significance. (So one researcher would make a Type II error.) ■

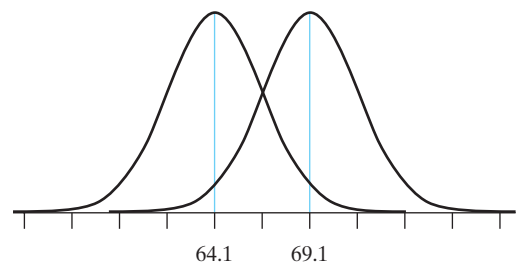


Figure 7.7.2 Height distributions for Example 7.7.2

As we have seen, in order to choose a sample size the researcher needs to specify not only the size of the effect she wishes to detect, but also how certain she wants to be of detecting it; that is, it is necessary to specify how much power is wanted. Since the power measures the protection against Type II error, the choice of a desired power level depends on the consequences that would result from a Type II error. If the consequences of a Type II error would be very unfortunate (for example, if a promising but risky cancer treatment is being tested on humans and a negative result would discredit the treatment so that it would never be tested again), then the researcher might specify a high power, say 0.95 or 0.99. But of course high power is expensive in terms of n. For much research, a Type II error is not a disaster, and a lower power such as 0.80 is considered adequate.

The following example illustrates a typical use of Table 5 in planning an experiment.

Example 7.7.3

Postpartum Weight Loss A group of scientists wished to investigate whether or not an Internet-based intervention program would help women lose weight after giving birth. One group of postpartum women was to be enrolled in an Internet-based program that provides weekly exercise and dietary guidance appropriate to their

time since giving birth, track their weight-loss progress, and establish an online forum for nutrition and exercise discussion with other recent mothers. Another group of postpartum women (the "control group") was to be given traditional written dietary and exercise guidelines by their doctors. The response variable for the study was to be the amount of weight lost at 12 months postpartum in kg. Previous studies have shown that at 12 months postpartum, the mean weight loss is about 3.6 kg with a standard deviation of 4.0 kg. (*Note*: A negative weight loss is a weight gain). The research team wanted to show at least a 50% improvement in weight loss for the Internet-intervention group; that is, they would like to show that the Internet-based program women lose at least 1.8 kg (50% of 3.6kg) more weight than the controls. They planned to conduct a one-tailed t-test at the 5% significance level. The team had to decide how many women (n) to put in each group.

The effect size that the team wanted to consider is

$$\frac{|\mu_1 - \mu_2|}{\sigma} = \frac{1.8}{4.0} = 0.45$$

For this effect size, and for a power of 0.80 with a one-tailed test at the 5% significance level, Table 5 yields $n = 62$, which means 62 women were needed in each group.

At this point, the research team had to consider questions, such as (1) Is it feasible to enroll 124 postpartum women (62 in each group) in the study? If not, then (2) Would they perhaps be willing to redefine the size of the difference between the groups that they considered to be important, in order to reduce the required n? With questions such as these, and repeated use of Table 5, they could finally decide on a firm value for n, or possibly decide to abandon the project because an adequate study would be too costly.

Normally the story ends here, but there was an extra wrinkle in the planning of this study: The research team knew from experience that about 20% of the women enrolled in these types of studies would drop out, for one reason or another, before the study ended. (There is no formula or table that tells one how many subjects will drop out of a study such as this. Here the only guide is experience.) In this case, the research team planned to enroll 150 women (a little more than 20% extra, 13 women in each group), in order to allow for some attrition and still end up with enough data so that they would have the power they wanted. [51] ■

Exercises 7.7.1–7.7.11

7.7.1 One measure of the meat quality of pigs is backfat thickness. Suppose two researchers, Jones and Smith, are planning to measure backfat thickness in two groups of pigs raised on different diets. They have decided to use the same number (n) of pigs in each group, and to compare the mean backfat thickness using a two-tailed t test at the 5% significance level. Preliminary data indicate that the SD of backfat thickness is about 0.3 cm.

When the researchers approach a statistician for help in choosing n, she naturally asks how much difference they want to detect. Jones replies, "If the true difference is 1/4 cm or more, I want to be reasonably sure of rejecting H_0." Smith replies, "If the true difference is 1/2 cm or more, I want to be very sure of rejecting H_0."

If the statistician interprets "reasonably sure" as 80% power, and "very sure" as 95% power, what value of n will she recommend

(a) to satisfy Jones's requirement?

(b) to satisfy Smith's requirement?

7.7.2 Refer to the brain NE data of Example 7.2.1. Suppose you are planning a similar experiment; you will study the effect of LSD (rather than toluene) on brain NE. You anticipate using a two-tailed t test at $\alpha = 0.05$. Suppose you have decided that a 10% effect (increase or decrease in mean NE) of LSD would be important, and so you want to have good power (80%) to detect a difference of this magnitude.

(a) Using the data of Example 7.2.1 as a "pilot study," determine how many rats you should have in each group. (The mean NE in the control group in Example 7.2.1 is 444.2 ng/g and the SD is = 69.6 ng/g.)

(b) If you were planning to use a one-tailed t test, what would be the required number of rats?

7.7.3 Suppose you are planning a greenhouse experiment on growth of pepper plants. You will grow n individually potted seedlings in standard soil and another n seedlings in specially treated soil. After 21 days, you will measure Y = total stem length (cm) for each plant. If the effect of the soil treatment is to increase the population mean stem length by 2 cm, you would like to have a 90% chance of rejecting H_0 with a one-tailed t test. Data from a pilot study (such as the data in Exercise 2.62) on 15 plants grown in standard soil give $\bar{y} = 12.5$ cm and $s = 0.8$ cm.

(a) Suppose you plan to test at $\alpha = 0.05$. Use the pilot information to determine what value of n you should use.

(b) What conditions are necessary for the validity of the calculation in part (a)? Which of these can be checked (roughly) from the data of the pilot study?

(c) Suppose you decide to adopt a more conservative posture and test at $\alpha = 0.01$. What value of n should you use?

7.7.4 Diastolic blood pressure measurements on American men aged 18–44 years follow approximately a normal curve with $\mu = 81$ mm Hg and $\sigma = 11$ mm Hg. The distribution for women aged 18–44 is also approximately normal with the same SD but with a lower mean: $\mu = 75$ mm Hg.[52] Suppose we are going to measure the diastolic blood pressure of n randomly selected men and n randomly selected women in the age group 18–44 years. Let E be the event that the difference between men and women will be found statistically significant by a t test. How large must n be in order to have $\Pr\{E\} = 0.9$

(a) if we use a two-tailed test at $\alpha = 0.05$?

(b) if we use a two-tailed test at $\alpha = 0.01$?

(c) if we use a one-tailed test (in the correct direction) at $\alpha = 0.05$?

7.7.5 Suppose you are planning an experiment to test the effect of a certain drug treatment on drinking behavior in the rat. You will use a two-tailed t test to compare a treated group of rats against a control group; the observed variable will be Y = one-hour water consumption after 23-hour deprivation. You have decided that, if the effect of the drug is to shift the population mean consumption by 2 ml or more, then you want to have at least an 80% chance of finding significant evidence for H_A at the 5% significance level.

(a) Preliminary data indicate that the SD of Y under control conditions is approximately 2.5 ml. Using this as a guess of σ, determine how many rats you should have in each group.

(b) Suppose that, because the calculation of part (a) indicates a rather large number of rats, you consider modifying the experiment so as to reduce σ. You find that, by switching to a better supplier of rats and by improving lab procedures, you could cut the SD in half; however, the cost of each observation would be doubled. Would these measures be cost-effective; that is, would the modified experiment be less costly?

7.7.6 Data from a large study indicate that the serum concentration of lactate dehydrogenase (LD) is higher in men than in women. (The data are summarized in Example 7.6.1.) Suppose Dr. Sanchez proposes to conduct his own study to replicate this finding; however, because of limited resources Sanchez can enlist only 35 men and 35 women for his study. Supposing that the true difference in population means is 4 U/l and each population SD is 10 U/l, what is the probability that Sanchez will be successful? Specifically, find the probability that Sanchez will reject H_0 with a one-tailed t test at the 5% significance level.

7.7.7 Refer to the painkiller study of Exercise 7.5.10. That study included 25 observations in each treatment group and showed an effect size of about 0.5. If this is the true population effect size, what is the (approximate) chance of finding a significant difference between the mean effectiveness of the two drugs in an experiment of this size (i.e., with samples of 25 each)?

7.7.8 Refer to the painkiller study of Exercise 7.5.10. In that study, the evidence favoring the drug was marginally significant ($0.025 < P < 0.05$). Suppose Dr. Williams is planning a new study on the same drug in order to try to replicate the original findings, that is, to show the drug to be effective. She will consider this study successful if she rejects H_0 with a one-tailed test at $\alpha = 0.05$. In the original study, the difference between the treatment means was about half a standard deviation [$(32 - 25)/13 \approx 0.5$]. Taking this as a provisional value for the effect size, determine how many patients Williams should have in each group in order for her chance of success to be

(a) 80% (b) 90%

(*Note:* This problem illustrates that surprisingly large sample sizes may be required to make a replication study worthwhile, especially if the original findings were only marginally significant.)

7.7.9 Consider comparing two normally distributed distributions for which the effect size of the difference is

(a) 3 (b) 1

In each case, draw a sketch that shows how the distributions overlap. (See Figure 7.2.1.)

7.7.10 An animal scientist is planning an experiment to evaluate a new dietary supplement for beef cattle. One group of cattle will receive a standard diet and a second group will receive the standard diet plus the supplement.

The researcher wants to have 90% power to detect an increase in mean weight gain of 20 kg, using a one-tailed t test at $\alpha = 0.05$. Based on previous experience, he expects the SD to be 17 kg. How many cattle does he need for each group?

7.7.11 A researcher is planning to conduct a study that will be analyzed with a two-tailed t test at the 5% significance level. She can afford to collect 20 observations in each of the two groups in her study. What is the smallest effect size for which she has at least 95% power?

7.8 Student's *t*: Conditions and Summary

In the preceding sections we have discussed the comparison of two means using classical methods based on Student's t distribution. In this section we describe the conditions on which these methods are based. In addition, we summarize the methods for convenient reference.

Conditions

The t test and confidence interval procedures we have described are appropriate if the following conditions* hold:

1. *Conditions on the design of the study*
   (a) It must be reasonable to regard the data as random samples from their respective populations. The populations must be large relative to their sample sizes. The observations within each sample must be independent.
   (b) The two samples must be independent of each other.
2. *Condition on the form of the population distributions*
   The sampling distributions of $\overline{Y}_1$ and $\overline{Y}_2$ must be (approximately) normal. This can be achieved via normality of the populations or by appealing to the Central Limit Theorem (recall Section 6.5) if the populations are nonnormal but the sample sizes are large, where "largeness" depends on the degree of nonnormality of the populations. In many practical situations, moderate sample sizes (say, $n_1 = 20, n_2 = 20$) are quite "large" enough. However, we always need to be aware that one or two extreme outliers can have a great effect on the results of any statistical procedure, including the t test.

Verification of Conditions

A check of the preceding conditions should be a part of every data analysis.

A check of condition 1(a) would proceed as for a confidence interval (Section 6.5), with the researcher looking for biases in the experimental design and verifying that there is no hierarchical structure within each sample.

Condition 1(b) means that there must be no pairing or dependency between the two samples. The full meaning of this condition will become clear in Chapters 8 and 9.

Sometimes it is known from previous studies whether the populations can be considered to be approximately normal. In the absence of such information, the normality requirement can be checked by making histograms, normal probability plots,

*Many authors use the word "assumptions" where we are using the word "conditions."

or Shaprio–Wilk normality tests for each sample separately. Fortunately, the t test is fairly robust against departures from normality.[53] Usually, only a rather conspicuous departure from normality (outliers, or long straggly tails) should be cause for concern. Moderate skewness has very little effect on the t test, even for small samples.

Consequences of Inappropriate Use of Student's t

Our discussion of the t test and confidence interval (in Sections 7.3–7.8) was based on the conditions (1) and (2). Violation of the conditions may render the methods inappropriate.

If the conditions are not satisfied, then the t test may be inappropriate in two possible ways:

1. It may be invalid in the sense that the actual risk of Type I error is larger than the nominal significance level α. (To put this another way, the P-value yielded by the t test procedure may be inappropriately small.)
2. The t test may be valid, but less powerful than a more appropriate test.

If the design includes hierarchical structures that are ignored in the analysis, the t test may be seriously invalid. If the samples are not independent of each other, the usual consequence is a loss of power.

One fairly common type of departure from the condition of normality is for one or both populations to have long straggly tails. The effect of this form of nonnormality is to inflate the SE, and thus to rob the t test of power.

Inappropriate use of confidence intervals is analogous to that for t tests. If the conditions are violated, then the confidence interval may not be valid (i.e., too narrow for the prescribed level of confidence), or it may be valid but wider than necessary.

Other Approaches

Because methods based on Student's t distribution are not always the most appropriate, statisticians have devised other methods that serve similar purposes. One of these is the Wilcoxon–Mann–Whitney test, which we will describe in Section 7.10. Another approach to the difficulty is to transform the data, for instance, to analyze $\log(Y)$ or $\ln(Y)$ instead of Y itself.

Example 7.8.1

Tissue Inflammation Researchers took skin samples from 10 patients who had breast implants and from a control group of 6 patients. They recorded the level of interleukin-6 (in pg/ml/10 g of tissue), a measure of tissue inflammation, after each tissue sample was cultured for 24 hours. Table 7.8.1 shows the data.[54] Parallel dotplots of these data shown in Figure 7.8.1(a) and normal probability plots shown in Figure 7.8.2(a) indicate that the distributions are severely skewed, so a transformation is needed before Student's t procedure can be used. Taking the base 10 logarithm of each observation produces the values shown in the right-hand columns of Table 7.8.1 and in Figure 7.8.1(b). The normal probability plots in Figure 7.8.2(b) show that the condition of normality is met after the data have been transformed to log scale. Thus, we will conduct an analysis of the data in log scale. That is, we will test

$$H_0: \mu_1 = \mu_2$$

Table 7.8.1 Interleukin-6 levels of breast implant patients and control patients

	Original data		Log scale	
	Breast implant patients	Control patients	Breast implant patients	Control patients
	231	35,324	2.364	4.548
	308,287	12,457	5.489	4.095
	33,291	8,276	4.522	3.918
	124,550	44	5.095	1.643
	17,075	278	4.232	2.444
	22,955	840	4.361	2.924
	95,102		4.978	
	5,649		3.752	
	840,585		5.925	
	58,924		4.770	
$\bar{y}$	150,665	9,537	4.549	3.262
s	259,189	13,613	0.992	1.111

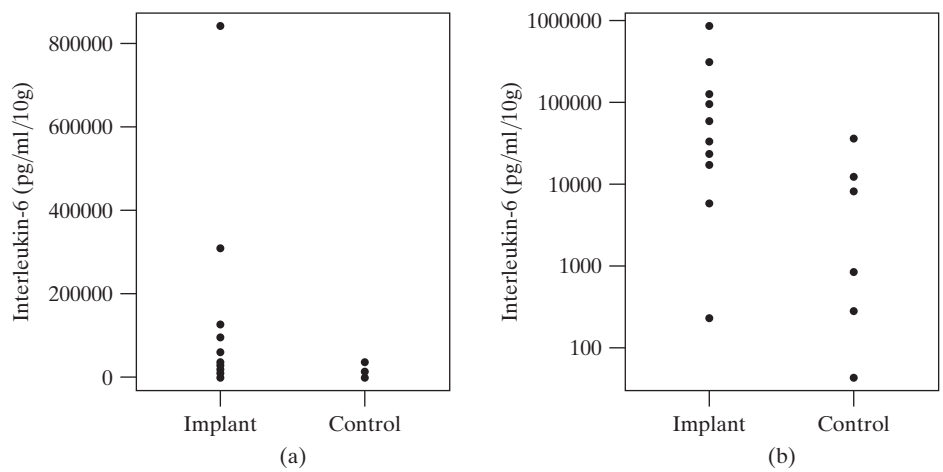


Figure 7.8.1 Dotplots of tissue inflammation data from Example 7.8.1 (a) in the original scale; (b) in log scale

against

$$H_A: \mu_1 \neq \mu_2$$

where μ_1 is the population mean of the log of interleukin-6 level for breast implant patients and μ_2 is the population mean of the log of interleukin-6 level for control patients. Suppose we choose $\alpha = 0.10$. The test statistic is

$$t_s = \frac{(4.549 - 3.262)}{0.553} = 2.33$$

Formula (6.7.1) yields df $= 9.7$. The P-value for the test is 0.045. Thus, we have evidence, at the 0.10 level of significance (and at the 0.05 level, as well), that the mean log interleukin-6 level is higher in the breast implant population than in the control population. ■

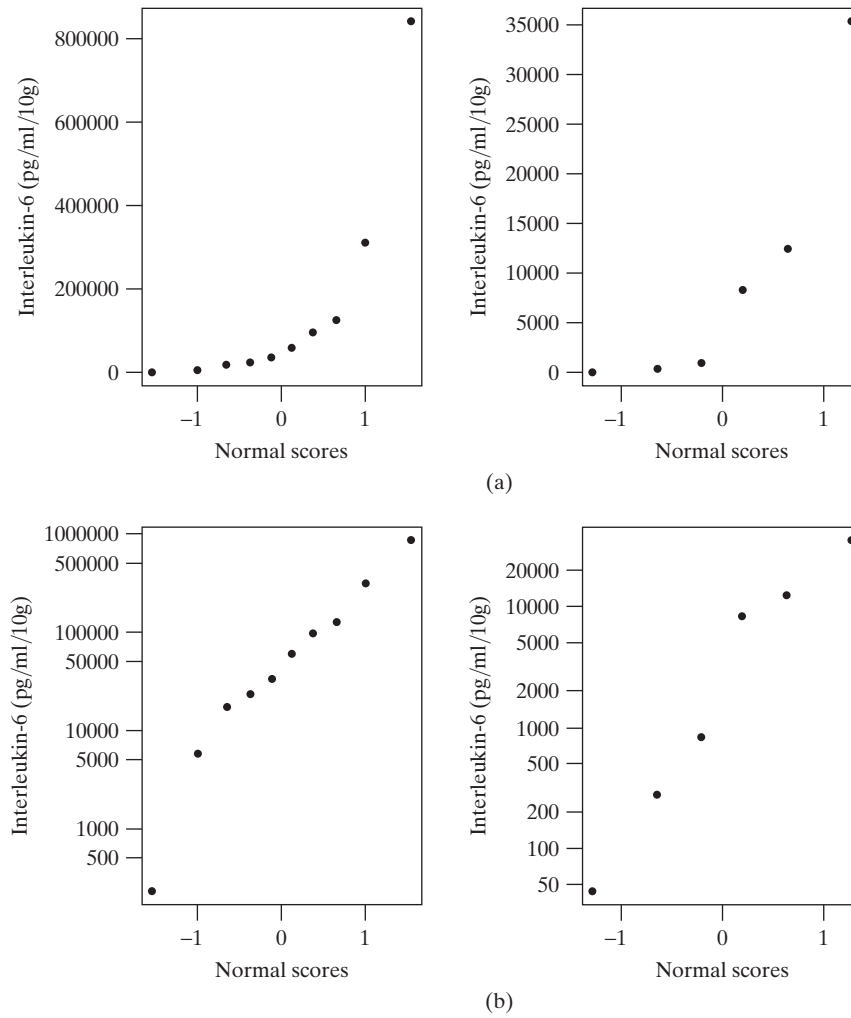


Figure 7.8.2 Normal probability plots of tissue inflammation data from Example 7.8.1 (a) in the original scale and (b) in log scale

Summary of *t* Test Mechanics

For convenient reference, we summarize the mechanics for Student's *t* test of equality of the means of independent samples.

t Test

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 \neq \mu_2 \text{ (nondirectional)}$$

$$H_A: \mu_1 < \mu_2 \text{ (directional)}$$

$$H_A: \mu_1 > \mu_2 \text{ (directional)}$$

$$\text{Test statistic: } t_s = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}}$$

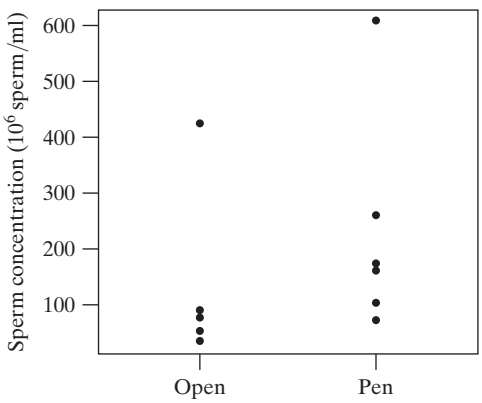
$$P\text{-value} = \text{tail area under Student's } t \text{ curve with}$$

$$\text{df} = \frac{(\text{SE}_1^2 + \text{SE}_2^2)^2}{\text{SE}_1^4/(n_1 - 1) + \text{SE}_2^4/(n_2 - 1)}$$

Nondirectional H_A: P-value = two-tailed area beyond t_s and $-t_s$
Directional H_A: Step 1. Check directionality.
Step 2. P-value = single-tail area beyond t_s
Decision: Significant evidence for H_A if P-value $\leq \alpha$

Exercises 7.8.1–7.8.2

7.8.1 To determine if the environment can affect sperm quality and production in cattle, a researcher randomly assigned 13 bulls to one of two environments. Six were raised in an open range environment while 7 were reared in a smaller penned environment. The following plot displays the sperm concentrations (millions of sperm/ml) of semen samples from the 13 bulls.[55]



(a) Using the preceding graph to justify your answer, would the use of Student's t method be appropriate to compare the mean sperm concentrations under these two experimental conditions?

(b) How would your answer to (a) change if the data consisted of 60 and 70 specimens rather than 6 and 7?

(c) The Shapiro–Wilk test of normality yields P-values of 0.0012 and 0.0139 for the Open and Pen data, respectively. How do these results support or refute your response to part (a)?

(d) How might a transformation help you analyze these data?

7.8.2 Refer to the serotonin data of Exercise 7.2.7. On what grounds might an objection be raised to the use of the t test on these data? (*Hint:* For each sample, calculate the SD and compare it to the sample mean.)

7.9 More on Principles of Testing Hypotheses

Our study of the t test has illustrated some of the general principles of statistical tests of hypotheses. In the remainder of this book we will introduce several other types of tests besides the t test.

A General View of Hypothesis Tests

A typical statistical test involves a null hypothesis H_0, an alternative hypothesis, or research hypothesis, H_A, and a test statistic that measures deviation or discrepancy of the data from H_0. The sampling distribution of the test statistic, under the assumption that H_0 is true, is called the **null distribution** of the test statistic. (If we are conducting a randomization test as in Section 7.1, then the null distribution is the distribution of all possible differences in sample means due to random assignment of observations to groups, such as that shown in Table 7.1.2; as another example, if we are conducting a t test, then the null distribution of the t statistic t_s is—under certain conditions—a Student's t distribution.) The null distribution indicates how much the test statistic can be expected to deviate from H_0 because of chance alone.

In testing a hypothesis, we assess the evidence against H_0 (and in favor of H_A) by locating the test statistic within the null distribution; the P-value is a measure of

this location, which indicates the degree of compatibility between the data and H_0. The dividing line between compatibility and incompatibility is specified by an arbitrarily chosen significance level α. The decision whether to claim there is significant evidence for H_A is made according to the following rule:

Reject H_0 if P-value $\leq \alpha$.

When a computer is not available, we will not be able to calculate the P-value exactly but will bracket it using a table of critical values. If H_A is directional, the bracketing of P-value is a two-step procedure.

Every test of a null hypothesis H_0 has its associated risks of Type I error (finding significant evidence for H_A when H_0 is true) and Type II error (not finding significant evidence for H_A when H_A is true). The risk of Type I error is always limited by the chosen significance level, α:

$\Pr\{\text{reject } H_0\} \leq \alpha$ if H_0 is true

Thus, the hypothesis testing procedure treats the Type I error as the one to be most stringently guarded against. By contrast, the power of a test can be quite low, and equivalently the risk of Type II error can be quite large, if the samples are small.

How Are H_0 and H_A Chosen?

A common difficulty when first studying hypothesis testing is figuring out what the null and alternative hypotheses should be. In general, the null hypothesis represents the status quo—what one would believe, by default, unless the data showed otherwise.* Typically the alternative hypothesis is a statement that the researcher is trying to establish; thus H_A is also referred to as the research hypothesis. For example, if we are testing a new drug against a standard drug, the research hypothesis is that the new drug is better than the standard drug, while the null hypothesis is that the new drug is no different than the standard—in the absence of evidence, we would expect the two drugs to be equally effective. The typical null hypothesis, H_0: $\mu_1 = \mu_2$, states that the two population means are equal and that any difference between the sample means is simply due to chance error in the sampling process. The alternative hypothesis is that there *is* a difference between the drugs, so that any observed difference in sample means is due to a real effect, rather than being due to chance error alone. We conclude that we have statistically significant evidence for the research hypothesis if the data show a difference in sample means beyond what can reasonably be attributed to chance.

Here are other examples: If we are comparing men and women on some attribute, the usual null hypothesis is that there is no difference, on average, between men and women; if we are studying a measure of biodiversity in two environments, the usual null hypothesis is that the biodiversities of the two environments are equal, on average; if we are studying two diets, the usual null hypothesis is that the diets produce the same average response.

Another Look at P-Value

In order to place P-value in a general setting, let us consider some verbal interpretations of P-value.

*This general rule is not always true; it is provided only as a guideline.

First we revisit the randomization test. For a nondirectional H_A the P-value is the proportion of all randomizations that results in a difference of sample means that is as large, or larger than, the difference that was observed in the actual study. Thus we can define the P-value as follows:

> The P-value of the data is the probability (assuming H_0 is true) of getting a result as extreme as, or more extreme than, the result that was actually observed.

To put this another way,

> The P-value is the probability that, if H_0 were true, a result would be obtained that would deviate from H_0 as much as (or more than) the actual data do.

Now consider the t test. For a nondirectional H_A, we have defined the P-value to be the two-tailed area under the Student's t curve beyond the observed value of t_s.

Actually, these descriptions of P-value are a bit too limited. The P-value actually depends on the nature of the alternative hypothesis. When we are performing a t test against a *directional* alternative, the P-value of the data is (if the observed deviation is in the direction of H_A) only a *single-tailed* area beyond the observed value of t_s. The more general definition of P-value is the following:

> The P-value of the data is the probability (assuming H_0 is true) of getting a result as deviant as, or more deviant than, the result actually observed—where deviance is measured as discrepancy from H_0 in the direction of H_A.

The P-value measures how easily the observed deviation could be explained as chance variation rather than by the alternative explanation provided by H_A. For example, if the t test yields a P-value of $P = 0.036$ for our data, then we may say that *if H_0 were true* we would expect data to deviate from H_0 as much as our data did only 3.6% of the time (in the meta-study).

Another definition of P-value that is worth thinking about is the following:

> The P-value of the data is the value of α for which H_0 would just barely be rejected, using those data.

To interpret this definition, imagine that a research report that includes a P-value is read by a number of interested scientists. The scientists who are quite skeptical of H_A might require very strong evidence before being convinced and thus would use a very conservative decision threshold, such as $\alpha = 0.001$; the scientists who are more favorably disposed toward H_A might require only weak evidence and thus use a liberal value such as $\alpha = 0.10$. The P-value of the data determines the point, within this spectrum of opinion, that separates those who find the data to be convincing in favor of H_A and those who do not. Of course, if the P-value is large, for instance $P = 0.40$, then presumably no reasonable person would reject H_0 and be convinced of H_A.

As the preceding discussion shows, the P-value does not describe all facets of the data, but relates only to a test of a particular null hypothesis against a particular alternative. In fact, we will see that the P-value of the data also depends on which statistical test is used to test a given null hypothesis. For this reason, when describing in a scientific report the results of a statistical test, it is best to report the P-value (exactly, if possible), the name of the statistical test, and whether the alternative hypothesis was directional or nondirectional.

We repeat here, because it applies to any statistical test, the principle expounded in Section 7.6: The P-value is a measure of the strength of the evidence against

H_0, but the P-value does *not* reflect the *magnitude* of the discrepancy between the data and H_0. The data may deviate from H_0 only slightly, yet if the samples are large, the P-value may be quite small. By the same token, data that deviate substantially from H_0 can nevertheless yield a large P-value. The P-value alone does *not* indicate whether a scientific finding is important.

Interpretation of Error Probabilities

A common mistake is to interpret the P-value as the probability that the null hypothesis is true. A related misconception is the belief that, if we find significant evidence for H_A (for example) at the 5% significance level, then the probability that H_0 is true is 5%. These interpretations are not correct.* This point can be illustrated by an analogy with medical diagnosis.

In applying a diagnostic test for an illness, the null hypothesis is that the person is healthy—this is what we will believe unless the medical test indicates otherwise. Two types of error are possible: A healthy individual may be diagnosed as ill (false positive) or an ill individual may be diagnosed as healthy (false negative). Trying out a diagnostic test on individuals *known* to be healthy or ill will enable us to estimate the proportions of these groups who will be misdiagnosed; yet this information alone will not tell us what proportion of all positive diagnoses are false diagnoses. These ideas are illustrated numerically in the next example.

Example 7.9.1

Medical Testing Suppose a medical test is conducted to detect an illness. Further, suppose that 1% of the population has the illness in question. If the test indicates that the disease is present, we reject the null hypothesis that the person is healthy. If H_0 is true, then this is a Type I error—a false positive. If the test indicates that the disease is not present, we have a lack of significant evidence for H_A (illness). Suppose that the test has an 80% chance of detecting the disease if the person has it (this is analogous to the power of a hypothesis test being 80%) and a 95% chance of correctly indicating that the disease is absent if the person really does not have the disease (this is analogous to a 5% Type I error rate). Figure 7.9.1 shows a probability tree for this situation, with bold lines indicating the two ways in which the test result can be positive (i.e., the two ways that H_0 can be rejected).

Now suppose that 100,000 persons are tested and that 1,000 of them (1%) actually have the illness. Then we would expect results like those given in Table 7.9.1, with 5,750 persons testing positive (which is like finding significant evidence for H_A 5,750 times). Of these, 4,950 are false positives. Put another way, the proportion of the time that H_0 is true, given that we found significant evidence for H_A, is $\frac{4{,}950}{5{,}750} \approx 0.86$, which is quite different from 0.05; this startlingly high proportion of false positives is due to the rarity of the disease. (The proportion of times that there is significant evidence for H_A, given that H_0 is true, is $\frac{4{,}950}{99{,}000} = 0.05$, as expected, but that is a different conditional probability. Pr{A given B} $\neq$ Pr{B given A}: The probability of rainfall, given that there are thunder and lightning, is not the same as the probability of thunder and lightning, given that it is raining.) ■

*In fact, the probability that H_0 is true cannot be calculated at all within the standard, "frequentist" approach to hypothesis testing. Pr {H_0 is true} *can* be calculated if one uses what are known as Bayesian methods, which are beyond the scope of this book.

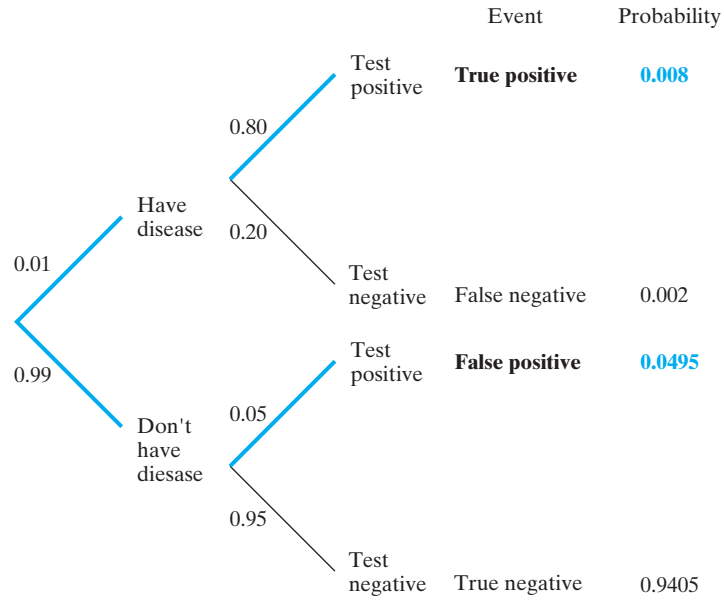


Figure 7.9.1 Probability tree for medical testing example

Table 7.9.1 Hypothetical results of medical test of 100,000 persons

		True situation		
		Healthy (H_0 true)	Ill (H_A true)	Total
TEST RESULT	Negative (lack of significant evidence for H_A)	94,050	200	94,250
	Positive (significant evidence for H_A)	4,950	800	5,750
	Total	99,000	1,000	100,000

The risk of Type I error is a probability computed *under the assumption that* H_0 *is true*; similarly, the risk of a Type II error is computed assuming that H_A is true. If we have a well-designed study with adequate sample sizes, both of these probabilities will be small. We then have a good test procedure in the same sense that the medical test is a good diagnostic procedure. But this does not in itself guarantee that most of the null hypotheses we reject are in fact false, or that most of those we do not reject are in fact true. The validity or nonvalidity of such guarantees would depend on an unknown and unknowable quantity—namely, the proportion of true null hypotheses among all null hypotheses that are tested (which is analogous to the incidence of the illness in the medical test scenario).

Perspective

We should mention that the philosophy of statistical hypothesis testing that we have explained in this chapter is not shared by all statisticians. The view presented here, which is called the **frequentist view**, is widely used in scientific research. An alternative view, the **Bayesian view**, incorporates not only the data observed in the study at hand, but also the information that the researcher has from previous, related studies.

In the past, many Bayesian techniques were not practical due to the complexity of the mathematics that they require. However, greater computing power and improved software have made Bayesian methods more popular in recent years.

Exercise 7.9.1

7.9.1 Suppose we have conducted a t test, with $\alpha = 0.05$, and the P-value is 0.04. For each of the following statements, say whether the statement is true or false and explain why.

(a) There is a 4% chance that H_0 is true.

(b) We reject H_0 with $\alpha = 0.05$.

(c) We should reject H_0, and if we repeated the experiment, there is a 4% chance that we would reject H_0 again.

(d) If H_0 is true, the probability of getting a test statistic at least as extreme as the value of the t_s that was actually obtained is 4%.

7.10 The Wilcoxon-Mann-Whitney Test

The **Wilcoxon-Mann-Whitney test** is used to compare two independent samples.* It is a competitor to the t test, but unlike the t test, the Wilcoxon-Mann-Whitney test is valid even if the population distributions are not normal. The Wilcoxon-Mann-Whitney test is therefore called a **distribution-free** type of test. In addition, the Wilcoxon-Mann-Whitney test does not focus on any particular parameter such as a mean or a median; for this reason it is called a **nonparametric** type of test.

Statement of H_0 and H_A

Let us denote the observations in the two samples by Y_1 and Y_2. A general statement of the null and alternative hypotheses of a Wilcoxon-Mann-Whitney test are

H_0: The population distributions of Y_1 and Y_2 are the same.

H_A: The population distribution of Y_1 is shifted from the population distribution of Y_2 (i.e., Y_1 tends to be either greater or less than Y_2).

In practice, it is more natural to state H_0 and H_A in words suitable to the particular application, as illustrated in Example 7.10.1.

Example 7.10.1

Soil Respiration Soil respiration is a measure of microbial activity in soil, which affects plant growth. In one study, soil cores were taken from two locations in a forest: (1) under an opening in the forest canopy (the "gap" location) and (2) at a nearby area under heavy tree growth (the "growth" location). The amount of carbon dioxide given off by each soil core was measured (in mol CO_2/g soil/hr). Table 7.10.1 contains the data.[56]

An appropriate null hypothesis could be stated as

H_0: The populations from which the two samples were drawn have the same distribution of soil respiration.

*The test presented here is was developed by Wilcoxon in a 1945 article. Mann and Whitney, in a 1947 article, elaborated on the test, which can be conducted in two mathematically equivalent ways. Thus, some books and some computer programs implement the test in a different fashion than the way it is presented here. Also note that some books refer to this as the Wilcoxon test, some as the Mann–Whitney test, and some (including this text) as the Wilcoxon-Mann-Whitney test.

Table 7.10.1 Soil respiration data (mol CO_2/g soil/hr) from Example 7.10.1

Growth	Gap
17 20 170 315 22 190 64	22 29 13 16 15 18 14 6

or, more informally, as

H_0: The gap and growth areas do not differ with respect to soil respiration.

A nondirectional alternative could be stated as

H_A: The distribution of soil respiration rates tends to be higher in one of the two populations.

or the alternative hypothesis might be directional, for example,

H_A: Soil respiration rates tend to be greater in the growth area than there are in the gap area. ■

Applicability of the Wilcoxon-Mann-Whitney Test

Figure 7.10.1 shows dotplots of the soil respiration data from Example 7.10.1; Figure 7.10.2 shows normal probability plots of these data. The growth distribution

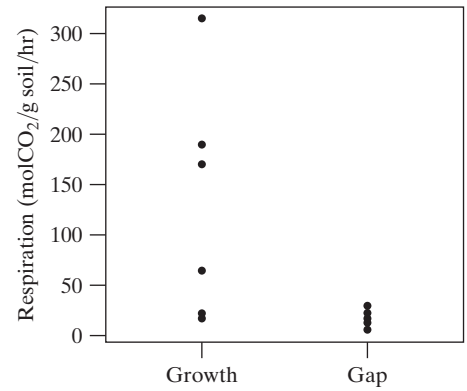


Figure 7.10.1 Dotplots of the soil respiration data from Example 7.10.1

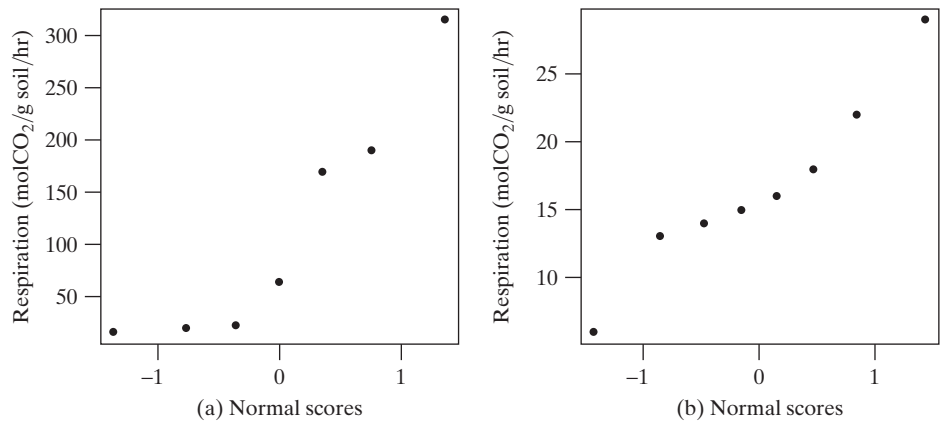


Figure 7.10.2 Normal probability plots of (a) the growth data and (b) the gap data from Example 7.10.1

is skewed to the right, whereas the gap distribution is slightly skewed to the left. If both distributions were skewed to the right, we could apply a transformation to the data. However, any attempt to transform the growth distribution, such as taking logarithms of the data, will make the skewness of the gap distribution worse. Hence, the t test is not applicable here. The Wilcoxon-Mann-Whitney test does not require normality of the distributions.

Method

The Wilcoxon-Mann-Whitney test statistic, which is denoted U_s, measures the degree of separation or shift between two samples. A large value of U_s indicates that the two samples are well separated, with relatively little overlap between them. Critical values for the Wilcoxon-Mann-Whitney test are given in Table 6 at the end of this book. The following example illustrates the Wilcoxon-Mann-Whitney test.

Example 7.10.2

Soil Respiration Let us carry out a Wilcoxon-Mann-Whitney test on the biodiversity data of Example 7.10.1.

1. The value of U_s depends on the relative positions of the Y_1's and the Y_2's. The first step in determining U_s is to arrange the observations in increasing order, as is shown in Table 7.10.2.
2. We next determine two counts, K_1 and K_2, as follows:
   (a) *The K_1 count* For each observation in sample 1, we count the number of observations in sample 2 that are smaller in value (that is, to the left). We count 1/2 for each tied observation. In the above data, there are five Y_2's less than the first Y_1, there are six Y_2's less than the second Y_1, there are six Y_2's less than the third Y_1 and one equal to it, so we count 6 1/2. So far we have counts of 5, 6, and 6.5. Continuing in a similar way, we get further counts of 8, 8, 8, and 8. All together there are seven counts, one for each Y_1. The sum of all seven counts is $K_1 = 49.5$.
   (b) *The K_2 count* For each observation in sample 2, we count the number of observations in sample 1 that are smaller in value, counting 1/2 for ties.

Table 7.10.2 Wilcoxon-Mann-Whitney calculations for Example 7.10.2

Number of gap observations that are smaller	Y_1 Growth data	Y_2 Gap data	Number of growth observations that are smaller
5	17	6	0
6	20	13	0
6.5	22	14	0
8	64	15	0
8	170	16	0
8	190	18	1
8	315	22	2.5
		29	3
$K_1 = 49.5$			$K_2 = 6.5$

This gives counts of 0, 0, 0, 0, 0, 1, 2.5, and 3. The sum of these counts is $K_2 = 6.5$.

(c) *Check* If the work is correct, the sum of K_1 and K_2 should be equal to the product of the sample sizes:

$$K_1 + K_2 = n_1 n_2$$
$$49.5 + 6.5 = 7 \times 8$$

3. The test statistic U_s is the larger of K_1 and K_2. In this example, $U_s = 49.5$.
4. To determine the P-value, we consult Table 6 with n = the larger sample size, and n' = the smaller sample size. In the present case, $n = 8$ and $n' = 7$. Values from Table 6 are reproduced in Table 7.10.3.

Table 7.10.3 Values from Table 6 for $n = 8$, $n' = 7$

40 *0.189*	**44** *0.093*	**46** *0.054*	**47** *0.040*	**48** *0.021*	**49** *0.014*	**50** *0.009*

Let us test H_0 against a nondirectional alternative at significance level $\alpha = 0.05$. From Table 7.10.3, we note that when $U_s = 49$, the P-value is 0.014 and when $U_s = 50$, the P-value is 0.009; since $49 < U_s < 50$, the P-value is between 0.009 and 0.014 and thus there is significant evidence for H_A. There is sufficient evidence to conclude that soil respiration rates are different in the gap and growth areas. ■

As Example 7.10.2 illustrates, Table 6 can be used to bracket the P-value for the Wilcoxon-Mann-Whitney test just as Table 4 is used for the t test. If the observed U_s value is not given, then one simply locates the values that bracket the observed U_s. One then brackets the P-value by the corresponding column headings.

Directionality For the t test, one determines the directionality of the data by seeing whether $\overline{Y}_1 > \overline{Y}_2$ or $\overline{Y}_1 < \overline{Y}_2$. Similarly, one can check directionality for the Wilcoxon-Mann-Whitney test by comparing K_1 and K_2: $K_1 > K_2$ indicates a trend for the Y_1's to be larger than the Y_2's, while $K_1 < K_2$ indicates the opposite trend. Often, however, this formal comparison is unnecessary; a glance at a graph of the data is enough.

Directional Alternative If the alternative hypothesis H_A is directional rather than nondirectional, the Wilcoxon-Mann-Whitney procedure must be modified. As with the t test, the modified procedure has two steps and the second step involves halving the nondirectional P-value to obtain the directional P-value.

Step 1 Check directionality—see if the data deviate from H_0 in the direction specified by H_A.

(a) If not, the P-value is greater than 0.50.

(b) If so, proceed to step 2.

Step 2 The P-value of the data is half as much as it would be if H_A were nondirectional.

To make a decision at a prespecified significance level α, one claims significant evidence for H_A if P-value $\leq \alpha$.

The following example illustrates the two-step procedure.

Example 7.10.3

Directional H_A Suppose $n = 8, n' = 7$, and H_A is directional. Suppose further that the data do deviate from H_0 in the direction specified by H_A. The values shown in Table 7.10.3 can be used to find the P-value as follows:

If $U_s = 40$, then P-value $= 0.189/2 = 0.0945$.

If $U_s = 46$, then P-value $= 0.054/2 = 0.027$.

If $U_s = 49.5$, then $0.009/2 < P$-value $< 0.014/2$ so $0.0045 < P$-value < 0.007.

If $U_s = 50$ (or larger), then P-value $< 0.009/2 = 0.0045$. ■

Rationale

Let us see why the Wilcoxon-Mann-Whitney test procedure makes sense. To take a specific case, suppose the sample sizes are $n_1 = 5$ and $n_2 = 4$, so that there are $5 \times 4 = 20$ comparisons that can be made between a data point in the first sample and a data point in the second sample. Thus, regardless of what the data look like, we must have

$$K_1 + K_2 = 5 \times 4 = 20$$

The relative magnitudes of K_1 and K_2 indicate the amount of overlap of the Y_1's and the Y_2's. Figure 7.10.3 shows how this works. For the data of Figure 7.10.3(a), the two samples do not overlap at all; the data are *least* compatible with H_0 and show the *strongest* evidence for H_A and thus U_s has its maximum value, $U_s = 20$. Similarly, $U_s = 20$ for Figure 7.10.3(b). On the other hand, the arrangement *most* compatible with H_0 and shows a lack of evidence for H_A is the one with maximal overlap, shown in Figure 7.10.3(c); for this arrangement $K_1 = 10, K_2 = 10$, and $U_s = 10$.

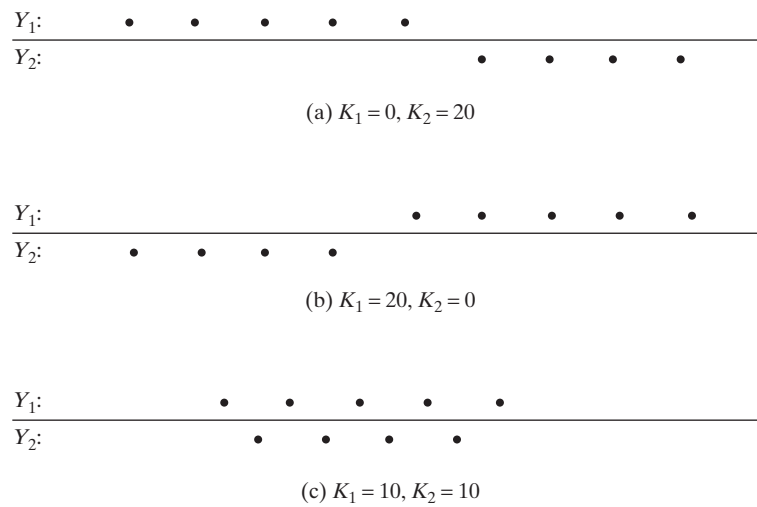


Figure 7.10.3 Three data arrays for a Wilcoxon-Mann-Whitney Test

All other possible arrangements of the data lie somewhere between the three arrangements shown in Figure 7.10.3; those with much overlap have U_s close to 10, and those with little overlap have U_s closer to 20. Thus, large values of U_s indicate evidence for the research hypothesis, H_A, or equivalently the incompatibility of the data with H_0.

We now briefly consider the null distribution of U_s and indicate how the critical values of Table 6 were determined. (Recall from Section 7.10 that, for any statistical test, the reference distribution for critical values is always the null distribution of the test statistic—that is, its sampling distribution under the condition that H_0 is true.) To determine the null distribution of U_s, it is necessary to calculate the probabilities associated with various arrangements of the data, assuming that all the Y's were actually drawn from the same population.* (The method for calculating the probabilities is briefly described in Appendix 7.2.)

Figure 7.10.4(a) shows the null distribution of K_1 and K_2 for the case $n = 5, n' = 4$. For example, it can be shown that, if H_0 is true, then

$$\Pr\{K_1 = 0, K_2 = 20\} = 0.008$$

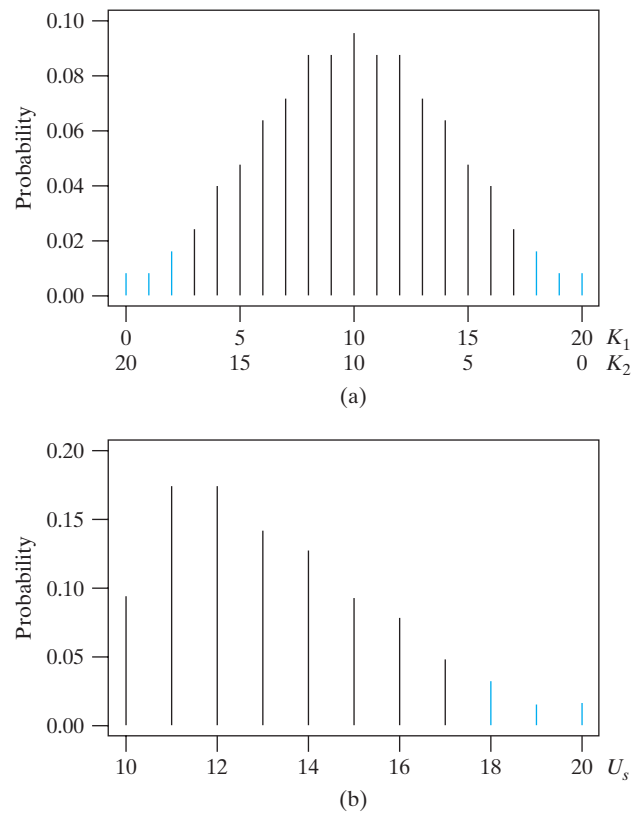


Figure 7.10.4 Null distributions for the Wilcoxon-Mann-Whitney test when $n = 5, n' = 4$. (a) Null distribution of K_1 and K_2; (b) Null distribution of U_s. Shading corresponds to the P-value when $U_s = 18$.

This is the first probability plotted in Figure 7.10.4(a). Note that Figure 7.10.4(a) is roughly analogous to a t distribution; large values of K_1 (right tail) represent evidence that the Y_1's tend to be larger than the Y_2's and large values of K_2 (left tail) represent evidence that the Y_2's tend to be larger than the Y_1's.

Figure 7.10.4(b) shows the null distribution of U_s, which is derived directly from the distribution in Figure 7.10.4(a). For instance, if H_0 is true, then

$$\Pr\{K_1 = 0, K_2 = 20\} = 0.008$$

*In calculating the probabitities used in this section, it has been assumed that the chance of tied observations is negligible. This will be true for a continuous variable that is measured with high precision. If the number of ties is large, a correction can be made; see Noether (1967).[57]

and

$$\Pr\{K_1 = 20, K_2 = 0\} = 0.008$$

so that

$$\Pr\{U_s = 20\} = 0.008 + 0.008 = 0.016$$

which is the rightmost probability plotted in Figure 7.10.4(b). Thus, both tails of the K distribution have been "folded" into the upper tail of the U distribution; for instance, the one-tailed shaded area in Figure 7.10.4(b) is equal to the two-tailed shaded area in Figure 7.10.4(a).

P-values for the Wilcoxon-Mann-Whitney test are upper-tail areas in the U_s distribution. For instance, it can be shown that the blue shaded area in Figure 7.10.4(b) is equal to 0.064; this means that if H_0 is true, then

$$\Pr\{U_s \geq 18\} = 0.064$$

Thus, a data set that yielded $U_s = 18$ would have an associated P-value 0.064 (assuming a nondirectional H_A).

The values in Table 6 have been determined from the null distribution of U_s. Because the U_s distribution is discrete, only a few P-values are possible for any given sample sizes n_1 and n_2. Table 6 shows selected values of U_s in bold type, with the P-value given in italics. For example, if the sample sizes are 5 and 4, then a U_s value of 17 gives a P-value of 0.111, a U_s value of 18 gives a P-value of 0.064, and a U_s value of 19 gives a P-value of 0.032. Thus, to achieve statistical significance at the $\alpha = 0.05$ level requires a test statistic (U_s) value of 19. The smallest possible P-value when the sample sizes are 5 and 4 is 0.016, when $U_s = 20$, which means that statistical significance at the $\alpha = 0.01$ level cannot be obtained with a nondirectional test.

Conditions for Use of the Wilcoxon-Mann-Whitney Test

In order for the Wilcoxon-Mann-Whitney test to be applicable, it must be reasonable to regard the data as random samples from their respective populations, with the observations within each sample being independent, and the two samples being independent of each other. Under these conditions, the Wilcoxon-Mann-Whitney test is valid no matter what the form of the population distributions, provided that the observed variable Y is continuous.[58]

The critical values given in Table 6 have been calculated assuming that ties do not occur. If the data contain only a few ties, then the P-values are approximately correct.*

The Wilcoxon-Mann-Whitney Test versus the *t* Test and the Randomization Test

While the Wilcoxon-Mann-Whitney test and the t test are aimed at answering the same basic question—Are the locations of the two population distributions different or does one population tend to have larger (or smaller) values than the other?—

*Actually, the Wilcoxon-Mann-Whitney test need not be restricted to continuous variables; it can be applied to any ordinal variable. However, if Y is discrete or categorical, then the data may contain many ties, and the test should not be used without appropriate modification of the critical values.

they treat the data in very different ways. Unlike the t test, the Wilcoxon-Mann-Whitney test does not use the actual values of the Y's but only their relative positions in a rank ordering. This is both a strength and a weakness of the Wilcoxon-Mann-Whitney test. On the one hand, the test is distribution free because the null distribution of U_s relates only to the various rankings of the Y's, and therefore does not depend on the form of the population distribution. On the other hand, the Wilcoxon-Mann-Whitney test can be inefficient: It can lack power because it does not use all the information in the data. This inefficiency is especially evident for small samples.

The randomization test is similar in spirit to the Wilcoxon-Mann-Whitney test in that it does not depend on normality, yet the power of the randomization test is often similar to that of the t test. Conducting a randomization test can be difficult, which is a primary reason that randomization tests were not more widely used until computing power became more prevalent.

None of the competitors—the randomization test, the t test, or the Wilcoxon-Mann-Whitney test—is clearly superior to the others. If the population distributions are not approximately normal, the t test may not even be valid. In addition, the Wilcoxon-Mann-Whitney test can be much more powerful than the t test, especially if the population distributions are highly skewed. If the population distributions are approximately normal with equal standard deviations, then the t test is best, but its properties are similar to those of the randomization test. For moderate sample sizes, the Wilcoxon-Mann-Whitney test can be nearly as powerful as the t test.[59]

There is a confidence interval procedure for population medians that is associated with the Wilcoxon-Mann-Whitney test in the same way that the confidence interval for $(\mu_1 - \mu_2)$ is associated with the t test. The procedure is beyond the scope of this book.

Exercises 7.10.1–7.10.9

7.10.1 Consider two samples of sizes $n_1 = 5$, $n_2 = 7$. Use Table 6 to find the P-value, assuming that H_A is nondirectional and that

(a) $U_s = 26$

(b) $U_s = 30$

(c) $U_s = 35$

7.10.2 Consider two samples of sizes $n_1 = 4$, $n_2 = 8$. Use Table 6 to find the P-value, assuming that H_A is nondirectional and that

(a) $U_s = 25$

(b) $U_s = 31$

(c) $U_s = 32$

7.10.3 In a pharmacological study, researchers measured the concentration of the brain chemical dopamine in six rats exposed to toluene and six control rats. (This is the same study described in Example 7.2.1.) The concentrations in the striatum region of the brain were as shown in the table.[4]

DOPAMINE (ng/gm)	
TOLUENE	CONTROL
3,420	1,820
2,314	1,843
1,911	1,397
2,464	1,803
2,781	2,539
2,803	1,990

(a) Use a Wilcoxon-Mann-Whitney test to compare the treatments at $\alpha = 0.05$. Use a nondirectional alternative.

(b) Proceed as in part (a), but let the alternative hypothesis be that toluene increases dopamine concentration.

7.10.4 In a study of hypnosis, breathing patterns were observed in an experimental group of subjects and in a control group. The measurements of total ventilation (liters of air per minute per square meter of body area) are shown.[60] (These are the same data that were summarized in Exercise 7.5.6.) Use a Wilcoxon-Mann-Whitney test to compare the two groups at $\alpha = 0.10$. Use a nondirectional alternative.

EXPERIMENTAL	CONTROL
5.32	4.50
5.60	4.78
5.74	4.79
6.06	4.86
6.32	5.41
6.34	5.70
6.79	6.08
7.18	6.21

7.10.5 In an experiment to compare the effects of two different growing conditions on the heights of greenhouse chrysanthemums, all plants grown under condition 1 were found to be taller than any of those grown under condition 2 (that is, the two height distributions did not overlap). Calculate the value of U_s and find the P-value if the number of plants in each group was

(a) 3
(b) 4
(c) 5

(Assume that H_A is nondirectional.)

7.10.6 In a study of preening behavior in the fruitfly *Drosophila melanogaster*, a single experimental fly was observed for three minutes while in a chamber with 10 other flies of the same sex. The observer recorded the timing of each episode ("bout") of preening by the experimental fly. This experiment was replicated 15 times with male flies and 15 times with female flies (different flies each time). One question of interest was whether there is a sex difference in preening behavior. The observed preening times (average time per bout, in seconds) were as follows:[61]

Male:	1.2, 1.2, 1.3, 1.9, 1.9, 2.0, 2.1, 2.2, 2.2, 2.3, 2.3, 2.4, 2.7, 2.9, 3.3

$\bar{y} = 2.127 \qquad s = 0.5936$

Female:	2.0, 2.2, 2.4, 2.4, 2.4, 2.8, 2.8, 2.8, 2.9, 3.2, 3.7, 4.0, 5.4, 10.7, 11.7

$\bar{y} = 4.093 \qquad s = 3.014$

(a) For these data, the value of the Wilcoxon-Mann-Whitney statistic is $U_s = 189.5$. Use a Wilcoxon-Mann-Whitney test to investigate the sex difference in preening behavior. Let H_A be nondirectional and let $\alpha = 0.01$.

(b) For these data, the standard error of $(\overline{Y}_1 - \overline{Y}_2)$ is SE = 0.7933 sec. Use a t test to investigate the sex difference in preening behavior. Let H_A be nondirectional and let $\alpha = 0.01$.

(c) What condition is required for the validity of the t test but not for the Wilcoxon-Mann-Whitney test? What feature or features of the data suggest that this condition may not hold in this case?

(d) Verify the value of U_s given in part (a).

7.10.7 Substances to be tested for cancer-causing potential are often painted on the skin of mice. The question arose whether mice might get an additional dose of the substance by licking or biting their cagemates. To answer this question, the compound benzo(a)pyrene was applied to the backs of 10 mice: Five were individually housed and 5 were group-housed in a single cage. After 48 hours, the concentration of the compound in the stomach tissue of each mouse was determined. The results (nmol/gm) were as follows:[62]

SINGLY HOUSED	GROUP-HOUSED
3.3	3.9
2.4	4.1
2.5	4.8
3.3	3.9
2.4	3.4

(a) Use a Wilcoxon-Mann-Whitney test to compare the two distributions at $\alpha = 0.01$. Let the alternative hypothesis be that benzo(a)pyrene concentrations tend to be higher in group-housed mice than in singly housed mice.

(b) Why is a directional alternative valid in this case?

7.10.8 Human beta-endorphin (HBE) is a hormone secreted by the pituitary gland under conditions of stress. An exercise physiologist measured the resting (unstressed) blood concentration of HBE in two groups of men: Group 1 consisted of 11 men who had been jogging regularly for some time, and group 2 consisted of 15 men who had just entered a physical fitness program. The results are given in the following table.[63]

	FITNESS PROGRAM
JOGGERS	ENTRANTS
39 40 32 60	70 47 54 27 31
19 52 41 32	42 37 41 9 18
13 37 28	33 23 49 41 59

Use a Wilcoxon-Mann-Whitney test to compare the two distributions at $\alpha = 0.10$. Use a nondirectional alternative.

7.10.9 (Continuation of 7.10.8) Below are normal probability plots of the HBE data from Exercise 7.10.8.

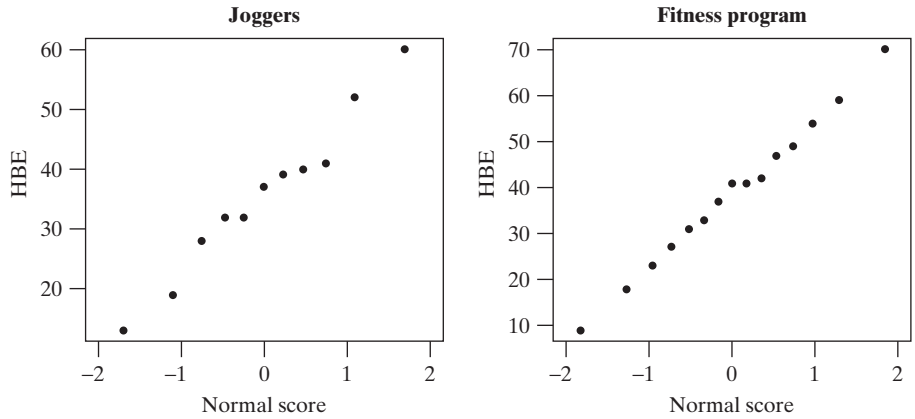
(a) Using the plots to support your answer, is there evidence of abnormality in either of the samples?

(b) Considering your answer to (a) and the preceding plots, should we conclude that the data are indeed normally distributed? Explain.

(c) If the data are indeed normally distributed, explain in the context of this problem what the drawback would be with using the Wilcoxon-Mann-Whitney test over the two-sample t test to analyze these data.

(d) If the data are not normally distributed, explain in the context of this problem what the drawback would be with using the two-sample t test over the Wilcoxon-Mann-Whitney test to analyze this data.

(e) Considering your answers to the above, argue which test should be used with these data. Note there is more than one correct answer.



7.11 Perspective

In this chapter we have discussed several techniques—confidence intervals and hypothesis tests—for comparing two independent samples when the observed variable is quantitative. In coming chapters we will introduce confidence interval and hypothesis testing techniques that are applicable in various other situations. Before proceeding, we pause to reconsider the methods of this chapter.

An Implicit Assumption

In discussing the tests of this chapter—the t test and the Wilcoxon-Mann-Whitney test—we have made an unspoken assumption, which we now bring to light. When interpreting the comparison of two distributions, we have assumed that the relationship between the two distributions is relatively simple—that if the distributions differ, then one of the two variables has a consistent tendency to be larger than the other. For instance, suppose we are comparing the effects of two diets on the weight gain of mice, with

$$Y_1 = \text{Weight gain of mice on diet 1}$$
$$Y_2 = \text{Weight gain of mice on diet 2}$$

Our implicit assumption has been that, if the two diets differ at all, then that difference is in a consistent direction for all individual mice. To appreciate the meaning

of this assumption, suppose the two distributions are as pictured in Figure 7.11.1. In this case, even though the mean weight gain is higher on diet 1, it would be an oversimplification to say that mice tend to gain more weight on diet 1 than on diet 2; apparently *some* mice gain *less* on diet 1. Paradoxical situations of this kind do occasionally occur, and then the simple analysis typified by the t test and the Wilcoxon-Mann-Whitney test may be inadequate.

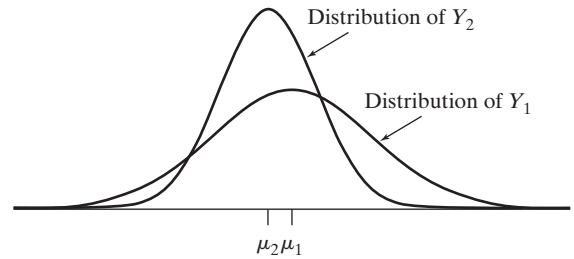


Figure 7.11.1 Weight gain distributions on two diets

It is relatively easy to compare two distributions that have the same general shape and similar standard deviations. However, if either the shapes or the SDs of two distributions are very different from one another, then making a meaningful comparison of the distributions is difficult. In particular, a comparison of the two means might not be appropriate.

Which Method to Use When

If we are comparing samples from two normally distributed populations, a t test can be used to infer whether the population means differ and a confidence interval can be used to estimate how much the two population means might differ, if at all. A confidence interval generally provides more information than does a test, since the test is restricted to a narrow question ("Might the difference between the sample be reasonably attributed to chance?"), whereas the confidence interval addresses a larger question ("How much larger is μ_1 than μ_2?").

Both the confidence interval and the t test depend on the condition that the populations are normally distributed. If this condition is not met, then a transformation might be used to make the distributions approximately normal before proceeding. If, despite considering transformations, the normality condition is questionable, then the Wilcoxon-Mann-Whitney test can be used. (Indeed, the Wilcoxon-Mann-Whitney test can be used if the data are normal, although it is less powerful than the t test). When in doubt, a good piece of advice is to conduct both a t test and a Wilcoxon-Mann-Whitney test. If the two tests give similar, clear, conclusions (i.e., if the P-values for the tests are similar and both are considerably larger than α or both are considerably smaller than α), then we can feel comfortable with the conclusion. However, if one test yields a P-value somewhat larger than α and the other gives a P-value smaller than α, then we might well declare that the tests are inconclusive.

Sometimes an outlier will be present in a data set, calling into question the result of a t test. It is not legitimate to simply ignore the outlier. A sensible procedure is to conduct the analysis with the outlier included and then delete the outlier and repeat the analysis. If the conclusion is unchanged when the outlier is removed, then we can feel confident that no single observation is having undue influence on the inferences we draw from the data. If the conclusion changes when the outlier is

removed, then we cannot be confident in the inferences we draw. For example, if the P-value for a test is small with the outlier present but large when the outlier is deleted, then we might state, "There is evidence that the populations differ from one another, but this evidence is largely due to a single observation." Such a statement warns the reader that not too much should be read into any differences that were observed between the samples.

Comparison of Variability

It sometimes happens that the variability of Y, rather than its average value, is of primary interest. For instance, in comparing two different lab techniques for measuring the concentration of an enzyme, a researcher might want primarily to know whether one of the techniques is more precise than the other, that is, whether its measurement error distribution has a smaller standard deviation. There are techniques available for testing the hypothesis H_0: $\sigma_1 = \sigma_2$, and for using a confidence interval to compare σ_1 and σ_2. Most of these techniques are very sensitive to the condition that the underlying distributions are normal, which limits their use in practice. The implementation of these techniques is beyond the scope of this book.

Supplementary Exercises 7.S.1–7.S.30

(*Note:* Exercises preceded by an asterisk refer to optional sections.)

Answers to hypothesis testing questions should include a statement of the conclusion in the context of the setting. (See Examples 7.2.4 and 7.2.5.)

7.S.1 For each of the following pairs of samples, compute the standard error of $(\overline{Y}_1 - \overline{Y}_2)$.

(a)

	SAMPLE 1	SAMPLE 2
n	12	13
$\bar{y}$	42	47
s	9.6	10.2

(b)

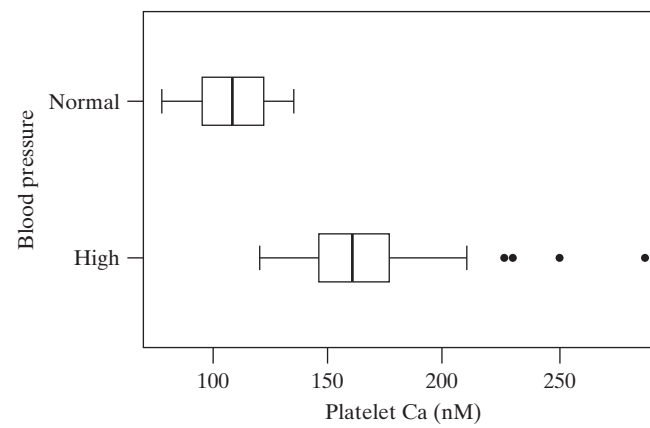
	SAMPLE 1	SAMPLE 2
n	22	19
$\bar{y}$	112	126
s	2.7	1.9

(c)

	SAMPLE 1	SAMPLE 2
n	5	7
$\bar{y}$	14	16
SE	1.2	1.4

7.S.2 To investigate the relationship between intracellular calcium and blood pressure, researchers measured the free calcium concentration in the blood platelets of 38 people with normal blood pressure and 45 people with high blood pressure. The results are given in the table and the distributions are shown in the boxplots.[64] Use the t test to compare the means. Let $\alpha = 0.01$ and let H_A be nondirectional. [*Note:* Formula (6.7.1) yields 67.5 df.]

	PLATELET CALCIUM (nM)		
BLOOD PRESSURE	n	MEAN	SD
Normal	38	107.9	16.1
High	45	168.2	31.7



7.S.3 Refer to Exercise 7.S.2. Construct a 95% confidence interval for the difference between the population means.

7.S.4 Refer to Exercise 7.S.2. The boxplot for the high blood pressure group is skewed to the right and includes outliers. Does this mean that the t test is not valid for these data? Why or why not?

7.S.5 In a study of methods of producing sheep's milk for use in cheese manufacture, ewes were randomly allocated to either a mechanical or a manual milking method. The investigator suspected that the mechanical method might irritate the udder and thus produce a higher concentration of somatic cells in the milk. The accompanying data show the average somatic cell count for each animal.[65]

	SOMATIC COUNT (10^{-3} × cells/ml)	
	MECHANICAL MILKING	MANUAL MILKING
	2,966	186
	269	107
	59	65
	1,887	126
	3,452	123
	189	164
	93	408
	618	324
	130	548
	2,493	139
n	10	10
Mean	1,215.6	219.0
SD	1,342.9	156.2

(a) Do the data support the investigator's suspicion? Use a t test against a directional alternative at $\alpha = 0.05$. The standard error of $(\overline{Y}_1 - \overline{Y}_2)$ is SE = 427.54 and formula (6.7.1) yields 9.2 df.

(b) Do the data support the investigator's suspicion? Use a Wilcoxon-Mann-Whitney test against a directional alternative at $\alpha = 0.05$. (The value of the Wilcoxon-Mann-Whitney statistic is $U_s = 69$.) Compare with the result of part (a).

(c) What condition is required for the validity of the t test but not for the Wilcoxon-Mann-Whitney test? What features of the data cast doubt on this condition?

(d) Verify the value of U_s given in part (b).

7.S.6 A plant physiologist conducted an experiment to determine whether mechanical stress can retard the growth of soybean plants. Young plants were randomly allocated to two groups of 13 plants each. Plants in one group were mechanically agitated by shaking for 20 minutes twice daily, while plants in the other group were not agitated. After 16 days of growth, the total stem length (cm) of each plant was measured, with the results given in the accompanying table.[66]

Use a t test to compare the treatments at $\alpha = 0.01$. Let the alternative hypothesis be that stress tends to retard growth. [*Note:* Formula (6.7.1) yields 23 df.]

	CONTROL	STRESS
n	13	13
$\bar{y}$	30.59	27.78
s	2.13	1.73

7.S.7 Refer to Exercise 7.S.6. Construct a 95% confidence interval for the population mean reduction in stem length. Does the confidence interval indicate whether the effect of stress is "horticulturally important," if "horticulturally important" is defined as a reduction in population mean stem length of at least

(a) 1 cm

(b) 2 cm

(c) 5 cm

7.S.8 Refer to Exercise 7.S.6. The observations (cm), in increasing order, are shown. Compare the treatments using a Wilcoxon-Mann-Whitney test at $\alpha = 0.01$. Let the alternative hypothesis be that stress tends to retard growth.

CONTROL	STRESS
25.2	24.7
29.5	25.7
30.1	26.5
30.1	27.0
30.2	27.1
30.2	27.2
30.3	27.3
30.6	27.7
31.1	28.7
31.2	28.9
31.4	29.7
33.5	30.0
34.3	30.6

7.S.9 One measure of the impact of pollution along a river is the diversity of species in the river floodplain. In one study, two rivers, the Black River and the Vermilion River, were compared. Random 50-m × 20-m plots were sampled along each river and the number of species of trees in each plot was recorded. The following table contains the data.[67]

VERMILION RIVER	BLACK RIVER
9 9 16 13 12	13 10 6 9
13 13 13 8 11	10 7 6 18
9 9 10	6

The Black River was considered to have been polluted quite a bit more than the Vermilion River, and this was expected to lead to lower biodiversity along the Black River. Conduct a Wilcoxon-Mann-Whitney test, with $\alpha = 0.10$, of the null hypothesis that the populations from which the two samples were drawn have the same biodiversity (distribution of tree species per plot) versus an appropriate directional alternative.

7.S.10 A developmental biologist removed the oocytes (developing egg cells) from the ovaries of 24 frogs *(Xenopus laevis)*. For each frog the oocyte pH was determined. In addition, each frog was classified according to its response to a certain stimulus with the hormone progesterone. The pH values were as follows:[68]

Positive response:

7.06, 7.18, 7.30, 7.30, 7.31, 7.32, 7.33, 7.34, 7.36, 7.36, 7.40, 7.41, 7.43, 7.48, 7.49, 7.53, 7.55, 7.57

No response:

7.55, 7.70, 7.73, 7.75, 7.75, 7.77

Investigate the relationship of oocyte pH to progesterone response using a Wilcoxon-Mann-Whitney test at $\alpha = 0.05$. Use a nondirectional alternative.

7.S.11 Refer to Exercise 7.S.10. Summary statistics for the pH measurements are given in the following table. Investigate the relationship of oocyte pH to progesterone response using a t test at $\alpha = 0.05$. Use a nondirectional alternative. [*Note:* Formula (6.7.1) yields 14.1 df.]

	POSITIVE RESPONSE	NO RESPONSE
n	18	6
$\bar{y}$	7.373	7.708
s	0.129	0.081

7.S.12 A proposed new diet for beef cattle is less expensive than the standard diet. The proponents of the new diet have conducted a comparative study in which one group of cattle was fed the new diet and another group was fed the standard. They found that the mean weight gains in the two groups were not statistically significantly different at the 5% significance level, and they stated that this finding supported the claim that the new cheaper diet was as good (for weight gain) as the standard diet. Criticize this statement.

***7.S.13** Refer to Exercise 7.S.12. Suppose you discover that the study used 25 animals on each of the two diets, and that the coefficient of variation of weight gain under the conditions of the study was about 20%. Using this additional information, write an expanded criticism of the proponents' claim, indicating how likely such a study would be to detect a 10% deficiency in weight gain on the cheaper diet (using a two-tailed test at the 5% significance level).

7.S.14 In a study of hearing loss, endolymphatic sac tumors (ELSTs) were discovered in 13 patients. These 13 patients had a total of 15 tumors (i.e., more patients had a single tumor, but two of the patients had 2 tumors each). Ten of the tumors were associated with the loss of functional hearing in an ear, but for 5 of the ears with tumors the patient had no hearing loss.[69] A natural question is whether hearing loss is more likely with large tumors than with small tumors. Thus, the sizes of the tumors were measured. Suppose that the sample means and standard deviations were given and that a comparison of average tumor size (hearing loss versus no hearing loss) was being considered.

(a) Explain why a t test to compare average tumor size is not appropriate here.

(b) If the raw data were given, could a Wilcoxon-Mann-Whitney test be used?

7.S.15 (Computer exercise) In an investigation of the possible influence of dietary chromium on diabetic symptoms, 14 rats were fed a low-chromium diet and 10 were fed a normal diet. One response variable was activity of the liver enzyme GITH, which was measured using a radioactively labeled molecule. The accompanying table shows the results, expressed as thousands of counts per minute per gram of liver.[70] Use a t test to compare the diets at $\alpha = 0.05$. Use a nondirectional alternative. [*Note:* Formula (6.7.1) yields 21.9 df.]

LOW-CHROMIUM DIET		NORMAL DIET	
42.3	52.8	53.1	53.6
51.5	51.3	50.7	47.8
53.7	58.5	55.8	61.8
48.0	55.4	55.1	52.6
56.0	38.3	47.5	53.7
55.7	54.1		
54.8	52.1		

7.S.16 (Computer exercise) Refer to Exercise 7.S.15. Use a Wilcoxon-Mann-Whitney test to compare the diets at $\alpha = 0.05$. Use a nondirectional alternative.

7.S.17 (Computer exercise) Refer to Exercise 7.S.15.

(a) Construct a 95% confidence interval for the difference in population means.

(b) Suppose the investigators believe that the effect of the low-chromium diet is "unimportant" if it shifts mean GITH activity by less than 15%—that is, if the population mean difference is less than about 8 thousand cpm/gm. According to the confidence interval of part (a), do the data support the conclusion that the difference is "unimportant"?

(c) How would you answer the question in part (b) if the criterion were 4 thousand rather than 8 thousand cpm/gm?

7.S.18 (Computer exercise) In a study of the lizard *Scelopons occidentalis*, researchers examined field-caught lizards for infection by the malarial parasite *Plasmodium*. To help assess the ecological impact of malarial infection, the researchers tested 15 infected and 15 noninfected lizards for stamina, as indicated by the distance each animal could run in two minutes. The distances (meters) are shown in the table.[71]

INFECTED ANIMALS		UNINFECTED ANIMALS	
16.4	36.7	22.2	18.4
29.4	28.7	34.8	27.5
37.1	30.2	42.1	45.5
23.0	21.8	32.9	34.0
24.1	37.1	26.4	45.5
24.5	20.3	30.6	24.5
16.4	28.3	32.9	28.7
29.1		37.5	

Do the data provide evidence that the infection is associated with decreased stamina? Investigate this question using

(a) a t test.

(b) a Wilcoxon-Mann-Whitney test.

Let H_A, be directional and $\alpha = 0.05$.

7.S.19 In a study of the effect of amphetamine on water consumption, a pharmacologist injected four rats with amphetamine and four with saline as controls. She measured the amount of water each rat consumed in 24 hours. The following are the results, expressed as ml water per kg body weight:[72]

AMPHETAMINE	CONTROL
118.4	122.9
124.4	162.1
169.4	184.1
105.3	154.9

(a) Use a t test to compare the treatments at $\alpha = 0.10$. Let the alternative hypothesis be that amphetamine tends to suppress water consumption.

(b) Use a Wilcoxon-Mann-Whitney test to compare the treatments at $\alpha = 0.10$, with the directional alternative that amphetamine tends to suppress water consumption.

(c) Why is it important that some of the rats received saline injections as a control? That is, why didn't the researchers simply compare rats receiving amphetamine injections to rats receiving no injection?

7.S.20 Nitric oxide is sometimes given to newborns who experience respiratory failure. In one experiment, nitric oxide was given to 114 infants. This group was compared to a control group of 121 infants. The length of hospitalization (in days) was recorded for each of the 235 infants. The mean in the nitric oxide sample was $\bar{y}_1 = 36.4$; the mean in the control sample was $\bar{y}_2 = 29.5$. A 95% confidence interval for $\mu_1 - \mu_2$ is $(-2.3, 16.1)$, where μ_1 is the population mean length of hospitalization for infants who get nitric oxide and μ_2 is the mean length of hospitalization for infants in the control population.[73] For each of the following, say whether the statement is true or false and say why.

(a) We are 95% confident that μ_1 is greater than μ_2, since most of the confidence interval is greater than zero.

(b) We are 95% confident that the difference between μ_1 and μ_2 is between –2.3 days and 16.1 days.

(c) We are 95% confident that the difference between $\bar{y}_1$ and $\bar{y}_2$ is between -2.3 days and 16.1 days.

(d) 95% of the nitric oxide infants were hospitalized longer than the average control infant.

7.S.21 Consider the confidence interval for $\mu_1 - \mu_2$ from Exercise 7.S.20: (–2.3, 16.1). True or false: If we tested H_0: $\mu_1 = \mu_2$ against H_A: $\mu_1 \neq \mu_2$, using $\alpha = 0.05$, we would reject H_0.

7.S.22 Researchers studied subjects who had pneumonia and classified them as being in one of two groups: those who were given medical therapy that is consistent with American Thoracic Society (ATS) guidelines and those who were given medical therapy that is inconsistent with ATS guidelines. Subjects in the "consistent" group were generally able to return to work sooner than were subjects in the "inconsistent" group. A Wilcoxon-Mann-

Whitney test was applied to the data; the P-value for the test was 0.04.[74] For each of the following, say whether the statement is true or false and say why.

(a) There is a 4% chance that the "consistent" and "inconsistent" population distributions really are the same.

(b) If the "consistent" and "inconsistent" population distributions really are the same, then a difference between the two samples as large as the difference that these researchers observed would only happen 4% of the time.

(c) If a new study were done that compared the "consistent" and "inconsistent" populations, there is a 4% probability that H_0 would be rejected again.

7.S.23 A student recorded the number of calories in each of 56 entrees—28 vegetarian and 28 nonvegetarian—served at a college dining hall.[75] The following table summarizes the data. Graphs of the data (not given here) show that both distributions are reasonably symmetric and bell shaped. A 95% confidence interval for $\mu_1 - \mu_2$ is $(-27, 85)$. For each of the following, say whether the statement is true or false and say why.

	n	MEAN	SD
Vegetarian	28	351	119
Nonvegetarian	28	322	87

(a) 95% of the data are between −27 and 85 calories.

(b) We are 95% confident that $\mu_1 - \mu_2$ is between −27 and 85 calories.

(c) 95% of the time $\overline{Y}_1 - \overline{Y}_2$ will be between −27 and 85 calories.

(d) 95% of the vegetarian entrees have between 27 fewer calories and 85 more calories than the average nonvegetarian entree.

7.S.24 Refer to Exercise 7.S.23. True or false (and say why): 95% of the time, when conducting a study of this size, the difference in sample means $(\overline{Y}_1 - \overline{Y}_2)$ will be within approximately $\dfrac{(85 - (-27))}{2} = 56$ calories of the difference in population means $(\mu_1 - \mu_2)$.

7.S.25 (Computer exercise) Lianas are woody vines that grow in tropical forests. Researchers measured liana abundance (stems/ha) in several plots in the central Amazon region of Brazil. The plots were classified into two types: plots that were near the edge of the forest (less than 100 meters from the edge) or plots far from the edge of the forest. The raw data are given and are summarized in the table.[76]

	n	MEAN	SD
Near	34	438	125
Far	34	368	114

NEAR			FAR		
639	601	600	470	339	384
605	581	555	309	395	393
535	531	466	236	252	407
437	423	380	241	215	427
376	362	350	320	228	445
349	346	337	325	267	451
320	317	310	352	294	493
285	271	265	275	356	502
250	450	441	181	418	540
436	432	420	250	425	590
419	407		266	495	
702	676		338	648	

(a) Make normal probability plots of the data to confirm that the distributions are mildly skewed.

(b) Conduct a t test to compare the two types of plots at $\alpha = 0.05$. Use a nondirectional alternative.

(c) Apply a logarithm transformation to the data and repeat parts (a) and (b).

(d) Compare the t tests from parts (b) and (c). What do these results indicate about the effect on a t test of mild skewness when the sample sizes are fairly large?

7.S.26 Androstenedione (andro) is a steroid that is thought by some athletes to increase strength. Researchers investigated this claim by giving andro to one group of men and a placebo to a control group of men. One of the variables measured in the experiment was the increase in "lat pulldown" strength (in pounds) of each subject after four weeks. (A lat pulldown is a type of weightlifting exercise.) The raw data are given below and are summarized in the table.[77]

	n	MEAN	SD
Control	9	14.4	13.3
Andro	10	20.0	12.5

CONTROL				ANDRO			
30	10	10	30	0	10	0	10
40	20	30	20	10	40	20	10
10	0			30			

(a) Conduct a t test to compare the two groups at $\alpha = 0.10$. Use a nondirectional alternative. [*Note:* Formula (6.7.1) yields 16.5 df.]

(b) Prior to the study it was expected that andro would increase strength, which means that a directional alternative might have been used. Redo the analysis in part (a) using the appropriate directional alternative.

7.S.27 The following is a sample of computer output from a study.[78] Describe the problem and the conclusion, based on the computer output.

```
Y = number of drinks in the previous 7 days

  Two-sample T for treatment vs. control:

                  n       Mean        SD
   Treatment      244     13.62      12.39
   Control        238     16.86      13.49

95% CI for mu1 − mu2:(−5.56, − 0.92)

T-test mu1 = mu2 (vs < ):
T = −2.74 P = .0031 DF = 474.3
```

7.S.28 In a controversial study to determine the effectiveness of AZT, a group of HIV-positive pregnant women were randomly assigned to get either AZT or a placebo. Some of the babies born to these women were HIV-positive, while others were not.[79]

(a) What is the explanatory variable?

(b) What is the response variable?

(c) What are the experimental units?

7.S.29 Patients suffering from acute respiratory failure were randomly assigned to either be placed in a prone (face down) position or a supine (face up) position. In the prone group, 21 out of 152 patients died. In the supine group, 25 out of 152 patients died.[80]

(a) What is the explanatory variable?

(b) What is the response variable?

(c) What are the experimental units?

7.S.30 A study of postmenopausal women on hormone replacement therapy (H.R.T.) reported that they had a reduced heart attack rate, but had even greater reductions in death from homicide and accidents—two causes of death that cannot be linked to H.R.T. It seems that the women on H.R.T. differ from others in many other aspects of their lives—for instance, they exercise more; they also tend to be wealthier and to be better educated.[81] Use the language of statistics to discuss what these data say about the relationships between H.R.T., heart attack risk, and variables such as exercise, wealth, and education. Use a schematic diagram similar to Figure 7.4.1 or Figure 7.4.2 to support your explanation.